INFORMATION-THEORETIC SENSOR NETWORK DESIGN AND RELIABLE ROUTE GUIDANCE: PLANNING AND ENRICHING ADVANCED TRAVELER INFORMATION SYSTEMS

by

Tao Xing

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Civil and Environmental Engineering

The University of Utah

May 2012

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Tao Xing**

has been approved by the following supervisory committee members:

Xuesong Zhou	, Chair	**10/18/2011** Date Approved
Peter Martin	, Member	**10/18/2011** Date Approved
Richard Porter	, Member	**10/18/2011** Date Approved
Harvey Miller	, Member	**10/18/2011** Date Approved
Rong-Rong Chen	, Member	**10/18/2011** Date Approved

and by **Paul Tikalsky**, Chair of

the Department of **Civil and Environmental Engineering**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

This dissertation aims to develop an innovative and improved paradigm for real-time large-scale traffic system estimation and mobility optimization. To fully utilize heterogeneous data sources in a complex spatial environment, this dissertation proposes an integrated and unified estimation-optimization framework capable of interpreting different types of traffic measurements into various decision-making processes.

With a particular emphasis on the end-to-end travel time prediction problem, this dissertation proposes an information-theoretic sensor location model that aims to maximize information gains from a set of point, point-to-point and probe sensors in a traffic network. After thoroughly examining a number of possible measures of information gain, this dissertation selects a path travel time prediction uncertainty criterion to construct a joint sensor location and travel time estimation/prediction framework.

To better measure the quality of service for transportation systems, this dissertation investigates the path travel time reliability from two perspectives: variability and robustness. Based on calibrated travel disutility functions, the path travel time variability in this research is represented by its standard deviation in addition to the mean travel time. To handle the nonlinear and nonadditive cost functions introduced by the quadratic forms of the standard deviation term, a novel Lagrangian substitution approach is introduced to

estimate the lower bound of the most reliable path solution through solving a sequence of standard shortest path problems. To recognize the asymmetrical and heavy-tailed travel time distributions, this dissertation proposes Lagrangian relaxation based iterative search algorithms for finding the absolute and percentile robust shortest paths. Moreover, this research develops a sampling-based method to dynamically construct a proxy objective function in terms of travel time observations from multiple days. Comprehensive numerical experiment results with real-world travel time measurements show that 10-20 iterations of standard shortest path algorithms for the reformulated models can offer a very small relative duality gap of about 2-6%, for both reliability measure models.

This broadly-defined research has successfully addressed a number of theoretically challenging and practically important issues for building the next-generation Advanced Traveler Information Systems, and is expected to offer a rich foundation beneficial to the model and algorithmic development of sensor network design, traffic forecasting and personalized navigation.

To

My Parents

TABLE OF CONTENTS

ABSTRACT iii

LIST OF TABLES viii

LIST OF FIGURES ix

ACKNOWLEDGMENTS xi

Chapter

1. INTRODUCTION 1

   1.1. Background 1
   1.2. Motivations and Challenges 4
   1.3. Overview of Proposed Methods 11
   1.4. Organization of the Dissertation 14

2. LITERATURE REVIEW 15

   2.1. Data Collection through Sensors 15
   2.2. Travel Time Estimation and Prediction 18
   2.3. Variability-oriented Routing Solution 20

3. HETEROGENEOUS SENSOR NETWORK DESIGN FOR ESTIMATING AND PREDICTING PATH TRAVEL TIME DYNAMICS 23

   3.1. Notation and Modeling Framework Overview 24
   3.2. System Process and Measurement Models for Estimating Historical Regular Patterns 38
   3.3. Travel Time Prediction under Nonrecurring Conditions 46
   3.4. Measure of Information for Historical Traffic Patterns 52
   3.5. Sensor Design Model and Beam Search Algorithm 56
   3.6. Complex Cases for Updating Historical Traffic Patterns 60
   3.7. Illustrative Example and Numerical Experiments 65

4. FINDING THE MOST RELIABLE PATH WITH AND WITHOUT LINK TRAVEL TIME CORRELATION 77

4.1. Problem Statement and Model Assumptions 78
4.2. Algorithm for Finding Most Reliable Path without Link Travel Time Correlation 85
4.3. Algorithm for Finding Most Reliable Path with Link Travel Time Correlation .. 96
4.4. Numerical Experiments 108

5. FINDING ABSOLUTE AND PERCENTILE ROBUST SHORTEST PATHS 119

5.1. Finding Absolute Robust Shortest Path 120
5.2. Finding α-Percentile Robust Shortest Path 129
5.3. Numerical Experiments 138

6. CONCLUSIONS 143

6.1. Research Highlights 143
6.2. Summary of Contributions 147
6.3. Future Research 148

REFERENCES 150

LIST OF TABLES

Table Page

1.1: List of estimation, optimization targets and model products for three problems 4

3.1: Calculation results of single-link example..... 45

3.2: Derivation of transition matrix L for different time periods 51

3.3: Critical path travel time estimation error under existing and optimized sensor location strategies..... 69

4.1: Path travel time data..... 95

4.2: Travel time samples of each path (min)..... 104

4.3: Results of first few iterations for Algorithm 4.2..... 106

4.4: Solution quality comparison over three methods..... 117

5.1: Travel time calculation of the illustrative example..... 132

5.2: Results of first few iterations for Algorithm 5.1..... 136

5.3: Results of first few iterations for Algorithm 5.2..... 136

LIST OF FIGURES

Figure Page

1.1: Travel time distribution for the eastbound lanes of State Route 520, Seattle, adapted from FHWA report (Cambridge Systematics, 2005) 9

3.1: Section-level travel time estimation 30

3.2: Conceptual framework and data flow 36

3.3: Single-link example 45

3.4: Cumulative flow count curve for a single bottleneck with reduced capacity due to an incident 49

3.5: A zoom-in view on capacity reduction 50

3.6: Illustration of time-dependent transition coefficient L 52

3.7: Steady-state travel time estimation variance 61

3.8: Two partially overlapping path between AVI readers at nodes a and c 63

3.9: Example network with three parallel paths 65

3.10: Example of locating AVI sensors on a linear corridor 66

3.11: Numerical experiment results for regular traffic pattern estimation 70

3.12: Comparison of different sensor location schemes (for regular traffic estimation) .. 73

3.13: Numerical experiment results for traffic prediction under recurring and nonrecurring traffic conditions 74

4.1: Feasible region of optimal path travel time variance y. 86

4.2: $L_y(\mu)$ as a function of y, and optimal value obtained at one of the extreme points. . 89

4.3: Subgradient algorithm for solving Lagrangian dual problems. 92

4.4: Solution results for independent distribution-based algorithm. 95

4.5: Network of illustrative example..... 103

4.6: Evolution of lower bound (*LB*) and upper bound (*UB*) in the first a few iterations of the example 107

4.7: Measurement coverage of travel time for Bay Area, California. Links with measurements are highlighted. A subcorridor on Bayshore Freeway from Mountain View to San Jose, California is enlarged. 109

4.8: Partial link correlations on selected subcorridor..... 110

4.9: Relative gap in model WITHOUT link travel time correlation. Beta is the travel time reliability coefficient. 113

4.10: Relative gap in model WITH link travel time correlation. Beta is the travel time reliability coefficient. 113

4.11: Standard deviation of relative gaps for 246 OD pairs under two models and reliability coefficients. 114

4.12: Solution quality comparison of three methods: random draws, individual days and proposed sampling method 117

5.1: Network of the illustrative example..... 132

5.2: Sensor data coverage for Bay Area, California. 139

5.3: Relative gap for ARSP and 95% PRSP 140

5.4: Upper and lower bounds evolution for ARSP (top) and 95% PRSP (bottom) 141

ACKNOWLEDGMENTS

Throughout my four-year study in the University of Utah, I received great help from so many people who deserve my special acknowledgment.

First, I am very much indebted to my mentor, Professor Xuesong Zhou. It was his strong and continuous support over the past four years that guaranteed the completion of my study. Every part of my work is influenced by his wise and patient tutorials and directions. His enthusiasm for research inspired me to move on the right track for my study and I cherish everything I learned from him when facing obstacles and opportunities. Especially, I would like to thank him for giving me the courage to think and work with foresight and innovation.

I would like to show my deep appreciation to Professor Peter Martin and Professor RJ Porter for their continuous support during my graduate career. From our faculties of the transportation group, I have been inspired and benefited not just in research but in every aspect of being an engineer. I would like to take this opportunity to give my special thanks to Professor Harvey Miller and Professor Rong-Rong Chen for their participation and constructive comments on my research.

I feel grateful as well to my dear colleagues and friends for being there through my ups and downs. I would like to give my special thanks to Dr. Mingxin Li and Hao Lei, who always gave me sound suggestions and advice on my research without any hesitation,

and in particular, Jeff Taylor, who helped me by reviewing almost all my work. I also want to thank Li Su, TJ, Dr. Yang, Milan, Ivana, Yunqi, Wen Deng and other team members in our department. The creative and cooperative atmosphere provided me the motivation in every step of the way. I feel lucky and thankful for being a member of this group.

I would also like to show my appreciation to all my co-workers from several projects I participated in. Special thanks to Shawn, Ken, Daniel, Chao, Chris, Claurissa, Dr. Zongwei, Jeff and Bin, from whom I learned a lot through all the collaboration and communication. Although I may not have named everyone who contributed to my studies at the University of Utah, they will always be on my mind and my deepest appreciation goes to those who played a positive role in my life.

Last, but not the least, I would like to dedicate this thesis to my parents whose love and support have kept me moving forward throughout my life.

CHAPTER 1

INTRODUCTION

1.1. Background

The goal of Intelligent Transportation Systems (ITS) is to improve the transportation safety, mobility and environment through a wide range of advanced information and communication technologies. Among many ITS applications, Advanced Traveler Information Systems (ATIS) work closely with drivers by providing accurate and real-time traffic information, and have been implemented in practice for decades. Most commonly applied ATIS systems include in-vehicle routing and navigation systems, advanced roadway guidance signs (e.g., Variable Message Signs (VMS)), and traffic information services (e.g., 511 traveler information systems).

It has been well recognized that traffic congestion in metropolitan areas is difficult to mitigate due to the lack of mechanisms to (1) reliably measure and estimate network-wide traffic patterns and (2) effectively inform and divert travelers to avoid recurring and nonrecurring congestion. From this point of view, the success of ATIS deployment relies on research and practical developments on the following three components: traffic observability, estimation/prediction, and information dissemination.

1.1.1. System-wide Traffic Observability

Based on the types of measurement data, traffic sensors can be categorized into three groups, namely point sensors, point-to-point sensors, and probe sensors. A small subset of freeway links is typically instrumented with in-pavement and road-side traffic detectors, which experience significant failure rates, to estimate travel time and traffic flow. Accurate travel time and traffic flow information on ramps and arterial corridors are critically needed, but very costly to collect with the nation's existing infrastructure. Overall, limited point measurements from the current traffic sensor infrastructure are unable to provide sufficient spatial and temporal coverage to measure complex traffic flow patterns in a traffic network.

An effective ATIS program should use data from multiple sources to enhance the system-wide observability for an entire traffic network. Many Automatic Vehicle Identification (AVI) and Automatic Vehicle Location (AVL) technologies, such as toll tags and Bluetooth signal reading, provide new possibilities for traffic monitoring to semi-continuously obtain detailed passing time and location information along individual vehicle trajectories on both freeway and arterial corridors. As the personal navigation market grows rapidly, probe data from in-vehicle Personal Navigation Devices (PND) and cell phones become more readily available for continuous travel time measurement. By estimating network-wide traffic states from multiple data sources, the data mining engines in ATIS applications can further extract useful traffic pattern information to provide accurate recurring and nonrecurring traffic congestion information.

1.1.2. Travel Time Estimation/Prediction Accuracy

Most travelers, especially commuters, are constrained by their arrival time. The typical departure-time strategy for a commuter is to subtract the expected trip time from the required arrival time, plus an additional amount to account for the uncertainty of that expectation. Network instability at the point of congestion causes a large variability of operation and high uncertainty in a travel time expectation. Thus, many commuters are forced to significantly increase the time they allot to their commute in order to reliably arrive on time. Traditionally, transportation modelers and policy makers declare success if the commuter arrives early, whether or not the commuter can be productive with that extra time. By considering this potential lost productivity, the accuracy of the predicted trip time becomes as important to most travelers as its magnitude, and most travelers would accept a somewhat longer average trip time if it came with a guarantee.

1.1.3. Intelligent Information Provision and Diversion Strategies

An effective ATIS should provide (1) Pretrip time-dependent travel time information based on specific origins, destinations and departure times for both recurring and nonrecurring congestion conditions, and (2) En-route travel time updates based on real-time traffic, weather, work zone, and incident data. In its current implementation, travelers, as the ultimate consumers, have not significantly benefited from the existing ATIS. Specifically, travelers lack the means to understand and estimate the impact of nonrecurring congestion on their travel time. For example, web-based map services from Google, Yahoo, and MapQuest only provide static information on routing and average traffic time – they do not supply any travel time reliability-related information for

travelers to make better route or departure time decisions. Transportation agencies are very limited by how they inform and deliver effective travel time and reliability information to travelers, primarily relying upon posting travel time messages on dynamic message signs. However, existing travel time messages usually have "fixed" origins and destinations with limited reliability information.

1.2. Motivations and Challenges

Focusing on improving the mobility and reliability for ATIS systems, this dissertation will discuss the following three practically important and theoretically challenging questions in information-driven sensor network design, traffic estimation and prediction, and route guidance applications. A list of estimation and optimization targets for each problem is shown in Table 1.1, along with the expected products for three models.

Table 1.1: List of estimation, optimization targets and model products for three problems

	Estimation Component	**Optimization Component**	**Model Product**
Sensor Network Design	Uncertainty propagation, quantify value of information	Scenario-based sensor network optimization	Information-oriented sensor placement plan
Traffic Prediction	Traffic estimation with heterogeneous data sources	Prediction error minimization	Accurate traffic prediction under nonrecurring congestion
Route Guidance	Travel time reliability with spatial correlation	Solve nonadditive and nonlinear objective functions with Lagrangian relaxation method	Desirable route that minimizes (1) the mean and standard deviation of travel time, (2) absolute and percentile robust travel time

1.2.1. Challenges on Sensor Network Design

How to determine what sensor investments should be made, as well as when, how, where and with what technologies, in a transportation network design application?

Traffic monitoring systems provide fundamental data inputs for public agencies to measure time-varying traffic network flow patterns and accordingly generate coordinated control strategies. This dissertation will focus on a series of critical and challenging modeling issues in traffic sensor network design, in particular, how to locate different types of detectors to improve path travel time estimation accuracy, as reliable end-to-end trip travel time information is critically needed in ATIS applications.

While significant progress has been made in formulating and solving the sensor location problem for travel time estimation, a number of challenging theoretical and practical issues remain to be addressed.

First, the optimization criteria used in the existing sensor location models typically differ from those used in travel time estimation. Due to the inconsistency between two models, the potential of scarce sensor resources might not be fully achieved in terms of maximizing information gain for travel time estimation. For example, a sensor location plan that maximizes sensor coverage does not necessarily yield the least end-to-end travel time estimation uncertainty. As a result, a unified travel time estimation model for utilizing different data sources is required as the underlying building block for the sensor network design problem.

Second, most of the existing studies do not explicitly take into account various uncertainty sources in the travel time estimation process, e.g., prior travel time mean estimate and the corresponding errors in a historical travel time database, and sensor

measurement errors that depend on the size of samples collected and sensor quality. To seamlessly integrate diverse sources of measurements, a desirable travel time estimation framework should be able to recognize error sources associated with individual sensors, and possible error correlation between new and existing sensors.

1.2.2. Challenges on Traffic Estimation and Prediction

How to estimate and predict real-time travel time at system bottlenecks using multiple types of measurement?

Traffic delays are usually categorized as recurring congestion and nonrecurring congestion. Recurring congestion is caused by excessive regular traffic volumes or limited capacity at bottlenecks, thus the resulting traffic delay is well perceived and can be reasonably estimated by commuters traveling at the same locations at similar times. On the other hand, travelers might lack a clear understanding of the source, magnitude and frequency of nonrecurring congestion, which are associated with short-term and unexpected events, such as incidents, special events, severe weather conditions and work zones. This study will focus on estimating and predicting the travel time caused by nonrecurring congestion at highway bottlenecks and try to address the following theoretical and practical challenges.

First, to provide high quality travel time estimation and prediction, the proposed approach will utilize heterogeneous measurements from different traffic monitoring sources. Several types of measurements are currently available in practice, including traffic volume counts, point speed and lane occupancy from loop detectors, point-to-point travel time from AVI and probe sensors with limited market penetration rates. How to

fully leverage available data sources to improve the network-wide travel time estimation and prediction becomes an emerging research challenge.

Second, nonrecurring congestion may cause challenging problems to the traffic prediction. More specifically, a reliable travel time prediction relies on the knowledge of future network demand and supply. For nonrecurring cases, changes in traffic volume and road capacity from various congestion sources need to be considered in order to provide accurate traffic prediction.

Third, the quality of travel time prediction should be well managed and evaluated in the proposed method. As real-world measurements usually contain errors and are limited in their spatial coverage, the uncertainty propagation within a prediction model could highly affect the final confidence level of the forecasting results. To quantify and reduce the error propagation, parsimonious models with fewer parameter inputs are considered in this study.

To address the above research challenges, a travel time estimation/prediction approach based on simplified queueing models is proposed in this study. From a queueing theory perspective, the impacts of various nonrecurring congestion sources can be integrated into demand (as incoming flow) and capacity (as outgoing discharge rate). Moreover, with the travel time, counts and queue length can be connected through a cumulative flow count curve diagram, traffic measurements from heterogeneous data sources are fully utilized in the travel time estimation/prediction.

1.2.3. Challenges on Route Guidance

How to rapidly find the most reliable routes in a large-scale regional network?

Travel time reliability has been widely recognized as an important element of a traveler's route and departure time scheduling. In recent years, operating agencies have begun to shift their focus more toward monitoring and improving the reliability of transportation systems through probe-based data collection, integrated corridor management and advanced traveler information provision. With a growing trend of incorporating trip time variability into traffic network analysis models, finding reliable path alternatives motivates substantial algorithmic development efforts.

For many common route finding criteria, such as physical distance and travel time, the (generalized) path cost functions are linear and additive across different links, so the resulting optimization problem can be directly solved by the standard label correcting or label setting shortest path algorithms. In an early study by Sen et al. (2001), the path travel time reliability is modeled as a linear combination of travel time mean and variance, and the resulting 0-1 quadratic integer program is solved by as a sequence of parametric subproblems. However, most end-to-end trip reliability measures, as discussed below, lead to nonlinear and nonadditive cost functions, which considerably increase the complexity and impose challenges for the path search procedures. A wide range of definitions and formulations have been proposed to measure travel time reliability, including (1) 90th- or 95th-percentile travel time, buffer and planning time index, (2) on-time arrival probability, (3) travel time variation expressed in terms of standard deviation or coefficient of variation. Figure 1.1, adapted from a recent FHWA report (Cambridge Systematics, 2005), shows a distribution of travel times on eastbound State Route 520 in

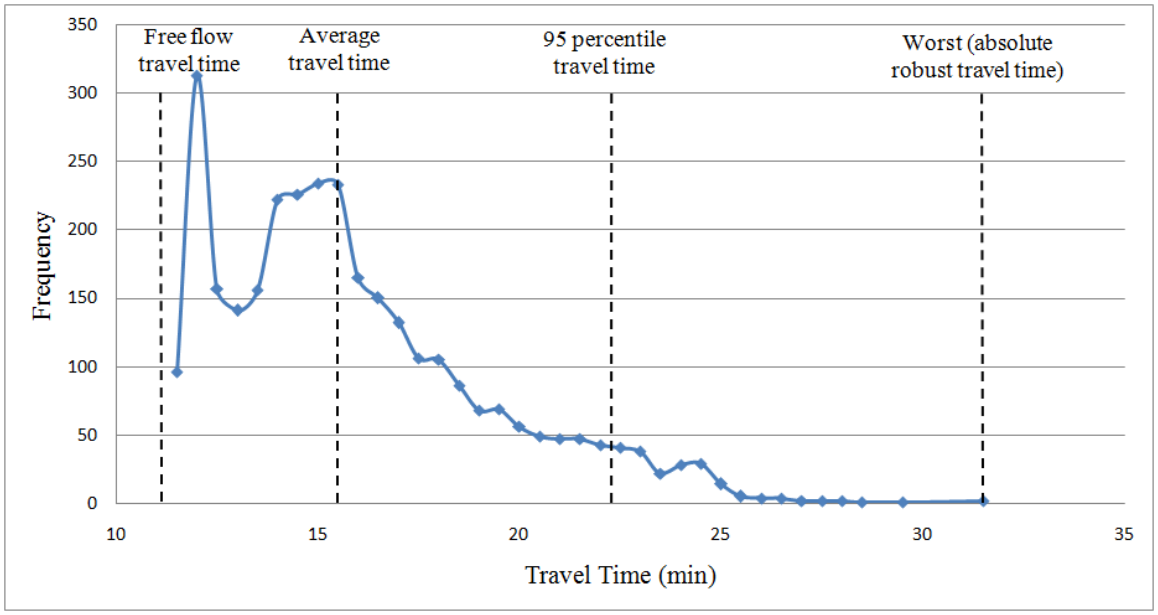


Figure 1.1: Travel time distribution for the eastbound lanes of State Route 520, Seattle, adapted from FHWA report (Cambridge Systematics, 2005)

Seattle, based on 3096 observation samples taken on weekdays between 4:00 to 7:00 pm. In the heavy-tailed travel time distribution along this 11.5-mile corridor, the mean and the standard deviation statistics (i.e., 15.9 min and 5.5 min) are insufficient to fully measure the extreme delay during the daily commutes, where contributing factors may include traffic crashes or severe weather conditions. In particular, the worst or absolute robust travel time is about 31.5 min (during the survey period of 4 months), while the 95% percentile travel time is around 22.5 min.

The first two definitions (1 and 2) are built on the probability distribution function of travel time. The absolute robust shortest path (ARSP) problem under consideration aims to find the path that minimizes the maximum path travel time over all samples. Similarly, the α-percentile robust shortest path (PRSP) problem is defined as the path that minimizes the travel time within which α-percentile trips in all samples are completed.

The on-time arrival probability measure, on the other hand, considers the percentage of trips that are completed within a reasonable buffered travel time (e.g., average travel time plus 20% buffer). In a study by Fan et al. (2005a), a path finding algorithm was proposed to minimize the probability of arriving at the destination later than a specified arrival time. Recently, Nie and Wu (2009a) developed solution algorithms with first-order stochastic dominance rules for the routing problem with on-time arrival reliability.

While ARSP emphasizes the extreme tail of the travel time distribution, PRSP is able to systematically balance the trade-off involving the overall reliability and low-probability events. The latter solution also provides a better statistical measure to avoid possible outliers in the real-world data sample set, and meets the needs for travelers with different degrees of risk-avoidance preferences. On the other hand, from a trip planning point of view, the PRSP problem highlights travel time guarantees over uncertain traffic situations, while the on-time arrival probability or travel time variation emphasizes the probability of later arrivals for a given preferred arrival time or a given buffer time index.

The third type of models, which characterizes the travel time reliability measure in terms of standard deviation, has been calibrated in various empirical studies (e.g., Small, 1982; Noland et al., 1998; Noland and Polak, 2002), and the corresponding utility function is also incorporated in dynamic traffic assignment models (e.g., Zhou et al., 2008). It is important to recognize that, within a Kalman filtering framework, which is the building block of real-time traffic state estimation and prediction systems (e.g., Ashok and Ben-Akiva, 1993; Zhou and Mahmassani, 2007), the variance of travel time estimates, and accordingly its standard deviation, can be analytically derived and calculated through a recursive estimation error propagation formula. In comparison, the

first two types of reliability measures must be assessed by relatively complicated numerical probabilistic methods.

1.3. Overview of Proposed Methods

To maintain the inherent consistency between the travel time estimation/prediction and its data measurement network, in Chapter 3, we jointly consider the sensor location problem with its underlying travel time estimation and prediction models. Expressly, by extending a Kalman filtering-based information theoretic approach proposed by Zhou and List (2010) for OD demand estimation applications, this research focuses on how to analyze the information gain for real-time travel time estimation and prediction problem with heterogeneous data sources. Since the classical information theory been proposed by Shannon (1948) on measuring information gain related to signal communications, the sensor location problem has been an important and active research area in the fields of electrical engineering and information science. Various measures have been used to quantify the value of sensor information in different sensor network applications, where the unknown system states (e.g., the position and velocity of targets studied by Hintz and McVey, 1991) can typically be directly measured by sensors. In comparison, sensing network-wide travel time patterns is difficult in its own right because point sensors only provides a partial coverage of the entire traffic state. Using AVI data involves complex spatial and temporal mapping from raw measurements, and AVL data are not always available on a fixed set of links, especially under an early sensor network deployment stage.

There are a wide range of time series-based methods for traffic state estimation, and many studies (e.g., Okutani and Stephanedes, 1984; Zhang and Rice, 2003; Stathopoulos and Karlaftis, 2003) have been devoted to travel time prediction using Kalman filtering and Bayesian learning approaches. To extract related statistics from complex spatial and temporal travel time correlations, a recent study by Fei et al. (2011) extends the structure state space model proposed by Zhou and Mahmassani (2007) to detect the structural deviations between the current and historical travel times and apply a polynomial trend filter to construct the transition matrix and predict future travel time. In this dissertation, we aim to present a unified Kalman filtering-based framework under both recurring and nonrecurring traffic conditions. More importantly, a spatial queue-based cumulative flow count diagram is introduced to derive the important transition matrix for modeling traffic evolution under nonrecurring congestions. Different from existing data-driven or time-series-based methods, this dissertation derives a series of point-queue-model-based analytical travel time transition equations, which lay out a core modeling building block for quantifying prediction uncertainty. In addition, a steady-state uncertainty formula is presented to fully capture day-to-day uncertainty evolution and convergence of the sensor network in a long-term horizon.

To allow further extensions in real-time traffic prediction and route guidance systems, two reliable path definitions are considered in this dissertation. In Chapter 4, this study considers the most reliable path problem with a linear disutility function of mean trip travel time and its standard deviation: $\min\left\{\text{mean}+\beta\sqrt{\text{var}}\right\}$. In particular, this research tries to address two fundamental challenges introduced by this special functional form. First, the standard deviation of path travel time is not a linear summation of the standard

deviation of related link travel times. Second, the square root transformation associated with the standard deviation term is, in fact, a concave function, so it is difficult to directly apply many convex programming techniques in this application. Based on the variable splitting approach in the Lagrangian reformulation framework, we first replace the complex quadratic portion of the objective function with equivalent equality constraint(s) to remove the nonadditivity, and the auxiliary constraint(s) can be further dualized to a simplified objective function that leads to easy subproblems. In particular, one integer subproblem involving linear link cost functions can be efficiently solved by standard shortest path algorithms, while another subproblem containing the concave square root objective function with a single variable can be solved analytically by checking the boundary values in the feasible region. The similar bounding technique was used by Larsson et al. (1994).

In Chapter 5, we will focus on the absolute robust shortest path and percentile robust shortest path problems. Using a sampling-based representation scheme, this research utilizes historical travel time records from multiple days of traffic measurements to capture day-by-day traffic dynamics and the complex spatial network correlations. Specifically, a scenario (corresponding to travel time samples on a day) is considered as a realization of random travel time distributions. In this research, we focus on how to efficiently find approximate solutions for the ARSP and PRSP problems, and a Lagrangian relaxation based algorithm is used to generate satisfactory feasible solutions and provide the corresponding quality evaluation on large-scale real-world networks. In particular, we adopt a variable splitting approach to reformulate the minimax objective function of the ARSP problem and the percentile definitional constraint of the PRSP

problem. The variable splitting approach was proposed by Joernsten and Naesberg (1986) and independently by Guignard and Kim (1987). To reformulate a complex objective function, auxiliary variables and additional constraints are introduced so that easy-to-solve subproblems can be constructed in a Lagrangian relaxation solution framework.

1.4. Organization of the Dissertation

This dissertation is organized as follows. Chapter 2 provides a comprehensive review and discussion on several topics related to sensor network design, traffic prediction and route guidance problems. In Chapter 3, a Kalman filtering based travel time estimation and prediction model is presented jointly with information measure models for the sensor location problem. Chapter 4 discusses the most reliable path problem using mean and standard deviation of path travel time as the disutility function. In particular, two different models are considered for the most reliable path problem: with and without link correlation. In Chapter 5, two models are proposed to evaluate the travel time robustness: absolute and α-percentile robust shortest path problems. Numerical experiments and results are presented following the methodologies and algorithms proposed in each of Chapters 3-5. Conclusions for all the proposed models under the integrated estimation-optimization approach are discussed in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews several topics relevant to system estimation and mobility optimization of the Advanced Traveler Information Systems. In Section 2.1, traffic surveillance and observation technologies are reviewed along with sensor location researches. Section 2.2 reviews major travel time estimation and prediction models. Section 2.3 provides a comprehensive review on the reliability and robustness related routing problems.

2.1. Data Collection through Sensors

Essentially, any application of real-time traffic measurements for supporting Advanced Traveler Information Systems and Advanced Traffic Management Systems (ATMS) functionalities involves the estimation and/or prediction of traffic states. Depending on underlying traffic process assumptions, the existing traffic state estimation and prediction models can be classified into three major approaches. (1) Approach purely based on statistical methods, focusing on travel time forecasting; (2) Approach based on macroscopic traffic flow models, focusing on traffic flow estimation on successive segments of a freeway corridor; (3) Approach based on dynamic traffic assignment

models, focusing on wide-area estimation of origin-destination trip demand and route choice probabilities so as to predict traffic network flow patterns for links with and without sensors. In this research, the researchers are interested in how to place different types of sensors to improve information gains for the first statistical method-based travel time estimation applications.

In sensor location models for the second approach, significant attention (e.g., Liu and Danczyk (2009), Danczyk and Liu (2010), and Leow et al. (2008)) has been devoted to placing point detectors along a freeway corridor to minimize the traffic measurement errors of critical traffic state variables, such as segment density and flow. The traffic origin-destination (OD) matrix estimation problem is also closely related to the travel time estimation problem under consideration. To determine the priority of point detector locations, there are a wide range of selection criteria, to name a few, "traffic flow volume" and "OD coverage" criteria proposed by Lam and Lo (1990), a "maximum possible relative error (MPRE)" criterion proposed by Yang et al. (1991) that aims calculate the greatest possible deviation from an estimated demand table to the unknown true OD trip demand.

Recently, based on the trace of the a posteriori covariance matrix produced in a Kalman filtering model, Zhou and List (2010) proposed an information-theoretic framework for locating fixed sensors in the traffic OD demand estimation problem. Related studies along this line include an early attempt by Eisenman et al. (2006) that uses a Kalman filtering model to minimize the total demand estimation error in a dynamic traffic simulator, and a recent study by Xiang and Mahmassani (2010) that considers additional criteria, such as OD demand coverage, within a multiobjective

decision making structure. This research will adapt and extend the information-theoretic framework from Zhou and List (2010) and further propose traffic measurement models and information quantification models specifically for path travel time estimation applications with heterogeneous data sources.

Chen et al. (2004) studied the AVI reader location problem for both travel time and OD estimation applications. They presented the following three location section criteria: minimizing the number of AVI readers, maximizing the coverage of OD pairs, and maximizing the number of AVI readings. To maximize the information captured with regard to the network traffic conditions under budget constraints, Lu et al. (2006) formulated the roadside servers locating problem as a two stage problem. The first stage was a sensitivity analysis to identify a subset of links on which the flows have large variability of travel demand, and more links were gradually selected to maximize the overall sensor network coverage in the second stage. Sherali et al. (2006) proposed a discrete optimization approach for locating AVI readers to estimate corridor travel times. They used a quadratic zero-one optimization model to capture travel time variability along specified trips.

This research adopts a Kalman filtering-based information theoretic approach to qualify the information gain for different sensor placement scenarios. The classical information theory proposed by Shannon (1948) aims to measure information gain related to signal communications. The sensor location problem is an important and very active research area in the fields of electrical engineering and information science. Various measures have been used to quantify the value of sensor information in different sensor network applications, and the unknown system states (e.g., the position and

velocity of targets studied by Hintz, 1991) typically can be directly measured by sensors. In comparison, sensing network-wide travel time patterns is difficult in its own right because point sensors only provides a partial coverage of the entire traffic state. Using AVI data involves complex spatial and temporal mapping from raw measurements, and AVL data are not always available on a fixed set of links, especially under an early sensor network deployment stage.

