LINK TRAVEL TIME ESTIMATION BASED ON NETWORK ENTRY/EXIT TIME STAMPS OF TRIPS

A Dissertation

by

WEN WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xiubin Bruce Wang
Committee Members,	Yunlong Zhang
	Luca Quadrifoglio
	Thomas Wehrly
Head of Department,	Robin Autenrieth

December 2015

Major Subject: Civil Engineering

ABSTRACT

This dissertation studies the travel time estimation at roadway link level using entry/exit time stamps of trips on a steady-state transportation network. We propose two inference methods based on the likelihood principle, assuming each link associates with a random travel time. The first method considers independent and Gaussian distributed link travel times, using the additive property that trip time has a closed-form distribution as the summation of link travel times. We particularly analyze the mean estimates when the variances of trip time estimates are known with a high degree of precision and examine the uniqueness of solutions. Two cases are discussed in detail: one with known paths of all trips and the other with unknown paths of some trips. We apply the Gaussian mixture model and the Expectation-Maximization (EM) algorithm to deal with the latter. The second method splits trip time proportionally among links traversed to deal with more general link travel time distributions such as log-normal. This approach builds upon an expected log-likelihood function which naturally leads to an iterative procedure analogous to the EM algorithm for solutions. Simulation tests on a simple nine-link network and on the Sioux Falls network respectively indicate that the two methods both perform well. The second method (i.e., trip splitting approximation) generally runs faster but with larger errors of estimated standard deviations of link travel times.

DEDICATION

To my parents, my husband Qing and daughter Sylvia

ACKNOWLEDGEMENTS

I would like to especially thank my advisor, Dr. Xiubin Bruce Wang, for his encouragement and guidance, and also Dr. Yunlong Zhang, Dr. Luca Quadrifoglio and Dr. Thomas Wehrly for serving on my committee and for their suggestions.

Thanks also go to my friends and colleagues in the transportation engineering division, for the happiness and encouragement they have brought to me. I have greatly enjoyed being with them at Texas A&M University. In particular, I would like to thank Kai Yin for many helpful discussions and pleasant collaboration.

I am grateful for all funding agencies that have supported me during my studies at Texas A&M University. Specifically, I would like to thank the Southwest Region University Transportation Center (SWUTC), which is funded by a grant from the U.S. Department of Transportation, University Transportation Centers Program. I would also gratefully acknowledge the kind support from the National Center for Freight and Infrastructure Research and Education (CFIRE) at the University of Wisconsin-Madison.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background and Motivation

Travel time is one of the most important factors when a traveler plans a route from an origin to a destination, and it is also critical to transportation planners and operators as a performance measure. Reducing travel time (e.g., through traffic congestion mitigation or tolling) is often considered as equivalent to improving mobility and network efficiency. Therefore, accurate travel time estimation on a transportation network is becoming an essential task and is being made possible now by widely available traffic data.

A regular way to obtain travel time data on a network is by means of traffic tracking. This can be done through probing phones (e.g., Bar-Gera [5], Ygnace et al. [98]), global positioning system (GPS) devices (Bertini and Tantiyanugulchai [7]), and vehicle ID readers (through either Bluetooth or vehicle plate identification and matching, Haghani et al. [34], Barcelo et al. [6], Chang et al. [12]). When data is sufficient all these methodologies work well. However, a main drawback is that these methods demand a huge volume of vehicular data. For example, in order to obtain the speed or travel time information of a specific roadway link, one would need to track vehicle movement at both ends of that particular link, which gives rise to the requirement for a huge amount of data, both across the network and over a period of time (e.g., during both peak hours and off-peak hours). A research question is: Is it necessary to have two points tracked for every link in order to obtain the link travel time? This research studies an alternative way to estimate link travel time based only on records of the travelers' start and end locations and time stamps of trips on a network. We refer to the start and end locations and time stamps of trips

as travelers' entry/exit time stamps on a network in this dissertation. By knowing both ends of trip itineraries for a sufficiently large amount of travelers, one is able to estimate the link travel times with an acceptable accuracy.

This research is first motivated by a practical application in which vehicles' entry/exit locations and time stamps are available on a toll road network. Toll road operators have a practical need to use this information for link travel time estimation and prediction. There may be many other similar applications with the public transit systems as well. A broader impact of this study is that through finding the mapping relationship between trip itinerary and link travel time, one may choose to archive the itinerary information in order to keep link travel time information, therefore to reduce the amount of data collected and archived for the transportation network performance measures.

1.2 Problem Statement

In this dissertation, a roadway network is represented by a graph, where nodes represent intersections and links (edges) represent road segments. A link connects either between two intersections on an urban arterial road or between two entry and exit ramps on a highway section. Each link associates with a random travel time that follows a certain distribution. A path is defined as an alternating sequence of links and nodes from an origin to a destination node (known as an OD pair). Each trip consists of a path, the entry (starting) time at the origin, and the exit (ending) time at the destination. Multiple trips may take place on the same path. Trips on the network are observed each with an OD pair associated entry/exit times. Paths may not be known for some trips. A trip time, the difference between entry/exit times, is the summation of link travel times along a path. With a sufficiently large number of trips observed, our goal is to estimate the parameters of link travel time

distribution by handling the unobserved routes if necessary.

We propose two inference methods based on the likelihood principle. The first method (Method I) considers independent and Gaussian distributed link travel times, using the additive property that trip time has a closed-form distribution as the summation of link travel times. To overcome the modeling challenge that random link times do not typically add up to a trip time with closed-form distribution, we develop another method (Method II) that can apply to the general case with arbitrary link travel time distribution. For each case, two versions of the study problem are examined respectively. First we address a simpler version in which each trip observation has a known route on the network. In a second step, we further study the problem in which the associated routes of some trip observations are unknown. The first version of the problem provides a basis for the study of the second one.

1.3 Outline

The remainder of this dissertation is organized as follows. Section 2 overviews the prior studies on travel time estimation techniques and relevant literature on statistical inference methods. Section 3 proposes the first method assuming that the trip time has a closed-form distribution, using Gaussian distribution for link travel time as an example. Section 4 develops a statistical framework of the trip splitting method to deal with a more general trip travel time distribution. Two cases are discussed in each model approach: one with known routes of all trips and the other with unknown routes of some trips. Section 5 tests the proposed methods with simulated data on a simple 9-link network and the Sioux Falls network, respectively. Section 6 discusses the advantages and disadvantages of both methods. Section 7 concludes this research.

2. LITERATURE REVIEW*

Several commercially available systems are capable of estimating roadway travel times on the real-time basis using varying sources of traffic data. Dion and Rakha [28] broadly classify the systems into three categories: spot speed measurement systems, spatial travel time systems, and probe vehicle technologies. The following reviews literature according to this classification, as in Yin et al. [99].*

2.1 Data Sources

2.1.1 Spot Speed Measurement Systems

Spot speed measurement system, specifically consisting of inductance loop detectors, has been a main source of traffic information in the past decades. The traditional single loop detectors consist of a single inductance loop that is able to generate a magnetic field and detect the passing of vehicles. These detectors are usually set in fixed points along a roadway, and they output traffic variables such as traffic flow (number of passing vehicles per hour), and occupancy (percentage of time that detector is occupied) at specific points. Substantial studies have focused on this indirect estimation of roadway travel times using each vehicle's speed observed at discrete points along a roadway. The spatial travel time over an entire trip can be calculated based on the space-mean-speed estimates.

The prominence of this spot speed measurement approach results from the large number of available traffic data provided by inductance loop detectors. Additional research efforts have also been made in improving the accuracy of spot speed estimation from single loop detectors (Coifman [17], Dailey [23], and Pushkar et al. [74]).

*Part of this section is reprinted with permission from "Link travel time inference using entry/exit information of trips on a network" by K. Yin, W. Wang, X.B. Wang, and T.M. Adams, 2015. *Transportation Research Part B: Methodological*, 80, 303-321, Copyright [2015] by Elsevier.

The use of double loop detectors can also help to obtain accurate estimates of spot speed. The double loop detectors consist of a pair of single loop detectors, which are set very close to each other. Other than the traffic flow and occupancy, this pair of sensors can also collect traffic speed and vehicle lengths using the obtained travel times of vehicles between two sensors (Leduc [50]).

These conventional sensors can provide high quality data and are not affected by external factors. They are usually widely deployed along a roadway. However, their installation and maintenance are expensive and complicated (Bar-Gera [5]). To resolve these issues, other evolving measurements have emerged such as infrared and radar technology as well as the video image detection method in recent years. For example, detectors can use video cameras and the image processing method to obtain vehicle counts and speeds at specific points along the road. The main drawbacks are that they are usually susceptible to external factors (for example, weather), and they may also need periodic maintenance (Leduc [50]).

2.1.2 Spatial Travel Time Systems

Different from indirect estimation using spot speed measurement, study on travel time estimation has also focused on direct measurement of the time interval that a particular vehicle takes to travel from one point to another.