2.2. Travel Time Estimation and Prediction

Travel time estimation and prediction problems have been extensively studied in past decades. A variety of models have been proposed and developed with different theoretical foundations. Most of the travel time prediction models fall into one of the following groups: (1) time-series methods, e.g., Auto-Regressive Integrated Moving Average (ARIMA) models (Box and Jenkins, 1970), (2) state space methods, e.g., Kalman Filtering (KF) technique, which is first applied in traffic volume prediction by Okutani and Stephanedes (1984), (3) nonparametric methods such as K Nearest Neighbor (KNN) (Davis and Nihan, 1991) and Neural Network (Clark et al., 1993) approaches, and (4) traffic flow based methods.

In traffic flow based estimation/prediction models, several methods have been widely used: cell transmission model (CTM), real-time simulation-based Dynamic Traffic Assignment (DTA) based model and Newell's model. The cell transmission model, proposed by Daganzo (1994), decomposed a road corridor into multiple cells and estimate/predict traffic using density, flow of each cell and boundary conditions. Unfortunately, the numerical implementations of CTM in current real-time traffic

estimation/prediction applications are not consistent to the theoretical derivation of the shock wave propagation behavior. Real-time DTA based models consider user rerouting behaviors and can capture system-wide travel time estimation. However, the DTA based models have too many network-wide parameters to be estimated (such as dynamic origin-destination demand matrix), which are difficult for decision support systems to maintain consistency between simulated states and reality.

Nonrecurring congestion has been well known as one of the key factors influencing travel time reliability (Cambridge Systematics, Inc, et al., 2003). The impacts of various nonrecurring congestion sources have been decomposed and studied by Kwon et al. (2010). This study will extend Newell's model to construct a real-time travel time estimation and prediction algorithm under nonrecurring congestion using heterogeneous data sources.

In 1993, Newell (1993a, b, and c) proposed a simplified theory based on the classical traffic wave theory. In this model he used the cumulative count curves instead of flows for most of the calculations and a triangular flow-density relation to describe traffic flow (i.e., forward wave and backward wave) propagation. Newell's model has been tested on freeway segments by Hurdle and Son (2000) and demonstrated the model's computational efficiency and prediction results on severely congested cases. Along this line, a more complicated, but also more general, two-detector problem has been studied by Daganzo (2001).

2.3. Variability-oriented Routing Solution

Several previous pioneering research efforts have been devoted to addressing computational issues caused by nonlinear, nonadditive or concave objective functions in the shortest path problem. The early work by Henig (1986) presented efficient approximate methods on the shortest path problem with two criteria, which are assumed to be quasiconcave or quasiconvex utility functions. Scott and Bernstein (1997) developed an iterative solution method for the shortest path problem where the value of time function is nonlinear and nondecreasing. In their algorithm, the search space is decomposed to a series of resource-constrained shortest path subproblems, which can be solved by the Lagrangian relaxation technique (Handler and Zang, 1980). Gabriel and Bernstein (2000) further proposed a path-finding heuristic algorithm for the nonadditive path problem using a linear approximation reformulation.

By dualizing hard constraints to the objective function (Fisher, 1981), the Lagrangian relaxation method is a well-known solution procedure for integer programming problems. To further introduce separability in Lagrangian reformulations, one important extension of Lagrangian relaxation is the variable splitting and Lagrangian decomposition approach proposed by Joernsten and Naesberg (1986) and independently by Guignard and Kim (1987). This approach aims to split original variable x into the pair (x, y), and then link the auxiliary variable y with x through a linking constraint Ax=y, which will be further relaxed to the objective function. Larsson et al. (1994) adapted this problem restatement approach to decompose a minimum cost network flow problem with a concave objective function to a standard linear minimum cost network flow subproblem and an easy-to-solve concave minimization problem. Along the same line, Tsaggouris and Zaroliagis

(2004) combined the Lagrangian relaxation and hull approach to solve the nonadditive shortest path problem with a nonlinear, convex and nondecreasing cost function. Readers interested in general Lagrangian relaxation methods are referred to the review paper written by Guignard (2003).

In stochastic routing problems, spatial and temporal dependences have been exclusively studied by a number of researchers, and interested readers are referred to the comprehensive studies by Miller-Hooks and Mahmassani (2000) and Nie and Wu (2009a) on the a priori time-varying least travel time problem. Considering spatial dependence in terms of congestion level and state transfer probability, Fan et al. (2005b) proposed a multistage adaptive feedback control process to address shortest path problem with correlated link costs. Recently, limited spatial and temporal dependences have been considered by Boyles and Waller (2007) for the nonlinear disutility shortest path problem, and by Nie and Wu (2009b) for the reliable routing problem. Specifically, the above studies characterize the randomness of link travel time by using certain probability density functions abstracted from a historical database, and incorporate limited spatial correlation through a Markovian model that considers the transition probabilities of link states.

The robust shortest path problem and its variants have been extensively studied in the last few decades. Murty and Her (1992) proposed a relaxation based label-correcting procedure to provide exact solutions for the ARSP problem. Specifically, two pruning techniques, namely one-row and Lagrangian-based relaxation, were used to improve the algorithm efficiency. Their approach was later enhanced by Bruni and Guerriero (2010) by using heuristic rules and evaluation functions to better guide the solution search

procedure. Yu and Yang (1998) studied both ARSP and robust deviation shortest path (RDSP) problems by using a set of scenarios to capture the uncertainty of travel time. They first proved that both ARSP and RDSP problems are NP-complete under limited scenarios and NP-hard for an unbounded number of scenarios, and then proposed dynamic programming algorithms with a pseudo-polynomial computational time and a few heuristic methods. Mainly focusing on the RDSP problem, Karasan et al. (2001) proposed a simple ARSP approximation algorithm by setting each link travel time to its upper bound over all scenarios/samples. Montemanni and Gambardella (2008) developed two algorithms for the ARSP problem, namely a Benders decomposition-based algorithm and a solution method by generating duality reformulation and solving through mixed integer linear programming techniques.

There are also a number of other definitions related to robust shortest paths. For example, Yu and Yang (1998) considered the robust deviation shortest path problem that minimizes the maximum deviation of the path length from the optimal path length of the corresponding scenario, and Sigal et al. (1980) suggested using the probability of being the shortest path as an optimality index.

CHAPTER 3

HETEROGENEOUS SENSOR NETWORK DESIGN FOR ESTIMATING AND PREDICTING PATH TRAVEL TIME DYNAMICS

This chapter proposes an information-theoretic sensor location model that aims to minimize information uncertainty from a set of point, point-to-point and probe sensors in a traffic network. Based on a Kalman filtering structure, the proposed measurement and information quantification models explicitly take into account several important sources of errors in the travel time estimation/prediction process, such as the uncertainty associated with prior travel time estimates, measurement errors and sampling errors. After thoroughly examining a number of possible measures of information gain, this dissertation selects a path travel time prediction uncertainty criterion to construct a joint sensor location and travel time estimation/prediction framework. The remainder of this chapter is organized as follows. The overall framework and notation are described in Section 3.1. In Sections 3.2 and 3.3, a Kalman filtering based travel time estimation and prediction model is presented for both recurring and nonrecurring traffic conditions. A comprehensive discussion of information measure models is presented in Section 3.4. Section 3.5 describes the beam-search based sensor design model and solution algorithms.

Finally, the proposed model is further extended to some complex cases considering AVI and AVL sensors in Section 3.6, followed by numerical experiment results on a test network shown in Section 3.7.

3.1. Notation and Modeling Framework Overview

We first introduce the notation used in the travel time prediction and sensor network design problems.

3.1.1. Notation and Problem Statement

Sets and Subscripts:

N = set of nodes.

A = set of links.

m = number of links in set A.

A' = set of links with point sensors (e.g., loop detectors), $A' \subseteq A$.

N'' = set of nodes with point-to-point sensors, $N'' \subseteq N$.

A''' = set of links with reliable probe sensor data, $A''' \subseteq A$.

$\overline{A}'$ = sets of links that have been equipped with point sensors, $\overline{A}' \subseteq A'$.

$\overline{N}''$ = sets of nodes that have been equipped with AVI sensors, $\overline{N}' \subseteq N'$.

n', n'', n''' = numbers of measurements, respectively, from point sensors, point-to-point sensors and probe sensors.

n = number of total measurements, $n = n' + n'' + n'''$.

t = time index for state variables.

h = travel time prediction horizon.

d = subscript for day index.

o = subscript for origin index, $o \in O$, O = set of origin zones.

s = subscript for destination index, $s \in S$, S = set of destination zones.

a,b = subscript for link index, $a,b \in A$.

i, j = subscript for node index, $i,j \in N$.

k, λ = subscript for path.

$p(i,j,k)$ = set of links belong to path k from node i to node j.

Estimation variables:

$t_{d,a}$ = travel time of link a on day d.

$t_{d,o,s,k}$ = travel time on path k from origin o to destination s, on day d, $t_{d,o,s,k} = \sum_{a \in p(o,s,k)} t_{d,a}$.

Measurements:

$y'_{d,a}$ = single travel time measurement from a point sensor on link a, on day d.

$y''_{d,i,j,k}$ = single travel time measurement from a pair of AVI readers on path k and day d from node i to node j, where the first and second AVI sensors are located at nodes i and node j, respectively.

$y'''_{d,a}$ = a set of travel time measurements from a probe sensor that contain map-matched travel time records on links a on path k and day d from node i to node j, where $a \in p(i,j,k)$.

Vector and matrix forms in Kalman filtering framework:

Y_d = sensor measurement vector on day d, consisting of n elements.

T_d = travel time vector on day d, consisting of m elements $t_{d,a}$.

T_d^- = *a priori* estimate of the mean values in the travel time vector on day d, consisting of m elements.

T_d^+ = *a posteriori* estimate of the mean values in the travel time vector on day d, consisting of m elements.

T_d^h = historical regular travel time estimates using data up to day d.

V_d = structural deviation on day d.

P_d^- = *a priori* variance covariance matrix of travel time estimate, consisting of ($m \times m$) elements.

P_d^+ = *a posteriori* error covariance matrix, i.e., conditional covariance matrix of estimation errors after including measurements.

Σ^- = *a priori* variance covariance matrix of structure deviation, consisting of ($m \times m$) elements.

Σ^+ = *a posteriori* variance covariance matrix of structure deviation.

$\overline{T}$ = vector of regular historical mean travel time estimates, consisting of m elements, $\overline{T} = T_0^h$.

$\overline{P}$ = error covariance matrix of historical travel time estimate, consisting of ($m \times m$) elements, $\overline{P} = P_0^-$.

H_d = sensor matrix that maps unknown travel times T_d to measurements Y_d, consisting of ($n \times m$) elements.

K_d = updating gain matrix, consisting of ($n \times m$) elements, on day d.

K_d^{NR} = updating matrix for nonrecurring traffic estimations on day *d*.

$L_d(t, t+h)$ = nonrecurring traffic transition matrix from time *t* to *t*+*h* on day *d*.

w_d = system evolution noise vector for link travel times, $w_d \sim N(0, Q_d)$.

Q_d = system evolution noise variance-covariance matrix, on day *d*.

μ_d = nonrecurring derivation evolution noise vector for link travel times, $\mu_d \sim N(0, Q_d^{NR})$.

Q_d^{NR} = nonrecurring derivation evolution noise variance-covariance matrix, on day *d*.

$q_{d,a}$ = systematic travel time variance on link *a*.

ε_d = combined measurement error term, $\varepsilon_d \sim N(0, R_d)$, on day *d*.

R_d = variance-covariance matrix for measurement errors, on day *d*.

Parameters and variables used in measurement and sensor design models:

$\phi_{i,j,k,a}$ = path-link incidence coefficient, $\phi_{i,j,k,a}$ =1 if path *k* from node *i* to node *j* passes through link *a*, and 0 otherwise.

$\tilde{\gamma}'''_{d,a}$ = stochastic link traversing coefficient for GPS probe vehicles, $\tilde{\gamma}'''_{d,a} = 1$ if GPS probe vehicles pass through link *a* on day *d*, and 0 otherwise.

$e_{d,o,s,k}$ = path travel time estimation error on path *k* from origin *o* to destination *s*.

$f_{o,s,k}$ = traffic flow volume on path *k* from origin *o* to destination *s*.

TU_d = total path travel time estimation uncertainty on day *d*.

α = market penetration rate for vehicles equipped with AVI sensors/tags.

β = market penetration rate for vehicles equipped with AVL sensors.

l = subscript of sensor design solution index.

X_l = l^{th} sensor design solution, represented by $X_l = [A', N'', A''', \alpha, \beta]$.

X^* = optimal sensor design solution.

$z(X_l)$ = overall information gain (i.e., performance function) for a given sensor design scenario X_l.

Consider a traffic network with multiple origins $o \in O$ and destinations $s \in S$, as well as a set of nodes connected by a set of directed links. We assume the following input data are available:

(1)The prior information on historical travel time estimates, including a vector of historical mean travel time estimates $\bar{T}$ and the corresponding variance-covariance matrix $\bar{P}$.

(2)The link sets with point sensor and point-to-point AVI sensor data, specified by $\bar{A}'$ and $\bar{N}''$.

(3)Estimated market penetration rate α for point-to-point AVI sensors.

(4)Estimated market penetration rate β for probe sensors, and set of links with accurate probe data A'''.

The sensor network to be designed and deployed will include additional point sensors and point-to-point detectors that lead to sensor location sets of A' and N'', where $\bar{A}' \subseteq A'$, $\bar{N}'' \subseteq N''$. In the new sensor network, through GPS map-matching algorithms, GPS probe data can be converted from raw longitude/latitude location readings to link travel time records on a set of links A'''. In this study, we assume that probe data will be available through a certain data sharing program (e.g., Herrera and Bayen, 2010), from vehicles equipped with Internet-connected GPS navigation systems or GPS-enabled

mobile phones. It should be noticed that, depending on the underlying map-matching algorithm and data collection mechanism, only a subset of links in a network, denoted by A''' can produce reliable GPS map-matching results. For example, it is very difficult to distinguish driving vs. walking mode on arterial streets through data from GPS-equipped mobile phones, so typically only travel time estimates on freeway links are considered to be reliable in this case.

One of the key assumptions in our study is that the historical travel time information can be characterized by the *a priori* mean vector $\bar{T}$ and the estimation error variance matrix $\bar{P}$. If point sensor or point-to-point data are available from sets $\bar{A}'$ and $\bar{N}''$, then we can construct the mean travel time vector $\bar{T}$, and estimate the variance of estimates in the diagonal elements of corresponding variance-covariance matrix $\bar{P}$. For links without historical sensor measurements, the travel time mean estimate can be approximated by using national or regional travel time index (e.g., 1.2) and set the corresponding variance to a sufficient large value or infinity. One can assume zero for the correlation of initial travel time estimates. In the case of a complete lack of historical demand information, we can set $(\bar{P})^{-1} = 0$.

It should be remarked that, measurements from a point sensor are typically instantaneous speed values observed at the exact location of the detector. Using a section-level travel time modeling framework (e.g., Lindveld et al., 2000), a homogenous physical link can be decomposed into multiple cells or sections, with the speed measurement directly reflecting only the section where the sensor is located. In some previous studies, the link or corridor speed can be estimated using the section based speed, while cells without sensors using approximated values from adjacent instrumented

sections. As shown in Figure 3.1, the travel times on section A and D are directly measured using sensors 1 and 2, respectively. Meanwhile, the travel times for sections B, C and E, as well as the entire corridor, are estimated using upstream and downstream sensors. There are a number of travel time reconstruction approaches, such as constant speed based methods and trajectory methods (Van Lint and Van der Zijpp, 2003).

In this study, for sections without point sensors, the above mentioned approximation error is modeled as prior estimation errors, which can be obtained through a historical travel time database by considering other related links such as adjacent links or links with similar characteristics. Furthermore, our proposed framework can be also easily generalized to a section-based representation scheme, where a section in Figure 3.1 can be viewed as a link in our link-to-path-oriented modeling structure.

3.1.2. Generic State Transition and Measurement Models

By adapting a structure state model for dynamic OD demand estimation by Zhou and Mahmassani (2007), this study decomposes a true travel time pattern into three modeling components:

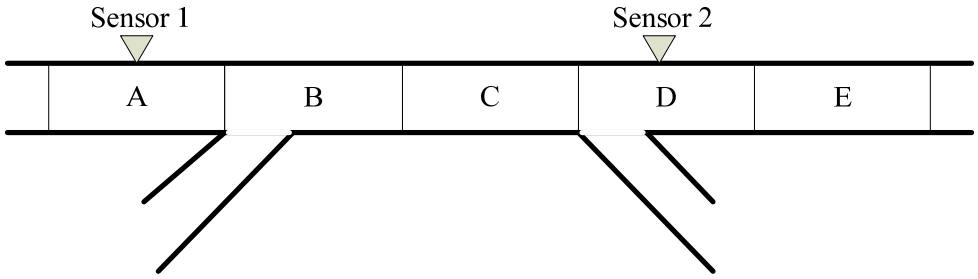


Figure 3.1: Section-level travel time estimation

true travel time = regular recurring pattern + structural deviations + random fluctuations.

Under this assumption, travel time estimation/prediction can be studied in two categories: recurring traffic conditions and nonrecurring conditions. For travel time prediction under recurring conditions, structural deviation is considered as zero, and the regular travel time patterns/profiles can be constructed based on historical data for recurring traffic. On the other hand, for travel time prediction under nonrecurring conditions, the structural deviation is further modeled in this study as a function of time-dependent capacity and time-dependent demand. Without loss of generality, this dissertation mainly focuses on time-dependent capacity reductions due to incidents, a major source of nonrecurring congestion.

To further jointly consider both recurring and nonrecurring traffic conditions in the sensor location problem, the overall system uncertainty under a certain sensor design is modeled as a probabilistic combination of recurring and nonrecurring uncertainty measures:

overall system prediction uncertainty =

$(1-\rho_{NR})\times$ uncertainty under recurring conditions + $\rho_{NR}\times$ uncertainty under nonrecurring conditions,

where ρ_{NR} represents the given probability of nonrecurring events.

The state transition model of the travel time is written as

$$T_d(t) = T_d^h(t) + V_d(t) + w_d, \quad w_d \sim N(0, Q_d) \tag{3.1}$$

In Eq. (3.1), the travel time for each link is represented as a combination of three components: regular pattern, structural deviation and random fluctuation. The regular pattern $T_d^h(t)$ is the time-dependent historical travel time average which is determined by the day-to-day regular traffic demand and capacity. For nonrecurring traffic conditions, a structural deviation $V_d(t)$ exists due to nonrecurring congestion sources such as incidents, work zones and severe weathers. Considering a stationary congestion pattern, this study assumes that w_d follows a normal distribution with zero-mean and a variance-covariance matrix Q_d. Q_d corresponds to random travel time variation magnitude, which is further determined by dynamics and stochasticity in the underlying traffic demand and road capacity supply. For example, the travel time variations are more significant on a congested freeway link with close-to-capacity demand flow volume, compared to a rural highway segment with low traffic volume and sufficient capacity where the speed limit could yield a good estimate most of the time. More specifically, $q_{d,a}$, the (diagonal) variance elements of the matrix Q_d, exhibit the travel time variability/uncertainty of each individual link, while the covariance elements should reveal the spatial correlation relationship (mostly due to queue spillbacks) between adjacent links in a network. We refer readers to a study by Min and Wynter (2011) for calibrating spatial correlations of link travel times.

In order to estimate the regular pattern $T_d^h(t)$ and structural deviation $V_d(t)$, a linear measurement model is constructed as:

$$Y_d(t) = H_d(t) \times T_d(t) + \varepsilon_d \text{, where } \varepsilon_d \sim N(0, R_d) \tag{3.2}$$

With the measurement model in Eq. (3.2), the travel times are estimated using the latest measurements $Y_d(t)$. Measurement vector Y_d is composed of travel time observations from point sensors, point-to-point sensors and probe sensors. The mapping matrix H_d, with ($n \times m$) elements, connects unknown link travel time T_d to measurement data Y_d. Particularly, each row in the mapping matrix H_d corresponds to a measurement and each column corresponds to a physical link in the network. For an element at u^{th} row and v^{th} column of the matrix H_d, a value of 1 indicates that the uth measurement covers or includes the travel time on the vth link of the network, otherwise it is 0. With the measurement equation, the historical recurring travel time pattern is then updated through the Kalman filtering process. A detailed discussion on the mapping matrix H and the measurement error term is provided in Section 3.2.

3.1.3. Uncertainty Analysis under Recurring and Nonrecurring Conditions

We now focus on the conceptual analysis of the uncertainty reduction and propagation. By assuming independence between different components in the structure state model (3.1), the total variance of the predicted travel time can be obtained by

$$\text{var}\left(T_d(t)\right) = \text{var}\left(T_d^h(t)\right) + \text{var}\left(V_d(t)\right) + Q_d \tag{3.3}$$

Under recurring congestion conditions, the structural deviation $V_d(t) = 0$, which leads to

$$\mathrm{var}\left(T_d(t)\right) = \mathrm{var}\left(T_d^h(t)\right) + Q_d \quad (3.4)$$

To reduce prediction error under recurring conditions (e.g., at the beginning of each day d for pretrip routing applications), we need to reduce the variance of the historical travel time estimates, $\mathrm{var}\left(T_d^h(t)\right)$, while the variance of inherent traffic process noises Q_d (due to traffic demand and supply variations) cannot be reduced and sets a limit for travel time prediction accuracy. Along this line, this article will first focus on updating the historical travel time pattern and the uncertainty reduction due to added sensors under recurring conditions. Under nonrecurring conditions, in addition to the above mentioned uncertainty elements $\mathrm{var}\left(T_d^h(t)\right)$ and Q_d, the total prediction variance is mainly determined by the structural travel time deviation $V_d(t)$. Through a simplified spatial queue model, a detailed discussion is provided in Section 3.3, and we will focus on capacity reductions due to incidents.

3.1.4. Conceptual Framework and Data Flow

Focusing on predicting end-to-end path travel time applications and considering future availability of GPS probe data on links A''', the goal of the sensor location problem is to maximize the overall information gain $X^* = \arg\min_l z(X_l)$ by locating point and point-to-point sensors in sets A' and N'', subject to budget constraints for installation and

maintenance. To systematically present our key modeling components in the proposed sensor design model, we will sequentially describe the following three modules.

3.1.4.1. Link travel time estimation and prediction module

Given prior travel time information T_d^- and P_d^-, with traffic measurement vector Y_d that includes $y'_{d,a}$, $y''_{d,i,j,k}$ *and* $y'''_{d,a}$ from sensor location sets A', N'' and A''', the link travel time estimation and prediction module seeks to update current link travel times T_d^+ and their variance-covariance matrix P_d^+.

3.1.4.2. Information quantification module

With prior knowledge on the link travel time estimates $\bar{T}$ and $\bar{P}$, the information quantification module aims to find the single-valued information gain $z(X_l)$ for the critical path travel times for a sensor design scenario X_l, represented by location sets A', N'', A''', as well as AVI and AVL market penetration rates α and β.

3.1.4.3. Sensor network design module

The sensor design module aims to find the optimal solution $X^* = \arg\min_l z(X_l)$, subject to budget constraints for installation and maintenance. For each candidate solution X_l, this module needs to call the information quantification module to calculate $z(X_l)$. The optimal solution X^* produces optimal location sets A' and N'', for a predicted AVI and AVL market penetration rates α and β, and predicted location set A''' with reliable travel time map-match results.

Figure 3.2 illustrates the conceptual framework and data flow for the proposed modules. From sensor network design plans in block 1, we need to extract three groups of critical input parameters: AVI/AVL market penetration rates α and β at block 2, measurement error variance-covariance R in block 3, and sensor location mapping matrix H in block 4. Location mapping matrix H is derived from the sensor location sets A', N'' and A'''.

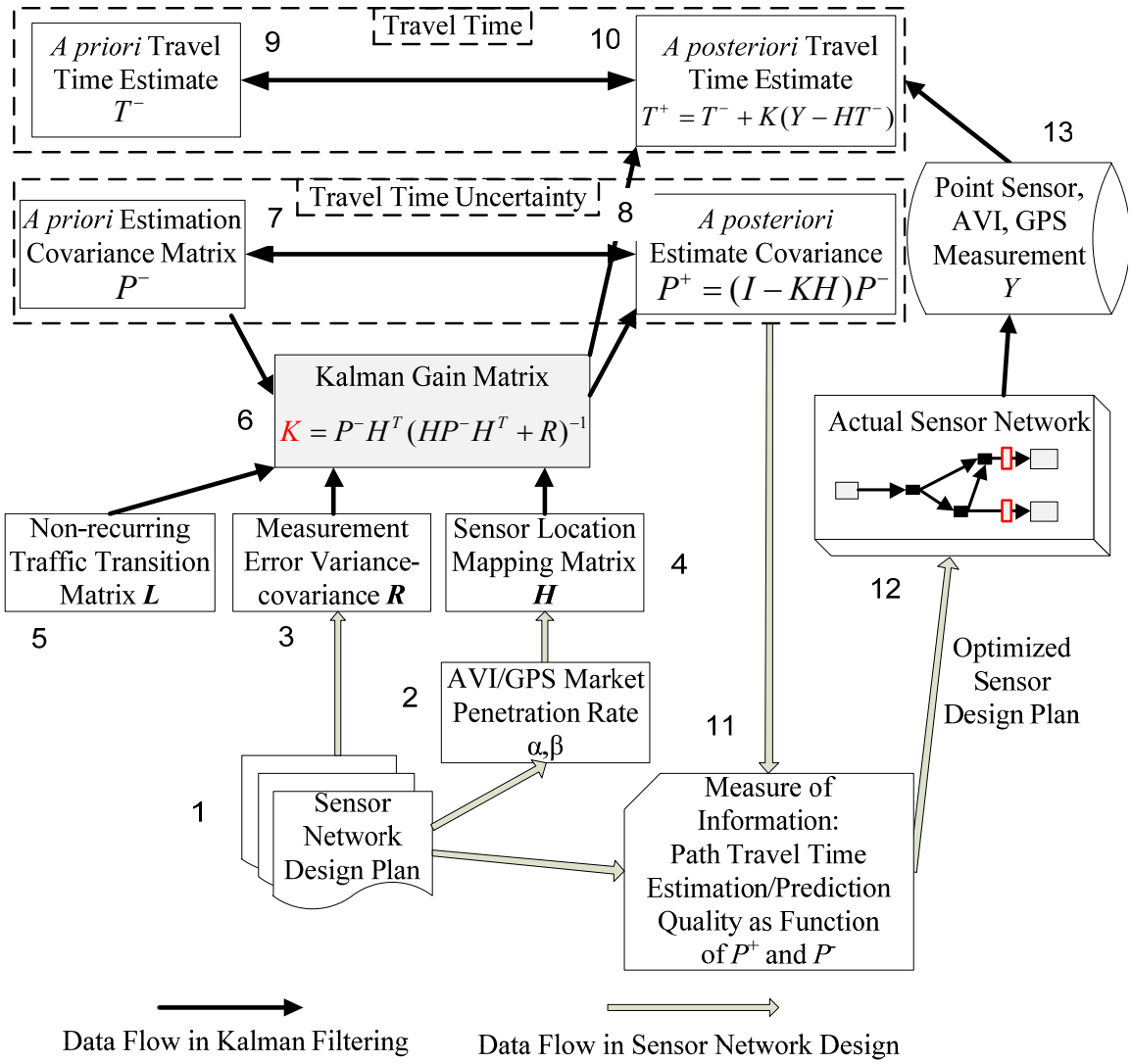


Figure 3.2: Conceptual framework and data flow

The link travel time estimation module uses a Kalman filtering model to iteratively update the travel time (blocks 9 and 10) and the corresponding error variance matrix (blocks 7 and 8), where the critical Kalman gain matrix K, calculated in block 6, is applied to the above two mean and variance propagation processes. Based on the estimation or prediction error variance statistics in blocks 7 and 8, the information quantification module derives the measure of information in block 11 by representing the path travel time estimation/prediction quality as a function of P^+ and P^-. By minimizing the network-wide path travel time estimation uncertainty, the sensor network design module finally selects and implements an optimized sensor plan so that point sensor, AVI, and AVL measurement data in block 13 can be produced from the actual sensor network illustrated by block 12.

One of the key features offered by the Kalman Filtering model is that although updating the travel time mean estimates from T^- in block 9 to T^+ in block 10 requires sensor measurements Y, the uncertainty propagation calculation from block 7 to 8 (i.e., updating P^+ from P^-) does not rely on the actual sensor data, as the uncertainty reduction formula in block 8 is a function of three major inputs: *a priori* uncertainty matrix P^-, measurement error range R, and sensor mapping matrix H. In other words, if a transportation analyst can reasonably prepare the above three input parameters, then he/she can apply the proposed analytical model to compute the information gain for a sensor design scenario and further assist the decision-maker to determine where and with what technologies sensor investments should be made in a traffic network.

3.2. System Process and Measurement Models for Estimating Historical Regular Patterns

This section first introduces the travel time estimation and prediction model as the building block for the (upper-level) sensor design model. In particular, we want to highlight how a classical Kalman filtering model can be used to estimate link travel time using data from AVI and AVL sources, and further used to analytically estimate the uncertainty propagation associated with the travel time mean estimates. Additionally, the discussion focuses on how sensor mapping matrix H and measurement error matrix R should be constructed, as they form the basis for the proposed information measuring model. There are two essential sets of equations within a Kalman filtering structure: stochastic process model, detailed in Section 3.2.1, and measurement model, described in Section 3.2.2.

3.2.1. Process Model of Day-varying Traffic System under Recurring Conditions

In this study, link travel times are characterized as random variables through stochastic linear process models. Two modeling approaches are available to capture travel time variations with different settings of time horizon and resolution: within-day dynamic and day-to-day dynamic. Specifically, the within-day model estimates current travel time based on the travel time at the previous time interval(s) on the same day, with a typical time resolution of 5 or 15 min. Without loss of generality, this study ignores the time-dependent travel time dimension in the estimation equation below, and will discuss the time-dependent state transition equation in Section 3.3.

3.2.2. Measurement Model

Shown in Eq. (3.2), a linear measurement model is used to map the measurement vector Y_d to the travel time vector T_d (as state variables) by taking into account measurement error term ε_d from variant sources.

The following three equations show how H_d is constructed for a specific type of measurements on each day d.

For a point sensor on link a,

$$y'_{d,a} = t_a + \varepsilon'_{d,a}, \forall a \in A'. \tag{3.5}$$

For a pair of point-to-point sensors that capture end-to-end travel time from node i to node j through path k,

$$y''_{d,i,j,k} = \sum_a \left(\phi_{i,j,k,a} t_a \right) + \varepsilon''_{d,i,j,k}, \forall i, j \in N''. \tag{3.6}$$

For an AVL sensor/probe on link a,

$$y'''_{d,a} = \tilde{\gamma}'''_{d,a} \times t_a + \varepsilon'''_{d,a}, \forall a \in A'''. \tag{3.7}$$

A measurement in this model might be referred to an average value of multiple raw samples within a certain time period (e.g., from 8:00 AM to 8:15AM). The error term ε

in the above equations is a combined error term that reflects the overall effect of errors from the data conversion, measurement reading and sampling processes.

Data conversion error: Typically, only time-mean speed data are available from a point sensor, and the travel time value (i.e., space-mean speed) needs to be inferred and approximated from a point speed reading. This introduces significant data conversion errors, depending on the placement of a point sensor on a link (e.g., the relative location with respect to the tail of a queue from the downstream node of a link). In addition, a single-loop detector has to use the observed occupancy and flow counts to calculate the point speed value, which leads to sensor measurement errors. GPS location data (in terms of longitude, latitude, point speed, bearing and timestamps) need to be map-matched to specific links in the study network to estimate corresponding link travel times. This map-matching process again brings data conversion errors to the final link travel time estimates.

Sensor reading error: Point sensors that are not carefully calibrated are more likely to generate large measurement noises. The detection rates of AVI readers are relatively low when vehicle tags are not powered by batteries. The data quality of GPS location readings depends on the number of satellites that a GPS receiver can find.

Sampling error: As a measurement can come from multiple readings, the variance of sampling error, e.g., for an AVI measurement that is aggregated from $g''_{d,i,j,k}$ samples, can be described as

$$\mathrm{var}(\varepsilon''_{d,i,j,k}) = \frac{\mathrm{var}(T_{d,i,j,k})}{g''_{d,i,j,k}} \tag{3.8}$$

If we assume there is no correlation between link travel times along path k from node i to node j, then

$$\mathrm{var}(T_{d,i,j,k}) = \sum_{a \in p(i,j,k)} \mathrm{var}(T_{d,a}) = \sum_{a \in p(i,j,k)} q_{d,a} \quad . \tag{3.9}$$

Assuming there are a total of $f_{d,i,j,k}$ vehicles traveling from node i to node j through path k, then the AVI market penetration rate α can be derived as $a = g''_{d,i,j,k} / f_{d,i,j,k}$. That is, when the market penetration rate increases, the size of samples also becomes larger, leading to a smaller sampling error and a more reliable travel time measurement.

In summary, the magnitude of the combined error ε_d is determined by a number of external factors, and there are also possible error correlations among different sensors depending on traffic conditions. For simplicity, the following analysis assumes the combined errors belong to a white normal probability distribution with zero-mean and a variance-covariance matrix R.

As point and AVL detectors are installed at fixed locations, the corresponding sensor mapping matrices, denoted as H'_d and H''_d, typically remain the same within the study horizon, that is $H'_d = H'$ and $H''_d = H''$. Even with more accurate vehicle based link travel time samples, the AVL-based sensors still have two major limitations: low market penetration rate and stochastic temporal coverage. Specifically, similar to the point-to-point AVI sensor, a large sampling error is introduced to probe sensor measurements under a low market penetration rate, as shown by $\mathrm{var}(\varepsilon'''_{d,a}) = \mathrm{var}(T_{d,a}) / g'''_{d,a}$, where

$g'''_{d,a}$ is the number of probe samples on link *a* on day *d*, and $\mathrm{var}(T_{d,a})$ is the systematic variance of link travel time on link *a* on day *d.* The same number of probe samples can generate smaller measurement errors on a link with low travel time variability compared to a link with highly dynamic traffic. Moreover, individual travelers with GPS probes can use different paths and links on different days, which leads to a day-varying and stochastic sensor mapping matrix H'''_d that consists of stochastic link traversing coefficient $\tilde{\gamma}'''_{d,a}$ for GPS probe vehicles.

3.2.3. Travel Time Estimation and Prediction Models

Given above process and measurement models, we are ready to derive a Kalman Filtering based estimation and prediction model to update link travel times with different types of measurements.

In the following discussion, we need to distinguish two states of each day *d*: (1) *a priori* state before the start of the current day (e.g., morning peak hour), corresponding to the predicted travel time T_d^- and uncertainty P_d^-, and (2) *a posteriori* state after the morning peak hour of current day, corresponding to estimated travel time T_d^+ and uncertainty P_d^+ after taking into account new measurements Y_d available on day *d*. We further define the *a priori* estimate error $\tilde{T}_d^-$ as the difference between the true travel time vector T_d and the *a priori* link travel time estimate T_d^-, where $\tilde{T}_d^- = T_d - T_d^-$. The *a posteriori* estimate error $\tilde{T}_d^+$ is the difference between the true travel time vector T_d and the *a posteriori* link travel time estimate T_d^+. Define $\tilde{T}_d^+ = T_d - T_d^+$, correspondingly, the

variance-covariance matrices of the *a priori* and *a posteriori* estimate error are expressed as

$$P_d^- = E[\tilde{T}_d^- \tilde{T}_d^{-\mathrm{T}}] = E[(T_d - T_d^-)(T_d - T_d^-)^{\mathrm{T}}] = \mathrm{cov}(T_d - T_d^-) \tag{3.10}$$

$$P_d^+ = E[\tilde{T}_d^+ \tilde{T}_d^{+\mathrm{T}}] = E[(T_d - T_d^+)(T_d - T_d^+)^{\mathrm{T}}] = \mathrm{cov}(T_d - T_d^+) \tag{3.11}$$

3.2.3.1. Travel time estimate updating

In Kalman filtering, the *a posteriori* travel time estimate T_d^+ is updated through a linear function of the *a priori* estimate T_d^- and a weighted difference $Y_d - HT_d^-$, which is the error of *a priori* estimate, otherwise known as the innovation residual or measurement residual.

$$T_d^+ = T_d^- + K(Y_d - HT_d^-) \tag{3.12}$$

3.2.3.2. Kalman gain factor and uncertainty propagation

By assuming the measurement error covariance R_d is uncorrelated to K_d and Y_d, a general formulation for the variance-covariance matrix of the *a posteriori* estimate error can be derived (see appendix for equation derivation).

$$P_d^+ = (I - K_d H_d) P_d^- (I - K_d H)^{\mathrm{T}} + K_d R K_d^{\mathrm{T}} \tag{3.13}$$

The above equation shows the propagation of the estimation error covariance for any given matrix K_d. In Eqs. (3.12) and (3.13), matrix K_d is used as the gain factor to update the *a posteriori* estimation T_d^+ and its error covariance P_d^+. In Kalman filtering, K_d is determined by minimizing the trace of *a posteriori* estimate error matrix, which is setting the first derivative of Eq. (3.13) to 0 as follows:

$$\frac{\partial \mathrm{trace}\left(P_d^+\right)}{\partial K_d} = -2\left(H_d P_d^-\right)^{\mathrm{T}} + 2K_d\left(H_d P_d^- H_d^{\mathrm{T}} + R_d\right) = 0 \tag{3.14}$$

The optimal form of K_d is then derived as

$$K_d = P_d^- H_d^{\mathrm{T}} (H_d P_d^- H_d^{\mathrm{T}} + R)^{-1} \tag{3.15}$$

Under the optimal formulation of the Kalman gain matrix in Eq. (15), a simplified expression for the estimation error covariance is derived as

$$P_d^+ = (I - K_d H) P_d^- \tag{3.16}$$

Other formulas of the estimation error covariance are available, for example,

$$P_d^+ = \left((P_d^-)^{-1} + H^T R^{-1} H\right)^{-1} \tag{3.17}$$

Consider a single link shown in Figure 3.3 where a link from *a* to *b* corresponding to sensor mapping matrix $H = 1$. In the historical travel time database, the estimated travel time follows a normal distribution with a mean of 15 min and a standard deviation of 5 min (i.e . $P_d^- = 25$). Given a new measurement $Y_d = 20$ min with a measurement error variance of 5 min, based on Eq. (3.15), we can calculate the optimal Kalman filtering gain factor as $K_d = 25/(25+5) = 5/6$. Then the travel time estimate is updated by Eq. (3.12), calculated as $T_d^+ = 15 + (5/6)(20-15) = 19.6$, and the posterior estimation variance is reduced to $P_d^+ = \left(1-(5/6)\right) \times 25 = 1.8$.

The calculation results are summarized in Table 3.1.

3.2.3.3. Travel time prediction

After updating travel time mean estimate and its covariance on day *d*, we now need to predict travel time and its uncertainty range for the same time interval of the next day,

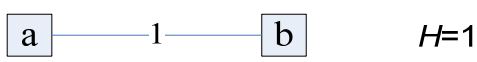


Figure 3.3: Single-link example

Table 3.1: Calculation results of single-link example

	Prior estimate	Measurement	Posterior estimate
Travel time (min)	$T_d^- = 15$	$Y_d = 20$	$T_d^+ = 19.6$
Variance (min^2)	$P_d^- = 25$	$R_d = 5$	$P_d^+ = 1.8$

d+1. According to the system process equation (3.1), the mean estimate can be simply extended across days under recurring traffic conditions.

$$T_{d+1}^{-} = T_{d}^{+} \tag{3.18}$$

However, by taking into account the unpredicted random realizations of traffic demand and capacity, characterized by the system error matrix Q_d, we have to increase the uncertainty estimate for T_{d+1}^{-}

$$P_{d+1}^{-} = P_{d}^{+} + Q_{d} \tag{3.19}$$

With the above measurement updating Eqs. (3.12-3.17) and prediction equations (18-19), the Kalman filtering based travel time estimation and prediction model is able to recursively correct the link travel time estimate from streaming traffic measurements and dynamically adjust the error covariance matrix that indicates the uncertainty range of the prediction results.

3.3. Travel Time Prediction under Nonrecurring Conditions

Under nonrecurring congestion conditions, the structural deviation $V_d(t)$ is considered under various demand/capacity changes. The process equation of the structural deviation in real-time prediction applications can be written as a state space model:

$$V_d(t+h) = L_d(t,t+h)V_d(t) + \mu_d(t+h), \quad \mu_d(t+h) \sim N(0, Q_d^{NR}) \tag{3.20}$$

In Eq. (3.20), the transition matrix L denotes the process matrix of the structure derivation. The process error μ_d is considered as a normally distributed random noise. Following we will discuss the nonrecurring traffic estimation and prediction with and without sensor coverage, and present two case studies taking incident as a demonstration example.

3.3.1. Traffic Estimation and Prediction Equations

For links with sensor coverage, a measurement Y_d is obtained for each time interval. Similar to the derivation for the regular pattern traffic, the estimation equations for the nonrecurring structural deviation and its uncertainty are

$$V_d^+(t) = V_d^-(t) + K_{NR}(Y_d(t) - H_d(t)T^h(t) - H_d(t)V_d^-(t)) \tag{3.21}$$

$$\Sigma_d^+ = \left((\Sigma_d^-)^{-1} + H_d^T(t)R^{-1}H_d(t)\right)^{-1} \tag{3.22}$$

The Kalman gain factor K_{NR} can be derived similar to the recurring traffic model.

The prediction equations for the structural deviation and its uncertainty are

$$V_d^-(t+h) = L_d(t,t+h)V_d^+(t) \tag{3.23}$$

$$\Sigma_d^-(t+h) = L_d(t,t+h)\Sigma_d^+(t)L_d^T(t,t+h) + Q_d^{NR} \tag{3.24}$$

With the derivation of the structural deviation, the predicted travel time variance under nonrecurring conditions is represented as

$$\begin{aligned} P_d^-(t+h) &= P_{d-1}^+(t+h) + \Sigma_d^-(t+h) + Q_d \\ &= P_{d-1}^+(t+h) + Q_d + L_d(t,t+h)\Sigma_d^+(t)L_d^T(t,t+h) + Q_d^{NR} \end{aligned} \tag{3.25}$$

Eq. (3.25) computes the prediction uncertainty at time (t+h). By comparing to the regular pattern uncertainty prediction in Eq. (3.19), we noticed that the uncertainty of nonrecurring conditions is considered as a linear combination of the regular pattern and the structural deviation.

For links without sensor coverage, no measurement is available of the estimation for the structure derivation V_d. Therefore, predicted values for both structure derivation $V_d(t)$ and link travel time $T_d(t)$ will be biased, and an extra error has to be considered into the system uncertainty estimation and prediction. In this study we use the maximum V_d across all links with sensors from the historical database as a way to estimate the potential bias magnitude on links without sensors. As a result, the corresponding elements in the prior structure derivation uncertainty matrix Σ^- have large values for links without sensor coverage, and relatively small values for links equipped with sensors.

3.3.2. Single Bottleneck Model with Incident

We now shift our focus on how to compute the essential transition matrix $L_d(t,t+h)$, with a single bottleneck case study. Newell's kinematic wave model (Newell, 1993) is

used in this research to capture forward and backward waves as results of bottleneck capacities. Its simplified form of traffic flow models is particularly attractive in establishing theoretically sound and practically operational traffic transition models on bottlenecks. Interested readers are referred to a number of related studies on Newell's kinematic wave model, e.g., the model calibration effort by Hurdle and Son (2000), extensions to node merge and diverge cases by Yperman et al. (2005) and Ni et al. (2006).

Considering Figure 3.4, a recurring congestion is assumed with a constant queue discharging rate C. An incident under consideration begins at time s and ends at time e with a reduced capacity C^R, and this capacity is restored back to C after time e. In order to derive the transition matrix L for updating the structural deviation $V(t)$, we further examine the additional delay in a detailed plot.

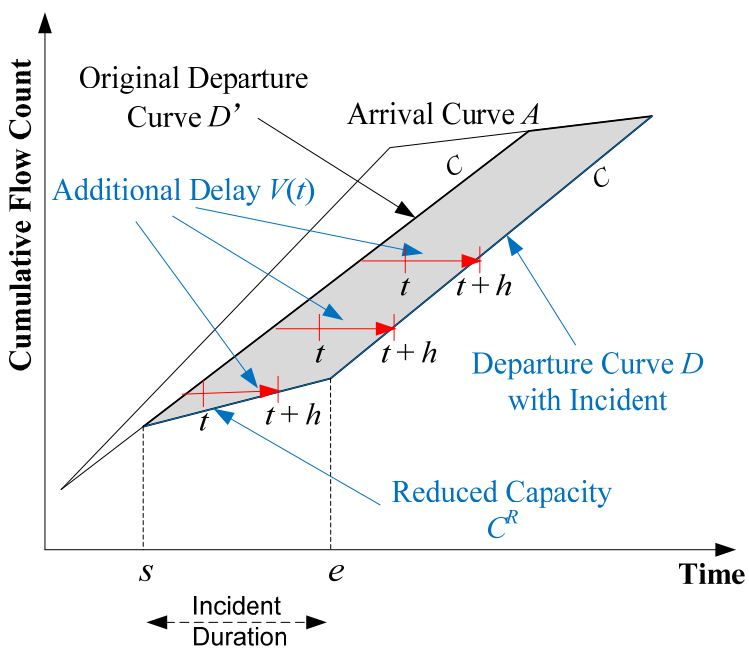


Figure 3.4: Cumulative flow count curve for a single bottleneck with reduced capacity due to an incident

Shown in Figure 3.5, $V(t) = t - t'$, where t' is the original leaving time from the queue under recurring congestion for the same vehicle. Δ is denoted as the number of vehicles can be discharged under recurring congestion from s to t', and we can derive $\Delta = C \times (t' - s) = C^R \times (t - s)$. That is, $t' - s = \Delta / C = (C^R / C)(t - s)$. Thus, the travel time structure deviation term can be determined as

$$V(t) = (t - s) - (t' - s) = (t - s) \times \left(1 - \frac{C^R}{C}\right). \tag{3.26}$$

After the incident ending time e, $V(t)$ becomes a constant value $(e - s) \times \left(1 - \frac{C^R}{C}\right)$.

We can further examine the transition matrix L in three cases (as shown in Figure 3.4), with a predefined prediction period h (e.g., 15 min). According to the prediction equation (3.23) for the structure derivation, the transition matrix L (a single value in this example) is derived as the ratio of structure deviation terms between current time t and future time $t+h$.

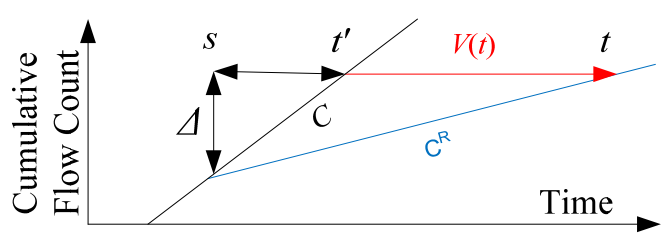


Figure 3.5: A zoom-in view on capacity reduction

$$L(t,t+h)=\frac{V(t+h)}{V(t)} \tag{3.27}$$

As shown in Table 3.2, with the prediction time stamp t is located in different time periods, the transition matrix L is derived in different forms. Figure 3.6 gives an illustrative example on how the transition coefficient L varies according to time t, by assuming the prediction period h = 15 min and a typical incident duration $e - s$ = 30 min.

The above example demonstrates how to derive the transition matrix L under nonrecurring conditions with short-term capacity drops. Similar matrices could be derived for severe weather and work zone cases. Obviously, $L(t,t+h)$ is a time-dependent and situation-dependent variable that needs to determine in a case-by-case basis in real-world travel time prediction applications. For the sensor location problem under consideration, we need to assume and use an aggregated transition matrix for simplicity. In our experiments for the sensor location problem, we consider the following typical case: average incident duration = (e-s) = 30 min, incident reporting period = (t-s) = 15min,

Table 3.2: Derivation of transition matrix L for different time periods

Scenarios	t	$t+h$	$V(t)$	$V(t+h)$	$L(t,t+h)$
Early Detection	$s < t < e-h$	$s+h < t+h < e$	$(t-s)\times\left(1-\frac{C^R}{C}\right)$	$(t+h-s)\times\left(1-\frac{C^R}{C}\right)$	$\frac{t+h-s}{t-s}$
Late Detection	$e-h < t < e$	$t+h > e$	$(t-s)\times\left(1-\frac{C^R}{C}\right)$	$(e-s)\times\left(1-\frac{C^R}{C}\right)$	$\frac{e-s}{t-s}$
Post-incident Detection	$t > e$	$t+h > e+h$	$(e-s)\times\left(1-\frac{C^R}{C}\right)$	$(e-s)\times\left(1-\frac{C^R}{C}\right)$	1

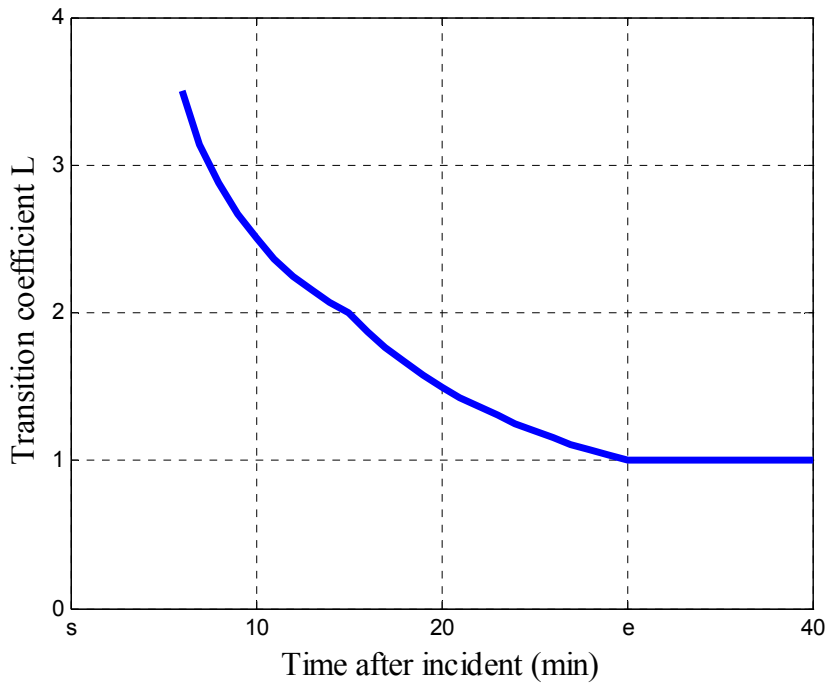


Figure 3.6: Illustration of time-dependent transition coefficient L

and this leads to a typical value $L = 2$ which will be used in the experiments in Section 3.7.

3.4. Measure of Information for Historical Traffic Patterns

One of the fundamental questions in sensor location problems is which criteria should be selected to drive the underlying optimization processes. Eqs. (3.13-3.17 & 3.19) in the above travel time estimation and prediction model offer an analytical model for quantifying the estimation/prediction error reduction due to additional measurements provided by new sensors. As the process variance-covariance matrix is assumed to be constant, the travel time uncertainty measure in this section uses the *a posterior* estimation error covariance P^+ as the basis to evaluate the information gain. A

challenging question then is how to select single-value information measures for a sensor design plan. To this end, we first examine two commonly used estimation criteria, namely, the mean-square error and entropy. We then propose total path travel time estimation variance as a new measure of information for end-to-end trip time prediction applications.

3.4.1. Trace and Entropy

As shown in Eq. (3.14), when selecting the gain factor K to utilize new measurements, the classic Kalman filter aims to minimize the mean-square error, i.e., the trace of P_d^+. The trace of the variance covariance matrix $tr\left(P_d^+\right)$ is the sum of the diagonals of the matrix, which is equivalent to the total variance of link travel time estimates for all links:

$$tr\left(P_d^+\right) = \sum_{a=1}^{m} \mathrm{cov}(t_{d,a}, t_{d,a}) = \sum_{a=1}^{m} \mathrm{var}(t_{d,a}) \tag{3.28}$$

While the trace does not consider the effects of correlation between travel times of adjacent links, an alternative measure of information is entropy which is commonly used in information theory applications. For a discrete variable, Shannon's original entropy is defined as the number of ways in which the solution could have arisen. For a continuously distributed random vector T, on the other hand, the entropy is measured by $-E(\ln f(T))$, where f is the joint density function for vector T. If travel time T in our study is assumed to follow a normal distribution, then its entropy is computed as

$\theta + \frac{1}{2}\ln(\det(P_d^+))$, where θ is a constant that depends on the size of *T*, the total number of links in our study network. The entropy measure is proportional to the log of the determinant of the covariance matrix. By ignoring the constant θ and the monotonic logarithm function, we can simplify the entropy-based information measure for the *a posteriori* travel time estimate as $\det(P_d^+)$. The determinant of the variance covariance matrix, as a measure of information, is also known as the generalized variance. Mathematically, the trace and determinant of the variance covariance matrix P_d^+ can be calculated from the sum and product, respectively, of the eigenvalues of P_d^+. Since the determinant considers the variance and covariance in the matrix, a smaller determinant is desirable because this indicates a more accurate estimate.

3.4.2. Total Path Travel Time Estimation Uncertainty

This study proposes a new measure of information to quantify the network-wide value of information, based on the travel time estimation quality of critical OD/paths.

The travel time estimation uncertainty of path *k* from origin *o* to destination *s* can be calculated from the posterior travel time estimate variance-covariance matrix P_d^+:

$$e_{d,o,s,k} = \sum_{a \in p(o,s,k)} \mathrm{var}(t_{d,a}) + 2 \times \sum_{a<b;a,b \in p(o,s,k)} \mathrm{cov}(t_{d,a}, t_{d,b}) \quad (3.29)$$

,

where var() and cov() are variance and covariance coefficients in the link travel time uncertainty matrix, respectively. Compared to trace or entropy based information

measures, the proposed path travel time based measure can better capture the possible correlation between traffic estimates along a path, with the covariance portion of the estimation error matrix.

A similar equation can be derived for travel time prediction based uncertainty measure, using the travel time estimate variance-covariance P_d^- matrix. For sensor location decisions that jointly consider recurring and nonrecurring conditions, an integrated uncertainty matrix can be generated from recurring travel time uncertainty P^- and nonrecurring structure derivation uncertainty Σ^-:

$$P_D^- = (1-\rho_{incident}-\rho_{workzone}-\rho_{weather})\times P_{recurring}^- + \rho_{incident}\Sigma_{incident}^- + \rho_{workzone}\Sigma_{workzone}^- + \rho_{weather}\Sigma_{weather}^-$$

Weighted by the path flow volume of different origin-destination pairs $f_{o,s,k}$, the overall estimation uncertainty of the network-wide traffic conditions on day d can be determined from the following equation:

$$TU_d = \sum_{o,s,k}\left(e_{d,o,s,k}\times f_{o,s,k}\right) \tag{3.30}$$

The above total path travel time estimation uncertainty measure includes three important components: (1) the sum of elements in the variance covariance matrix for link travel time estimates; (2) the sum of the travel time variance for each feasible or critical path in the network; and (3) weights of path flow volume for different paths. As the path travel time estimation accuracy (as opposed to individual link travel times) is the ultimate

information quality requirement by commuters traveling on various routes, this measure of information can capture the high-level monitoring performance of a sensor network. In relation to the trace and entropy measures, the total path travel time estimation uncertainty can be viewed as a more appropriate indicator for system-wide information gains.

3.5. Sensor Design Model and Beam Search Algorithm

The proposed sensor network design model is essentially a special case of the discrete network design problem, so an integer programming model, shown below, can be constructed to find the optimal sensor location solution.

Min *TU*

Subject to:

(1) Budget constraint:

$$\pi' \times \sum_{a} x'_a + \pi'' \times \sum_{i} x''_i \leq \pi \tag{3.31}$$

(2) Traffic pattern uncertainty propagation constraints under recurring and nonrecurring conditions (Eqs. 3.17, 3.19, 3.22, 3.24);

(3) Sensor mapping matrix constraint:

$$H_d = function\left(A' = \left[x'_a\right], N'' = \left[x''_i\right], \tilde{\gamma}'''_{d,a}, \alpha, \beta\right) \qquad d = 0,1,...,D \tag{3.32}$$

D = a sufficiently large day number for measure of information to reach convergence.

x'_a = 1 if a point sensor is installed on link a, 0 otherwise.

x''_i =1 if an AVI sensor (point-to-point sensor) is installed on node i, 0 otherwise.

π', π'' = installation and maintenance costs for point sensors and point-to-point sensors.

π = total available budget for building or extending the sensor network.

In the above objective function, the overall system uncertainty matrix P_D^- is calculated as a probabilistic combination of recurring and nonrecurring traffic variances. Structure derivations from different nonrecurring traffic conditions are considered in the total system uncertainty with corresponding probabilities. For links with sensors, the structure derivation uncertainty is aggregated and averaged from historical measurements (e.g., 95% or 2σ for normal distribution). For links without sensor, we will take the maximum of the structure derivation from limited historical database.

H_d is determined by the sensor location set $A'=[x'_a]$ and $N''=[x''_i]$, randomly generated link traversing coefficient for GPS probe vehicles $\tilde{\gamma}'''_{d,a}$, and AVI and AVL market penetration rates α and β.

Essentially, the goal of the above sensor location model is to add sensor information from spatially distributed measurements to minimize the weighted uncertainty associated with the path travel time estimates. In this study, a branch-and-bound search procedure can be used to solve the integer programming problem. To reduce the computational complexity, a beam search heuristic algorithm is implemented in this study.

Given prior information on the link travel time vector and its estimation error covariance from historical database, the proposed algorithm tries to find the best sensor location scenario from a set of candidates under particular budget constraints. Based on a breadth-first node selection mechanism, the beam search algorithm branches from the nodes level by level. At each level, it keeps only φ promising nodes, and prunes the other nodes permanently to limit the total number of nodes to be examined. φ is typically referred to as the beam width, and the total computational time of the beam search algorithm is proportional to the selected beam width.

Algorithm 3.1: Beam search algorithm

Step 1: Initialization

Generate candidate link set *LC* and candidate node set *NC* for point and point-to-point sensors, respectively.

Set the active node list $ANL = \varnothing$. Create the root node *u* with $A'(u) = \varnothing, N''(u) = \varnothing$, search level $sl(u) = 0$, where *u* is search node index. Insert the root node into *ANL*.

Step 2: Stopping criterion

Terminate and output the best-feasible solution under one of the following conditions:

(1) If all of the active nodes in *ANL* have been visited,

(2) The number of active nodes in memory is exceeded.

Step 3: Node generation and evaluation

For each node *u* at search level *sl* in ANL, remove it from ANL, and generate child nodes;

Scan through the candidate sets *LC*, if a link *a* is not in *A′*(*u*), generate a new child node *v'* where $A'(v') = A'(u) \bigcup a, N''(v') = N''(u), sl(v') = sl(u) + 1$;

Scan through the candidate sets *NC*, if a node *i* is not in *N″*(*u*), generate a new child node *v''* where $N''(v'') = N''(u) \bigcup i, A'(v'') = A'(u), sl(v'') = sl(u) + 1$;

For each newly generated node *v*, calculate the objective function through P_D^- in Eq. (30). If the budget constraint is satisfied for a newly generated node, add it into the *ANL*.

Step 4: Node filtering

Select *φ* best nodes from the *ANL* in the search tree, and go back to Step 2.

In the above beam search algorithm, the total computational time is determined by the number of nodes to be evaluated, which depends on the beam width *φ* and the size of the candidate sensor links/nodes. For each node in the tree search process, the complexity is determined by the evaluation of the objective function, which can be decomposed into three major steps: (1) calculating P^+ from $H^T R^{-1} H$, (2) calculating the inverse of the covariance matrix $(P^+)^{-1}$, and (3) calculating the path travel time uncertainty as a function of P_D^-. The first step involves two matrix multiplications: $H^T R^{-1}$ and $(H^T R^{-1})H$. Because *H* is an (*n*×*m*) matrix and *R* is an (*n*×*n*) matrix, the first step has a worst-case complexity of O(m^2n), and calculating the inverse of matrix leads to an O(m^3) operation if the Gaussian elimination method is used.

For a large-scale sensor network design application, we can adopt three strategies to reduce the size of the problem and therefore the computational time. First, one can focus on critical OD pairs with significant volumes. Second, one can aggregate original OD

demand zones into a set of super zones within a manageable size, with this strategy being especially suitable for a subarea analysis where many OD zones outside the study area can be consolidated together. Third, we can reduce the size of candidate AVL sensor nodes and point sensor links in order to decrease the number of search nodes to be evaluated.

3.6. Complex Cases for Updating Historical Traffic Patterns

3.6.1. Quantifying Steady State Information Gain

To consider long-term information gains of a sensor network in monitoring the travel time dynamics, the following discussion aims to derive the steady-state results of uncertainty reduction associated with a fixed sensor network design plan. Considering both point and AVL sensors, we first assume constant Q, R and H across different days, the travel time estimation error covariance updating equation as seen in Eq. (3.33), which was combined from Eqs. (3.16) and (3.19),

$$P_d^- = (I - K_{d-1}H)P_{d-1}^- + Q \tag{3.33}$$

Under steady state conditions, the travel time estimation error covariance will achieve a constant state as $P = P_d^- = P_{d-1}^-$ after a number of updates. By applying the optimal formulation of Kalman gain K in Eq. (3.15), the steady estimation error covariance P is rewritten as

$$PH^T(HPH^T + R)^{-1}HP = Q \tag{3.34}$$

or $$P = (I - PH^T(HPH^T + R)^{-1}H)P + Q \tag{3.35}$$

Eq. (3.35) is known as Algebraic Riccati Equation. When numerically solving this equation, the steady-state travel time estimation error covariance matrix for a long-term sensor location problem is obtained.

Figure 3.7 illustrates a day-by-day time series of the travel time estimation variance. Due to the presence of system evolution noise Q, the estimation variance always increases when we make a travel time prediction from day d to day d+1, that is, $P_d^- = P_d^+ + Q_d$. After receiving traffic measurement available every day, the uncertainty associated travel time estimates is reduced through $P_d^+ = (I - K_d H)P_d^-$. The uncertainty reduction and the resulting information gain are very dramatic after the first few days of sensor deployment. After 5 or 6 days, this zig-zag pattern reaches a stable state when

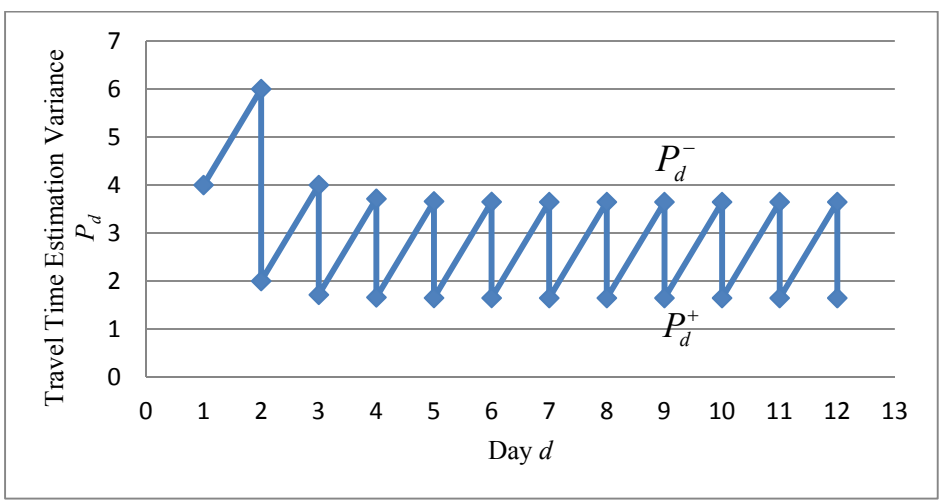


Figure 3.7: Steady-state travel time estimation variance

$P_d^- = P_{d-1}^-$ (corresponding to the upper portion of the time series) and $P_d^+ = P_{d-1}^+$ (corresponding to the lower portion of the time series).

Due to the stochastic coverage characteristic of AVL sensor data, we can use a sample-based iterative computation scheme to compute the stable-state *posterior* estimation covariance matrix P^+. In particular, representative samples of $\tilde{\gamma}'''_{d,a}$ can be first generated for each day, and then applied into the update equations (3.16) and (3.19) over multiple days to check if det(P^+) converges to a constant value .

3.6.2. AVI Extension with Multiple Paths

In the previous discussions, we assume that all AVI-equipped travelers use only a single path between each pair of AVI sensors. In the following discussion, we shift our focus from a single path case to a more complex but realistic situation with multiple used paths between a pair of AVI sensors. For simplicity, we assume the route choice probabilities for those paths can be computed from a deterministic or stochastic traffic assignment program.