Many researchers have proposed smart use of loop detector data by matching the particular vehicles in consecutive loop detectors based on their characteristic lengths (Coifman and Cassidy [19]; Coifman and Ergueta [20]; Coifman and Krishnamurthy [21]), or particular inductive signature on the detectors (Abdulhai and Tabib [1], and Sun et al. [86]). However, these techniques require the upgraded hardware and/or software loop configurations, thus they have not been widely put into practice for highway operation.

Other than the loop detector data, deployment of Intelligent Transportation Systems (ITS) in the last decade has brought the chance to use more suitable traffic data to directly measure travel times (Turner et al. [89]). The merging spatial travel time measurement systems use equipment at fixed locations to automatically identify and track vehicles in the traffic stream. Spatial travel time estimates can be computed by matching vehicle identifications at different reader locations. This is the case with data obtained from the Automated Vehicle Identification (AVI) systems, which can be of various types, such as toll collection systems (e.g., Al-Deek et al. [3]), video cameras and license plate recognition techniques (e.g., Kazagli and Koutsopoulos [48]), and also the recent Bluetooth-based detection systems (e.g., Haghani et al. [34]). The AVI systems can detect and match vehicles on both ends of a road section, thus the travel times can be directly computed if the clocks at different locations are properly synchronized.

The TranStar system in Houston (Houston TranStar [41]), and the Transmit system in the New York metropolitan area (Mouskos et al. [65]), estimate link travel times by tracking the passage times at specific locations among those vehicles equipped with electronic tags of automatic toll collection system. And the TransGuide system in San Antonio (Southwest Research Institute [87]) collects travel time information from voluntary vehicles equipped with electronic transponder tags for research purposes. These AVI systems monitor vehicles' movements using tag readers that are typically installed every 1 to 5 miles along highway segments.

2.1.3 Probe Vehicle Technologies

Another approach to measuring travel times is to use probe vehicle technologies, which are capable of tracking a sample of probe vehicles as they travel through a transportation network. The use of probe vehicles can provide the information of

vehicles' trajectories, and travel times between two points can be easily derived. The emerging technologies include smart phones, global positioning systems (GPS), and automatic vehicle location (AVL) systems. Those probe vehicles act as mobile traffic sensors equipped with tracking devices (e.g., GPS or mobile phones), and send location, direction and speed information every few seconds or minutes. They are being used to collect network-wide traffic information such as instantaneous speeds and travel times at any network location without the need of roadside equipment.

In order to accurately represent realistic traffic conditions, the sample size needed is generally quite large, especially in the case of probe vehicle systems (Turner and Holdener[90]; Chen and Chien [15]). Even the increasing GPS tracking of taxis, buses, and other vehicles has resulted in a large number of equipped vehicles traveling through an urban transportation network, prior research using probe vehicle data have examined the number of probe vehicles needed to reflect realistic traffic conditions.

Sanwal and Walrand [79] suggest the use of vehicles as sensors, considering the insufficient amount of sensors available for traffic surveillance. Their simulation results show that probe vehicles, accounting for approximately 4% of total traffic, are necessary for desirable travel speed estimation. Srinivasan and Jovanis [85] indicate that the number of probe vehicles required increases non-linearly as the reliability criterion is made more stringent. They also conclude that the number of probes required increases with the desired proportion of link coverage on the network, or with shorter travel time measurement periods. And with a fixed number of probes, a larger proportion of freeway links can be reliably covered than that of a major arterial. Zou et al. [104] propose a method for arterial speed estimation by utilizing taxi GPS data from 100 vehicles in Guangzhou, China. Their study shows that the number of probe vehicles accounting for 3% of total traffic result in significantly lower errors

for travel speed estimation. Lorkowski et al. [57] discuss the potential applications of probe vehicle data, such as dynamic routing and automatic congestion detection, using GPS data from 700 taxis in Stuttgart, Germany. Their results indicate that probe vehicles accounting for about 1% of total traffic are required to estimate traffic conditions.

Several studies have also been conducted to deal with the route inference in map-matching processes for probe vehicle data. Yokota and Tamagawa [100] develop a map-matching and route identification algorithm based on dynamic programming, using GPS probe data from freight vehicles. The experiment results demonstrate that their method can analyze the tour of freight vehicles along highway, and effectively detect vehicles' on and off ramp trajectories. Rahmani and Koutsopoulos [75] also propose a simultaneous map-matching and path inference method for low-frequency GPS probe data on urban network based only on available information of geo-locations and time stamps. Their case study in Stockholm indicate that the proposed method is robust with respect to the frequency of probe data, and appropriate for off-line and real-time applications.

2.2 Review on Travel Time Estimation Models

Two main issues concerning travel time are estimation and prediction. These two concepts are different with respect to objectives and dynamism (Mori et al. [64]). Travel time estimation calculates the travel times of vehicles' trajectories that have already ended based on data obtained during the trip. It aims to provide a reasonable value of travel time that gives a general idea of traffic conditions on a certain roadway section and within a certain time interval. In contrast, travel time prediction aims to forecast the travel time for a vehicle's trajectory that will start right away in future intervals, by using traffic data currently available as well as historical data from the

past.

Both estimation and prediction have been extensively studied in literature. This dissertation focuses on the estimation of link travel times on a transportation network. Therefore, a comprehensive review is provided on model approaches in terms of travel time estimation, in order to build a complete background analysis of available models and algorithms. The following summarizes the relevant literature on the estimation methods using data from various measurement systems.

2.2.1 Travel Time Estimation Using Loop Detector Data

The loop detector is able to output the traffic flow and the occupancy at the fixed point of detection. A significant body of literature has developed travel time estimation approaches using loop detector data, including traffic theory-based and data-based methods (Mori et al. [64]).

The traffic theory-based methods utilize relations between traffic variables based on the conventional traffic flow theory. Nam and Drew [66] propose a method to estimate freeway travel times in real time directly from flow measurements. Their model approach is essentially based on the stochastic queuing theory, flow conservation and propagation principles. The analysis results indicate that the estimates are consistent with empirical data.

Long et al. [56] develop link travel time models based on the piecewise-linearized profiles of link cumulative flows. They prove that the proposed models preserve the first-in-first-out (FIFO) principle and the continuity of travel times with respect to flows.

The advantage of traffic theory-based methods is that they are capable of capturing the dynamic characteristics of traffic, by applying the realistic relations between traffic variables. However, traffic flow needs extra monitoring if an entry/exit ramp

exists between two point detectors, in order to obtain accurate cumulative inflow and outflow profiles for a study link.

The data-based methods use statistical and machine learning approaches to find underlying structures that relate traffic flow, occupancy of detector, and travel times using empirical data, for example, time series analysis with cross correlation techniques (Dailey [22]), polynomial regression model (Sisiopiku and Rouphail [81]), stochastic model assuming travel times of vehicles arriving at a detector in a given interval follow a distribution (Petty et al. [73]), and application of artificial neural networks (Palacharla and Nelson [70]). The main drawbacks of these data-based methods are that they require a large amount of quality data and only apply to specific sites.

Even with accurate spot speed estimates obtained from point detectors, travel time estimates can still be flawed as extrapolating spot measurements to a roadway section. Different traffic conditions may exist along a roadway. It is noted that this issue particularly arises on a roadway with low density of detection sites. As suggested by Hopkin et al. [40], one detector site every 500 meters of highway is desirable to provide accurate travel time estimates.

Several approaches have also been developed to overcome the issue and avoid the enormous cost of intensive loop surveillance, such as the identification of vehicle trajectories between loop detectors (Coifman [18] and Li et al. [51]), and development of sensor deployment methods for reliable travel time estimation (Hu el al. [42], Li and Ouyang [52]). In addition, it shall be taken into account that in the traffic situations of stop and go, the loop speed estimates may not represent the space mean speed of traffic stream. Therefore, indirect estimation of roadway travel times using spot speed measurement systems still has limitations to generate accurate travel time estimates.

2.2.2 Travel Time Estimation Using AVI Data

The AVI systems can provide real-time travel time information to travelers within Advanced Traveler Information Systems (ATIS). The collected AVI data need to be cleaned and filtered prior to their use in ATIS or other applications. For example, the outlier observations exist when a tracked vehicle makes a stop for refueling, or detours between successive detection stations. Some research efforts have been made to address this issue (Ohba et al. [69]; Dion and Rakha [28]; Tam and Lam [88]).

Dion and Rakha [27] develop a method to estimate the roadway link travel times using AVI data by designing a robust data-filtering procedure to identify valid observations. Their method deals with both steady state and transient traffic conditions, and can be applied to the roadway segments with low levels of AVI penetration. The case study using travel time data from the San Antonio AVI system demonstrates the validity of proposed method, and its ability to track sudden travel time changes even with a small sample.