This study adapts a multivariate normal (MVN) distribution to represent the route choice behavior. In particular, each route choice decision from an individual traveler can be considered as an independent Bernoulli trial from one of the K possible outcomes (i.e., paths) with probabilities $p_1, \ldots, p_k, \ldots, p_K$.

Consider an example network shown in Figure 3.8, where two AVI sensors are located at nodes a and c. Travelers can take either of two routes through link sequence: (path k = 1)1→3 or (path k = 2) 2→3, where link 3 is shared by these two paths. We now

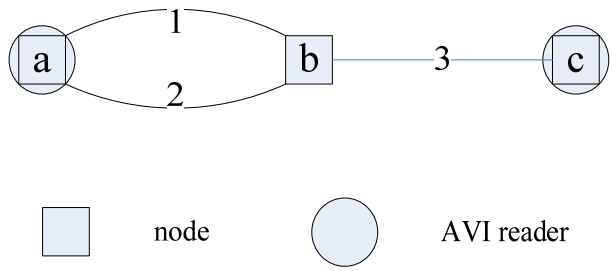


Figure 3.8: Two partially overlapping path between AVI readers at nodes *a* and *c*

extend AVI measurement equation (3.6) from a single path to the following with two paths (subscripts *d*, *i*,*j* are omitted for notation simplicity).

$$y'' = H''T + \varepsilon'' = \begin{bmatrix} p_1 & p_2 & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} + \eta + \delta = p_1 t_1 + p_2 t_2 + t_3 + \eta + \delta \tag{3.36}$$

where p_1 and p_2 represent the route choice probability for link 1 and 2, with $p_1 + p_2 = 1$,

y'' is the average travel time from travelers using two routes, the contribution from travelers using route 1 is $p_1(t_1 + t_3)$ and the contribution from route 2 is $p_2(t_1 + t_3)$,

η is the error term introduced by sampling variations due to multiple paths,

δ is the error term associated with using the sample average value to approximate the population mean value *T*, as described in Eq. (3.8).

The variance of multichoice sampling variation η can be calculated by

$$\mathrm{var}(\eta) = \mathrm{var}(p_1 t_1 + p_2 t_2 + t_3) = t_1^2 \mathrm{var}(p_1) + t_2^2 \mathrm{var}(p_2) + 2 t_1 t_2 \mathrm{cov}(p_1, p_2) \tag{3.37}$$

Assuming there are g AVI samples observed between this AVI sensor pair, and x_k use path k, we can derive $\mathrm{E}(p_k) = \mathrm{E}(x_k/g) = p_k$, $\mathrm{var}(p_k) = \mathrm{var}(x_k)/g^2 = p_k(1-p_k)/g$, and $\mathrm{cov}(h_k, h_\lambda) = -p_k p_\lambda / g$.

Thus, Eq. (3.37) is reduced to

$$\mathrm{var}(\eta) = (t_2 - t_1)^2 \frac{p_1 p_2}{g} \tag{3.38}$$

First, it is clear that the overlapping portion, link 3, does not affect the variance associated with the multichoice sampling error. Under perfect deterministic user equilibrium conditions, all of the used routes have the same travel time, so the $\mathrm{var}(\eta)$ further reduces to zero. Under a more realistic stochastic user equilibrium assumption, the travel time difference between different used routes will increase the range of the combined error $R = \mathrm{var}(\varepsilon'') = \mathrm{var}(\eta + \delta)$.

We can further extend the two-path case to consider multiple paths on a corridor with K parallel nonoverlapping routes, shown in Figure 3.9. In general, the variance of the combined error increases as there are more used paths with significant travel time differences.

$$\mathrm{var}(\eta) = \mathrm{var}(p_1 t_1 + \cdots + p_k t_k + \ldots + p_K t_K) = \frac{1}{g} \sum_{k,\lambda,k>\lambda} (t_k - t_\lambda)^2 p_k p_\lambda \tag{3.39}$$

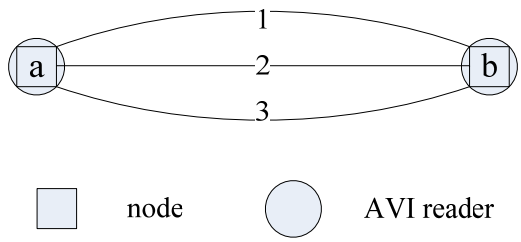


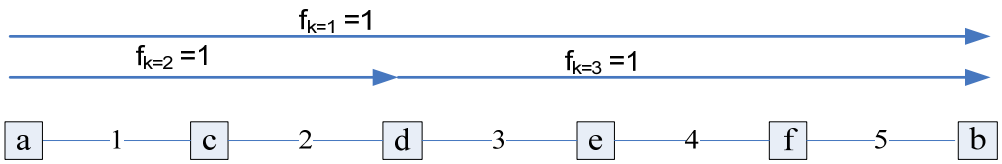
Figure 3.9: Example network with three parallel paths

3.7. Illustrative Example and Numerical Experiments

3.7.1. Illustrative Example for Locating AVI Sensors

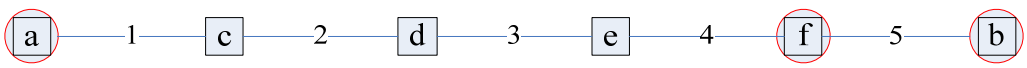
In Figure 3.10, we present an illustrative example with a 6-node hypothetical transportation network to demonstrate how the proposed measures of information can systematically evaluate the trade-offs between the accuracy and placement of individual AVI sensors for path travel time estimation reliability. In Figure 3.10, subscript day d is omitted for simplicity. As shown in the base case, there are three traffic analysis zones at nodes a, d and b, and three major origin-to-destination trips: (1) a to b, (2) a to d and (3) d to b, each with a unit of flow volume. P^- (e.g., obtainable from a historical travel time database with point detectors) leads to a trace of 12 and a determinant of 48. Among the 5 links in the corridor, link 5 from node f to b has the highest uncertainty in terms of link travel time estimation variance. We can view node b as a downtown area, and the incoming flow from the other two zones creates dramatic traffic congestion and travel time uncertainty, first on link 5 and then on link 4. For the base case, we can calculate the variance of path travel time estimates for these three OD pairs, respectively, as 12, 3 and 9, leading to a total path travel time estimation uncertainty (TU) as $TU = 24$.

In both cases (I) and (II), two AVI sensors are first installed at nodes a and b. In



$$P^{-} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} \qquad Tr = 12 \qquad Det = 48 \qquad TU = 12 + 3 + 9 = 24$$

Base case



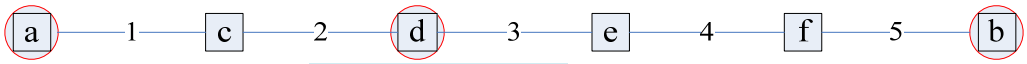
$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad H = \begin{array}{c} \text{Link } 1 \quad 2 \quad 3 \quad 4 \quad 5 \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Measurement
a->f
f->b

$$H^T R^{-1} H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad P^{+} = \begin{bmatrix} 0.89 & -0.22 & -0.22 & -0.33 & 0 \\ -0.22 & 1.56 & -0.44 & -0.67 & 0 \\ -0.22 & -0.44 & 1.56 & -0.67 & 0 \\ -0.33 & -0.67 & -0.67 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}$$

$$Tr = 6.8 \qquad Det = 1.07 \qquad TU = 1.71 + 2.01 + 3.02 = 6.74$$

(I) Locate additional AVI reader to reduce highest link travel time uncertainty



$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad H = \begin{array}{c} \text{Link } 1 \quad 2 \quad 3 \quad 4 \quad 5 \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Measurement
a->d
d->b

$$H^T R^{-1} H = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \qquad P^{+} = \begin{bmatrix} 0.75 & -0.5 & 0 & 0 & 0 \\ -0.5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1.6 & -0.6 & -0.8 \\ 0 & 0 & -0.6 & 2.1 & -1.2 \\ 0 & 0 & -0.8 & -1.2 & 2.4 \end{bmatrix}$$

$$Tr = 7.85 \qquad Det = 1.2 \qquad TU = 1.65 + 0.75 + 0.9 = 3.3$$

(II) Locate additional AVI reader to match origin-to-destination trip pairs

Figure 3.10: Example of locating AVI sensors on a linear corridor

case (I), an additional AVI sensor is located at node *f* so that we can obtain two pairs of end-to-end travel time measurements: from node *a* to node *f*, and from node *f* to node *d*. The second measurement directly monitors travel time dynamics on link 5. In this particular example along the linear corridor, the end-to-end travel time statistics from *a* to *b* can be explicitly determined from the above two *mutually exclusive* observations. In order to avoid double-counting the information gain for the same data sources, the information quantification module in this study only considers two raw measurements: from *a* to *f*, and from *f* to *d*, to update the link travel time variance covariance matrix from P^- to P^+. To do so, the measurement error matrix is assumed to be $R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and the mapping matrix $H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$, where the first measurement from *a* to *f* covers links 1,2,3 and 4, and the second measurement from *f* to *b* covers link 5. As link 5, with the highest travel time uncertainty, is directly measured from AVI readings, its link travel time estimate variance is reduced from 4 to 0.8, but the resulting P^+ contains a large amount of correlation in its link travel time estimates for links 1 to 4. All the path travel time uncertainties for the three OD pairs have been reduced, and $TU = 6.74$.

In case (II), the third AVI sensor is installed at node *d* to match the nature OD trip demand pattern, which produces sensor mapping matrix $H = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$. The resulting P^+ still contains two clusters of correlations corresponding to two individual measurements from *a* to *d* and from *d* to *b*. The path travel time estimate variances for the OD pairs from *a* to *d* and from *d* to *b* are dramatically reduced to 0.75 and 0.9. Although

link 5 still has a relatively large estimate variance of 2.4, its overall estimation error measure, total travel time estimation uncertainty *TU* is now 3.3, which is much lower than *TU* = 8.5 in case (I).

In comparison, by locating and spacing AVI sensors to naturally match the spatial trip patterns of commuters, case (II) is able to systematically balance the trade-off between the needs for monitoring local traffic variations and end-to-end trip time dynamics. It is also important to notice that, both cases (I) and (II) have the same network coverage and generate the same number of measurements every day, but they provide different information gains from a commuter/road user perspective. Thus, simple measures of information, such as traffic network coverage and the number of measurements, might not be able to quantify the system-wide uncertainty reduction and information gain for traveler information provision applications.

3.7.2. Sensor Location Design for Traffic Estimation with Recurring Conditions

In this study, we examine the performance of the proposed modeling approach through a set of experiments on a simplified Irvine, California network, which is comprised of 16 zones, 31 nodes and 80 directed link. This study considers a single path between each OD pair in this simple network.

All the experiments are performed on a computer system equipped with an Intel Core Duo 1.8GHz CPU and 2 GB memory. Shown in Table 3.3, a set of critical OD pairs with large flow is selected to estimate the network-wide path travel time based uncertainty.

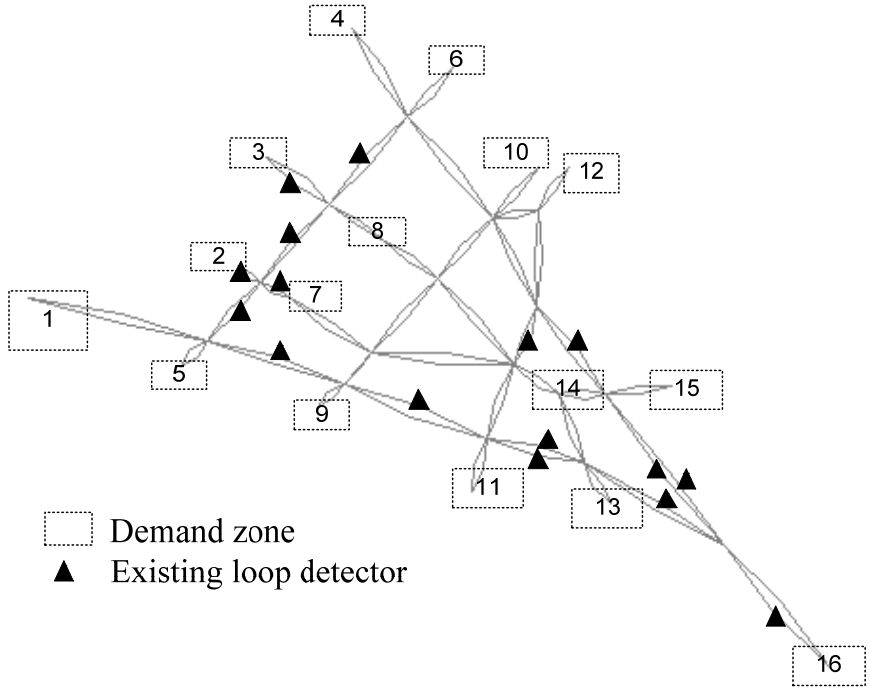
Table 3.3: Critical path travel time estimation error under existing and optimized sensor location strategies

Origin	Destination	Hourly volume	Prior path travel time estimation variance without sensor	Posterior path travel time estimation variance with existing sensors	% reduction in variance due to exiting sensors	Posterior path travel time estimation variance with optimized sensor locations	% reduction in variance due to optimized sensor locations
1	16	4000	5.87	5.14	12.44%	3	48.89%
16	4	6820	5.24	3.94	24.81%	2.8	46.56%
12	4	1152	1.85	1.85	0	1.32	28.65%
4	16	2480	5.23	3.54	32.31%	3.19	39.01%
16	12	832	4.91	3.61	26.48%	3	38.90%
15	4	880	2.81	2.6	7.47%	2.07	26.33%
12	16	680	4.9	3.21	34.49%	3.21	34.49%
16	1	4800	5.86	4.28	26.96%	3	48.81%
4	15	604	2.81	2.81	0	2.47	12.10%
4	12	444	1.85	1.85	0	1.5	18.92%

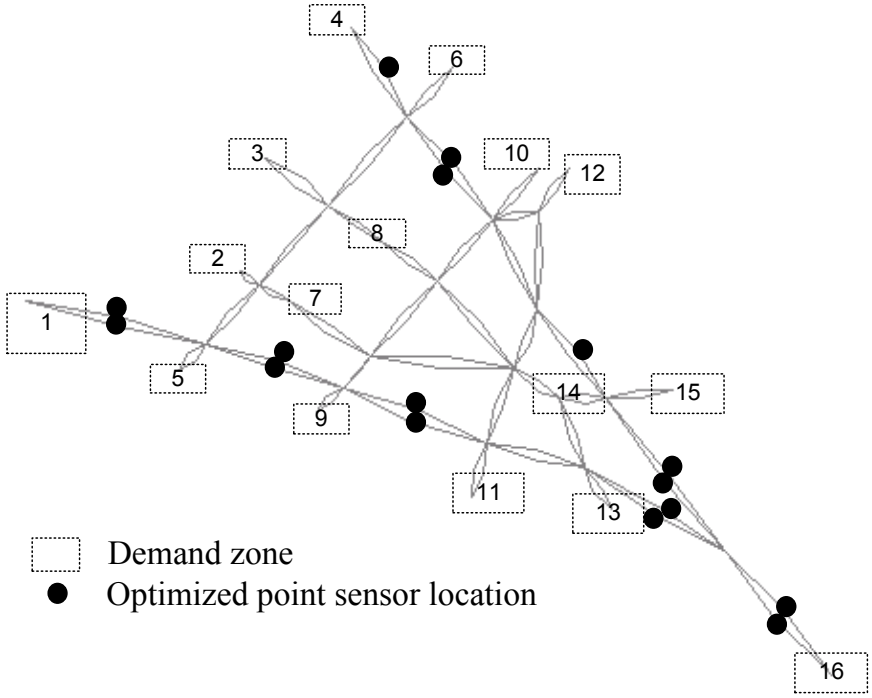
Additionally, a beam search width of 10 is used in the beam search algorithm to reduce the computational complexity. The total number of nodes in the search tree is the number of additional sensors times the beam search width. In our experiments, with standard Matlab matrix calculation functions, it takes about 30 min to compute 160 nodes in the beam search tree for this small-scale network.

In this section, we examine the proposed information measure model and sensor location algorithm for the estimation of recurring traffic conditions. With given OD flow and prior uncertainty information, three scenarios of sensor location plan are designed to compare with current sensor network.

We first conduct experiments to compare the existing point sensor network (Figure 3.11a) and an optimized point sensor network plan (Figure 3.11b), both with the same number (i.e., 16) of point sensors. Table 3.3 shows the critical path travel time estimation

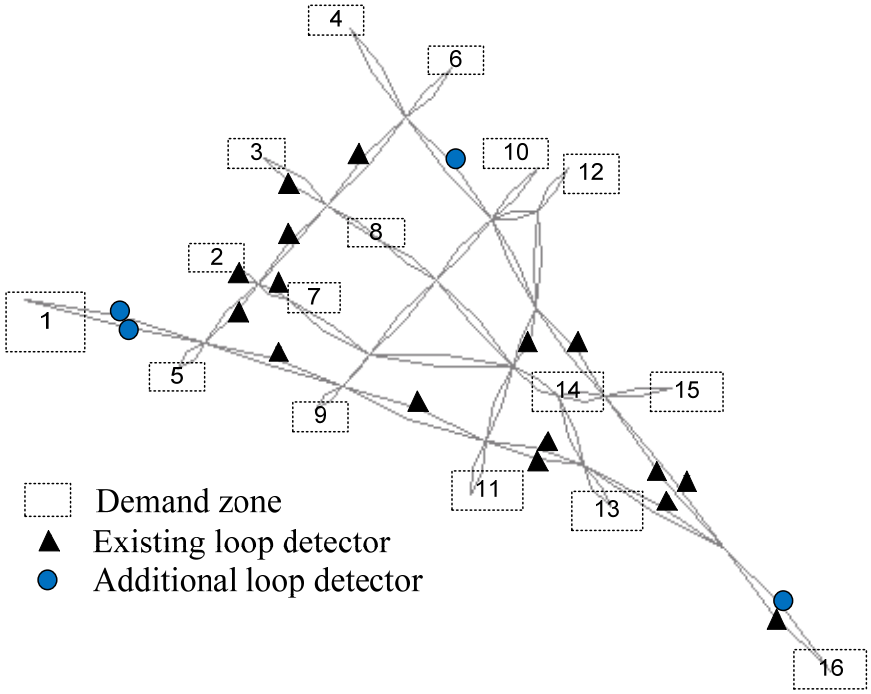


(a) Existing point sensor locations

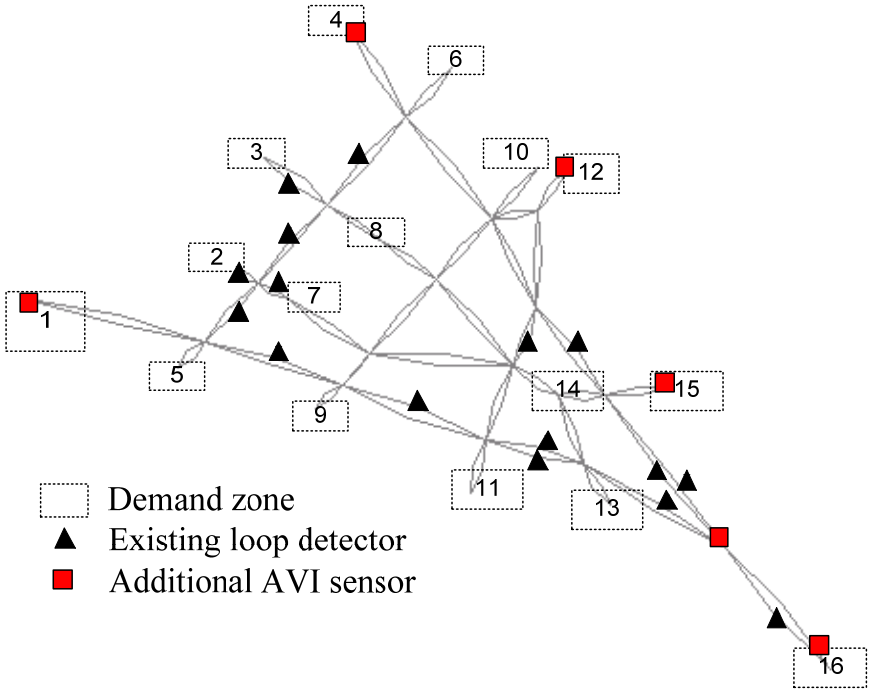


(b) Optimized sensor network locations
(start from a zero-sensor case)

Figure 3.11: Numerical experiment results for regular traffic pattern estimation



(c) Sensor network with 4 point sensors in addition to existing 16 point sensors



(d) Sensor network with 6 AVI sensors in addition to existing 16 point sensors

Figure 3.11, continued

errors under those two scenarios. The results show that the proposed optimization model can reduce the path travel time estimation variance by an average of 34.3%, while the existing sensor plan only reduces the same measure by about 16.5%. By factoring in the OD demand volume (shown in the third column), we can compute the proposed measure of information: the total path travel time estimation variance. The base case with zero sensor produces TU_zero_sensor=114855, the existing locations reduce TU to 88878 (77.3% of TU_zero_sensor), and the optimized sensor location scenario using the proposed model further decreases the system-wide uncertain to TU= 63586 (55.3% of TU_zero_sensor). This clearly demonstrates the advantage of the proposed model in terms of improving end-to-end travel time estimation accuracy.

In the next set of numerical experiments, we compare two scenarios with additional sensors on top of the existing sensors.

1) Add 4 point sensors on uncovered links still with large travel time variance, leading to a total 16+4=20 point sensors, shown in Figure 3.11(c);

2) Add 6 AVI readers on major zones with large volume, leading to a network with 16 point sensors and 6 AVI readers, shown in Figure 3.11(d).

Figure 3.12 further compares the measure of uncertainty at different stages. It is interesting to observe that compared to the optimized scenario (from the scratch) with 16 sensors, this additional point sensors scenario (with 20 sensors) does not offer a superior uncertainty reduction performance for different OD pairs. On the other hand, compared to adding four point sensors to cover highly dynamic links, installing additional 6 AVI sensors does not further improve the estimation performance dramatically.

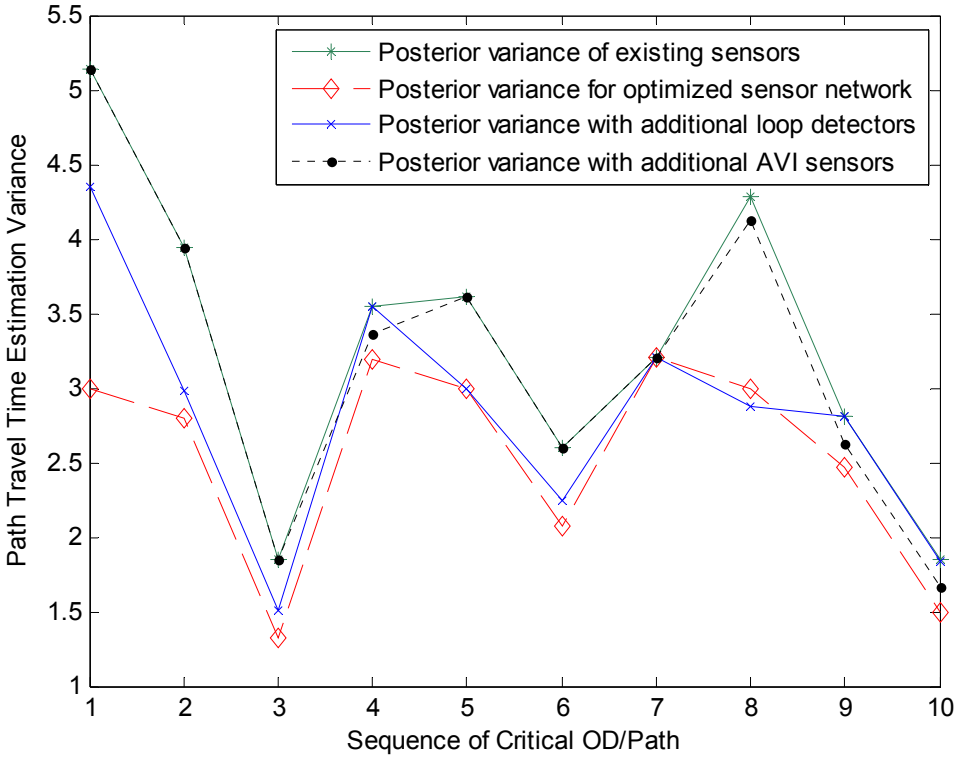
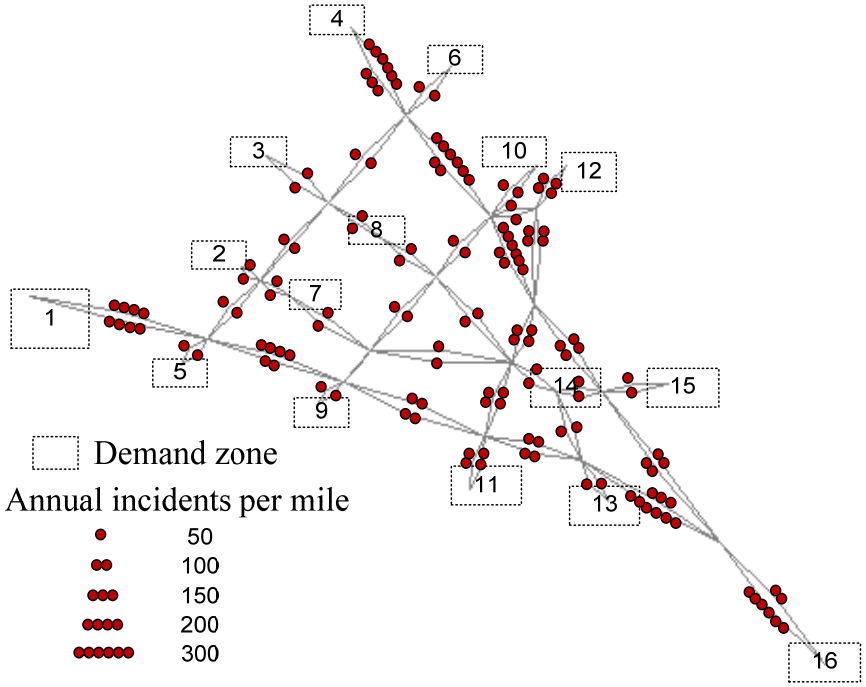


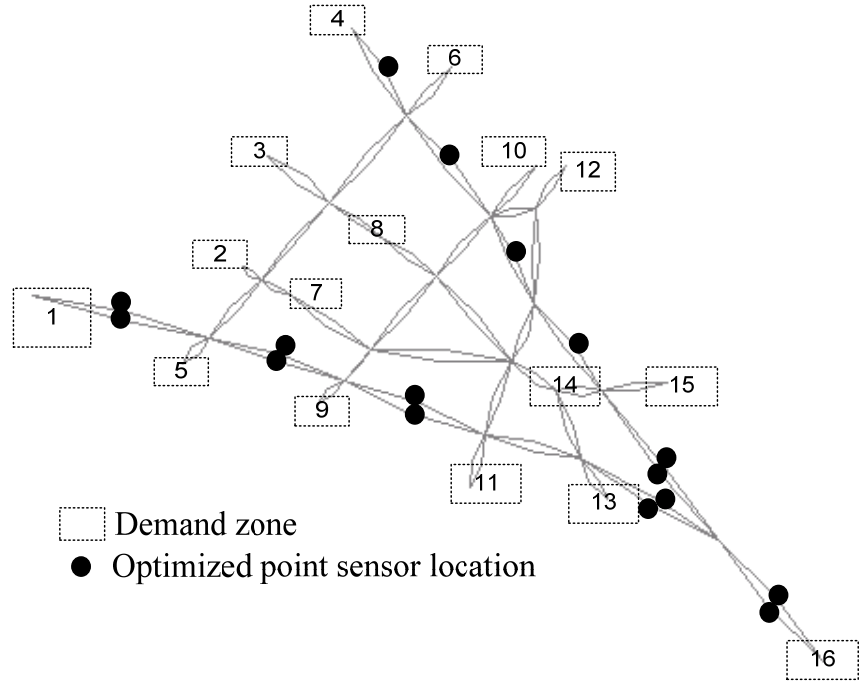
Figure 3.12: Comparison of different sensor location schemes (for regular traffic estimation)

3.7.3. Sensor Location Design for Traffic Prediction with Recurring and Nonrecurring Conditions

Now we perform the proposed algorithm by considering both regular and nonrecurring traffic conditions. As discussed in Section 3.4, the total traffic prediction uncertainty is computed as a probabilistic combination of recurring and nonrecurring uncertainties. In this numerical experiment, we take the incident as a demonstration example, with the link based incident rates shown in Figure 3.13(a). The proposed sensor location algorithm is applied in three scenarios: (1) optimized sensor network with 16 point sensors, (2) current network with additional four point sensors, and (3) current

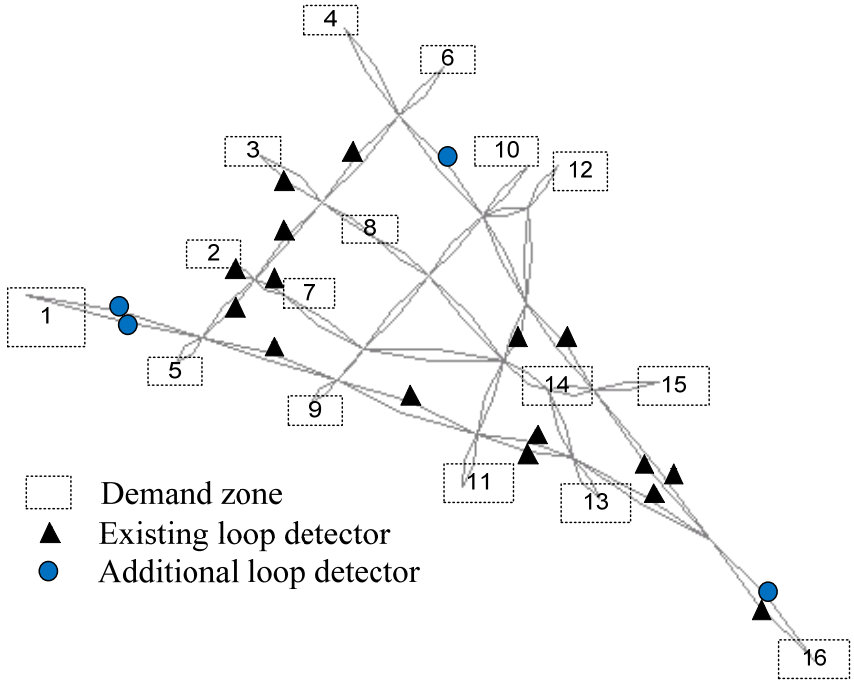


(a) Annual incident rate

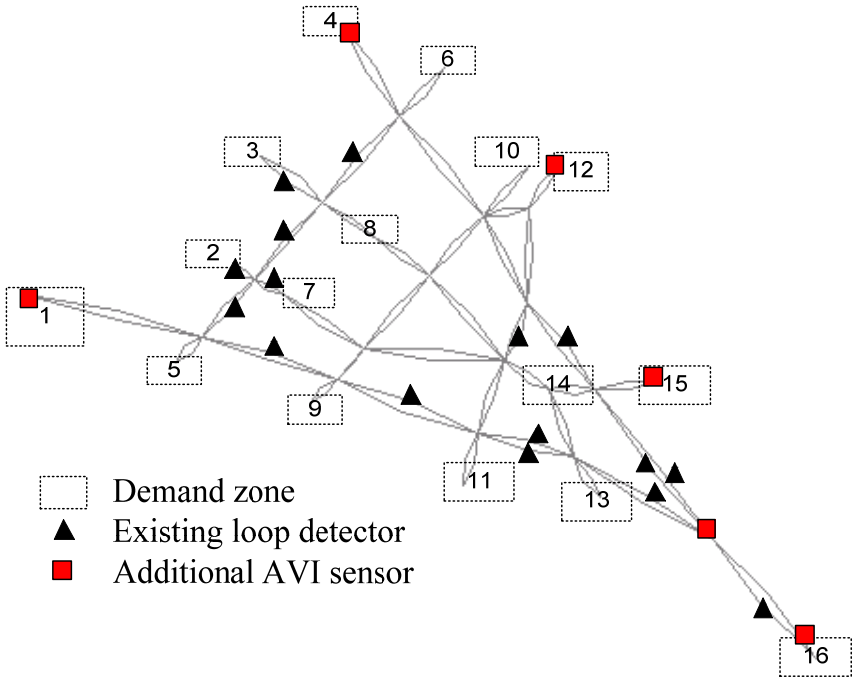


(b) Optimized sensor network locations
(start from a zero-sensor case)

Figure 3.13: Numerical experiment results for traffic prediction under recurring and nonrecurring traffic conditions



(c) Sensor network with 4 point sensors in addition to existing 16 point sensors



(d) Sensor network with 6 AVI sensors in addition to existing 16 point sensors

Figure 3.13, continued

network with additional 6 AVI sensors. Consequentially, these three sensor network design results are plotted in Figures 3.13(b-d). It is interesting to note that when considering nonrecurring traffic conditions (incidents), the optimized sensor locations (Figure 3.13b) are more focused on links with higher incident rates, compared to the regular pattern estimation based planning result in Figure 3.11(b).

CHAPTER 4

FINDING THE MOST RELIABLE PATH WITH AND WITHOUT LINK TRAVEL TIME CORRELATION

This dissertation investigates a fundamental problem of finding the most reliable path under different spatial correlation assumptions, where the path travel time variability is represented by its standard deviation. To handle the nonlinear and nonadditive cost functions introduced by the quadratic forms of the standard deviation term, a Lagrangian substitution approach is adopted to estimate the lower bound of the most reliable path solution through solving a sequence of standard shortest path problems. A subgradient algorithm is used to iteratively improve the solution quality by reducing the optimality gap. To characterize the link travel time correlation structure associated with the end-to-end trip time reliability measure, this research develops a sampling-based method to dynamically construct a proxy objective function in terms of travel time observations from multiple days. In specific, Section 4.1 provides the formal problem statement and briefly discusses two different models for the most reliable path problem: with and without link correlation. Focusing on each of the two different models, Sections 4.2 and 4.3 present the theoretical derivations and algorithmic development in detail, followed by

illustrative examples. Finally, Section 4.4 evaluates the performance of proposed algorithms through numerical experiments on a large-scale network.

4.1. Problem Statement and Model Assumptions

The following notation is used to represent variables in the problem formulation.

β = reliability coefficient

N = set of nodes

A = set of links

p = path index

m = link index in a path

c_p = travel time of path p

$\bar{c}_p$ = mean travel time of path p

i, j = subscripts for node index

l = subscript for the index of a link in a path, $l = 1, \ldots, m$

a_l = link in path p, with index l

a_{ij} = directed link from node i to j

c_l = travel time of link a_l

c_{ij} = travel time of link a_{ij}

$\bar{c}_l$ = mean travel time of link a_l

$\bar{c}_{ij}$ = mean travel time of link a_{ij}

$f(c_l)$ = probability distribution function of c_l

$f(c_{ij})$ = probability distribution function of c_{ij}

σ_{ij}^2 = variance of link travel time c_{ij}

x_{ij} = binary variable that indicates link a_{ij} is included in path solution if $x_{ij} = 1$

X = set of binary variables $\{x_{ij} \mid ij \in A\}$

D = set of travel time measurement samples

n = number of samples in set D

d = subscript for samples, $d = 1,\ldots,n$

$c_{p,d}$ = travel time of path p in sample d

$c_{l,d}$ = travel time of link l in sample d

$c_{ij,d}$ = travel time of link (i, j) in sample d

4.1.1. Problem Statement

Let G (N, A) represent a transportation network, where N is the set of nodes and A is the set of links. Each link can be denoted as either a directed link a_{ij} from node i to j, or an indexed link a_l in a path p with m links. Accordingly, the travel time of each link is denoted as c_{ij} or c_l. The travel time of each link (i, j) can be described as a random variable with probability distribution function $f(c_{ij})$ which has a mean of $\bar{c}_{ij}$ and a variance of σ_{ij}^2. Generally, the mean and variance of link travel times evolve considerably depending on the time of day and underlying traffic congestion levels. For simplicity, this study focuses on the static shortest path problem (to be used in static traffic assignment) and considers link travel times as time-invariant parameters in the underlying study horizon (e.g., morning peak hours).

Consider binary variable $x_{ij} \in \{0,1\}$ that indicates the selection of link a_{ij} for the optimal path solution, the least mean travel time problem for a prespecified OD pair (o, d) is described as

$$z^* = \min \sum_{ij \in A} \bar{c}_{ij} x_{ij} \tag{4.1}$$

Subject to the following flow balance constraints:

$$\sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = \begin{cases} 1 & i = o \\ 0 & i \in N - \{o, d\} \\ -1 & i = d \end{cases} \tag{4.2}$$

The above integer linear program (4.1) can be solved using regular label correcting or label setting shortest path algorithms (Ahuja et al., 1993).

In this dissertation, we formulate the most reliable path problem (P) as the following nonlinear integer programming problem by combining mean path travel time and its standard deviation in the objective function:

$$z^* = \min \sum_{ij \in A} \bar{c}_{ij} x_{ij} + \beta \sqrt{\mathrm{var}(c_p)}\ , \tag{4.3}$$

subject to

$$\sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b \tag{4.4}$$

where $b = \begin{cases} 1 & i = o \\ 0 & i \in N - \{o, d\} \\ -1 & i = d \end{cases}$ represents the flow status for each node i in the network, and β is the reliability coefficient, which reflects the significance of travel time variability. It can be derived as the ratio of Value of Reliability (VOR) and Value of Time (VOT). The reliability coefficient could also vary across different travelers and different trip purposes (e.g., business trip vs. recreational trip), and a typical value can be 1.27, calibrated by Noland et al. (1998).

Given link travel time statistics, the mean and variance of path travel time can be derived as

$$\begin{aligned} \bar{c}_p &= \sum_{l=1}^{m} \bar{c}_l \\ \mathrm{var}(c_p) &= \int_0^{+\infty} \left(c_p - \bar{c}_p\right)^2 f(c_p) dc_p \\ &= \int_{c_m=0}^{+\infty} \cdots \int_{c_2=0}^{+\infty} \int_{c_1=0}^{+\infty} \left(\sum_{l=1}^{m} c_l - \sum_{l=1}^{m} \bar{c}_l \right)^2 f(c_1, c_2, \ldots, c_m) dc_1 dc_2 \ldots dc_m \end{aligned} \tag{4.5}$$

Obviously, it is computationally intractable to obtain the multidimension probability distribution function $f(c_1, c_2, \ldots, c_m)$ for each path p, especially when spatial correlation exists. Along this line, two different approximation modeling approaches are proposed below to calculate the path travel time variance, with and without link travel time correlation assumptions.

4.1.2. Finding Most Reliable Path without Link Travel Time

Correlation: Independent Distribution based Model

To consider travel time variance in the most reliable path problem, one simplifying approach is to assume that there is no spatial correlation among travel times on different links. That is, by assuming the link travel time distributions are independent, we can reduce Eq. (4.5) to

$$\begin{aligned}\mathrm{var}(c_p) &= \sum_{l=1}^{m} \int_{c_l=0}^{+\infty} \left(c_l - \overline{c}_l\right)^2 f(c_1, c_2, \ldots, c_m) dc_l \\ &= \sum_{l=1}^{m} \mathrm{var}(c_l)\end{aligned} \tag{4.6}$$

so that the variance of the path travel time is now expressed as the sum of independent link travel time variances, which are relatively easy to measure based on historical traffic databases with multiday observations. Given independent link travel time distributions, the most reliable path problem (P) is then rewritten as

$$\begin{aligned}& z_1{*} = \min \sum_{ij \in A} \overline{c}_{ij} x_{ij} + \beta \sqrt{\sum_{ij \in A} \sigma_{ij}^2 x_{ij}} \\ & \text{s.t.} \quad \sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b\end{aligned} \tag{4.7}$$

4.1.3. Finding Most Reliable Path with Link Travel Time

Correlation: Sampling-based Model

In reality, travel times among different links could be highly correlated, e.g., due to the propagation of traffic congestion from a lane drop or merge bottleneck to its upstream links along a freeway or arterial corridor. In order to explicitly consider the link correlation in path travel time variable calculation, a sampling-based approximation method is adapted in this research to formulate the most reliable path problem.

According to the Monte Carlo method, a continuous stochastic distribution can be approximated as a discrete function by taking *n* samples from its population:

$$\begin{aligned}\mathrm{var}(c_p) &\approx \frac{1}{n-1}\sum_{d=1}^{n}\left(c_{p,d}-\overline{c}_p\right)^2 \\ &= \frac{1}{n-1}\sum_{d=1}^{n}\left(\sum_{l=1}^{m}c_{l,d}-\sum_{l=1}^{m}\overline{c}_l\right)^2\end{aligned} \tag{4.8}$$

That is to say, one can take *n* days' samples from a multiday historical traffic database and use them to directly calculate the sample variance and sample standard deviation of path travel time. By doing so, the inherent correlation among link travel times has been automatically represented by the sample set without explicitly requiring the variance-covariance matrix.

According to the sampling approach in Eq. (4.8), the most reliable path problem (P) that allows link travel time correlation is reformulated as

$$
\begin{aligned}
z_2^* = \min & \sum_{ij \in A} \bar{c}_{ij} x_{ij} + \beta \sqrt{\frac{1}{n-1} \sum_{d=1}^{n} \left(\sum_{ij \in A} c_{ij,d} x_{ij} - \sum_{ij \in A} \bar{c}_{ij} x_{ij} \right)^2} \\
\text{s.t.} \quad & \sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b
\end{aligned}
\tag{4.9}
$$

where $\bar{c}_{ij} = \frac{1}{n} \sum_{d=1}^{n} c_{ij,d}$ is the sample mean of link travel time.

Compared to the model assuming independent link travel time distributions, the proposed sampling-based method obviously requires a much larger number of measurements across different days to reduce sampling error and capture any possible spatial correlation. This approach fully recognizes link travel time correlation in calculating path travel time variance, but it also considerably complicates the path search process, especially when an acceptable level of approximation accuracy is needed. For real-world applications lacking sufficient link travel time measurements, the first model without link travel time correlation is still a viable option, but we also need to recognize that it might not find the most reliable path (i.e., optimal solution) in a network with possible spatial correlation.

For solving the two models proposed above, two separate lower bound algorithms are addressed with the similar Lagrangian relaxation-based approach in the next two sections. In particular, the complex quadratic parts of the two problems in Eq. (4.7) and (4.9) are first replaced with auxiliary variables and equivalent equality constraints. Then, based on Lagrangian substitution, the auxiliary constraints are further dualized into the simplified objective functions and lead to easy-to-solve and variable-independent subproblems. Specifically, the new set of subproblems contains one integer program with

linear and additive cost functions, one single-variable concave function that is solvable by checking the boundary values in the feasible region, and a set of convex subproblems particularly for the sampling-based model. Finally, the subgradient method is applied on both algorithms to update the lower bound.

4.2. Algorithm for Finding Most Reliable Path without Link Travel Time Correlation

In this section, an algorithm is presented for finding the most reliable path without assuming spatial dependency of link travel times. The Lagrangian relaxation-based model reformulation and derivation are first described in detail, followed by the subgradient method implementation and illustrative example demonstration.

4.2.1. Model Reformulation Using Lagrangian Substitution Method

As shown in the optimization program (4.7), the standard deviation of path travel time is a nonlinear, concave and nonadditive function of link travel time variance. The nonadditivity violates Bellman's Principle of Optimality, which forms the basis for standard label setting or label correcting shortest path algorithms. The overarching goal of the following model reformulation is to approximate the complicating nonadditive objective function by a linear additive cost function that is suitable for the regular shortest path algorithm. To remove the nonadditivity on decision variable x, we first introduce a nonnegative auxiliary variable y to the program (4.7) to convert the variance term to an equality constraint (4.11):

$$\min \sum_{ij \in A} \bar{c}_{ij} x_{ij} + \beta \sqrt{y} \tag{4.10}$$

$$\text{s.t.} \quad \sum_{ij \in A} \sigma_{ij}^2 x_{ij} = y \tag{4.11}$$

$$\sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b$$

$$0 \le y \le y'$$

where y' is the path travel time variance of the least expected travel time path.

Lemma 1: The feasible interval of optimal path travel time variance is $[0, y']$, i.e., between 0 and the variance of the least expected travel time path y'.

Proof: Shown in Figure 4.1, for any optimal solution to the most reliable path problem, its mean travel time is no less than that of the least expected travel time path. Moreover, it cannot have a path travel time variance greater than the variance of the least

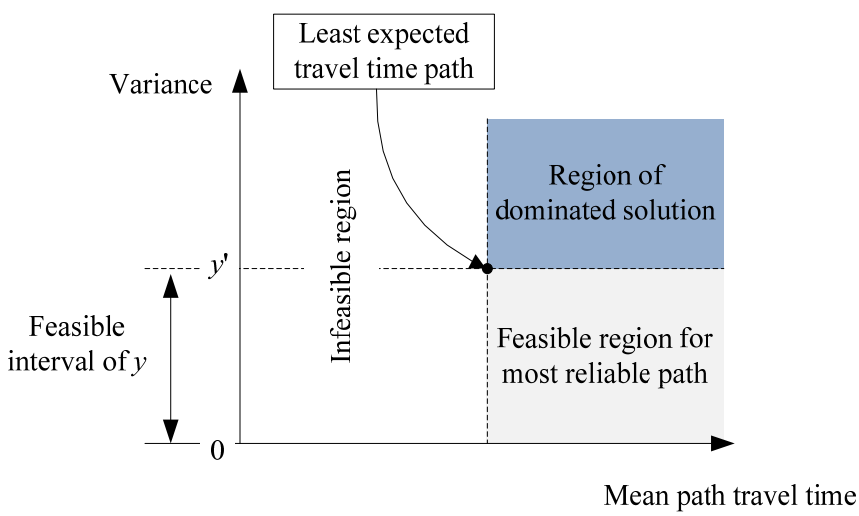


Figure 4.1: Feasible region of optimal path travel time variance y.

expected travel time path. Otherwise, this path has worse mean travel time and travel time variance compared to the least expected travel time path, which means that it is dominated by the least expected travel time path and should not be the optimal solution for optimization problem $\min\left\{\text{mean}+\beta\sqrt{\text{var}}\right\}$. Since the variance of path travel time is always greater than or equal to zero, the feasible region of y is an interval between 0 and the variance of the least expected travel time path y'.

After the reformulation in Eqs. (4.10) and (4.11), the original optimization program is transferred to a constrained shortest path problem with a linear cost function $\sum_{ij\in A}\bar{c}_{ij}x_{ij}$ and a univariate function $U(y)=\beta\sqrt{y}$ which is concave and monotonically increasing.

In a Lagrangian relaxation modeling framework, we can further relax the equality constraint (11) to an inequality (with a larger feasible region for x):

$$\sum_{ij\in A}\sigma_{ij}^2 x_{ij} \le y \tag{4.12}$$

and then introduce a Lagrangian multiplier $\mu \ge 0$ to bring the definitional linear constraint back to the original objective function (4.10):

$$\min \sum_{ij\in A}\bar{c}_{ij}x_{ij} + \beta\sqrt{y} + \mu\left(\sum_{ij\in A}\sigma_{ij}^2 x_{ij} - y\right) \tag{4.13}$$

$$\text{s.t.} \sum_{j:ij\in A} x_{ij} - \sum_{j:ji\in A} x_{ji} = b$$

$$0 \le y \le y'$$

By further regrouping variables, we can obtain the following Lagrangian dual function of the original primal problem (4.7),

$$L(\mu) = \min\left\{\sum_{ij\in A}\left(\bar{c}_{ij} + \mu\sigma_{ij}^2\right)x_{ij} + \beta\sqrt{y} - \mu y : \sum_{j:ij\in A} x_{ij} - \sum_{j:ji\in A} x_{ji} = b, 0 \le y \le y'\right\} \tag{4.14}$$

The Lagrangian function (4.14) is called a dual problem, in contrast to the primal problem (4.7), which can be divided and solved by two independent subfunctions:

$$L(\mu) = L_x(\mu) + L_y(\mu) \tag{4.15}$$

The first subfunction $L_x(\mu)$ involves a linear combination of primal variables x with a new link cost function $\bar{c}_{ij} + \mu\sigma_{ij}^2$, and the resulting additive shortest path problem can be solved efficiently using label correcting or label setting algorithms (Ahuja et al., 1993).

$$L_x(\mu) = \min\left\{\sum_{ij\in A}\left(\bar{c}_{ij} + \mu\sigma_{ij}^2\right)x_{ij} : \sum_{j:ij\in A} x_{ij} - \sum_{j:ji\in A} x_{ji} = b\right\} \tag{4.16}$$

The second subfunction $L_y(\mu)$ is a univariate concave minimization problem with respect to auxiliary variable y, and it can be solved through Lemma 1 by selecting the boundary value y'. The least expected travel time path can be obtained by solving the optimization problem with objective function (4.1) and constraint set (4.2).

$$L_y(\mu) = \min\left\{\beta\sqrt{y} - \mu y : 0 \le y \le y'\right\} \tag{4.17}$$

The optimal value of the concave function is $\min\left\{0, \beta\sqrt{y'} - \mu y'\right\}$ because the optimal value of a concave function is attained at one of the extreme points of the feasible region (a similar modeling technique was used in Larsson et al., 1994), as shown in Figure 4.2. Note that, for a general multidimensional concave minimization problem, its optimal solution is found by enumerating a large number of extreme points, while our proposed solution algorithm takes advantage of the fact that the simple concave function in Eq. (4.17) only involves a single variable.

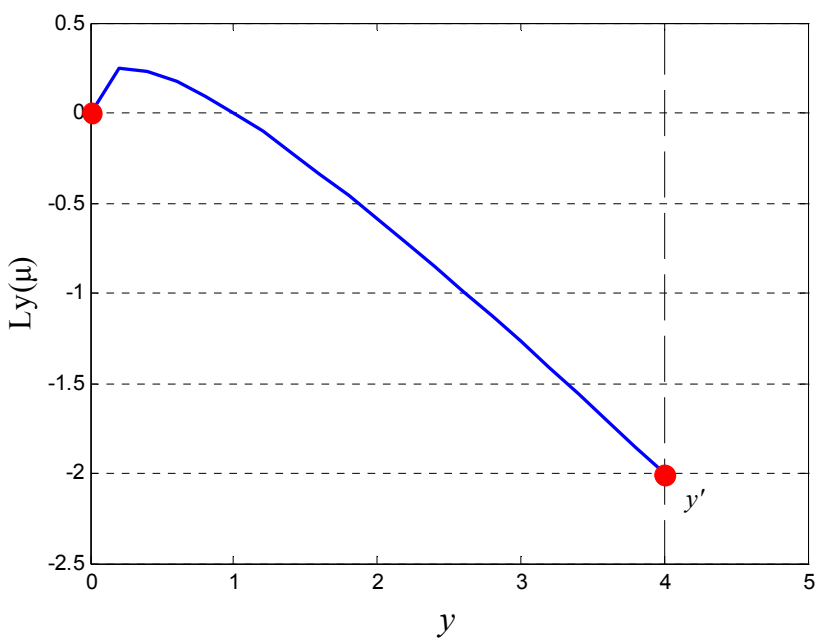


Figure 4.2: $L_y(\mu)$ as a function of y, and optimal value obtained at one of the extreme points.

In the Lagrangian dual problem in Eq. (4.14), essentially, the original nonadditive and nonlinear integer program (4.7) is approximated (lower-bounded) by a shortest path problem with linear and additive cost functions and another concave problem. In this new problem, the Lagrangian multiplier μ corresponds to the weight of the travel time variance in its reconstructed objective function, and its optimal value can be identified by some iterative search methods to be described below.

4.2.2. Subgradient Method

For each positive value of the Lagrangian multiplier μ, the corresponding value of the Lagrangian function $L(\mu)$ provides a lower bound to the optimal objective function value z_1* of the primal problem (4.7). Let us denote L* to be the maximum value of $L(\mu)$ according to μ:

$$L^* = \max_{\mu} L(\mu) \tag{4.18}$$

Let us define ε as the gap or tolerance level between the lower bound L* and the upper bound of the optimal solution, while the upper bound UB can be derived based on a feasible solution. Since the true optimal solution to the primal problem must have an objective value within the lower bound L* and the upper bound, the approximation error of the current best solution (corresponding to the upper bound UB) is no larger than the gap ε with respect to the primal optimal value z_1*. To reduce approximation error (UB – L*), the Lagrangian multiplier μ and the auxiliary variable y can be determined iteratively through the following subgradient method.

The search direction of μ is typically calculated from the subgradient of $L(\mu)$:

$$\nabla L(\mu) = \sum_{ij \in A} \sigma_{ij}^2 x_{ij} - y \tag{4.19}$$

Let k denote the iteration number in the subgradient method. Starting from any feasible initial choice of the Lagrangian multiplier, to update the Lagrangian multiplier μ^k at iteration k, the subfunctions (4.16) and (4.17) must be solved first, with solutions denoted as x_{ij}^k and y^k, respectively. With solutions at iteration k, we calculate the value of the Lagrangian multiplier as follows:

$$\mu^{k+1} = \mu^k + \theta^k \left(\sum_{ij \in A} \sigma_{ij}^2 x_{ij}^k - y^k \right) \tag{4.20}$$

A heuristic algorithm can be used to update the step size θ^k.

$$\theta^k = \frac{\lambda^k \left[UB - L(\mu^k) \right]}{\left\| \sum_{ij \in A} \sigma_{ij}^2 x_{ij}^k - y^k \right\|^2} \tag{4.21}$$

In this expression, UB is computed as the current best objective function value z_1* in the primal problem and can be updated iteratively to speed up the optimization process. λ^k is a scalar chosen between 0 and 2 to adjust the step-size of the process and make sure no negative value appears in the cost function. It should be noted that negative

Lagrangian multiplier values are not acceptable, and we can adjust the θ^k term in Eq. (4.20) to ensure the resulting multipliers are nonnegative.

As illustrated in Figure 4.3, the overall algorithm for solving the most reliable path problem without link travel time correlation is described below.

Algorithm 4.1:

Step 1: Initialization

Choose an initial Lagrangian multiplier $\mu > 0$;

Initialize iteration number $k = 0$;

Solve the least expected travel time path problem, set its objective function value as UB, and set the variance of the least travel time path as y'.

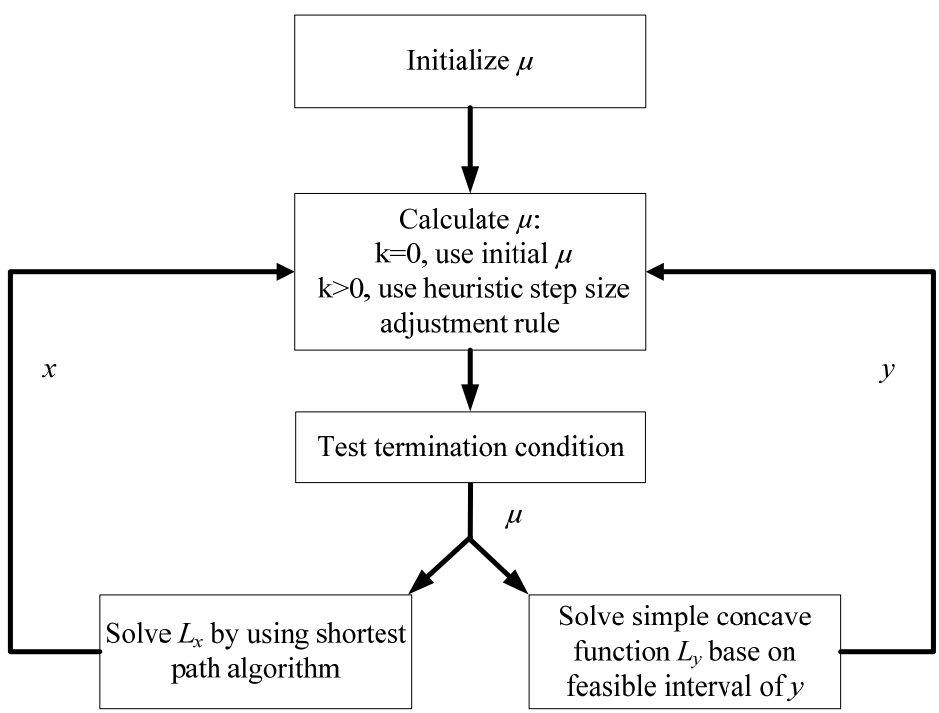


Figure 4.3: Subgradient algorithm for solving Lagrangian dual problems.

Step 2: Solve decomposed dual problems

Solve $L_x(\mu)$ using a standard shortest path algorithm;

Solve $L_y(\mu) = \min\left\{0, \beta\sqrt{y'} - \mu y'\right\}$;

Calculate primal, dual and gap values.

Step 3: Update Lagrangian multiplier

Calculate Lagrangian multiplier μ with Eqs. (4.20-4.21)

Step 4: Termination condition test

If $k > K_{\max}$ or the gap is smaller than the predefined toleration gap ε, terminate the algorithm, otherwise go back to Step 2. Here, $K_{\max}$ is a predefined number for the maximum of iterations.

4.2.3. Relative Gap and Optimal Solution

The duality gap is defined as the difference between the primal optimal value z_1* and the dual optimal value L*, and it offers an important metric for evaluating the performance of solution methods. In the proposed algorithm, the duality gap is the maximum approximation error range in the linear approximation approach taken by the LR lower bound method. This means that the distance between the optimal solution and the best solution found in the proposed algorithm is no larger than the duality gap. Specifically, the duality gap may contain two types of approximation error: (1) error of solution quality (distance between the true optimal solution and the proposed best solution) and (2) approximation error due to the limitation of LR. If the duality gap is equal to zero, then the primal solution corresponds to an optimal solution.

In practice, the real primal and dual optimal values cannot be achieved directly. Therefore, at each iteration of the above searching process, we update the tight upper bound of the primal problem with the shortest path solution of the linear integer subproblem in Eq. (4.16). Upon the termination of the algorithm, the minimal value of the primal problem serves as the tightest upper bound (*UB*), while the maximal value of the dual problem serves as the best lower bound (*LB*). To evaluate the solution quality, we can define a relative optimality measure as

$$\varepsilon' = \frac{UB - LB}{UB}, \tag{4.22}$$

while $\frac{UB - LB}{UB} \geq \frac{z_1 * - L *}{z_1 *}$, meaning that this relative gap is no less than the real gap between the optimal primal and dual values. With a reasonably small relative gap, we provide a satisfied solution quality guarantee on the suggested reliable routes. It is important to notice that, due to the nature of the approximation from the Lagrangian relaxation-based lower bound estimation method, there could still be a positive gap even if the optimal solution of the primal problem has been achieved.

4.2.4. Illustrative Example

Consider a single origin-destination pair with three parallel links/paths with the following path travel time data in Table 4.1. By simply comparing the values of objective function, it is obvious that path C is the most reliable path for objective function:

Table 4.1: Path travel time data

Path	Mean Travel Time	Travel Time Variance	Travel Time Standard Deviation	Value of Objective Function (β = 1)
A	35	0	0	35
B	29	49	7	36
C(opt)	**31**	**4**	**2**	**33**

$\text{mean} + \beta\sqrt{\text{var}}$. Figure 4.4 shows the relationship between the Lagrangian multiplier and the value of objective function.

When the Lagrangian multiplier is equal to 0.142, the best lower bound is found at 31.57, while the tightest upper bound is found at 33 for the primal problem. Path C, identified as the best solution by the proposed algorithm (with results in Figure 4.4), has a

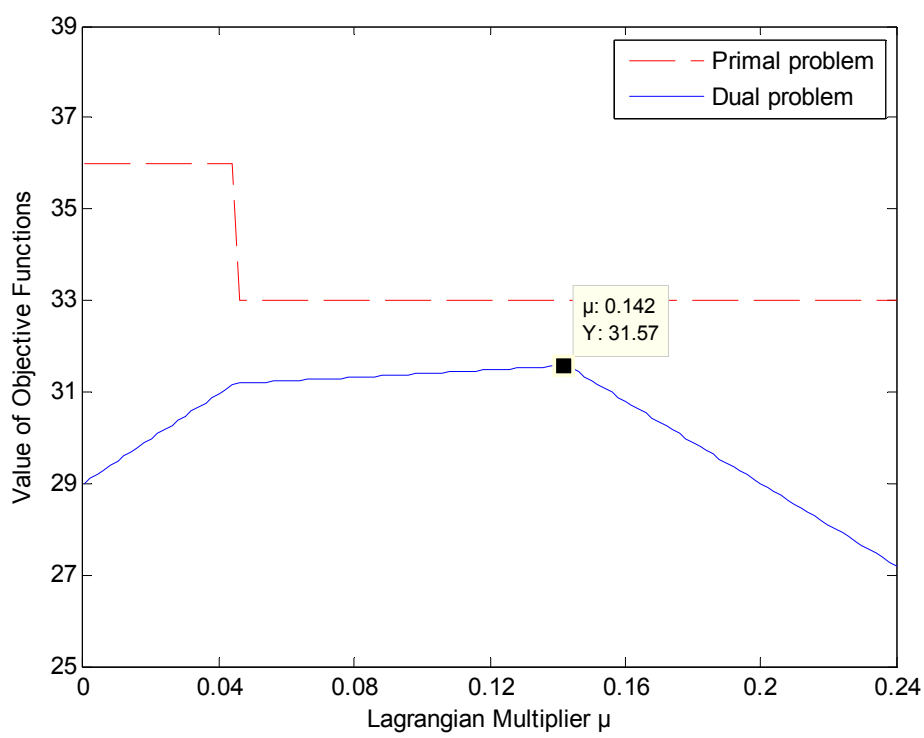


Figure 4.4: Solution results for independent distribution-based algorithm.

relative gap of 4.33% to the best lower bound (31.57). As mentioned previously, although path C is in fact the optimal solution in this example, there still exists a positive duality gap between primal and dual solutions, as the Lagrangian relaxation-based lower bound estimation method is, essentially, an approximation.

4.3. Algorithm for Finding Most Reliable Path with Link Travel Time Correlation

In last section, we presented the model reformulation and solution algorithm for the most reliable path problem without travel time correlation. In order to take the link travel time correlation into account, a Monte Carlo-based approximation method is used to propose the sampling-based model in Eq. (4.9). Along this line, given the same transportation network G (N, A), we construct a sample set D with n travel time measurements from the same time at the same day-of-week. The sample domain is denoted with the subscript d for variables. Similar to the independent distribution-based model, because of the nonlinear and nonadditive characteristics of the objective function, a Lagrangian substitution method is adapted here to solve the sampling-based model.

4.3.1. Lagrangian Substitution

In order to approximate the minimization problem (4.9) with a linear optimization problem, we implement a two-step Lagrangian relaxation approach with two sets of auxiliary variables:

$$\sum_{ij\in A} c_{ij,d} x_{ij} - \sum_{ij\in A} \bar{c}_{ij} x_{ij} = w_d \qquad \forall d \in D \tag{4.23}$$

$$\frac{1}{n-1}\sum_{d=1}^{n} w_d^2 = y \tag{4.24}$$

After relaxing the equality constraints in Eq. (4.23) and (4.24) with inequality constraints, the minimization problem can be reformulated as

$$z_2^* = \min \sum_{ij \in A} \bar{c}_{ij} x_{ij} + \beta\sqrt{y} \tag{4.25}$$

s.t.

$$\sum_{ij \in A} c_{ij,d} x_{ij} - \sum_{ij \in A} \bar{c}_{ij} x_{ij} \le w_d \qquad \forall d \in D \tag{4.26}$$

$$\frac{1}{n-1}\sum_{d=1}^{n} w_d^2 \le y \tag{4.27}$$

$$\sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b$$

$$0 \le y \le y'$$

In this reformulation, n+1 auxiliary variables are introduced. The variable w_d for each sample d represents the difference between the mean path travel time and the path travel time on sample d. The variable y corresponds to the average path travel time deviation between samples and the sample mean, which defines the path travel time variance.

Lemma 1 still holds when considering link travel time correlations for the most reliable path problem with time-invariant travel times. As the mean travel time of the

least expected travel time path is always the minimum among all the paths, the variance of the least expected travel time path still serves as an upper bound for the reliability component of the objective function. It should be remarked that, in order to approximate the unknown (true) spatial travel time correlation structure, this study uses multiday samples from the historical database and Eq. (4.8) to calculate the sample variance of the path travel time, instead of using the variance-covariance structure shown in Eq. (4.5).