The concept of using AVI data from toll collection systems to directly measure highway travel times is first proposed by Davies et al. [25]. A large number of literature mainly deal with the usage of Electronic Toll Collection (ETC) data to measure travel time. The systems can identify the vehicles through on-vehicle electronic tags and roadside equipment on highway segments. However, the basic problems of this configuration include the level of market penetration of electronic toll tags, and time periods in order to obtain a continuous measurement of travel times when only small samples are available (see Dion and Rakha [28]).

Additional research has been conducted on travel time measurement using the typical configuration of a closed toll system, which has been widely extended in Europe and Asia (Ohba et al. [69]). For a closed toll system, the toll a particular

vehicle is charged varies depending on its origin and destination, and the individual toll is approximately proportional to the traveled distance along highway. Soriguera et al. [83] present a new approach for measuring travel times on closed toll highways. Considering tollbooths are located on the entry/exit ramps with each vehicle charged a fee depending on its origin and destination, the data from toll collection system are filtered and fused in a statistical way in order to extract valid itinerary travel time information. The proposed method allows estimating travel times on single sections of highway using itineraries covering different pairs of origin and destination.

2.2.3 Travel Time Estimation Using Probe Vehicle Data

A significant body of literature focuses on model-based and data-driven methods to estimate travel times or link-based travel speeds with probe vehicle data for traffic monitoring or planning purposes.

A mathematical model by Jula et al. [47] estimates link travel times and arrival times at nodes on a real-time, stochastic network. Hellinga et al. [36] propose an analytical model to decompose partial link or route travel time from a probe vehicle into individual link travel times along urban arterial, utilizing real traffic conditions on arterial network. Their evaluation suggests that the proposed method outperforms the benchmark (deterministic) method.

Different from the conventional loop detector data, probe vehicle data does not provide direct information about flow, density, and average speeds that are usually the inputs for analytical models. Instead, data-driven methods are used for travel time estimation. The existing data-driven methods include regression models (Chan et al. [11]), and neural network based models. Zheng and Van Zuylen [102] propose a three-layer neural network model to estimate link travel times for individual probe vehicles. The results with simulated data demonstrate that their model outperforms

the analytical model. However, the many required parameters associated with those models limit their applicability in practice.

Instead of exclusively utilizing probe vehicle data, Bhaskar et al. [8] propose a model to incorporate probe vehicle data into traditional cumulative plots in order to estimate the average travel times on a urban network. Zhang and Rice [101] use data from both probe vehicles and double loop detectors to develop a linear model for travel time prediction on freeways. Sananmongkhonchai et al. [78] combine the real-time taxi data with the historical hourly speed profiles. Their results display an improvement in travel speed estimation. In addition, to address the issue of sparse truck GPS data available, Morgul et al. [63] present an empirical method for truck travel time estimation, using taxi GPS data to supplement the limited truck GPS data on the Manhattan network. Their results indicate that the taxi GPS data supplement the sparse truck data well.

2.3 Statistical Approaches in Relevant Literature

While the associated methods to estimate roadway travel time range from regression model (Chan et al. [11]), machine learning approach (Zheng and Van Zuylen [102]) to analytical model dealing with traffic conditions (Hellinga et al. [36]), many required parameters limit their applicability in practice and a lack of general model approaches has been identified when it comes to a network-wide travel time estimation problem. To date, valid statistical analysis becomes increasingly important as the data becomes widely available (Fan et al. [30]). In this section, we provide a detailed review on the statistical techniques and approaches for network travel time modeling and analysis.

2.3.1 Travel Time Distribution and Reliability

Many existing models relate the link travel time to traffic volume or signal timing information (Davidson [24]; Spiess [84]; Skabardonis and Dowling [82]; Xie et al. [96]), but they can only provide the average travel time for all the traveling vehicles along a roadway section and are generally used for planning purposes. In reality, it is important to consider the uncertainty associated with roadway travel time, due to the unexpected road conditions, different driver behavior, impact of traffic signals on arterial roads, etc. The estimation and prediction of travel time probability distribution can be more valuable than a deterministic estimate of travel time. Even though the common objective in literature is to provide the mean travel time (e.g., using the length of road section divided by the obtained space mean speed) for a study roadway, providing an estimation or prediction of travel time distribution is more informative and reliable to guide vehicles traveling through that roadway. It can also be used for risk-averse routing, fleet vehicle decision support of on-time delivery, or reporting travel time reliability to a traveler (Liu et al. [55]; Samaranayake et al. [77]; Chen et al. [13]).

Modeling travel time reliability on traffic networks has attracted substantial attention in literature (Noland and Polak [68]; Chen et al. [14]; Clark and Watling [16]; Al-Deek and Emam [2]; Li et al. [53]). It is increasingly important to accurately estimate and predict the range of possible variations in travel times and the associated probabilities for the use of roadway travelers and traffic system operators. Extensive studies on this topic have proposed various parametric probability density functions to characterize the travel time distributions based on historical travel time data. The traditional models that are commonly used in literature include Gaussian, lognormal, Gamma and Weibull distributions (Emam and Ai-Deek [29]; Arroyo and

Kornhauser [4]; Rakha et al. [76]).

Recent research on travel time data analysis and travel time distribution modeling has benefited from the application of advanced statistical techniques. One of the promising approaches in this context is the use of finite mixture models, which is considered a useful extension of classical statistical models.

Jintanakul et al. [46] apply a hierarchical Bayesian mixture model to the travel time distribution along freeway sections based on small samples of vehicle probe data. The model uses two normal components to capture the heterogeneity in the travel time observations and various distribution shapes such as the skewed or multimodal distributions. The results of their simulation study demonstrate that the proposed model can well approximate the true travel time distribution for each roadway section during each interval.

Kazagli and Koutsopoulos [48] develop a log-normal mixture model approach to identify valid observations in the processing of traffic data from an Automatic Number Plate Recognition (ANPR) system. Their model takes into account that ANPR observations have a significant amount of noise and need to be filtered due to vehicles stopping along the route, taking detours, mismatched license plates, etc.

Guo et al. [33] propose mixture distributions to model travel time reliability. Their model captures the multi-modality in travel time distributions considering the travel time data collected under multistate traffic conditions. The simulation study and field data analysis based on San Antonio AVI travel time data show the superiority of using the two-component normal mixture model over the traditional single-mode probability distributions.

Kim and Mahmassani [49] also propose a two-component Gamma-Gamma mixture model to capture the vehicle-to-vehicle and day-to-day variability of travel times on a traffic network. They compare the distribution fitting using both the proposed

model and the standard one-component Gamma-Gamma model. The results indicate that the mixture model provides a better fit to the travel delay observations.

Considering the availability of mixture modeling techniques described above, this dissertation addresses the trip observations with unlabeled routes to be on different possible paths with probabilities, such that the observed travel times of unlabeled trips with the same OD pair are thought of as a sample drawing from a multimodal distribution, where each of modality represents the random travel time on a possible path. Under the assumption of Gaussian distributed link travel times, this research formulates the multimodal distribution as the classical Gaussian mixture model(Bishop [10]; Bickel and Doksum [9]).

2.3.2 Maximum-Likelihood Method and Bayesian Approach

In what follows, we review two categories of statistical approaches to address network travel time estimation and prediction in relevant literature: the traditional maximum-likelihood method and Bayesian approach.

Among the scant literature that focuses directly on this topic, Hunter et al. [44] formulate a maximum-likelihood problem to estimate link travel time distributions on an arterial network. Their model takes into account that an unknown trajectory observation may incur the path uncertainty. They present the Expectation-Maximization (EM) algorithm to simultaneously learn the likely paths by probe vehicles as well as the travel time distributions on the network. They assume that the travel times on different links are independent, and briefly report the estimation results in their case study using San Francisco taxi data.

In order to extract travel time distributions from sparse, noisy GPS measurements collected in real-time from vehicles on a large network, Hunter et al. [43] also present a probabilistic model of travel times on the arterial network along with an online EM

algorithm for learning the model parameters. Their framework can accommodate a wide variety of travel time distributions proposed in prior studies (Hellinga and Fu [37]; Hofleitner et al. [39]; Lin et al. [54]). Although it is common to use Gaussian random variables because of closed-form solutions, they use Gamma distributions considering positive valued distributions with heavy tails, and present algorithms to sample and compute densities for Gamma distributed link travel times. Their EM algorithm has no closed-form expression, and requires sampling and nonlinear optimization techniques. But it can estimate travel times on a large urban network (e.g., the San Francisco bay area) by processing tens of thousands of observations per second, with a latency of a few seconds.

Instead of the assumption of independent link travel times, Jenelius and Koutsopoulos [45] present a statistical model for travel time estimation on an urban road network based on the vehicle trajectories from low frequency GPS probe data. They consider the correlation between travel times on different links, and capture the correlation using a moving average specification for link travel times. The specific information of link attributes (such as speed limit and roadway functional class) and trip conditions (such as day-of-week, time-of-day, and weather condition) are incorporated as explanatory variables in the model. The model is estimated using maximum-likelihood method, and it is applied to estimate travel times for a particular route of the Stockholm network in Sweden. Their case study results highlight the potential of using sparse probe vehicle data for monitoring the performance of urban transport system.

In contrast to traditional maximum-likelihood method, some relevant studies apply the Bayesian approach to travel time distribution prediction. Hofleitner et al. [38, 39] propose a dynamic Bayesian network for unobserved traffic conditions on links, and model link travel time distributions conditional on traffic states. Their

method is from the traffic flow perspective, and is applied to a San Francisco road network to predict travel times using taxi data.

Feng et al. [31] propose two approaches to estimate link travel time distributions on urban arterial roadways. They model the link travel time distributions as mixtures of normal densities. Their first approach applies the EM algorithm to empirical estimates of the travel times when prior travel time data is available. The second approach estimates the travel time distributions based on signal timing information and arterial geometry. The GPS data is utilized to update the parameters of the travel time distributions using the Bayesian approach. They conduct the case studies using both the Peachtree Street (in Atlanta, GA) data and Washington Avenue (in Minneapolis, MN) GPS data. The comparison results from the Bayesian update and EM algorithm indicate that overall, the EM algorithm fit the data better. However, the Bayesian approach can still reflect the real world situation for some scenarios with missing data.

Westgate et al. [93] also propose a Bayesian model to estimate the distribution of ambulance travel times on road segments in Toronto. They apply a multinomial Logit model to formulate the path choices for ambulance trips, and perform the path inference and travel time estimation simultaneously using a Bayesian approach. They also assume that the link travel times are independent and log-normally distributed. The parameters are estimated using Markov Chain Monte Carlo (MCMC) methods. Instead of modeling travel time at the link level in the previous work, Westgate et al. [94] model the ambulance travel times at trip level. They propose a regression approach for estimating the ambulance travel time distribution along an arbitrary route on a road network, and use a Bayesian formulation to estimate the model parameters. The advantage of applying the Bayesian approach is that it can utilize expert knowledge as prior information to represent estimates as a conditional

distribution, and it can also tackle many complicated problems that traditional statistical approaches find difficult to analyze. However, the implementation relies on computationally expensive methods such as MCMC.

2.4 Objectives and Contributions of this Research Compared with Literature

This research aims to develop inference methods for link travel time estimation on a steady state network, given that each link is associated with a random travel time. We estimate network-wide link travel times by only using vehicle start and end locations and time of trips, referred to as traveler entry/exit time stamps in this dissertation. This type of data is available nowadays when discrete points of a trip are recorded. Sparse vehicle trajectories reported by GPS-equipped probe vehicles or smart phones (Wang et al. [91]) can also be regarded as a particular case of traveler entry/exit trip information on a network. Specially, this research is motivated by a practical application on a toll road network, in which traveler entry/exit time stamps are recorded at tollbooths and the toll road authority has a practical need to use travel time inference results to evaluate the toll systems. Other potential applications include using public transit data for network performance analysis when passenger entry/exit information is recorded at fare boxes (Ma et al. [59]).

We start with the assumption of independent and Gaussian distributed link travel times, and present the EM algorithm to address the trips with unknown routes, as Hunter et al. [44] and Siripirote et al. [80]. However, different from relevant literature, we focus on exploring the analytical properties of fundamental model framework from the statistical perspective. We examine the impact of errors in trip variance estimates on mean link travel time estimates, and investigate the uniqueness of solutions in the algorithm. We also provide the calculation of confidence intervals for mean link times. Furthermore, we provide a statistical method of trip splitting

approximation to mainly address a technical situation in which the summation of random link travel times for a route does not have a closed-form probability distribution. The basic idea of decomposing trip travel time already has been seen in practical applications (Hellinga et al. [36]), but without appropriate justification and investigation. The proposed trip splitting method can apply to arbitrary distributions, and is statistically justified for the network estimation problem. Its potential application appears more promising if more traffic information is available.

3. METHOD I: ESTIMATION USING TRIP TIME DISTRIBUTIONS*

In this section, we study the problem where the trip travel time has a closed-form distribution as the summation of link travel times, as in Yin et al. [99].*

A key technical challenge is regarding the randomness of the link travel time and the specific distributions to represent it. A nice feature of the Gaussian distribution is that the sum of random variables that follow Gaussian distributions still follows a Gaussian distribution. Because the Gaussian distribution is often representative of reality, in Method I that follows, we develop models assuming link travel times follow Gaussian distributions. Note that we generally assume all link travel times are independent in our study unless specified otherwise.

3.1 Link Time Estimation Using Trips with Known Routes

We first study the basic case in which all the observed trips have known routes. In other words, each OD observation has a specific set of links on which the itinerary trip takes place. link travel times are estimated according to a specific time interval of the day, although the time interval may be wide such as half an hour or longer.

We let A be the set of road links and n be the total number of links. Let I be the total number of observations and x_i denote the observed travel time of trip i. We assume that I is larger than n throughout the rest of this study. The set of observations is represented by D, i.e., $D = \{x_1, x_2, \ldots, x_I\}$. As all the trips have known routes, we denote by $\delta_{i,a}$ an incidence indicator, which is equal to 1 when link a is on trip i and 0 otherwise. Let the corresponding incidence matrix be $\Delta = [\delta_{i,a}]_{I \times n}$. In addition, we denote the mean travel time on link a by μ_a and the corresponding

*Part of this section is reprinted with permission from "Link travel time inference using entry/exit information of trips on a network" by K. Yin, W. Wang, X.B. Wang, and T.M. Adams, 2015. *Transportation Research Part B: Methodological*, 80, 303-321, Copyright [2015] by Elsevier.

standard deviation by σ_a, where $a \in A$. Let μ and σ be the n-by-1 vectors of μ_a and σ_a, respectively. Following Rakha et al. [76] and Wen et al. [92], we make the following assumption:

Assumption 1 All link travel times on the study network are independently and normally distributed, as denoted by $\mathcal{N}(\mu_a, \sigma_a)$ for each link a.

If, however, the link travel times are correlated, as long as a joint distribution of link travel times is available, the problem can still be technically modeled. Here, we maximize the following likelihood for the trip observations:

The likelihood function of the observations is described as follows:

$$\text{Maximize} \quad \mathcal{LL}(\eta, \tau \mid D) \quad = \quad \sum_i \log\left(\frac{1}{\sqrt{2\pi}\tau_i} e^{-\frac{(x_i-\eta_i)^2}{2\tau_i^2}}\right), \tag{3.1}$$

where η_i and τ_i denote respectively the mean travel time and the standard deviation of trip i. η and τ denote vectors of η_i and τ_i, respectively.

As we assume the link travel time distributions are independent, the following equations hold:

$$\eta_i = \sum_a \delta_{i,a}\mu_a, \tag{3.2}$$

$$\tau_i = \sqrt{\sum_a \delta_{i,a}\sigma_a^2}. \tag{3.3}$$

The objective function (3.1) is equivalent to a minimization function as follows

$$\text{Minimize} \quad \mathcal{W}(\mu, \sigma \mid D) = \sum_i \left(\log\left(\sum_a \delta_{i,a}\sigma_a^2\right) + \frac{1}{\sum_a \delta_{i,a}\sigma_a^2}(x_i - \sum_a \delta_{i,a}\mu_a)^2\right) \tag{3.4}$$

The objective leads to the following equations, by setting the partial derivative to

zero for a specific link a with respect to its parameters μ_a and σ_a, respectively,

$$\sum_i \frac{\delta_{i,a}(\sum_b \delta_{i,b}\mu_b)}{\sum_b \delta_{i,b}\sigma_b^2} = \sum_i \frac{\delta_{i,a}x_i}{\sum_b \delta_{i,b}\sigma_b^2}, \tag{3.5}$$

$$\sum_i \frac{\delta_{i,a}}{\sum_b \delta_{i,b}\sigma_b^2} = \sum_i \left(\frac{\delta_{i,a}}{(\sum_b \delta_{i,b}\sigma_b^2)^2}(x_i - \sum_b \delta_{i,b}\mu_b)^2 \right). \tag{3.6}$$

Equations (3.5) and (3.6) are nonlinear but one may refer to Newton–Raphson's method for solutions. To solve Equations (3.5) and (3.6), an iterative practical approach can be designed as follows: First we observe that if σ_a^2 are determined, μ_a can be solved easily by Equation (3.5) due to the resulting linear system in terms of μ_a . Then based on the obtained μ_a, we solve for σ_a^2 by Equation (3.6) using traditional techniques for nonlinear system. This process iterates until convergence.

3.1.1 Matrix Representation

It is convenient to format Equations (3.5) and (3.6) in matrix to simplify the further analysis. Two approaches are available: one through the observation–link incidence matrix and the other through the itinerary–link matrix. While the former appears more natural, the latter has a more compact form that will be useful for practical implementation. We present the first approach in this section and present the second in Appendix A.1.