To further remove constraints in Eq. (4.26) and (4.27), a set of positive Lagrangian multipliers, denoted as μ_d and ν sequentially, is introduced in order to move the explicit inequality constraints into the objective function (4.25):

$$\begin{aligned} &L(\mu_1,...,\mu_n,\nu) \\ &= \min\left\{\sum_{ij\in A}\overline{c}_{ij}x_{ij} + \beta\sqrt{y} + \sum_{d=1}^{n}\mu_d\left(\sum_{ij\in A}c_{ij,d}x_{ij} - \sum_{ij\in A}\overline{c}_{ij}x_{ij} - w_d\right) + \nu\left(\frac{1}{n-1}\sum_{d=1}^{n}w_d^2 - y\right)\right. \\ &\qquad \left. : \sum_{j:ij\in A}x_{ij} - \sum_{j:ji\in A}x_{ji} = b, 0 \le y \le y'\right\} \end{aligned} \tag{4.28}$$

By regrouping the variables, we will have a clearer view on the components of the dual problem:

$$\begin{aligned} &L(\mu_1,...,\mu_n,\nu) \\ &= \min\left\{\sum_{ij\in A}\left[\left(1-\sum_{d=1}^{n}\mu_d\right)\overline{c}_{ij} + \sum_{d=1}^{n}\mu_d c_{ij,d}\right]x_{ij} + \sum_{d=1}^{n}\left(\frac{1}{n-1}\nu w_d^2 - \mu_d w_d\right) + \beta\sqrt{y} - \nu y\right. \\ &\qquad \left. : \sum_{j:ij\in A}x_{ij} - \sum_{j:ji\in A}x_{ji} = b, 0 \le y \le y'\right\} \end{aligned} \tag{4.29}$$

The dual function (4.29) has a linear objective function corresponding to the primal variable *X*. For each link, the cost function is a combination of weighted sample travel times on different days. By adjusting the Lagrangian multipliers μ_d and ν, we may iteratively construct an optimal linear cost function that will maximize the dual function and, more specifically, best approximate the nonlinear objective function in the primal problem.

4.3.2. Dual Function Decomposition

We decompose the dual function (4.29) into a set of subfunctions:

$$L_x(\mu_1,...,\mu_n,\nu) = \min\left\{\sum_{ij\in A}\left[\left(1-\sum_{d=1}^{n}\mu_d\right)\overline{c}_{ij} + \sum_{d=1}^{n}\mu_d c_{ij,d}\right]x_{ij} : \sum_{j:ij\in A} x_{ij} - \sum_{j:ji\in A} x_{ji} = b\right\} \quad (4.30)$$

$$L_{w_d}(\mu_1,...,\mu_n,\nu) = \min\left\{\frac{1}{n-1}\nu w_d^2 - \mu_d w_d\right\} \qquad \forall d \in D \quad (4.31)$$

$$L_y(\mu_1,...,\mu_n,\nu) = \min\left\{\beta\sqrt{y} - \nu y : 0 \le y \le y'\right\} \quad (4.32)$$

The first subfunction (4.30) can be easily solved using shortest path algorithms. Notice that in this subshortest path problem, the cost function for each link is a weighted combination of mean travel time and travel time of each day. In other words, the set of Lagrangian multiplier μ_d indicates the weights of the travel time variance for each sample. By adjusting the set of multipliers, the influences of different samples on the overall path travel time reliability measure are evaluated and assessed systematically at each iteration.

The second subfunction set (4.31) contains one convex minimization problem for each auxiliary variable w_d and can be solved using the first-order gradient:

$$\frac{\partial L_{w_d}(\mu_1,...,\mu_n,\nu)}{\partial w_d} = \frac{2}{n-1}\nu w_d - \mu_d = 0$$
$$w_d = \frac{\mu_d(n-1)}{2\nu} \tag{4.33}$$

The third subfunction (32) is a concave minimization problem for variable y. Since y represents the variance of the path travel time, the feasible region is between zero and the variance of the path with least travel time, and the minimization point locates at one of the extreme points of the feasible region, i.e., $L_y(\mu_1,...,\mu_n,\nu) = \min\left\{0, \beta\sqrt{y'} - \nu y'\right\}$.

4.3.3. Subgradient Method

The subgradient method is also used in the second algorithm considering spatial correlation. The search direction for each Lagrangian multiplier is found using the following equations.

$$\nabla L(\mu_1,...,\mu_n,\nu) = \left(\sum_{ij\in A}\left(c_{ij,1} - \overline{c}_{ij}\right)x_{ij} - w_1,...,\sum_{ij\in A}\left(c_{ij,n} - \overline{c}_{ij}\right)x_{ij} - w_n, \frac{1}{n-1}\sum_{d=1}^{n} w_d^2 - y\right) \tag{4.34}$$

$$\mu_d^{k+1} = \mu_d^k + \theta_{\mu_d}^k\left(\sum_{ij\in A}\left(c_{ij,d} - \overline{c}_{ij}\right)x_{ij}^k - w_d^k\right) \qquad \forall d \in D \tag{4.35}$$

$$\nu^{k+1} = \nu^k + \theta_\nu^k\left(\frac{1}{n-1}\sum_{d=1}^{n}\left(w_d^k\right)^2 - y^k\right) \tag{4.36}$$

The step size of each iteration k is calculated using heuristic algorithms:

$$\theta_{\mu_d}^k = \frac{\lambda_{\mu_d}^k \left[L_{UB}(\mu_1,...,\mu_n,\nu) - L(\mu_1^k,...,\mu_n^k,\nu^k) \right]}{\left\| \sum_{ij\in A} \left(c_{ij,d} - \overline{c}_{ij} \right) x_{ij}^k - w_d^k \right\|^2} \qquad \forall d \in D \tag{4.37}$$

$$\theta_{\nu}^k = \frac{\lambda_{\nu}^k \left[L_{UB}(\mu_1,...,\mu_n,\nu) - L(\mu_1^k,...,\mu_n^k,\nu^k) \right]}{\left\| \frac{1}{n-1} \sum_{d=1}^{n} \left(w_d^k \right)^2 - y^k \right\|^2} \tag{4.38}$$

Similar to Algorithm 4.1, an iterative algorithm for solving the most reliable path problem with link correlations are written as the following.

Algorithm 4.2:

Step 1: Initialization

Choose initial values for the set of positive Lagrangian multipliers, μ_d and ν;

Initialize iteration number k = 0;

Solve the least expected travel time path problem, set its objective function value as *UB*, set variance of the least travel time path as y'.

Step 2: Solve decomposed dual problems

Solve the first subfunction (4.30) using a standard shortest path algorithm;

Solve the second subfunction set (4.31) using Eq. (4.33);

Solve the third subfunction (4.32) with $L_y(\mu_1,...,\mu_n,\nu) = \min\left\{0, \beta\sqrt{y'} - \nu y'\right\}$;

Calculate primal, dual and gap values.

Step 3: Update Lagrangian multipliers

Calculate Lagrangian multipliers with Eqs. (4.35-4.38)

Step 4: Termination condition test

If $k > K_{\max}$ or the gaps are smaller than the predefined toleration gap, terminate the algorithm, otherwise go back to Step 2.

It is possible that the link cost term $\left[\left(1-\sum_{d=1}^{n}\mu_d\right)\overline{c}_{ij}+\sum_{d=1}^{n}\mu_d c_{ij,d}\right]$ in Eq. (29) becomes negative for certain conditions of μ_d. Although the label correcting algorithm can handle negative link costs, in order to avoid detecting and handling possible negative cycles in the resulting shortest path problem, our implementation uses the following rule for simplicity: when a negative link cost occurs, the values of the Lagrangian multiplier step size θ in Eq. (4.35) are adjusted proportionally until all link costs in the network are nonnegative.

This proposed algorithm (for adjusting multipliers to force a nonnegative cost) can be viewed as a special version of the subgradient projection method (Bertsekas, 1999), where the Lagrangian multipliers are projected to a feasible search direction in order to maintain the nonnegative link costs. Different from a regular subgradient projection method which only adjusts variables with infeasible values, the proposed problem is quite complex as each link cost is associated with all the Lagrangian multipliers μ (in Eq. 4.30). Therefore, our implementation uses the above heuristic rule to adjust the values of the Lagrangian multiplier step size θ to avoid negative link costs while maintaining the original search direction for each Lagrangian multiplier.

This adjustment method cannot guarantee to generate iterates with decreasing function values of the dual model at each iteration, but descent is more likely with the diminishing step size rule. Interested readers are referred to the dissertation by Nedic (2002) for an overview of the subgradient projection method and related convergence analysis in a Lagrangian optimization framework.

Furthermore, to avoid getting stuck in a suboptimal solution under a certain set of Lagrangian multipliers, we also use a multistart global optimization technique to randomly generate a new set of Lagrangian multipliers to restart the search process.

4.3.4. Illustrative Example

Consider a single origin-destination pair with three parallel paths, as shown in Figure 4.5. All three paths share a common link A, with different spatial correlations within each path.

As shown in Table 4.2, Path 1 is the least travel time path (3.75 min), with a positive correlation of 0.125 between two links; Path 3 is the most reliable path with negative correlation -0.25 between two links. In comparison, the two links on Path 2 have the

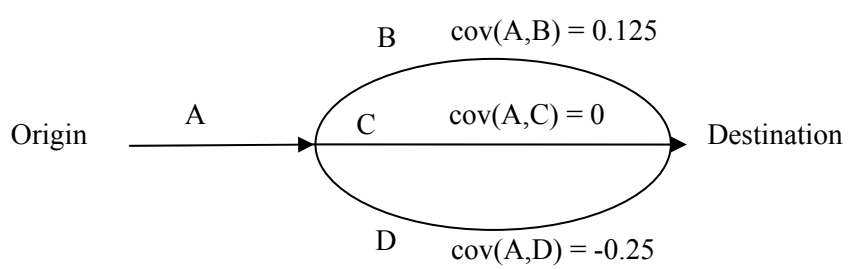


Figure 4.5: Network of illustrative example

Table 4.2: Travel time samples of each path (min)

Path	Travel Time	Day 1	Day 2	Day 3	Day 4	Ave.	Var.	Cov.	Path Var. No Corr	Path Var. with Corr., Covariance matrix based method	Path Var. with Corr. Sampling based method	Utility with Corr.
Path 1: positive corr., least travel time path	A	2	3	2	3	2.5	0.25	0.125	0.44	0.25 + 0.19 + 2 × 0.125 = 0.69	0.69	4.58
	B	1	1	1	2	1.25	0.19					
	Path total	3	4	3	5	3.75	0.69					
Path 2: no corr.	A	2	3	2	3	2.5	0.25	0	0.5	0.25 + 0.25 = 0.5	0.5	4.71
	C	2	2	1	1	1.5	0.25					
	Path Total	4	5	3	4	4	0.5					
Path 3: negative corr., most reliable path	A	2	3	2	3	2.5	0.25	-0.25	0.5	0.25 + 0.25 –2 × 0.25 = 0	0	4
	D	2	1	2	1	1.5	0.25					
	Path Total	4	4	4	4	4	0					

same mean and variance as those on Path 3, but with a zero spatial correlation. Now we systematically compare the following three approaches to computing the path variance.

(1) Without considering spatial correlation, the path variance is calculated directly from the summation of link variances, e.g., $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)$. This approach finds Path 1 as the most reliable path.

(2) Using the link covariance matrix, the path variance is calculated with Eq. (4.39). The best path found with the consideration of link travel time correlation is Path 3.

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\,\mathrm{cov}(X,Y) \tag{4.39}$$

(3) Using the sampling approach proposed in this dissertation, the path variance is calculated through sample path travel times (Eq. 4.8). This method finds Path 3 to be the most reliable path. Notice that, in this example, we consider the four sample days as the entire population, so the path variance is calculated in terms of population variance, e.g., $\text{PathVar} = \frac{1}{4}\sum_{day=1}^{4}\left(\text{PathTT}_{day} - \text{mean}\right)^2$. In practice, in order to achieve an unbiased estimation of the population, sample variance and sample standard deviation statistics should be calculated as shown in Eq. (4.8).

Table 4.2 shows that the path standard deviation computed through the proposed sampling approach leads to the same result as the method based on the analytical link variance-covariance matrix. In other words, the spatial correlation can be incorporated into the path travel time standard deviation measure directly from samples. It should be pointed out that the covariance matrix-based method may require a large amount of memory to store the link-to-link correlation values, and, more importantly, it is difficult to be directly embedded into standard shortest path algorithms.

Now we apply the proposed sampling-based approach in Algorithm 4.2 to find the most reliable path in the sample network. Table 4.3 shows some key intermediate computational results in the first few iterations of the search procedure.

In this example, by disseminating different weights μ on different samples, the proposed approach successfully uncovers the optimal solution (Path 3) for the most reliable routing problem. Specifically, starting with uniform distributed values (1/4 = 0.25), the weight set are adjusted (Eqs. 4.34, 4.35, 4.37) to improve the lower bound of the optimal solution.

Table 4.3: Results of first few iterations for Algorithm 4.2

K	μ				Lx			Lx	Lw	Ly	L	LB	UB	ε	ε' (%)
	1	2	3	4	1	2	3								
1	0.25	0.25	0.25	0.25	3.75	4.00	4	3.75	-1.88	0	1.88	1.88	4.58	2.70	59.05
2	0.19	0.17	0.19	0.14	3.69	3.98	4	3.69	-0.80	0	2.89	2.89	4.58	1.69	36.99
3	0.11	0.05	0.11	0.01	3.61	3.94	4	3.61	-0.13	0	3.47	3.47	4.58	1.11	24.18
4	0.05	0.01	0.05	0.11	3.81	3.96	4	3.81	-0.05	0	3.76	3.76	4.58	0.82	17.86
5	0.01	0.32	0.01	0.20	4.06	4.31	4	4.00	-0.17	0	3.83	3.83	4.00	0.17	4.22
6	0.01	0.30	0.01	0.21	4.07	4.29	4	4.00	-0.15	0	3.85	3.85	4.00	0.15	3.78
7	0.01	0.29	0.01	0.22	4.08	4.28	4	4.00	-0.14	0	3.86	3.86	4.00	0.14	3.43

Notations in Table 4.3:

K: number of iterations.

μ: a set of Lagrangian multipliers to approximate the original nonadditive and concave objective function with a linear combination, generated iteratively with Eqs. (4.35) & (4.37).

Lx1, Lx2, Lx3: subLagrangian function values for each path (Eq. 4.30). *Lx* is the cost of the shortest path using a linear link cost function.

Lw: a summation of the equation set Eq. (4.31).

Ly: the calculated value from Eq. (4.32).

L: the value of the dual problem for each iteration (Eq. 4.29).

LB: lower bound of the solution, obtained from the best dual value *L*.

UB: upper bound of the solution, generated from the best primal value among the paths uncovered by *Lx*; the corresponding path is the most reliable path found at current iteration.

ε : the gap between *UB* and *LB*.

ε' : relative gap, as defined in Eq. (4.22).

As shown in Table 4.3, the linear cost function-based shortest path problem *Lx* finds Path 1 during the first 4 iterations, while the lower bound *LB* is continually improving. Starting from iteration 5, *Lx3* achieves the minimum value and the algorithm finds Path 3 as the best solution with the lowest upper bound value, although the lower bound is still slightly increasing and the relative solution quality gap remains under 4%.

Figure 4.6 shows the evolution of *LB* and *UB* in the first few iterations of the above example. Notice that, although the optimal solution was found, there still exists a quality gap between upper bound and lower bound. This is due to the approximate nature of the Lagrangian Relaxation method.

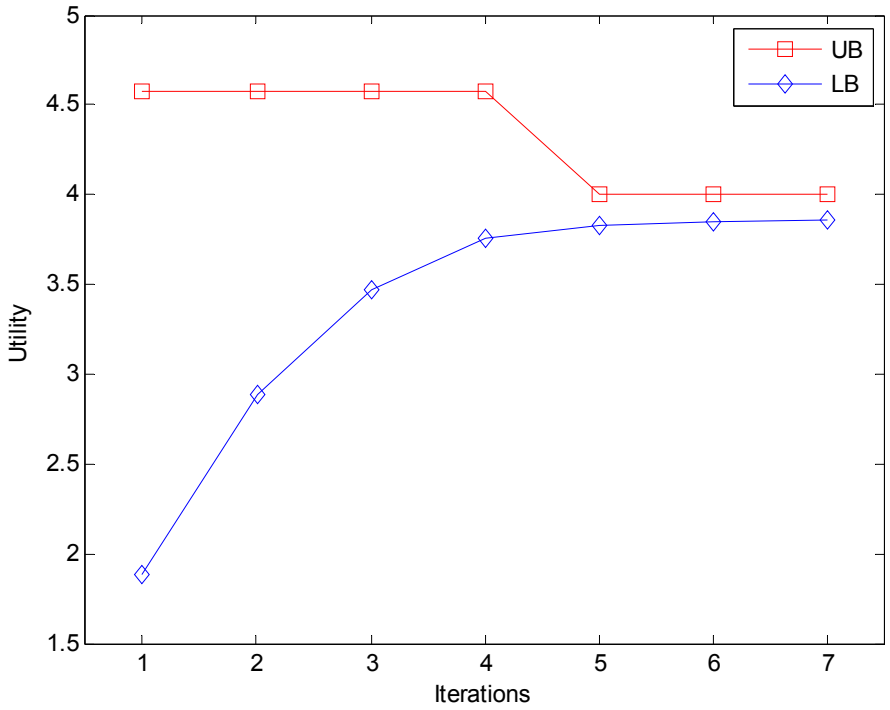


Figure 4.6: Evolution of lower bound (*LB*) and upper bound (*UB*) in the first a few iterations of the example

It is interesting to note that, in this illustrative example, the *Lx* value for Path 3 does not change when the μ values are modified. This is because that path travel times for different days on Path 3 are the same (4 min), which result in a zero variance in the linear cost function *Lx*.

4.4. Numerical Experiments

4.4.1. Test Network Overview

In this section, numerical experiments are conducted on a large-scale real-world transportation network for the Bay Area, California, which is comprised of 53,124 nodes and 93,900 links (Figure 4.7). Specifically, 8,511 links (9.1% of links) of the entire network are freeways with a total length of 1,774.8 miles (i.e., 15.8% of the total mileage); while 85,389 links (90.9%) are arterial roads with a total length of 9,431.8 miles (84.2%). The algorithm is implemented in C# on the Windows Vista platform and evaluated on a personal computer with an Intel Core Duo 1.8GHz CPU and 2 GB memory.

For the first model without considering link travel time correlation, the mean travel time and travel time variance for each link are calculated from available historical travel time records from the NAVTEQ traffic database, and the data used in this study (mainly from freeway segments) cover about 4.1% of the total mileage in the Bay Area as in Figure 4.7. Note that the underlying network includes a large number of major and minor arterial streets which do not have temporally continuous traffic observations for us to calculate variability statistics. For the second sampling-based model that recognizes spatial correlation, 73 days of travel time measurements between November 2009 and

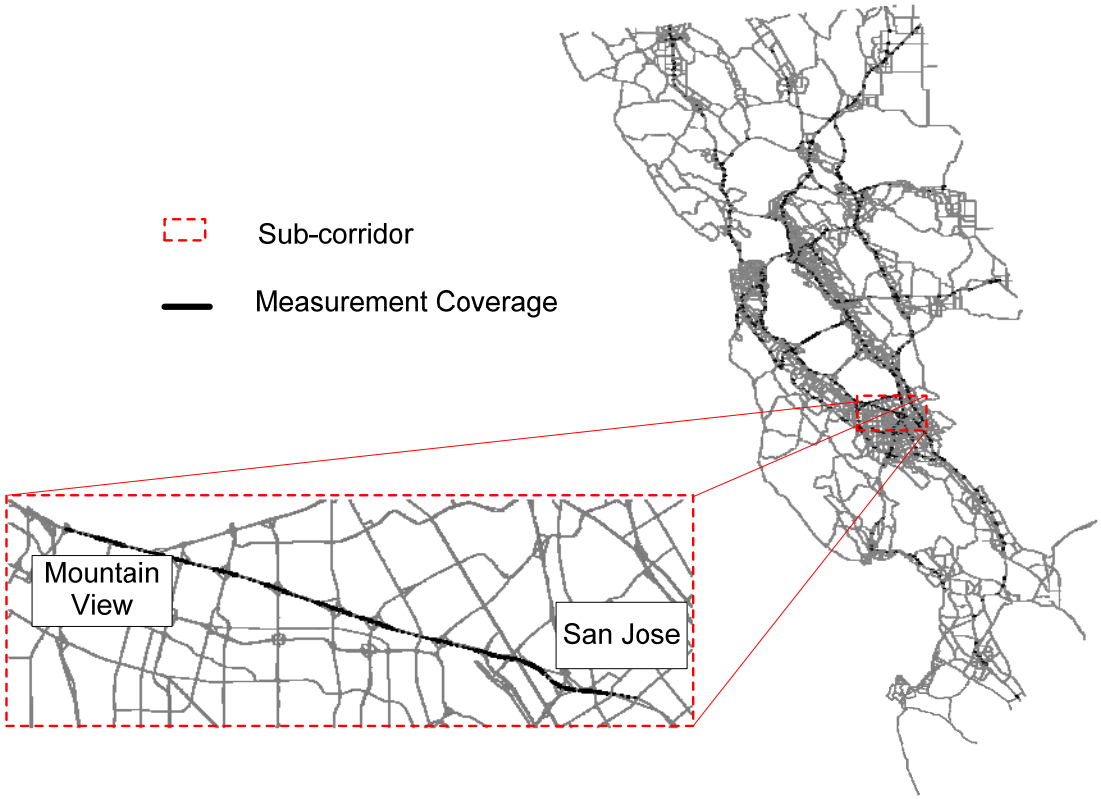


Figure 4.7: Measurement coverage of travel time for Bay Area, California. Links with measurements are highlighted. A subcorridor on Bayshore Freeway from Mountain View to San Jose, California is enlarged.

February 2010 are collected for the time interval of 9:00 AM to 9:15 AM of each sample day. For simplicity, this research does not remove traffic data from weekend days and holidays.

4.4.2. Spatial Correlation Analysis

This study extracts and examines the data from a segment of the Bayshore Freeway between Mountain View and San Jose, California. On this 11-mile freeway corridor, 6 miles of links have travel time measurements, as enlarged in Figure 4.7. The spatial link correlation is visualized in Figure 4.8, where the planar xy coordinates represent the link

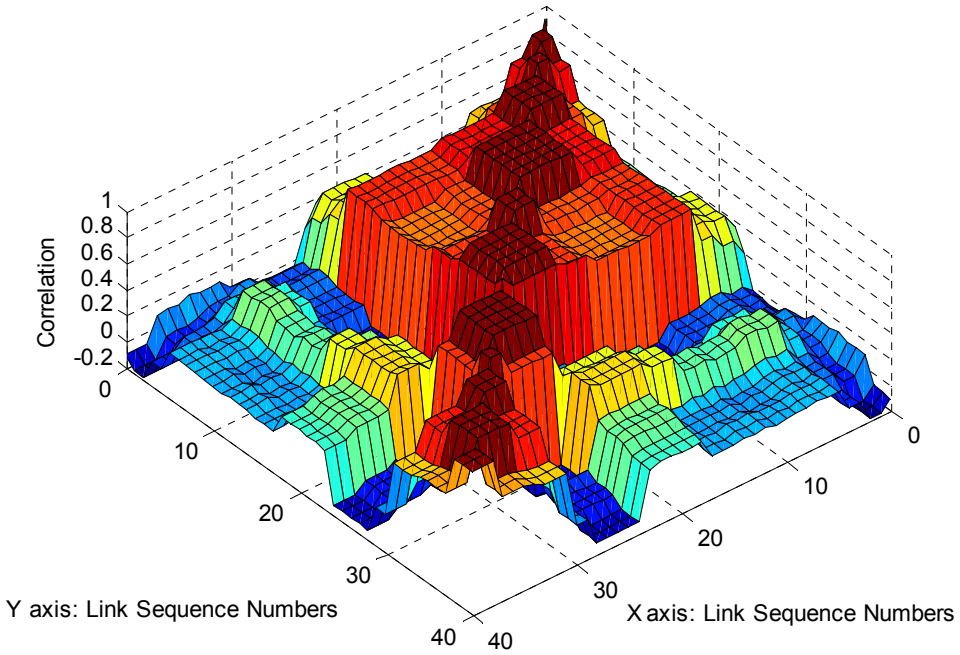


Figure 4.8: Partial link correlations on selected subcorridor

sequence numbers along the corridor and the vertical z axis shows the value of correlation. As a simple verification test, the reliable routing algorithm for the model with link correlation is carried out from the origin to the destination of this subcorridor, leading to a relative gap of 3.5%.

In our experiments, in order to correctly model the randomness of link travel time for links without measurements, random travel time values from a Normal distribution are generated in order to incorporate their variance. It is important to recognize that, if only single mean travel time values are used for links without measurements, then the calculated path travel time reliability measure would assume zero variability for those links, leading to potentially biased solutions. In this research, the travel time index

(calculated as the travel time divided by the free flow travel time of each link) is calculated first to capture the variance of the travel time index for links with measurements. The corresponding average variance of the travel time index is then applied to links without measurements to generate random travel times for each sampling day. It should be also noted that the randomly generated travel times in the above procedure are inherently independent and uncorrelated across different links, so the calculated path travel time reliability without complete data coverage is likely to be lower than the actual end-to-end travel time reliability measure experienced by commuters.

4.4.3. Numerical Performance and Solution Quality Analysis

As short-distance OD pairs might be covered by no or inadequate raw observations, and they typically have very limited alternative routes to examine, this study particularly imposes the following rules to select OD pairs to be tested: (1) the average path travel time is larger than 45 minutes, and (2) the measurement coverage on the least expected travel time path is larger than 30% in distance. As a result, an OD-pair set S containing s = 246 random OD pairs is generated from the Bay Area network.

The performance of our approach is assessed using the average relative gap, which is calculated as the average value of all 246 OD pairs under a predefined maximum number of iterations $K_{\max}$ and reliability coefficient β in Eq. (3), e.g., $\bar{\varepsilon} = \frac{\sum_{od \in S} \varepsilon_{od,\beta,K_{\max}}}{s}$. To consider the impact of the reliability coefficients on the experiment results, this study considers two sets of reliability coefficients: β = 1.27 as suggested in the travel time reliability study by Noland et al. (1998), and β = 4 as a comparatively large value for

testing purposes. The travel time reliability coefficient represents how much travel time reliability (in terms of standard deviation) the trip-maker is willing to trade for unit time saving.

In this dissertation we adopt a utility function of travel time reliability from the study by Noland et al. (1998). They performed a state-preference survey that used a sample set of more than 700 commuters in the Los Angeles region to empirically estimate the user preferred value of travel time reliability. Specifically, among the 543 valid questionnaires collected in their survey, each respondent was ask nine stated preference questions each with two commute routes under different distributions of travel times and departure times. A value of 1.27 was calibrated in their study for the travel time reliability coefficient β. Additionally, we also test $\beta = 4$ in our study, by considering that some road users, such as commercial fleet companies, have larger values of travel time reliability, compared to regular commuters.

As shown in Figures 4.9 and 4.10, the average gap decreases along with the increase of the predefined maximum number of iterations $K_{\max}$. To characterize the statistics distribution of the optimality gap measure, the standard deviation of relative gaps under variant $K_{\max}$ and β are calculated in Figure 4.11. Overall, the relative gap for the reliable routing algorithm without considering link travel time correlation is dramatically lower than the gap for the algorithm with correlation consideration, e.g., 1.7% vs. 5.4% when $\beta = 1.27$. It should be remarked that this difference does not indicate that the first algorithm is superior to the second one, as these two algorithms use different primal and dual objective functions (with vs. without spatial correlation).

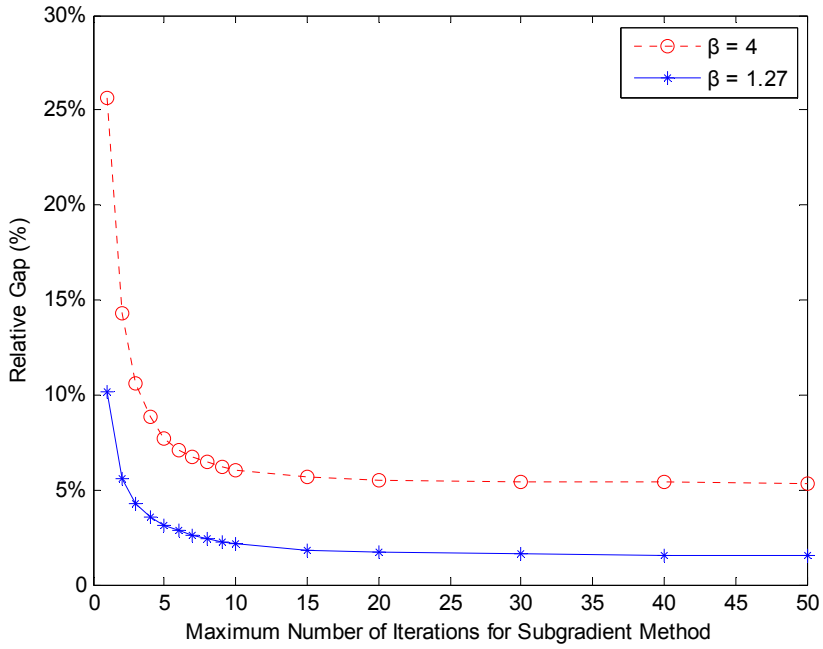


Figure 4.9: Relative gap in model WITHOUT link travel time correlation. Beta is the travel time reliability coefficient.

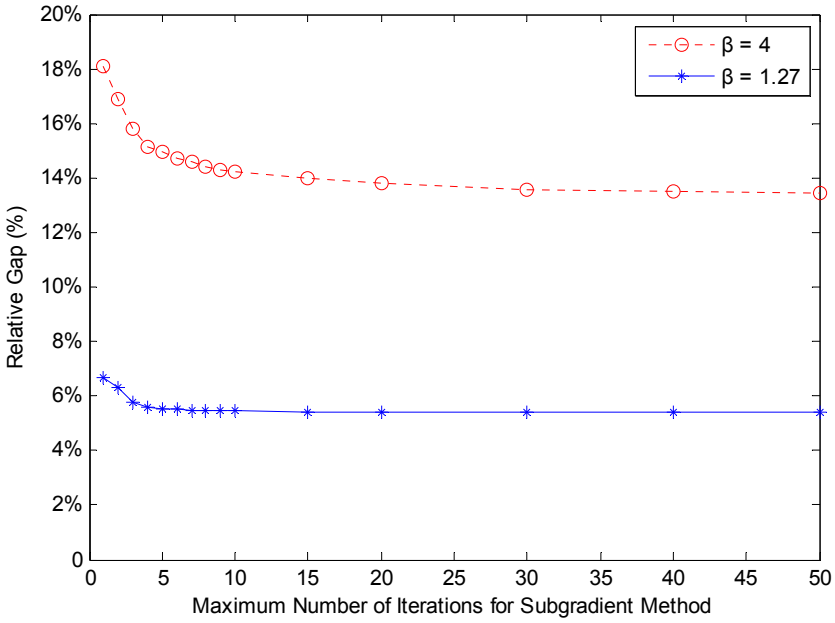


Figure 4.10: Relative gap in model WITH link travel time correlation. Beta is the travel time reliability coefficient.

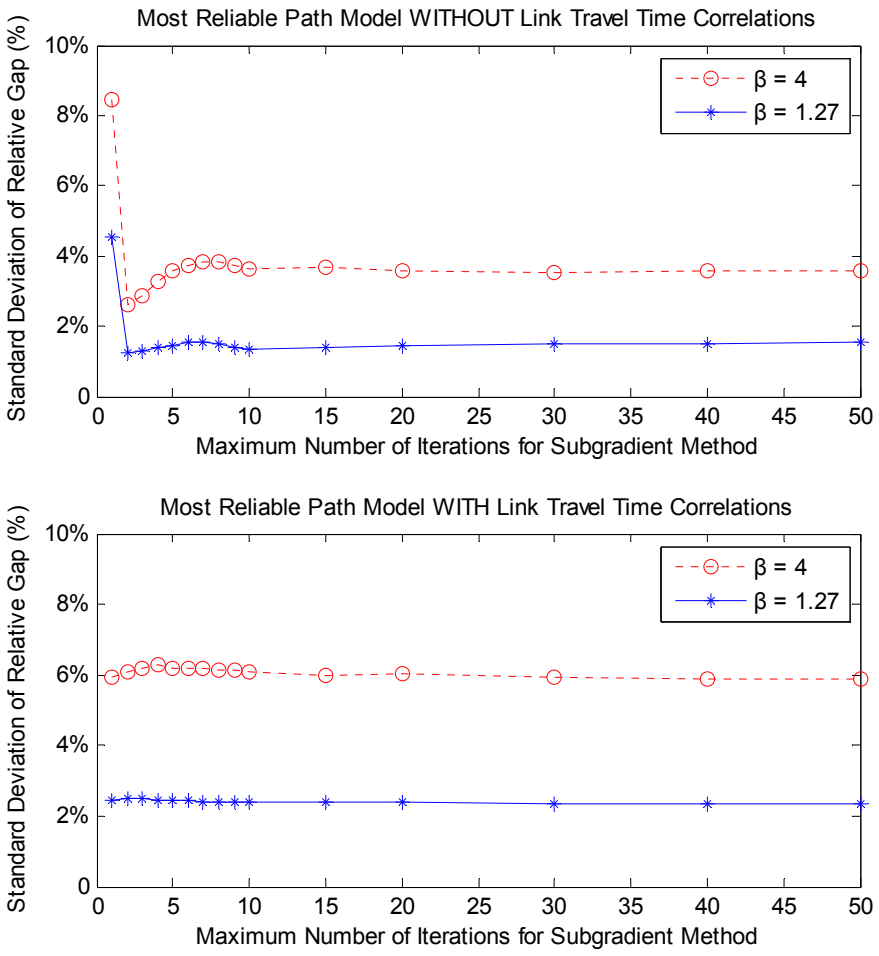


Figure 4.11: Standard deviation of relative gaps for 246 OD pairs under two models and reliability coefficients.