Let X be a I-by-1 vector with the i-th element being x_i and Σ be the $n \times n$ covariance matrix of link travel times. Since the link travel times are assumed to be independent, Σ is a diagonal matrix here with the element $\Sigma_{a,a} = \sigma_a^2$. We denote by Λ a $I \times I$ diagonal matrix with $\Lambda_{i,i} = \sum_b \delta_{i,b}\sigma_b^2$. In fact, we have the representation $\Lambda = \text{diag}(\Delta\Sigma \cdot \mathbf{1})$, where $\mathbf{1}$ is an n-by-1 vector with 1 as its element, and $\text{diag}(\cdot)$ denotes the transformation of a vector to a diagonal matrix. In this representation, the operator $\cdot$ emphasizes that the multiplication is taken between a matrix and

a vector. If we let $\tilde{\Delta} = \Lambda^{-1}\Delta$, i.e., $\tilde{\Delta}$ being the incidence matrix Δ scaled by $(\sum_b \delta_{i,b}\sigma_b^2)^{-1}$ for all $\delta_{i,a}$ in the row i, Equation (3.5) can be written as $\tilde{\Delta}^T\Delta \cdot \mu = \tilde{\Delta}^T \cdot X$. We note that the matrix Δ is of the same rank with the matrix $\tilde{\Delta}$. This result actually implies that a unique solution exists as long as the incidence matrix Δ is of full rank and all σ_a are known. Under this condition, Equation (3.5) has the solution $\mu = (\tilde{\Delta}^T\Delta)^{-1}\tilde{\Delta}^T \cdot X$, which is the weighted least squares estimation. If Δ does not have full rank, $(\tilde{\Delta}^T\Delta)^{-1}$ is considered as a generalized inverse. Moreover, Equation (3.6) can be written as $\tilde{\Delta}^T \cdot \mathbf{1} = \Delta^T \cdot [(\Lambda^{-1}(X - \Delta \cdot \mu)) \circ (\Lambda^{-1}(X - \Delta \cdot \mu))]$, where $\circ$ denotes the element-wise product.

3.1.2 Analysis of Mean Estimates: Impact of Errors in Variance Estimates

Equation (3.5) can be reduced to a series of linear equations regarding link time mean estimates, given the values of variance estimates. It can be shown that if the trip variance values are predetermined within a certain range of estimate errors, it would be computationally easy to solve for the mean link time estimates with reasonable errors. We illustrate this point below.

We let $\hat{\sigma}_b^2$ be the variance estimate used in Equation (3.5) and let σ_b^2 be its real value. For convenience, we assume that there is a disturbance ϵ_b in the variance estimates, i.e., $\hat{\sigma}_b^2 = \sigma_b^2 - \epsilon_b^2$ in the following analysis. A similar analysis can be applied to the case $\hat{\sigma}_b^2 = \sigma_b^2 + \epsilon_b^2$ as well as the general case $\hat{\sigma}_b^2 = (\sigma_b - \epsilon_b)^2$.

We denote by $\hat{\mu}$ the vector solution to Equation (3.5) with $\hat{\sigma}_b^2$. The matrix $\tilde{\Delta}$ is the same as defined before with σ_b^2, i.e., $\tilde{\Delta} = \Lambda^{-1}\Delta$ with $\Lambda_{i,i} = \sum_b \delta_{i,b}\sigma_b^2$. We also use $\|\cdot\|$ to denote the norm of matrices or vectors. Let Λ_ϵ be a $I \times I$ diagonal matrix with the i-th element in diagonal being $\frac{\sum_b \delta_{i,b}\epsilon_b^2}{\sum_b \delta_{i,b}\sigma_b^2}$, then we have the following.

Proposition 1 $\|\mu - \hat{\mu}\|$ is sufficiently small provided $\|\Lambda_\epsilon\| \ll 1$, where $\mu = (\tilde{\Delta}^T\Delta)^{-1}\tilde{\Delta}^T \cdot X$.

Proof. We have the following equation:

$$
\begin{aligned}
(\sum_b \delta_{i,b}\hat{\sigma}_b^2)^{-1} &= \left(\sum_b \delta_{i,b}\sigma_b^2\left(1-\frac{\sum_b \delta_{i,b}\epsilon_b^2}{\sum_b \delta_{i,b}\sigma_b^2}\right)\right)^{-1} \\
&= \left(\sum_b \delta_{i,b}\sigma_b^2\right)^{-1}\left(1+\frac{\sum_b \delta_{i,b}\epsilon_b^2}{\sum_b \delta_{i,b}\sigma_b^2}+...\right), \qquad (3.7)
\end{aligned}
$$

provided that $\frac{\sum_b \delta_{i,b}\epsilon_b^2}{\sum_b \delta_{i,b}\sigma_b^2} \ll 1$ and the higher order terms are omitted. Then the Left Hand Side (LHS) and the Right Hand Side (RHS) of Equation (3.5) for all links become

$$
\begin{aligned}
\text{LHS} &= \sum_i \frac{\delta_{i,a}(\sum_b \delta_{i,b}\hat{\mu}_b)}{\sum_b \delta_{i,b}\sigma_b^2} + \sum_i \frac{\delta_{i,a}(\sum_b \delta_{i,b}\hat{\mu}_b)(\sum_b \delta_{i,b}\epsilon_b^2)}{(\sum_b \delta_{i,b}\sigma_b^2)^2} + ... \,, \\
&= \tilde{\Delta}^T\Delta\cdot\hat{\mu} + (\Lambda_\epsilon\tilde{\Delta})^T\Delta\cdot\hat{\mu} + ... \,; \qquad (3.8) \\
\text{RHS} &= \sum_i \frac{\delta_{i,a}x_i}{\sum_b \delta_{i,b}\sigma_b^2} + \sum_i \frac{\delta_{i,a}x_i(\sum_b \delta_{i,b}\epsilon_b^2)}{(\sum_b \delta_{i,b}\sigma_b^2)^2} + ... \,, \\
&= \tilde{\Delta}^T\cdot X + (\Lambda_\epsilon\tilde{\Delta})^T\cdot X + ... \,, \qquad (3.9)
\end{aligned}
$$

where the omitted terms are of a higher order of $\frac{\sum_b \delta_{i,b}\epsilon_b^2}{\sum_b \delta_{i,b}\sigma_b^2}$. Note that the second lines in Equations (3.8) and (3.9) are understood as the matrix representation for all links. Then we have the following by omitting all higher order terms:

$$
\left[\tilde{\Delta}+\Lambda_\epsilon\tilde{\Delta}\right]^T\Delta\cdot\hat{\mu} = \left[\tilde{\Delta}+\Lambda_\epsilon\tilde{\Delta}\right]^T\cdot X, \qquad (3.10)
$$

and, assuming all inverse of matrices can be performed properly [1], we have

$$
\begin{aligned}
\hat{\mu} &= \left(\left[\tilde{\Delta}+\Lambda_\epsilon\tilde{\Delta}\right]^T\Delta\right)^{-1}\left[\tilde{\Delta}+\Lambda_\epsilon\tilde{\Delta}\right]^T\cdot X, \\
&= (\tilde{\Delta}^T\Delta)^{-1}\tilde{\Delta}^T\cdot X - (\tilde{\Delta}^T\Delta)^{-1}(\Lambda_\epsilon\tilde{\Delta})^T\Delta(\tilde{\Delta}^T\Delta)^{-1}\tilde{\Delta}^T\cdot X + ... \qquad (3.11)
\end{aligned}
$$

[1] We use $(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + \dots$, provided $\|A^{-1}B\| < 1$ where A and B are matrices. Such inverse in Equation (3.11) is guaranteed by assumption of $\|\Lambda_\epsilon\| \ll 1$.

Since the norm of Λ_ϵ is far less than 1, i.e., $\|\Lambda_\epsilon\| \ll 1$, the norm of all matrices from the second term in Equation (3.11) is less than or of higher order of $\|\Lambda_\epsilon\|\|(\tilde{\Delta}^T\Delta)^{-1}\tilde{\Delta}^T \cdot X\| \ll \|\mu\|$, i.e., $\|\mu - \hat{\mu}\|$ being sufficiently small. $\square$

This proposition indicates a network property that the ratio of estimate errors to the mean link time estimates has the same order with the ratio of total errors to the trip variance estimates along a route. In other words, even if errors of some link variance estimates are relatively large, the accuracy of mean estimates is still ensured as long as the trip variance estimates are with reasonable errors. This finding can help compute the mean estimates easily, by solving linear equations given that the predetermined link time variance values are with reasonable errors. Otherwise it would be difficult to solve all the derivative equations due to the nonlinear part in terms of travel time variances.

3.1.3 Relationship with Ordinary Least Squares

To illustrate the relationship between the objective (3.4) and least squares, let us look at some special cases below.

It can be seen that Equation (3.4) is generally an objective function in terms of both the variance and mean of travel time on each link. Therefore, if the variance of travel time on each link is considered as constant, the objective function would actually become *weighted* least squares, and the weight of each trip observation is equal to the reciprocal of total travel time variance along that trip itinerary.

Moreover, from the mathematical point of view, if the variance σ_a^2 are also the same for all links, one can let $\sum_a \delta_{i,a} = N_i$, which denotes the number of links that trip i traverses along its itinerary. Then Equation (3.5) becomes

$$\sum_i \frac{\delta_{i,a}(\sum_b \delta_{i,b}\mu_b)}{N_i} \;=\; \sum_i \frac{\delta_{i,a}x_i}{N_i}, \text{ for all } a. \tag{3.12}$$

The above equation is a *weighted* least squares, with the weight of each trip observation equal to the reciprocal of total number of traversed links.

A special case for Equation (3.12) that N_i are the same for all trips, would essentially lead to the Ordinary Least Square results, i.e., $\Delta^T\Delta \cdot \mu = \Delta^T \cdot X$. In other words, if the total travel time variance along each trip itinerary (i.e., $\sum_a \delta_{i,a}\sigma_a^2$ for each trip i) is the same, Equation (3.5) would definitely lead to the resulting estimates equal to the ones by solving the ordinary least squares. Therefore, under the strong mathematical assumptions with respect to the variance of travel times on links, solving the maximum likelihood estimates can be converted to the ordinary least squares.

3.1.4 Discussion on the Rank Issue

It is already demonstrated in Section 3.1.1 that there would exist an unique solution of mean estimates in Equation (3.5) as long as the incidence matrix Δ is of full rank. We further analyze the possible rank issue of incidence matrix in this section.

Mathematically, if the incidence matrix has the issue of deficient rank, we can still solve for the generalized inverse to get the mean estimates. Moreover, for practical applications, we may also identify those *co-existent* links and make appropriate allocation of travel time estimates among them.

The definition of co-existent links is illustrated in the following example. As shown in Figure 3.1, if links c and d are considered as two distinct sections of roadways, and if all trips coming from link a or b also cover both links c and d, it would be impossible to uniquely estimate the link travel times on link c and d. One may get an estimate for the total travel time on sections c and d, but any split of this total between them would result in a feasible solution for the under-determined system.

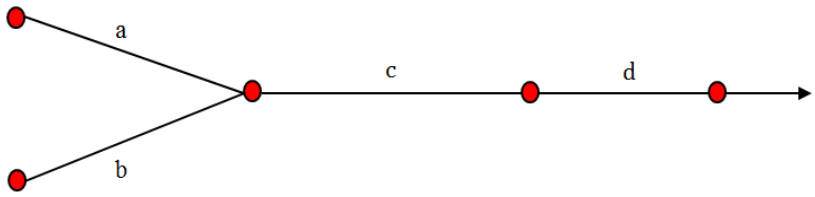

Figure 3.1: Illustrative example of co-existent links.

Therefore, we refer to such pair of links as co-existent links.

We propose the following procedure to identify those sets of co-existent links: Take two links from the set of arcs, and extract the corresponding two columns in the incidence matrix Δ as a sub-matrix. If the sub-matrix has a rank of one (i.e. two columns are exactly the same), the two chosen links are co-existent. Splitting the travel time estimates between the two links in different ways is always feasible. Besides, considering the spatial nature of network and physical connectivity of links, only connected links may be examined for unique estimation.

Proposition 2 If links a and b are co-existent, and so are b and c, then links a and c are also co-existent. It implies that the mean estimates of travel times for the entire sections through a to c can be split among a, b, and c in any way as feasible solution to the under-determined system.

This proposition is straightforward according to the definition of co-existent links. Considering the co-existent links are always the adjacent sections along a route, we can first regard them as a single link such that the incidence matrix can be reduced. After we get the mean estimate of this link, the allocation among them will be conducted to obtain the individual estimates of each co-existent link. A simple allocation method could be splitting the whole travel time estimate among

co-existent links in proportion to their free flow travel times (or travel distances). If extra information is available regarding traffic conditions or geometric features of links, further adjustment may be made for this allocation of travel time estimates.

3.2 Solution Framework Considering Unknown Route Trips

This section extends to the estimation problem where routes of some trip observations are unknown. The routes of travelers across the network need to be inferred. Finding the actual trajectory of a vehicle (path inference) can be challenging especially in dense urban areas, since multiple paths may exist that are consistent with a trip observation. Given observations $D = \{x_i\}$ that consist of some trips with *labeled* (or known) routes and a portion of trips with *unlabeled* (or unknown) routes, we can divide the entire trip observations $D = \{x_i\}$ into two subsets: D^l represents those labeled trips, and D^u denotes those trips with unlabeled route information, i.e., $D = D^l \cup D^u$.

In this case, we need to simultaneously infer the routes of recorded trips, with the objective of maximizing the total likelihood over all trip observations. One can easily imagine an iterative mechanism that once the path assignment is conducted, the resulting link travel time estimates would be affected, then in return, one can adjust the path assignment accordingly. Therefore, the critical challenge here is to examine the convergence condition for maximizing total likelihood function with adjusted trip assignment at iterations. In other words, a meaningful question is if we can derive a path assignment mechanism that assures the non-decrease of resulting total likelihood at iterations?

3.2.1 An Algorithm for Hard Assignment of Unknown Route Trips

To estimate the distribution parameters for each link, we maximize the total likelihood based on the sample of observed trip travel times:

$$\begin{aligned}\max_{\mu,\sigma,\pi} \mathcal{LL}(\mu,\sigma,\pi \mid D) &= \sum_{i\in D^l} \log\left(\mathcal{N}(x_i \mid \sum_a \delta_{i,a}\mu_a, \sum_a \delta_{i,a}\sigma_a^2)\right) \\ &+ \sum_{i\in D^u} \log \sum_{k\in K_i} r_{ik}\mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2). \quad (3.13)\end{aligned}$$

Note that the second term of function represents the likelihood for trips with unknown routes, and K_i denotes the set of possible paths that trip i may traverse. For the *hard* assignment of each data point x_i, we introduce a corresponding set of binary indicator variables $r_{ik} \in \{0,1\}$, such that $\sum_{k\in K_i} r_{ik} = 1$, for any $i \in I^u$.

A straightforward iterative mechanism can be designed as follows.

Step 1: Initialize the indicator variables r_{ik} for each unknown-route trip.

Step 2: Solve for the MLE of distribution parameters according to the proposed derivations for trips with known routes.

Step 3: Adjust the path assignment r_{ik} for those trip observations with unknown trajectory: For any $i \in I^u$, compare its resulting likelihood among candidate paths $k \in K_i$, based on the currently estimated distribution parameters, and reassign it onto the path with maximum one (i.e. its most likely path obtained at current iteration). More formally, this can be expressed as

$$r_{ik} = \begin{cases} 1 & \text{if } k = \arg\max_{k\in K_i} \mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2) \\ 0 & \text{otherwise} \end{cases} \quad (3.14)$$

Step 4: Repeat Steps 2 and 3 until convergence of either the estimated parameters

or the total log likelihood.

Intuitively, Step 3 would always improve or non-decrease the resulting total log-likelihood function (even with the same solution as obtained in Step 2 for distribution parameters). As long as we can solve Step 2 at each iteration, this procedure can guarantee the convergence generally. However, it may converge to a local rather than global optimum. We may try to randomly assign those unknown-trajectory trips onto their corresponding candidate paths initially in Step 1, and run the procedure multiple times in order to obtain the best convergence results.

This iterative procedure essentially applies the K-means clustering algorithm (Hastie et al. [35]), which is often used to identify clusters of data. We briefly introduce this algorithm in Appendix A.2. In the context of our travel time estimation problem here, we use such K-means algorithm to cluster unlabeled trips with the same OD pair based on their candidate paths. For each iteration, every unlabeled trip is assigned uniquely to a path, which may be considered as *hard* assignment in contrast to the model in the next section. However, there may be data points that lead to roughly similar likelihoods on different candidate paths. In that case, it is not clear that the hard assignment would be the most appropriate. Furthermore, the algorithm cannot guarantee the convergence. Therefore, we adopt a probabilistic approach next, known as *soft* assignment, for the unknown route trips.

3.2.2 Gaussian Mixture Model and EM Algorithm for Soft Assignment of Unknown Route Trips

We adopt a probabilistic point of view for the assignment of unlabeled trips. Instead of mapping it to a unique route, we consider each unlabeled trip to be on different possible paths with probabilities. The observed travel times of unlabeled trips with the same OD pair are thought of as a sample drawing from a multimodal

distribution, where each of modality represents the random travel time on a possible path. Such view may also be known as the *soft* assignment (Bishop [10]). Under the assumption of independent and normally distributed link times, it is natural to formulate the multimodal distribution as the classical Gaussian mixture model (Bishop [10]; Bickel and Doksum [9]).