For the proposed two algorithms, a more systematic comparison scheme should be the following: first, extract the route solutions (in terms of node sequence) from these two different algorithms, and then evaluate their solution quality in terms of the same travel time utility function with spatial correlation. Conceptually, possible solution quality loss occurs for those models that ignore the underlying correlations. This problem can be viewed as a special case of "price of correlation" in the field of stochastic optimization, as investigated recently by Agrawal et al. (2010).

From Figures 4.9 and 4.10, we find that 20 maximum iterations will be sufficient for the model without correlation to achieve a solution with relatively small gap value, and the decreasing trends for the model considering spatial dependencies begin to slow down after 10 iterations. As mentioned previously, a small duality gap can still exist despite a considerably large number of iterations, mainly caused by the inherent limitation of Lagrangian lower bound estimation techniques. As expected, a larger travel time reliability coefficient, representing higher weight on the standard derivation, could result in larger relative gaps for both models.

Clearly, the most computational consuming step of the proposed algorithms is the shortest path calculation in each iteration of the searching process. As a result, the average computational complexity of the algorithms is determined by the number of shortest path calculations. For the model without correlations with a maximum number of 20 iterations (as the termination condition), the average computing time for finding the most reliable path for a single OD pair on our test network is 1.2 seconds (including 20 shortest path calculations). For the sampling-based model, about 0.7 seconds is needed

with 10 shortest path calculations (to gain significant optimality improvements), and 1.4 seconds for at most 20 shortest path calculations.

To further evaluate the solution quality of the proposed sampling approach, we compare the results (β = 1.27) with two computationally intensive path enumeration methods: (1) random draws and (2) least travel time path of individual day. For the first path generation method, 1,000 random draws are extracted from a normal distribution for each link, with mean and variance equal to the link travel time (Prato, 2009). In other words, to enumerate variant paths, 1,000 shortest path calculations are performed, each with different randomized link costs. The second path enumeration method finds the least travel time path of each individual day using day-specific travel time measurements. For instance, the numerical experiment in this dissertation uses 73 days of travel time measurements, corresponding to 73 day-specific shortest path calculations applied to generate a path set for evaluation purposes.

We conduct the solution quality comparison by applying the three methods to all 246 OD pairs randomly selected in the numerical experiment. As shown in Figure 4.12, the best solutions found in the proposed sampling-based LR method are generally close to or better than the results of both path enumeration methods.

In particular, the sampling-based LR method finds the best available upper bound for 78.5% of the OD pairs. Notice that only 10 shortest path calculations are used here for each OD pair in the sampling-based LR method, compared with computationally consuming 1,000 shortest path calculations for the random draws approach and 73 calculations for the day-specific sampling approach. Table 4.4 shows the performance comparison of three methods over different comparison terms. Overall, the results show

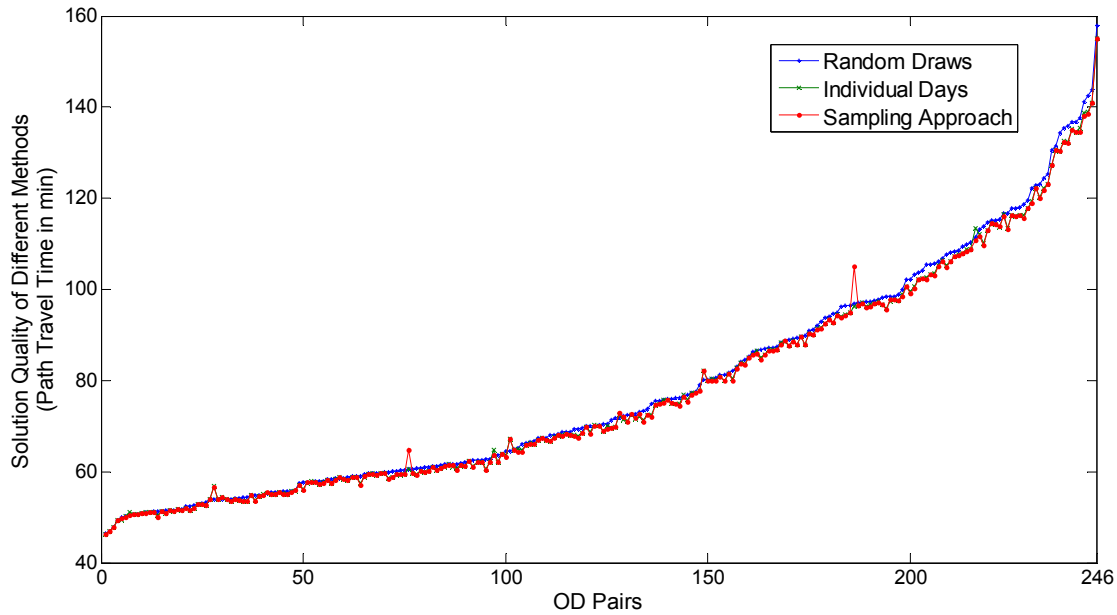


Figure 4.12: Solution quality comparison of three methods: random draws, individual days and proposed sampling method

Table 4.4: Solution quality comparison over three methods

Comparison Terms	**Sampling based LR**	**Random Draws (10 Draws)**	**Random Draws (1000 Draws)**	**Day-specific Sampling**
Percentage of OD pairs found best available solution	78.5%	0.8%	10.2%	49.2%
Average objective function values for all OD pairs	77.70	80.37	78.60	77.76
Average relative gap comparing to best available solutions (average value of best available solutions is 77.59)	0.14%	3.55%	1.30%	0.22%
Computational costs in shortest path calculations	10	10	1000	73

that the proposed sampling-based LR method is able to find the most reliable path solution with satisfactory quality and much lower computational costs. It is interesting to notice that, if the random draws approach only uses 10 draws to reduce the overall calculation efforts, then only 0.8% of OD pairs can be found with the best available solutions and the average relative gap is 3.55% compared to the best available solutions.

It should be remarked that, for the real-world network we used in the experiments (with 53,124 nodes and 93,900 links), a full path enumeration is nearly impossible. Therefore, we try to use the two stochastic path enumeration methods described above to generate a partial, but sufficiently complete, path set. To further illustrate the performance of the proposed algorithms, we extract a subnetwork with 312 nodes from the Bay Area highway system, and use a search tree-based algorithm to enumerate all simple paths of a certain OD pair (while the number of nodes in any simple path is no more than 312 in this case). A small-scale experiment with 7 OD pairs shows that the relative gaps between the proposed best solutions and the true optimal solutions are extremely small, as the resulting relative gap has an average of 0.29% and a maximum value of 4.7%.

CHAPTER 5

FINDING ABSOLUTE AND PERCENTILE ROBUST SHORTEST PATHS

To model driver route choice behavior under traffic dynamic and uncertainty, and further provide better traveler information with travel time reliability, this dissertation proposed two models to evaluate the travel time robustness: absolute and α-percentile robust shortest path problems. To meet the computational challenges of the ARSP and PRSP problems, a Lagrangian relaxation based two-bound approximation approach is proposed in this dissertation. Expressly, to reformulate the minimax objective function in the ARSP problem, we applied a variable splitting and relaxation approach to generate a dual problem that provides tight lower bounds for the optimal solution. Furthermore, a subgradient method is adopted in the solution procedure algorithm to iteratively improve both upper and lower bounds of the original problem. Following the same modeling framework, the α-percentile robust shortest path problem formulated the selection of samples by introducing a set of auxiliary variables into the objective function. The remainder of this chapter is structured as follows. Sections 5.1 and 5.2 provide formal problem statements, theoretical derivations and algorithmic development for both absolute and percentile robust shortest path problems. Section 5.3 evaluates the

performance of proposed algorithms through numerical experiments on a large-scale network with real-world observation data.

5.1. Finding Absolute Robust Shortest Path

5.1.1. Problem Statement

Consider a directed, connected transportation network $G(N, A)$ consisting of a set of nodes N and a set of links A. In this study, we assume a set of link travel time samples or measurements is available for the same time period of D days, for example, for the peak hour from 8am to 9am over $d = 1, \ldots, D = 60$ weekdays during a 3-month period, where d is the index of random scenarios (in a stochastic optimization framework) or the index of data collection days (from a traffic data mining perspective).

With a sufficiently large sample set, the calculated path travel time measure is able to capture the inherent spatial and temporal correlation of link travel times. Interested readers are referred to a discussion by Xing and Zhou (2011) on different models for representing spatial correlation for link travel times. For notational simplicity, this study considers time-invariant travel times during the analysis time period of individual days, but the presented solution framework can be extended to a space-time expanded network by adding unique physical path constraints, as illustrated in a recent study by Li et al. (2011).

We further denote the link from node i to node j as a paired index of ij, and accordingly the travel time of each link ij at the sample day d is expressed as $c_{ij,d}$. For a given OD pair (r, s), a set of binary variables $X = \{x_{ij} \mid ij \in A\}$ represents the selection of

links on a path (i.e., a path solution). The travel time for a path X at sample day d is then written as:

$$T_d = \sum_{ij \in A} c_{ij,d} x_{ij} \tag{5.1}$$

subject to a flow balance constraint

$$\sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b \tag{5.2}$$

where $b = \begin{cases} 1 & i = r \\ 0 & i \in N - \{r, s\} \\ -1 & i = s \end{cases}$ represents the flow status for each node i in the network.

Given a day-dependent sample set, the robust shortest path problem aims to find a single path solution X that satisfies a certain robustness criterion over all realized samples/scenarios from different days. Specifically, two types of criteria are considered for the evaluation of the path travel time robustness: absolute robust shortest path and α-percentile robust shortest path. The ARSP problem is mathematically expressed as

Problem *P*0: $z_0 = \min_x \max_d \sum_{ij \in A} \left(c_{ij,d} x_{ij} \right)$ (5.3)

subject to the flow balance constraint (5.2).

5.1.2. Variable Splitting-Based Model Reformulation

To reformulate the proposed minimax objective function (5.3), a variable splitting approach is adopted in this study. Expressly, we first introduce an auxiliary variable y into the objective function,

$$y = \max_{d} \sum_{ij \in A} \left(c_{ij,d} x_{ij} \right) \tag{5.4}$$

so the minimax problem is transformed to a standard minimization problem format as z = min y. The auxiliary variable y is defined as the maximum path travel time for the path X from r to s over all samples, and y also corresponds to the absolute robust travel time for a path solution X.

Further, the maximization subproblem in Eq. (5.4) can also be equivalently expressed as a set of inequality constraints for y over different days d=1, …, D:

$$y \geq \sum_{ij \in A} \left(c_{ij,d} x_{ij} \right), \forall d \tag{5.5}$$

Consequently, the ARSP problem $P0$ is formulated as $P1$

Problem *P*1: $z_1 = \min y$ (5.6)

subject to constraints (5.2) and (5.5).

To further iteratively find solutions for $P1$ efficiently, a Lagrangian relaxation based approach is implemented. That is, we introduce a set of nonnegative Lagrangian multipliers μ_d to relax and dualize the inequality constraint set (5.5) into the objective function (5.6).

$$\min y + \sum_d \mu_d \left(\sum_{ij \in A} \left(c_{ij,d} x_{ij} \right) - y \right) \tag{5.7}$$

By regrouping variables in Eq. (5.7), we now consider a Lagrangian problem:

$$\textbf{Problem } \boldsymbol{L1}\ L(\mu_1, \mu_2, ..., \mu_D) = \min \sum_{ij \in A} \left(\sum_d \mu_d c_{ij,d} \right) x_{ij} + \left(1 - \sum_d \mu_d \right) y \tag{5.8}$$

subject to constraint (5.2).

For any feasible (nonnegative) value set of the Lagrangian multipliers μ_d, the objective function value of the Lagrangian dual problem $L(\mu_1, \mu_2, ..., \mu_D)$ provides a lower bound to the optimal value z_1* of the original problem $P1$. By iteratively adjusting the Lagrangian multiplier set for $L1$, we want to maximize the dual objective function in Eq. (5.8) and therefore improve the lower bound estimate of the primal problem. Additionally, we will use paths generated through solving the dual problem to discover better solutions, which can also reduce the upper bound to the optimal value z_1* of the primal problem.

5.1.3. Lagrangian Decomposition

In the dual problem $L1$, Eq. (5.8) can further be decomposed into and solved by two independent subproblems for primal variables x and auxiliary variable y, respectively.

$$L(\mu_1, \mu_2, ..., \mu_D) = L_x(\mu_1, \mu_2, ..., \mu_D) + L_y(\mu_1, \mu_2, ..., \mu_D) \tag{5.9}$$

The subproblem $L_x(\mu_1, \mu_2, ..., \mu_D)$ is a binary integer programming problem for the primal variable set x, and it can be solved efficiently using standard label correcting or label setting algorithms (Ahuja et al., 1993) for new link cost values of $\left(\sum_d \mu_d c_{ij,d}\right)$.

Subproblem *SP*1:

$$L_x(\mu_1, \mu_2, ..., \mu_D) = \min\left\{\sum_{ij \in A}\left(\sum_d \mu_d c_{ij,d}\right) x_{ij} : \sum_{j:ij \in A} x_{ij} - \sum_{j:ji \in A} x_{ji} = b\right\} \tag{5.10}$$

The second part of the dual problem in Eq. (5.9) is a linear minimization problem for the single variable y. As a linear function, the optimal value of L_y is achieved at one extreme point of the feasible range of y.

Subproblem *SP*2: $$L_y(\mu_1, \mu_2, ..., \mu_D) = \min\left(1 - \sum_d \mu_d\right) y \tag{5.11}$$

To find the optimal solution for subproblem $SP2$, we need to consider a feasible range $\left[y^{LB}, y^{UB}\right]$ for y and propose a solution procedure for L_y:

Proposition 1: Depending on the value of $1-\sum_d \mu_d$, the variable y is selected at one extreme point of its feasible range for the optimal value of L_y, e.g.,:

$$y = \begin{cases} y^{LB} & 1-\sum_d \mu_d \geq 0 \\ y^{UB} & 1-\sum_d \mu_d < 0 \end{cases} \tag{5.12}$$

By referring back to the definitional Eq. (5.4) for variable y, for any feasible path solution x of the primal problem, it corresponds to an upper bound on the optimal objective function. For example, we can find the shortest path (with a cost function as the path distance), and use the corresponding maximum day-specific travel time as the upper bound y^{UB}. In the iterative search process to be presented below, the upper bound y^{UB} can be updated once a new path is discovered with a lower value of the maximum day-specific travel time compared to the current y^{UB}.

Essentially, y^{LB} should provide a lower bound to the maximum day-specific travel time on the optimal path. In this study, we first compute $c_{ij}^{\min} = \min_d \{c_{ij,d}\}$ as the least travel time of link (ij) across different days, and then use $\sum_{ij \in A} c_{ij}^{\min} x_{ij}$ as the objective function to find the least travel time path and the corresponding path travel time $T^{\min}$. As $c_{ij}^{\min}$ is the least possible travel time of each link, $T^{\min} \leq z_1$* for sure, for both ARSP and PRSP problems.

5.1.4. Subgradient Method

Let us denote L^* to be the maximum value of $L(\mu_1, \mu_2, ..., \mu_D)$ over different Lagrangian multiplier sets:

$$L^* = \max_{\mu_1, \mu_2, ..., \mu_D \geq 0} L(\mu_1, \mu_2, ..., \mu_D) \tag{5.13}$$

In order to find a tighter lower bound for the primal problem, we adopt a subgradient approach to iteratively search the Lagrangian multiplier set and the corresponding values of x and y.

The search directions of μ are typically calculated as the subgradient of L:

$$\nabla L(\mu_1, \mu_2, ..., \mu_D) = \left(\sum_{ij \in A} c_{ij,1} x_{ij} - y, \sum_{ij \in A} c_{ij,2} x_{ij} - y, ..., \sum_{ij \in A} c_{ij,D} x_{ij} - y \right) \tag{5.14}$$

Let us use k to denote the number of iterations. Starting from any feasible initial value set, we first find solutions x_{ij}^k and y^k for subproblems $SP1$ and $SP2$, respectively. Then the values of the Lagrangian multipliers μ_d^{k+1} at iteration k+1are updated using the following subgradient equation:

$$\mu_d^{k+1} = \mu_d^k + \theta_d^k \left(\sum_{ij \in A} c_{ij,d} x_{ij}^k - y^k \right) \tag{5.15}$$

where the step-size set θ_d^k can be updated by using the following heuristic algorithm

$$\theta_d^k = \lambda_d^k \left[z_1^{UB} - L(\mu_1^k, \mu_2^k, ..., \mu_D^k) \right]. \tag{5.16}$$

In Eq. (5.16), z_1^{UB} is the current best objective function value for feasible solutions in the primal problem and can be updated when a tighter upper bound is found. A scalar λ_d^k chosen between 0 and 2 is used in this study to adjust the step-size θ_d^k of the search process and ensure nonnegativity of Lagrangian multipliers.

5.1.5. Solution Procedure

The overall algorithm for solving the absolute robust shortest path problem is described below.

Algorithm 5.1:

Step 1: Initialization

Set iteration number $k = 0$;

Choose positive values to initialize the set of Lagrangian multipliers μ_d;

Select initial values for y^{UB} and y^{LB}. The upper bounds and lower bounds of the original problem $P1$ are $z_1^{UB} = y^{UB}$ and $z_1^{LB} = y^{LB}$

Step 2: Solve decomposed dual problems

Solve subproblem $SP1$ using a standard shortest path algorithm and find a path solution x;

Solve subproblem *SP*2 with Eq. (5.12) in Proposition 1 and find a value for *y*;

Calculate primal, dual and gap values, and update the upper and lower bound of *y*.

Step 3: Update Lagrangian multipliers

Update Lagrangian multipliers with Eqs. (5.14-5.16)

Step 4: Termination condition test

If $k > K_{\max}$ or the gap is smaller than a predefined toleration gap, terminate the algorithm, otherwise go back to Step 2. $K_{\max}$ is a predetermined maximum iteration value.

5.1.6. Solution Quality Measurement

To measure the path solution quality, we define $\varepsilon = z^{UB} - z^{LB}$ as the duality gap between the lower bound z^{LB} and the upper bound z^{UB} of the optimal solution. As a result, the gap between the optimal value z^* and the objective function value of the current best solution z^{UB} is no larger than the gap ε. To normalize the duality gap for comparison purposes, we can define a relative optimality measure as

$$\varepsilon' = \frac{z^{UB} - z^{LB}}{z^{UB}} \geq \frac{z^* - L^*}{z^*}. \tag{5.17}$$

With a reasonably small relative gap, we provide a satisfied solution quality guarantee on the suggested absolute robust path. It is important to notice that, due to the approximate nature of the Lagrangian relaxation estimator, there could still be a positive gap even if the optimal solution of the primal problem has been achieved.

The above proposed algorithm has a complexity of $O(|A|K)$ where K is the number of iteration (e.g., 10-20 for our experiments in Section 5.3), while the complexity of Yu and Yang's heuristic solution algorithm is $O(|A|D)$ with D being the number of scenarios/days. Our algorithm is comparatively more efficient when a large number of scenarios (say D=50) is required to achieve a low sampling error in capturing the network travel time stochasticity and dynamics.

5.2. Finding α-Percentile Robust Shortest Path

An α-percentile robust shortest path problem aims to minimize the α-percentile path travel time among all feasible paths. For instance, given α = 0.9, each feasible path solution x of the given OD pair has a path reliability measure $y(x)$ corresponding to 90th-percentile travel time. This measure ensures that, over all D sample days, 90% of those days have day-specific path travel times less than $y(x)$. Among all feasible paths, the path solution x^* with a minimum 90th-percentile travel time is then considered as the 90th-percentile robust shortest path. As a special case, the absolute robust shortest path problem can be viewed as the 100th-percentile robust shortest path, where the path reliability measure $y(x)$ is minimized and $y(x) = \max_d \sum_{ij \in A} \left(c_{ij,d} x_{ij} \right)$

5.2.1. Problem Formulation

The α-percentile represents the value below which α percent of the observations (in ascending order) may be found. To represent the α-percentile travel time over a finite number of days d=1, …, D, let us first denote the sorted path travel times of a path on

different days as $T_1 \le T_2 \le T_3 \le \cdots \le T_D$. The percentage α then corresponds to the nth value and T_n, where $n = \alpha \times D$. For example, when considering α = 0.9 and D = 100 sample days, n=90. If $\alpha \times D$ turns out to be a floating point number, especially when the sample size D is small, then the rank n can be obtained by rounding to the nearest integer of the value of $\alpha \times D$.

To model the α-percentile robust shortest path problem, we introduce the following formulation:

Problem *P2*: $z_2 = \min y$

subject to

$$\sum_{ij \in A} \left(c_{ij,d} x_{ij} \right) - y \le M w_d, \forall d \tag{5.18}$$

$$\sum_{d} w_d \le (1-\alpha) D \tag{5.19}$$

where M is a sufficiently large number and w_d is a binary variable for sample d.

When w_d is 0, then Eq. (5.18) reduces to

$$y \ge \sum_{ij \in A} \left(c_{ij,d} x_{ij} \right), \tag{5.20}$$

and this active inequality should be held for all day-specific path travel times T_d less than the α-percentile travel time. When w_d=1, Eq. (5.18) leads to an always-feasible and inactive constraint

$$\sum_{ij \in A} \left(c_{ij,d} x_{ij} \right) - y \le M \tag{5.21}$$

To make sure Eq. (5.21) is valid for those sample days d on which T_d is greater than the final y^* (i.e., the optimal α-percentile travel time), the parameter M should be sufficiently large. Furthermore, to ensure the robust path travel time measure y is larger than for a certain percentage of days, the variable set w is constrained by Eq. (5.19). For example, for α = 0.9 and D = 100, $\sum_d w_d \le (1-0.9)\times 100 = 10$, so there are a total of 10 inactive constraints, and 90 active constraints. Because problem $P2$ needs to minimize the variable of y, the optimization result needs to select the $n = \alpha \times D$ active constraints for travel time values on different days ranking from smallest to largest.

5.2.2. Illustrative Example

Consider a single origin-destination pair with two parallel paths, as shown in Figure 5.1. In this illustrative network, both paths share a common link A.

Table 5.1 shows the link and path travel times (min) over D=4 sample days. As calculated in this table, for the ARSP problem, path AB (along links A and B) has a minimax travel time of 12 min over four sample days. On the other hand, for a 75th-percentile robust shortest path problem, each path can only have $(1-75\%)\times 4 = 1$ day of sampled travel time greater than the variable of y. For path AB, only day 4 has a w =1, leading to its 75th-percentile travel time as 11 min. Path AC needs to set w =1 on day 3, leading to its 75th-percentile travel time as 10 min and therefore the optimal solution to the PRSP problem.

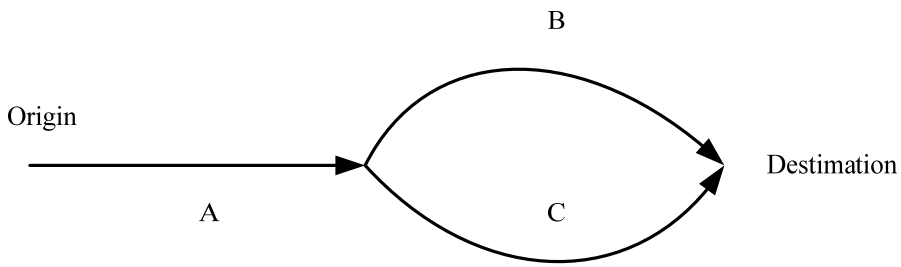


Figure 5.1: Network of the illustrative example

Table 5. 1: Travel time calculation of the illustrative example

Travel time	Link A	Link B	Link C	Path AB	*w* for Path AB	Path AC	*w* for Path AC
Day1	3	5	6	8	0	9	0
Day2	4	7	6	11	0	10	0
Day3	5	6	8	11	0	13	1
Day4	4	8	6	12	1	10	0
100th percentile				**12 (ARSP solution)**		13	
75th percentile				11		**10 (PRSP solution)**	

5.2.3. Lagrangian Relaxation and Decomposition

Following the same Lagrangian relaxation modeling framework for the ARSP problem, we introduce a set of nonnegative Lagrangian multipliers μ_d and v to relax the inequality constraint set (5.18) and (5.19) into the objective function:

$$\min y + \sum_d \mu_d \left(\sum_{ij \in A} c_{ij,d} x_{ij} - y - M w_d \right) + v \left(\sum_d w_d - (1-\alpha) D \right) \tag{5.22}$$

By regrouping variables in Eq. (5.22), a Lagrangian dual problem is constructed with three sets of independent variables:

Problem *L*2:

$$L(\mu_1,...,\mu_D,\nu) = \min \sum_{ij\in A}\left(\sum_d \mu_d c_{ij,d}\right)x_{ij} + \left(1-\sum_d \mu_d\right)y + \sum_d \left(\nu - M\mu_d\right)w_d - \nu\left(1-\alpha\right)D \quad (5.23)$$

We then divide the dual function into three independent subproblems. For notational convenience, we denote a constant variable $h = -\nu\left(1-\alpha\right)D$.

$$L(\mu_1,...,\mu_D,\nu) = L_x(\mu_1,...,\mu_D) + L_y(\mu_1,...,\mu_D) + \sum_d L_{w_d}(\mu_1,...,\mu_D,\nu) + h \quad (5.24)$$

The first two subproblems in the dual function are identical to subproblems *SP*1 and *SP*2 in the ARSP problem, and both can be solved efficiently. The third part of the dual problem in Eq. (5.24) is a number of binary integer problems, each corresponding to a day *d* and a single variable w_d.

Subproblem *SP*3: $L_{w_d}(\mu_1,...,\mu_D,\nu) = \left(\nu - M\mu_d\right)w_d, \quad \forall d$ (5.25)

Proposition 2: for each univariate linear programming problem in subproblem *SP*3, the optimal value of binary variable w_d is determined according to the given values of μ_d and *v*, i.e.:

$$w_d = \begin{cases} 0 & \nu - \mu_d M \geq 0 \\ 1 & \nu - \mu_d M < 0 \end{cases} \tag{5.26}$$

5.2.4. Subgradient Method

Similarly, we need to improve the upper and lower bounds of the primal problem by iteratively maximizing the dual problem in Eq. (5.23). The subgradient method is implemented here with two sets of Lagrangian multipliers μ and ν.

$$\nabla L(\mu_1,...,\mu_D,\nu) = \left(\sum_{ij\in A} c_{ij,1}x_{ij} - y - Mw_1,..., \sum_{ij\in A} c_{ij,D}x_{ij} - y - Mw_D, \sum_d w_d - (1-\alpha)D \right) \tag{5.27}$$

$$\mu_d^{k+1} = \mu_d^k + \theta_d^k \left(\sum_{ij\in A} c_{ij,d}x_{ij}^k - y^k - Mw_d^k \right) \quad \forall d \tag{5.28}$$

$$\nu^{k+1} = \nu^k + \theta_\nu^k \left(\sum_d w_d^k - (1-\alpha)D \right) \tag{5.29}$$

A heuristic algorithm is used to update the step-size set θ_d^k and θ_ν^k

$$\theta_d^k = \lambda_d^k \left[y^{UB} - L(\mu_1^k,...,\mu_D^k,\nu^k) \right] \quad \forall d \tag{5.30}$$

$$\theta_\nu^k = \lambda_\nu^k \left[y^{UB} - L(\mu_1^k,...,\mu_D^k,\nu^k) \right] \tag{5.31}$$

5.2.5. Solution Procedure

The overall algorithm for solving the α-percentile robust shortest path problem is described below.

Algorithm 5.2

Step 1: Initialization

Set iteration number $k = 0$;

Choose positive values to initialize the set of Lagrangian multipliers, μ_d and ν;

Select initial values for M, y^{UB} and y^{LB}.

Step 2: Solve decomposed dual problems

Solve Subproblem $SP1$ using a standard shortest path algorithm and find a solution x;

Solve Subproblem $SP2$ with Eq. (5.12) in Proposition 1 and find a value for y;

Solve Subproblem $SP3$ with Eq. (5.26) in Proposition 2 and find values for w_d;

Calculate primal, dual and gap values, and update the upper and lower bounds of the optimization problem $P2$.

Step 3: Update Lagrangian multipliers

Update Lagrangian multipliers with Eqs. (5.27-5.31)

Step 4: Termination condition test

If $k > K_{\max}$ or the gaps are smaller than the predefined toleration gap, terminate the algorithm, otherwise go back to Step 2.

5.2.6. Illustrative Numerical Examples

Now we apply the proposed Lagrangian relaxation approach in Algorithms 1 and 2 to find the ARSP and 75th-percentile PRSP in the sample network (Figure 5.1). Tables 5.2 and 5.3 show some key intermediate computational results in the first few iterations of the search procedure.

Table 5.2: Results of first few iterations for Algorithm 5.1

K	μ1	μ2	μ3	μ4	y	Lx1	Lx2	Lx	Ly	L	LB	UB	ε	ε' (%)
1	0.25	0.25	0.25	0.25	8	10.5	10.5	10.5	0	10.5	10.5	13	2.5	19
2	0.5	0.75	1.5	0.75	13	37.75	39	37.75	-32.5	5.25	10.5	12	1.5	12.5
3	0.01	0.45	1.2	0.6	12	25.43	26.19	25.43	-15.12	10.31	10.5	12	1.5	12.5
4	0.01	0.34	1.09	0.6	12	22.96	23.6	22.96	-12.42	10.54	10.54	12	1.465	12
5	0.01	0.25	0.99	0.6	12	21.02	21.58	21.02	-10.31	10.71	10.71	12	1.2892	11
6	0.01	0.19	0.94	0.6	12	19.6	20.1	19.6	-8.76	10.84	10.84	12	1.16	10
7	0.01	0.14	0.89	0.6	12	18.51	18.95	18.51	-7.57	10.94	10.94	12	1.06	9

Table 5.3: Results of first few iterations for Algorithm 5.2

K	μ1	μ2	μ3	μ4	v	M	y	Lx1	Lx2	Lx	Ly	w				Lw	L	LB	UB	ε	ε' %
												1	2	3	4						
1	0.25	0.25	0.25	0.25	1.5	4	8	10.5	10.5	10.5	0	0	0	0	0	-1.5	9	9	11	2	18
2	0.25	0.85	0.85	1.05	1.3	4	11	33.3	32.3	32.3	-22	0	1	1	1	-8.4	1.9	9	10	1	10
3	0.05	0.35	0.65	0.55	1.5	4	10	18	17.9	17.9	-6	0	0	1	1	-3.3	8.6	9	10	1	10
4	0.01	0.35	0.58	0.25	1.58	4	10	13.3	13.6	13.3	-1.9	0	0	1	0	-2.3	9.1	9.1	10	0.89	9
5	0.01	0.40	0.41	0.36	1.58	4	10	13.4	13.1	13.1	-1.9	0	1	1	0	-1.7	9.5	9.5	10	0.46	5
6	0.01	0.31	0.39	0.36	1.6	4	10	12.1	11.9	11.9	-0.7	0	0	0	0	-1.6	9.6	9.6	10	0.44	4
7	0.01	0.31	0.45	0.36	1.58	4	10	12.7	12.6	12.6	-1.3	0	0	1	0	-1.8	9.54	9.6	10	0.44	4

Additional notations in Tables 5.2 & 5.3:

K: number of iterations.