The Gaussian mixture model is a parametric probability density function represented as a weighted sum of Gaussian component densities. The mixture weight for each component is usually called the mixing coefficient. In our context, the probability density for those unlabeled trip travel times with a distinct OD pair, i.e., with a distinct set of candidate paths, is a Gaussian mixture model, where every mixture component corresponds to a candidate path that has normally distributed travel times.

Our objective is to maximize the likelihood function based on the sample of observed trip travel times as below: [2]

$$\begin{aligned}\max_{\mu,\sigma,\pi} \mathcal{LL}(\mu,\sigma,\pi \mid D) &= \sum_{i\in D^l} \log\left(\mathcal{N}(x_i \mid \sum_a \delta_{i,a}\mu_a, \sum_a \delta_{i,a}\sigma_a^2)\right) \\ &+ \sum_{i\in D^u} \log\left(\sum_{k\in K_i} \pi_k \mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2)\right) \quad (3.15)\end{aligned}$$

subject to

$$\sum_{k\in K_i} \pi_k = 1, \text{for path set } K_i \text{ with a distinct OD pair, } i \in D^u, \quad (3.16)$$

$$0 \le \pi_k \le 1. \quad (3.17)$$

where K_i denotes the set of possible paths that trip i may traverse; μ_a and σ_a denote the estimated mean travel time and the standard deviation on link a; π_k is the mixing

coefficient and the sum of all π_k for the corresponding path set is equal to 1; and $\delta_{k,a}$ by abuse of notation is equal to 1 if the path k of trip i contains the link a for $k \in K_i$, otherwise 0. Also note that $K_i = K_j$ if unlabeled trips i and j have the same OD pair.

The objective (3.15) leads to the following equations, by setting the partial derivative to zero for a specific link a with respect to its parameters μ_a and σ_a, respectively,

$$0 = \sum_{i \in D^l} \frac{\delta_{i,a}(x_i - \sum_b \delta_{i,b}\mu_b)}{\sum_b \delta_{i,b}\sigma_b^2} + \sum_{i \in D^u} \sum_{k \in K_i} \frac{\delta_{k,a}\gamma_k(x_i)(x_i - \sum_b \delta_{k,b}\mu_b)}{\sum_b \delta_{k,b}\sigma_b^2}, \quad (3.18)$$

$$\begin{aligned} 0 &= \sum_{i \in D^l} \left(\frac{\delta_{i,a}}{\sum_b \delta_{i,b}\sigma_b^2} - \frac{\delta_{i,a}}{(\sum_b \delta_{i,b}\sigma_b^2)^2}(x_i - \sum_b \delta_{i,b}\mu_b)^2 \right) \\ &+ \sum_{i \in D^u} \sum_{k \in K_i} \frac{\delta_{k,a}\gamma_k(x_i)\left[2(\sum_b \delta_{k,b}\sigma_b^2) - (x_i - \sum_b \delta_{k,b}\mu_b)^2\right]}{(\sum_b \delta_{k,b}\sigma_b^2)^2}, \end{aligned} \quad (3.19)$$

where $\gamma_k(x_i)$ represents the probability that the component (or candidate path) k takes for explaining the trip observation i:

$$\gamma_k(x_i) = \frac{\pi_k \mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2)}{\sum_{j \in K_i} \pi_j \mathcal{N}(x_i \mid \sum_a \delta_{j,a}\mu_a, \sum_a \delta_{j,a}\sigma_a^2)}, \quad (3.20)$$

where $\mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2)$ is used by abuse of notation to represent the probability density function of Gaussian distribution at x_i with parameters mean $\sum_a \delta_{k,a}\mu_a$ and variance $\sum_a \delta_{k,a}\sigma_a^2$. Equation (3.20) provides another perspective on mixing coefficients π_k and $\gamma_k(\cdot)$. We can think of π_k as the prior probability of taking the path k for trips between a OD pair and $\gamma(\cdot)$ as the posterior probability after observing a particular trip time.

[2]Note that the second term of Equation (3.15) essentially classifies the unknown-route trips with distinct OD pairs, and each class corresponds to a Gaussian mixture model with associated mixing coefficients to be determined. We ignore the summation over OD pairs here for the convenience of notations.

We also maximize the objective (3.15) with respect to the mixing coefficients π_k, taking into account the constraint (3.16) that requires the mixing coefficients summing up to one for unknown-route trips with a distinct OD pair. By incorporating Lagrange multipliers, we can solve for π_k by setting its partial derivative equal to zero:

$$\pi_k = \frac{\sum_{i \in D^u_{rs}} \gamma_k(x_i)}{\mid D^u_{rs} \mid}. \tag{3.21}$$

where D^u_{rs} denotes the set of unknown-route trips with a distinct OD pair rs.

We apply the expectation-maximization (EM) algorithm to solve for the parameter estimates, which leads to a MLE of the model if it exists. The algorithm iterates between performing an Expectation (E) step that creates a function for the expectation with respect to the latent variables (trip routes in our context) of the log-likelihood evaluated using current estimates, and a Maximization (M) step that updates the parameter estimates by maximizing the expected log-likelihood from the E-step. The detailed discussion on Gaussian mixture model and EM algorithm can be found in Dempster et al. [26], McLachlan and Krishnan [61], and Bickel and Doksum [9]. The EM algorithm is applied here as:

Step 1: Initialize μ_a, σ_a for all links, and mixing coefficients π_k for all mixture models (each model corresponds to unknown route trips with the same OD pair), and evaluate the initial value of the total log likelihood.

Step 2 (E-step): Evaluate the probabilities $\gamma_k(x_i)$ using the current parameter values based on Equation (3.20).

Step 3 (M-step): Re-estimate the parameters μ_a and σ_a sequentially using the current probabilities $\gamma_k(x_i)$: First keep current σ_a fixed, and update μ_a based on Equation (3.18), then update σ_a based on Equation (3.19).

Also update π_k accordingly: for those trips with the same OD pair, the mixing coefficients π_k are updated based on Equation (3.21).

Step 4: Evaluate the log likelihood as Equation (3.15), and check for the convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied, return to Step 2.

It is noted that updating σ_a values in Step 3 may be challenging due to the complicated nonlinear Equation (3.19) in terms of travel time variances. Referring to Proposition 1 of mean estimates, for simplicity we may compute the mean estimates by following the proposed EM algorithm given the constant variances, and then update and maximize the total likelihood function to solve for variance estimates.

The proposed EM algorithm can guarantee the improvement of total log likelihood at iterations, and lead to the local convergence. The detailed proof is given in Section 3.2.4.

3.2.3 Properties of the Mean Estimates

This section examines whether Equation (3.18) has unique solution with known $\gamma_k(x_i)$ and σ_a in each iteration. For simplicity, we answer by only considering the case that σ_a is identical for all links. Let $\sum_b \delta_{k,b} = N_k,\ k \in K_i, i \in D^u$, denoting the number of links on the possible path k in the set K_i of unlabeled trip i. Also let $\sum_b \delta_{i,b} = N_i,\ i \in D^l$. Then Equation (3.18) turns to

$$\begin{aligned}&\sum_{i\in D^l} \frac{\delta_{i,a}(\sum_b \delta_{i,b}\mu_b)}{N_i} + \sum_{i\in D^u}\sum_{k\in K_i} \frac{\delta_{k,a}\gamma_k(x_i)(\sum_b \delta_{k,b}\mu_b)}{N_k} \\ &= \sum_{i\in D^l} \frac{\delta_{i,a}x_i}{N_i} + \sum_{i\in D^u}\sum_{k\in K_i} \frac{\delta_{k,a}\gamma_k(x_i)x_i}{N_k},\ \text{for any link } a. \end{aligned} \tag{3.22}$$

To explain Equation (3.22) in the matrix form, we define an *augmented* incidence matrix Δ^* as the combination of all labeled and unlabeled trips. The observation

index in Δ^* is arranged beginning with those in D^l followed by those in D^u, so that a labeled trip $i \in D^l$ corresponds to a unique row of $\delta_{i,a}$ in Δ^*, while an unlabeled trip $i \in D^u$ corresponds to multiple rows of $\delta_{k,a}$, $k \in K_i$, in Δ^* (the number of corresponding rows is the cardinality of K_i). The augmented incidence matrix Δ^* differs from the original incidence matrix for including all the possible routes for each trip in D^u. Let Δ^{**} denote a matrix after $\delta_{i,a}$ in Δ^* is scaled by $\frac{1}{N_i}$, for $i \in D^l$ and $\delta_{k,a}$ is scaled by $\frac{\gamma_k(x_i)}{N_k}$ for $k \in K_i$ and $i \in D^u$. These two matrices are illustrated as follows.