Lx1, Lx2: objective function values of subproblem *SP*1 for paths AB and AC, respectively.

Lx: the cost of the shortest path found in subproblem *SP*1.

Ly: optimal value of subproblem *SP*2 at each iteration.

Lw: optimal value of subproblem *SP*3 at each iteration.

L: the value of the dual problem for each iteration.

LB: lower bound of the solution, obtained from the best dual value *L*.

UB: upper bound of the solution, generated from the best primal value among the paths uncovered up to the current search iteration.

ε: the gap between *UB* and *LB*.

ε': relative gap: as defined in Eq. (5.17).

In the above two examples, by iteratively configuring weights μ on different samples, the proposed approach successfully uncovered the optimal solutions (Path AB for ARSP and Path AC for PRSP). Specifically, starting with uniform distributed values (1/4 = 0.25), the Lagrangian multipliers are adjusted to improve the lower bound of the optimal solution even after the optimal upper bounds have been achieved.

It should be remarked that, although the optimal solution was found in both problems, a relative gap still exists due to the approximation nature of the Lagrangian relaxation method.

In the proposed subproblem *SP*3, the parameter *M* is included in the dualized objective function (5.25). Given its corresponding negative sign, a larger value of *M* could lead to a lower value in the final optimal function for (5.25) and therefore a looser lower bound estimator. Thus, we need to select a value for parameter *M* which is not only sufficiently large enough to make inequality (5.21) valid, but also small enough to construct a tight lower bound for function (5.25). In the illustrative example, *M* is selected to be 4 minutes so that it is larger than the maximum value of the gap between 100% and 75% travel times, which is 3 minutes for Path AC.

5.3. Numerical Experiments

5.3.1. Test Network Overview

In this section, we conducted numerical experiments based on a large-scale real-world transportation network of the Bay Area, California. The selected network is comprised of 53,124 nodes and 93,900 links. In this area, the freeway system has 8,511 links (9.1% of links), and covers 15.8% (1,774.8 miles) of the entire network, while 85,389 links (90.9%) are arterial roads with a total length of 9,431.8 miles (84.2%). The proposed algorithms are implemented in C# on a Windows platform and evaluated on a personal computer with an Intel Core Duo 1.8GHz CPU and 2 GB memory.

The samples of link travel time are calculated based on available historical records from the NAVTEQ traffic database. In particular, 73 days of travel time measurements between November 2009 and February 2010 are collected for the time interval of 9:00 AM to 9:15 AM of each sample day. As the observation data used in this study (mainly from freeway segments) cover about 4.1% of the total mileage in the Bay Area, random sample travel times are generated for links without data coverage. For simplicity, this research does not remove traffic data from weekend days and holidays. Figure 5.2 shows the test network and its sensor data coverage.

As short-distance OD pairs might be covered by no or inadequate raw observations, and they typically have very limited alternative routes to examine, this study imposes the following rules to select OD pairs to be tested: (1) the average path travel time is larger than 45 minutes, and (2) the measurement coverage on the least expected travel time path is larger than 30% in distance. As a result, an OD-pair set U containing u = 246 random OD pairs is generated from the Bay Area network. The performance of the proposed

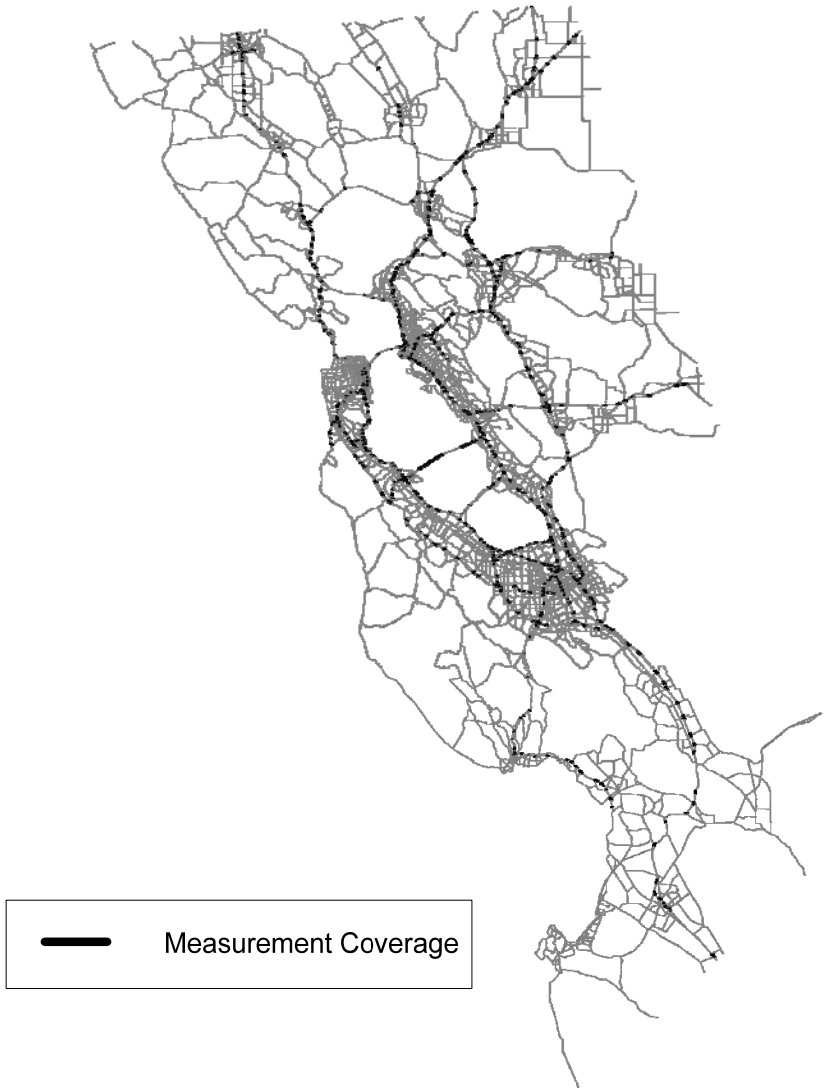


Figure 5.2: Sensor data coverage for Bay Area, California.

algorithms is assessed using the average relative gap, which is calculated as the average value of the relative gaps for all 246 OD pairs under a predefined maximum number of iterations $K_{\max}$, e.g., $\bar{\varepsilon} = \frac{\sum_{(r,s)\in U} \varepsilon'_{(r,s),K_{\max}}}{u}$. Additionally, the average objective function value of primal and dual problems among all OD pairs are also used to demonstrate the improvement of the solution quality over the iterative procedure.

5.3.2 Numerical Performance and Solution Quality Analysis

As shown in Figure 5.3, the average gap decreases along with the increase of the predefined maximum number of iterations $K_{\max}$. Figure 5.4 illustrates that, for both models, after about 5 iterations, the reduction of upper bound becomes very slow, while the lower bound keeps improving. The average gap of the 95% percentile robust path problem is larger than that of the absolute robust path problem, which can be explained by the difference in the constructed dual objective functions, and in particular the additional complexity in turning the parameter M for subproblem $SP3$.

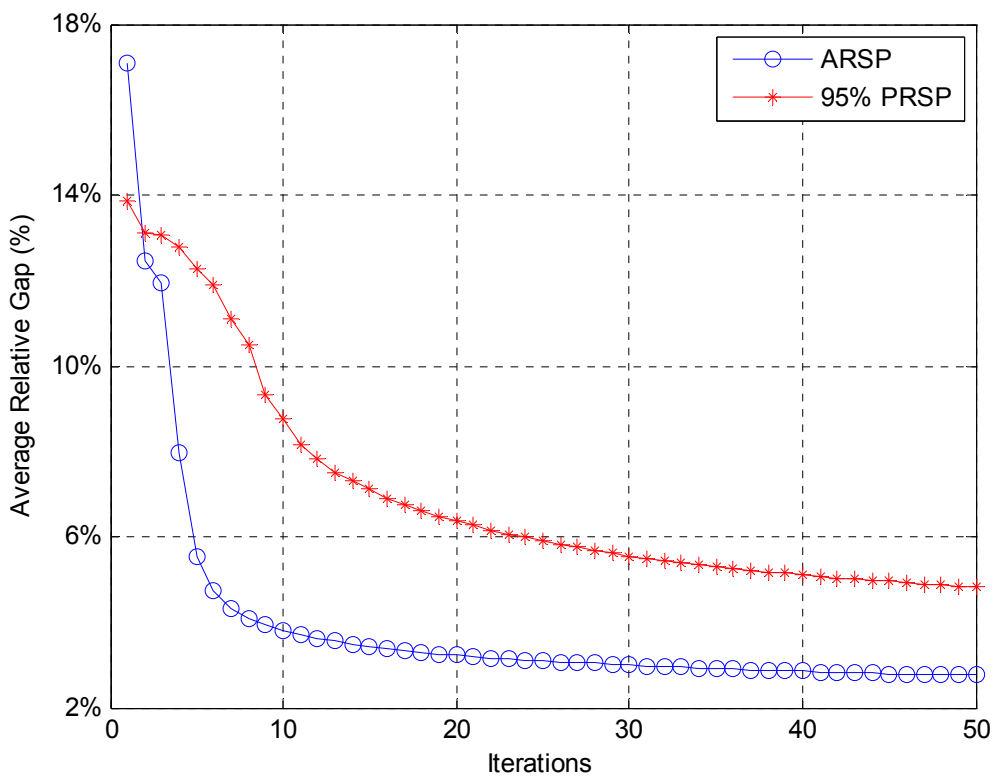


Figure 5.3: Relative gap for ARSP and 95% PRSP

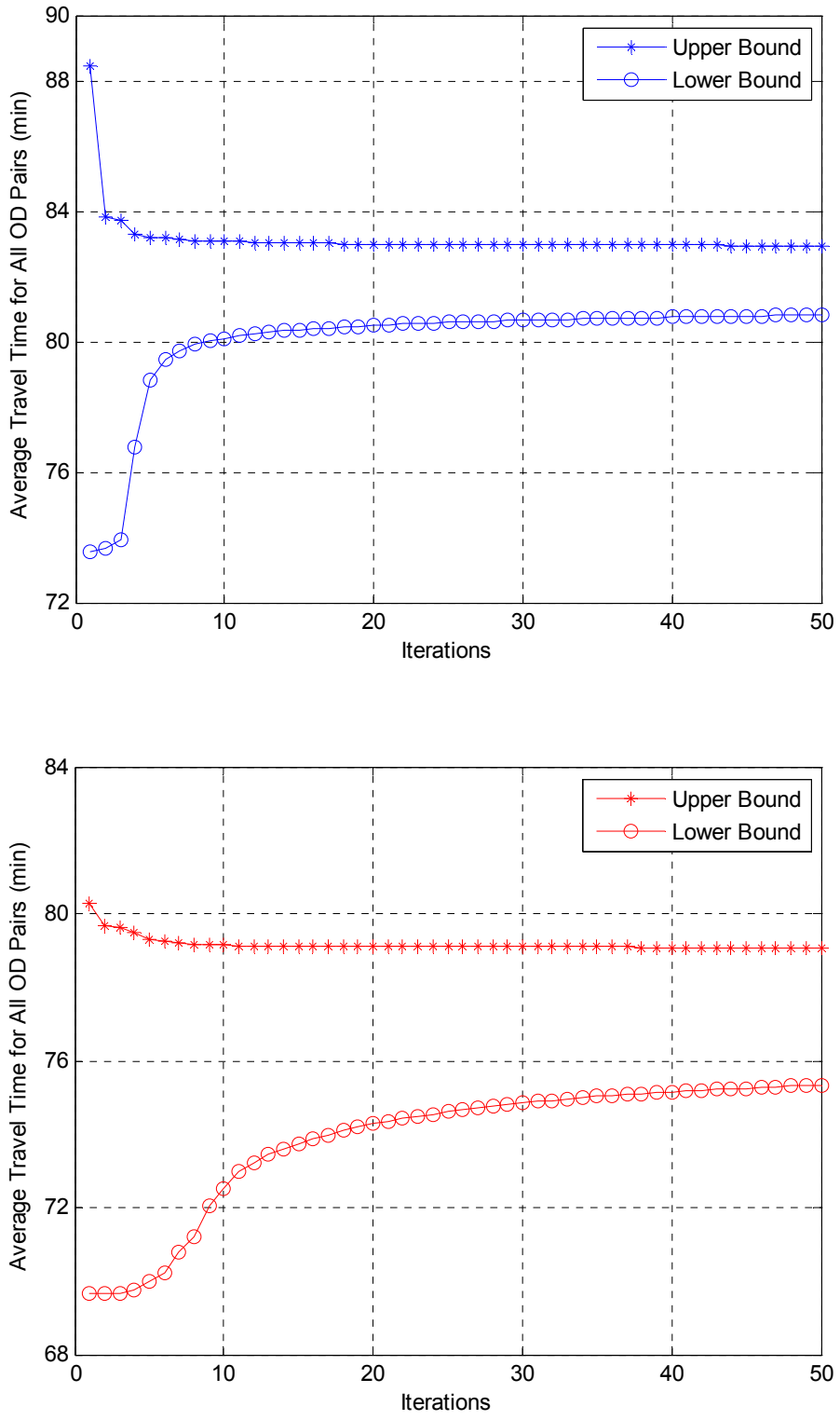


Figure 5.4: Upper and lower bounds evolution for ARSP (top) and 95% PRSP (bottom)

Overall, our experiments indicate that 20 iterations are sufficient for both models to achieve relatively small gap values, and the solution quality improvement begins to diminish after 10 iterations. It should be mentioned that a small duality gap can still exist even when an optimal solution is found, mainly due to the inherent limitation of Lagrangian lower bound estimation techniques.

CHAPTER 6

CONCLUSIONS

This chapter summarizes the major methodologies, algorithms and results proposed in this dissertation. In specific, Section 6.1 presents summaries and highlights for the topics discussed in this study, followed by the contributions of this research to the state of the art of Advanced Traveler Information Systems applications. Section 6.3 discusses further extensions and directions for future research in this area.

6.1. Research Highlights

This dissertation discusses a series of emerging issues in ATIS by proposing an integrated and unified estimation-optimization framework, including (1) design efficient and high-performance traffic monitoring network with heterogeneous sensors to accurately estimate and predict traffic under recurring and nonrecurring congestions in highway systems; and (2) provide route guidance with travel time reliability and variability information.

6.1.1. Heterogeneous Sensor Network Design for Travel Time Estimation and Prediction

To provide effective congestion mitigation strategies, transportation engineers and planners need to systematically measure and identify both recurring and nonrecurring traffic patterns through a network of sensors. The collected data are further processed and disseminated for travelers to make smart route and departure decisions. There are a variety of traditional and emerging traffic monitoring techniques, each with ability to collect real-time traffic data in different spatial and temporal resolutions. This study proposes a theoretical framework for the heterogeneous sensor network design problem. In particular, we focus on how to better construct network-wide historical travel time databases, which need to characterize both mean and estimation uncertainty of end-to-end path travel time in a regional network.

A unified Kalman filtering based travel time estimation and prediction model is first proposed in this research to integrate heterogeneous data sources through different measurement mapping matrices. Specifically, the travel time estimation model starts with the historical travel time database as prior estimates. Point-to-point sensor data and GPS probe data are mapped to a sequence of link travel times along the most likely travelled path. Through an analytical information updating equation derived from Kalman filtering, the variances of travel times on different links are estimated for possible sensor design solutions with different degree of sampling or measurement errors. The variance of travel time estimates for spatially distributed links are further assembled to calculate the overall path travel time estimation uncertainty for the entire network as the single-valued information measure. The proposed information quantification model and beam search

solution algorithm can assist decision-makers to select and integrate different types of sensors, as well as to determine how, when, where to integrate them in an existing traffic sensor infrastructure.

6.1.2. Providing Reliable Route Guidance under Stochastic Traffic Conditions

To meet the emerging needs for modeling travel time reliability, especially in the area of spatial network analysis, this dissertation proposes two models for the standard deviation based most reliable path problem, each with a different spatial dependency assumption. The path travel time reliability measure in this research is expressed through the standard deviation of path travel time. The computational challenges introduced by this reliability functional form stem from the nonlinearity and nonadditivity of standard deviation, as well as the concave characteristics of the corresponding square root transformation.

To tackle the above modeling and computational challenges, this study proposes two new approximation methods for solving the reliable path searching problem. First, focusing on the nonadditive and concave characteristics of the original reliability representation, a Lagrangian substitution-based lower bound approach is introduced to quantify the quality of solutions found by an iterative search process. More specifically, with efficient evaluation of feasible solutions and their dual problem results, a tight lower bound is achieved and a close-to-optimal solution can be obtained with a guaranteed level-of-service. Second, to incorporate the spatial correlation among link travel times, this dissertation constructs a sampling-based solution algorithm. Instead of relying on the

(independent or limited correlated) probability density functions of link travel times, a set of individual historical measurements are utilized to explicitly capture the inherent spatial correlation. Comprehensive experiment results on a large-scale network show that 10-20 iterations of standard shortest path algorithms for the reformulated models can offer a very small duality gap of about 2-6%.

6.1.3. Efficient Algorithms for Finding Absolute and Percentile Shortest Paths

To model driver route choice behavior under inherent traffic system stochasticity and dynamics, and further provide better route guidance with travel time reliability guarantees, this dissertation proposes two models to evaluate the travel time robustness: absolute and α-percentile robust shortest path problems. A Lagrangian relaxation approach and a scenario-based representation scheme are uniquely integrated to develop efficient solution algorithms for the ARSP and PRSP problems. To reformulate the complex minimax objective function in the ARSP problem, we applies a variable splitting and relaxation technique to generate a dual problem that provides tight lower bounds for the optimal solution. Furthermore, a subgradient method is adopted in the solution procedure algorithm to iteratively improve both upper and lower bounds of the original problem. Along this line, the α-percentile robust shortest path problem is reformulated as a set of easy-to-solve subproblems by introducing auxiliary variables and additional definitional constraints. The comprehensive experiment results on a large-scale network with real-world travel time measurements demonstrates that 10 to 20 iterations of standard shortest

path algorithms for the reformulated models can offer a very small relative duality gap of about 3-6%.

6.2. Summary of Contributions

In the study of heterogeneous sensor network design, a Kalman filtering-based information-theoretic sensor location model that aims to minimize information uncertainty from a set of point, point-to-point and probe sensors in a traffic network. Major contributions of this study include: (1) the sensor location problem was jointly considered with its underlying travel time estimation and prediction model to maintain the inherent consistency of these two closely related problems. (2) a unified modeling framework was developed to consider the uncertainty reduction and propagation in a heterogeneous sensor network with point, point-to-point and probe sensor observations, as well as possible error correlations between new and existing sensors. (3) a new measure of information based on the travel time uncertainty of critical origin-to-destination/paths was proposed to capture the network-wide end-to-end estimation/prediction quality, and (4) a series of close-form formulas was derived to quantify the information loss under both recurring and nonrecurring traffic conditions, and derived analytical travel time transition equations for nonrecurring traffic conditions.

For the study of route guidance based on travel time reliability and variability, this dissertation proposes to seamlessly incorporate and significantly enhance several modeling/algorithmic components from several previous studies. Based on the variable splitting approach in the Lagrangian reformulation framework, this study improves the solution quality of reliable or robust routing problems with Lagrangian relaxation based

upper and lower bound estimation. Another significant contribution of this research is an efficient and practical incorporation of spatial network correlations. Different from traditional methods focusing on the probability density functions of link travel times, this dissertation proposes a sampling-based algorithm to consider the spatial dependencies among links. By directly utilizing readily available historical travel time measurements from traffic monitoring systems, the proposed approach can systematically incorporate the inherent spatial correlation into the reliable route searching process.

6.3. Future Research

For the research on network design problems, we plan to expand the research in the following ways. First, this study only focuses on the sensor design problem for estimating the *mean* of path travel time, and a natural extension is to assist sensor design decisions for other network-wide traffic state estimation domains, such as measuring and forecasting point-to-point travel time *reliability*, and incident detection *probability*. Second, under assumptions of normal distributions for most error terms, the proposed sensor location model is specifically designed for the minimum path travel time estimation variance criterion, and our future work should consider other crucial factors for real-world sensor network design, such as allowing log-normally distributed error terms and minimizing maximum estimation errors. Furthermore, the offline model developed in this study could be extended to a real-time traffic state estimation and prediction framework with mobile and agile sensors. The numerical experimental results (for a small-scale network) in this study also demonstrate computational challenges (due to heavy-duty matrix operations) in applying the proposed information-theoretic sensor

location strategy in large-scale real-world networks, and these challenges call for more future research for developing efficient heuristic and approximation methods.

In the study of reliability-oriented route guidance problems, future research interests will cover four major extensions. (1) Expand the realm of application of these models from static travel times to time-varying travel times, and jointly consider temporal and spatial correlation in finding the most reliable path. (2) Extend current reliable routing algorithms from single OD case to the one-origin to all-destination application. Such one-to-all most reliable path problem may serve an important role in the dynamic traffic assignment. (3) Incorporate proposed reliable and robust routing models into the route choice component for network-wide dynamic traffic assignment and flow management problems. (4) Apply distributed computing techniques such as cloud computing to improve the computational efficiency.

REFERENCES

Agrawal, S., Ding, Y., Saberi, A., Ye, Y., 2010. Price of correlations in stochastic optimization, working paper, http://www.stanford.edu/~shipra/Publications/POC.pdf (accessed June 02, 2011).

Ahuja, R.K., Magnanti, T.L., Orlin J.B., 1993. *Network Flow: Theory, Algorithms and Applications*. Prentice Hall, N.J. Ch 4 and 5, 93-156.

Ashok, K., Ben-Akiva, M., 1993. Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. In: C. Daganzo ed. *Proceedings of the 12th International Symposium on Transportation*, Elsevier.

Ban, X., Herring, R., Margulici, J.D., Bayen, A., 2009. Optimal sensor placement for freeway travel time estimation. In: W.H.K. Lam, S.C. Wong, H.K. Lo ed. *Transportation and Traffic Theory*, Springer, Ch. 34, 697-721.

Bertsekas, D.P., 1999. *Nonlinear Programming*. Cambridge, MA.: Athena Scientific. Ch 2.3, 223-241.

Boyles, S.D., Waller, S.T., 2007. Online routing with nonlinear disutility functions with arc cost dependencies. Presented at the *Sixth Triennial Symposium on Transportation Analysis* (TRISTAN VI), Phuket, Thailand.

Bruni, M.E., Guerriero, F., 2010. An enhanced exact procedure for the absolute robust shortest path problem, Intl. Trans. in Op. Res. 17, 207–220.

Cambridge Systematics, 2005. *Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation*. Report for Federal Highway Administration, Washington, D.C. http://ops.fhwa.dot.gov/congestion_report/congestion_report_05.pdf (accessed Sep. 23, 2011).

Chen, A., Chootinan, P., Pravinvongvuth, S., 2004. Multiobjective model for locating automatic vehicle identification readers. *Transportation Research Record*, 1886, 49-58.

Coifman, B., Beymer, D., Mc Lauchlan, P., Malik, J., 1998. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research Part C*, 6(4), 271-288.

Danczyk, A., Liu, H., 2011. A mixed-integer linear program for optimizing sensor locations along freeway corridors. *Transportation Research Part B*, 45(1), 208-217.

Eisenman, S.M., Fei, X., Zhou, X., Mahmassani, H.S., 2006. Number and location of sensors for real-time network traffic estimation and prediction: a sensitivity analysis. *Transportation Research Record*, 1964, 260-269.

Fan, Y.Y., Kalaba, R.E., Moore, J.E., 2005a. Arriving on time. *Journal of Optimization Theory and Applications* 127(3), 497-513.

Fan, Y.Y., Kalaba, R.E., Moore, J.E., 2005b. Shortest paths in stochastic networks with correlated link costs. *Computers and Mathematics with Applications* 49(9-10), 1549-1564.

Fei, X., Mahmassani, H.S., 2011. Structural analysis of near-optimal sensor locations for a stochastic large-scale network. *Transportation Research Part C* 19(3), 440-453.

Fei, X., Lu, C-C., Liu, K., 2011. A Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C*, In Press, Corrected Proof, Available online 1 April 2011, ISSN 0968-090X, DOI: 10.1016/j.trc.2010.10.005.

Fisher, M.L., 1981. The Lagrangian relaxation method for solving integer programming problems. *Management Science* 27(1), 1-18.

Gabriel, S.A., Bernstein, D., 2000. Nonadditive shortest paths: subproblems in multi-agent competitive network models. Computational & Mathematical Organization Theory 6(1), 29-45.

Guignard, M., Kim, S., 1987. Lagrangean decomposition: a model yielding stronger Lagrangean bounds. *Mathematical Programming* 39(2), 215-228.

Haas, C., Mahmassani, H.S., Khoury, J., Haynes, M., Rioux, T., Logman, H., 2001. *Evaluation of automatic vehicle identification for San Antonio's transguide for incident detection and advanced traveler information systems*. Center for Transportation Research, University of Texas at Austin, Research Report No. 4957-1.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., Tarnoff, P., 2010. Data collection of freeway travel time ground truth with Bluetooth sensors. *Transportation Research Record*, 2160, 60-68.

Handler, G.Y, Zang, I., 1980. A dual algorithm for the constrained shortest path problem. *Networks* 10(4), 293-309.

Henig, M.I., 1986. The shortest path problem with two objective functions. *European Journal of Operational Research* 25(2), 281-291.

Herrera, J.-C., Bayen, A., 2010. Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transportation Research Part B*, 44(4), 160-481.

Hintz, K.J., McVey, E.S., 1991. Multi-Process constrained estimation. *IEEE Transactions on Systems, Man and Cybernetics*, 21(1), 237-244.

Hurdle, V. F., Son, B., 2000. Road test of a freeway model. *Transportation Research Part A*, 34(7), 537-564.

Joernsten, K., Naesberg, M., 1986. A new Lagrangian relaxation approach to the generalized assignment problem. *European Journal of Operational Research* 27(3), 313-323.

Karasan, O.E., Pinar, M.C., Yaman, H., 2001. The robust shortest path problem with interval data, *Computers and Operations Research*, to appear.

Lam, W.H.K., Lo, H.P., 1990. Accuracy of O-D estimates from traffic counts. *Traffic Engineering and Control*, 31(6), 358-367.

Larsson, T., Migdalas, A., Ronnqvist, M., 1994. A Lagrangean heuristic for the capacitated concave minimum cost network flow problem. *European Journal of Operational Research* 78(1), 116-129.

Leow, W.L., Ni, D., Pishro-Nik, H., 2008. A sampling theorem approach to traffic sensor optimization. *IEEE Transactions on Intelligent Transportation Systems*, 9(2), 369-374.

Li, M., Zhou, X., Nagui, R., 2011. Quantifying benefits of traffic information provision under stochastic demand and capacity conditions: A multi-day traffic equilibrium approach (I). Accepted for presentation at *14th International IEEE Annual Conference on Intelligent Transportation Systems*, Washington DC, USA.

Li, X., Ouyang, Y., 2011. Reliable sensor deployment for network traffic surveillance. *Transportation Research Part B*, 45(1), 218-231.

Lindveld, Ch.D.R., Thijs, R., Bovy, P.H.L., Van der Zijpp, N.J., 2000. Evaluation of on-line travel time estimators and predictors, *Transportation Research Record*, 1719, 45-53.

Liu, H., Danczyk, A., 2009. Optimal sensor locations for freeway bottleneck identification. *Computer-Aided Civil and Infrastructure Engineering*, 24(8), 535-550.

Lu, J., Tang, H., Lu, S., Ran, B., 2006. Locating roadside servers for advanced traveler information systems. *Transportation Research Board 85th Annual Meeting*, CD 06-1544, Washington DC, USA, January 2006.

Miller-Hooks, E.D., Mahmassani, H.S., 2000. Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science* 34(2), 198-215.

Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations, *Transportation Research Part C*, 19(4), 606-616.

Montemanni, R., Gambardella, L.M., 2008. *Robust shortest path problems with uncertain costs* (Technical Report No. IDSIA-03-08). Switzerland: IDSIA / USI-SUPSI.

Murty, I., Her, S.S., 1992. Solving min–max shortest path problems on a network. Naval Research Logistics 39, 669–683.

Nedic, A., 2002. *Subgradient methods for convex minimization*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Newell, G.F., 1993. A simplified theory on kinematic waves in highway traffic, part I: general theory. *Transportation Research Part B*, 27(4), 281-287.

Ni, D., Leonard, J.D., Williams, B.M., 2006. The network kinematic waves model: a simplified approach to network traffic. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 10(1), 1-14.

Nie, Y., Wu, X., 2009a. Shortest path problem considering on-time arrival probability. *Transportation Research Part B* 43(6), 597-613.

Nie, Y., Wu, X., 2009b. Reliable a priori shortest path problem with limited spatial and temporal dependencies. *Proceedings of the 18th International Symposium on Transportation and Traffic Theory*, 169-196.

Noland, R.B., Polak, J.W., 2002. Travel time variability: a review of theoretical and empirical issues. *Transportation Reviews* 122(1), 39-54.

Noland, R.B., Small, K.A., Koskenoja, P.M., Chu, X., 1998. Simulating travel reliability. *Regional Science and Urban Economics* 28(5), 535-564.

Okutani, I., Stephanedes, Y., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B*, 18(1), 1-11.

Prato, C.G., 2009. Route choice modeling: past, present and future research directions, *Journal of Choice Modelling*, 2(1), 65-100.

Scott, K., Bernstein, D., 1997. Solving a best path problem when the value of time function is nonlinear. preprint 980976 of the *Transportation Research Board*.

Sen, S., Pillai, R., Joshi, S., Rathi, A.K., 2001. A mean-variance model for route guidance in advanced traveler information systems. *Transportation Science* 35(1), 37-49.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423.

Sherali, H.D., Desai, J., Rakha, H., 2006. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. *Transportation Research Part B*, 40(10), 857-871.

Sigal, C.E., Alan, A., Pritsker, B., Solberg, J.J., 1980. The stochastic shortest route problem. *Operations Research* 28 (5), 1122–1129.

Sivakumar, R.A., Batta, R., 1994. The variance-constrained shortest path problem. *Transportation Science* 28(4), 309 - 316.

Small, K.A., 1982. The scheduling of consumer activities: work trips. *American Economic Review* 72(3), 467-479.

Stathopoulos, A., Karlaftis, M.G., 2003. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C*, 11(2), 121–135.

Tsaggouris, G., Zaroliagis, C., 2004. Nonadditive shortest paths, *Algorithms - ESA 2004*, LNCS 3221 (Springer-Verlag), 822-834.

Turner, S.M., Eisele, W.L., Benz, R.J., Holdener, D.J., 1998. *Travel time data collection handbook*, Federal Highway Administration, Office of Highway Information Management, Washington D.C. Report FHWA-PL-98-035.

Van Lint, J.W.C., Van der Zijpp, N.J., 2003. Improving a travel-time estimation algorithm by using dual loop detectors. *Transportation Research Record*, 1855, 41-48.

Wasson, J.S., Sturdevant, J.R., Bullock, D.M., 2008. Real-time travel time estimates using MAC address matching. *Institute of Transportation Engineers Journal*, 78(6), 20-23.

Work, D., Tossavainen, O.-P., Blandin, S., Bayen, A., Iwuchukwu, T., Tracton, K., 2008. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. *Proceedings of the 47th IEEE Conference on Decision and Control*, Cancun, Mexico, Dec. 9-11, 2008, 5062-5068.

Xing, T., Zhou, X., 2011. Finding the most reliable path with and without link travel time correlation: A Lagrangian substitution based approach, *Transportation Research Part B*, Vol. 45(10), pp. 1660-1679.

Yang, H., Iida, Y., Sasaki, T., 1991. An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts. *Transportation Research Part B*, 25(5), 351-363.

Yperman I., Logghe S., Immers B., 2005. The link transmission model: an efficient implementation of the kinematic wave theory in traffic networks. *Proceedings of the 10th EWGT Meeting and 16th Mini-Euro Conference*, Poznan, Poland, Sep. 2005, 122-127.

Yu, G., Yang, J., 1998. On the robust shortest path problem. *Computers and Operations Research* 25 (6), 457–468.

Zhang, X., Rice, J.A., 2003. Short-term travel time prediction. *Transportation Research Part C*, 11(3-4), 187-210.

Zhou, X., List, G.F., 2010. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. *Transportation Science*, 44(2), 254-273.

Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin–destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B* 41(8), 823-840.

Zhou, X, Mahmassani, H.S., Zhang, K., 2008. Dynamic micro-assignment modeling approach for integrated multimodal urban corridor management. *Transportation Research Part C* 16(2), 167-186.