$$
\Delta^* = \begin{array}{c} \\ 1 \\ \vdots \\ k \\ k_1^1 \\ \vdots \\ k_m^1 \\ \vdots \\ k_1^j \\ \vdots \\ k_q^j \end{array}
\begin{array}{c}
\begin{array}{ccccc} a_1 & a_2 & \dots & \dots & a_n \end{array} \\
\left(\begin{array}{ccccc}
\delta_{1,a_1} & \delta_{1,a_2} & \dots & \dots & \delta_{1,a_n} \\
\dots & \dots & \ddots & \ddots & \dots \\
\delta_{k,a_1} & \delta_{k,a_2} & \dots & \dots & \delta_{k,a_n} \\
\delta_{k_1^1,a_1} & \delta_{k_1^1,a_2} & \dots & \dots & \delta_{k_1^1,a_n} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\delta_{k_m^1,a_1} & \delta_{k_m^1,a_2} & \dots & \dots & \delta_{k_m^1,a_n} \\
\vdots & \vdots & \ddots & \ddots & \\
\delta_{k_1^j,a_1} & \delta_{k_1^j,a_2} & \dots & \dots & \delta_{k_1^j,a_n} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\delta_{k_q^j,a_1} & \delta_{k_q^j,a_2} & \dots & \dots & \delta_{k_q^j,a_n}
\end{array}\right)
\end{array}
\tag{3.23}
$$

$$\Delta^{**} = \begin{array}{c} \\ 1 \\ \vdots \\ k \\ k_1^1 \\ \vdots \\ k_m^1 \\ \vdots \\ k_1^j \\ \vdots \\ k_q^j \end{array} \begin{array}{c} \begin{array}{cccccc} & & & & & \end{array} \\ \begin{pmatrix} \frac{\delta_{1,a_1}}{N_1} & \frac{\delta_{1,a_2}}{N_1} & \cdots & \cdots & \frac{\delta_{1,a_n}}{N_1} \\ \cdots & \cdots & \ddots & \ddots & \cdots \\ \frac{\delta_{k,a_1}}{N_k} & \frac{\delta_{k,a_2}}{N_k} & \cdots & \cdots & \frac{\delta_{k,a_n}}{N_k} \\ \frac{\gamma_{k_1^1}(x_{k^1})\delta_{k_1^1,a_1}}{N_{k_1^1}} & \frac{\gamma_{k_1^1}(x_{k^1})\delta_{k_1^1,a_2}}{N_{k_1^1}} & \cdots & \cdots & \frac{\gamma_{k_1^1}(x_{k^1})\delta_{k_1^1,a_n}}{N_{k_1^1}} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\gamma_{k_m^1}(x_{k^1})\delta_{k_m^1,a_2}}{N_{k_m^1}} & \frac{\gamma_{k_m^1}(x_{k^1})\delta_{k_m^1,a_2}}{N_{k_m^1}} & \cdots & \cdots & \frac{\gamma_{k_m^1}(x_{k^1})\delta_{k_m^1,a_n}}{N_{k_m^1}} \\ \vdots & \vdots & \ddots & \ddots & \\ \frac{\gamma_{k_1^j}(x_{k^j})\delta_{k_1^j,a_1}}{N_{k_1^j}} & \frac{\gamma_{k_1^j}(x_{k^j})\delta_{k_1^j,a_2}}{N_{k_1^j}} & \cdots & \cdots & \frac{\gamma_{k_1^j}(x_{k^j})\delta_{k_1^j,a_n}}{N_{k_1^j}} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\gamma_{k_q^j}(x_{k^j})\delta_{k_q^j,a_1}}{N_{k_q^j}} & \frac{\gamma_{k_q^j}(x_{k^j})\delta_{k_q^j,a_2}}{N_{k_q^j}} & \cdots & \cdots & \frac{\gamma_{k_q^j}(x_{k^j})\delta_{k_q^j,a_n}}{N_{k_q^j}} \end{pmatrix} \end{array} \tag{3.24}$$

where $1, ..., k \in D^l$ denotes the row index for each labeled trip with known route, and $k_j^i \in K_{k^i}$ denotes the row index for unlabeled trip $k^i \in D^u$ with trip time x_{k^i}. In the presentation (3.24) of matrix Δ^{**}, for example, $k_1^1, ..., k_m^1$ indicate that there are m possible routes for trip time x_{k^1}. In general, $k_1^j, ..., k_q^j$ indicate that there are q possible routes for the trip time x_{k^j}. Note that matrix Δ^* is of the same rank with matrix Δ^{**}.

Equation (3.22) is therefore rewritten as

$$(\Delta^{**})^T \Delta^* \cdot \mu = (\Delta^{**})^T \cdot X. \tag{3.25}$$

where X by abuse of notation denotes the column vector of trip times with proper arrangement and augmentation, i.e., x_{k^i} has $|K_{k^i}|$ duplications in X. Equation (3.25) and prior analysis imply the following proposition.

Proposition 3 There is a unique solution to Equation (3.22) provided that the augmented incidence matrix Δ^* is of full rank, and $\gamma_k(x_i)$ and σ_a are known.

3.2.4 Proof of Convergence

In this section, we prove that the derived EM algorithm can guarantee the non-decrease of total log likelihood and lead to the local convergence, by referring to the classic proof on the convergence of EM algorithm (Dempster et al. [26], Wu [95], Bishop [10]).

For notation convenience, we denote $\Theta = \{\mu, \sigma, \pi\}$ as parameters to be estimated for the log likelihood function (3.15), and represent the total log likelihood of all trip observations as:

$$\begin{aligned} \mathcal{LL}(\Theta) &= \log p(\{x_i, i \in D^l\} \mid \Theta) + \log p(\{x_i, i \in D^u\} \mid \Theta) \\ &= \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \log p(x_i \mid \Theta) \end{aligned} \tag{3.26}$$

As for the second term in Equation (3.26), we denote the hidden variables Y for the route choices of those unlabeled route trips. Therefore, the second term is essentially the *marginal* log likelihood for the observed trip data X. Then, we can convert it as

$$\begin{aligned} \mathcal{LL}(\Theta) &= \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \log p(x_i \mid \Theta) \\ &= \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \log \sum_{k \in K_i} p(x_i, y = k \mid \Theta) \\ &= \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \log \sum_{k \in K_i} p(y = k \mid \Theta) \cdot p(x_i \mid y = k, \Theta) \end{aligned} \tag{3.27}$$

Note that compared to Equation (3.15), here $p(x_i \mid \Theta)$ corresponds to $\mathcal{N}(x_i \mid \sum_a \delta_{i,a}\mu_a, \sum_a \delta_{i,a}\sigma_a^2)$ for any $i \in D^l$, $p(y = k \mid \Theta)$ corresponds to π_k as the probability

of choosing path k, and $p(x_i \mid y = k, \Theta)$ corresponds to $\mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma_a^2)$ for any $i \in D^u, k \in K_i$.

Since there is a summation inside the log for the second term, there would be no longer a nice closed form solution if we maximize the total log likelihood by setting the gradient to zero. The EM algorithm essentially constructs a easy-to-optimize lower bound at each iteration based on the currently obtained parameters.

According to the *Jensen's inequality* $\log \sum_i p_i f_i \geq \sum_i p_i \log f_i$, where p_i forms a probability distribution (i.e., non-negative and sum up to 1), we have

$$\begin{aligned} \mathcal{LL}(\Theta) &= \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \log \sum_{k \in K_i} p(y = k \mid x_i, \Theta^{(t)}) \cdot \frac{p(x_i, y = k \mid \Theta)}{p(y = k \mid x_i, \Theta^{(t)})} \\ &\geq \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \sum_{k \in K_i} p(k \mid x_i, \Theta^{(t)}) \log \frac{p(x_i, k \mid \Theta)}{p(k \mid x_i, \Theta^{(t)})} \end{aligned} \tag{3.28}$$

where $\Theta^{(t)}$ is the estimated parameters at current iteration t, and it is noted that the introduced probability distribution $p(y = k \mid x_i, \Theta^{(t)})$ actually equals the computed $\gamma_k(x_i)$ value in E step to evaluate the posterior probabilities or responsibilities the path k takes to explain trip x_i.

We denote the lower bound of $\mathcal{LL}(\Theta)$ as

$$\mathcal{Q}(\Theta, \Theta^{(t)}) = \sum_{i \in D^l} \log p(x_i \mid \Theta) + \sum_{i \in D^u} \sum_{k \in K_i} p(y = k \mid x_i, \Theta^{(t)}) \log \frac{p(x_i, k \mid \Theta)}{p(k \mid x_i, \Theta^{(t)})} \tag{3.29}$$

Then, in M-step, the lower bound $\mathcal{Q}(\Theta, \Theta^{(t)})$ is actually maximized by setting its gradient to zero, and the obtained closed form solution is indeed Equations (3.18), (3.19) and (3.21). Here, we denote the new estimated parameters as $\Theta^{(t+1)}$. Since $\Theta^{(t+1)}$ maximizes the lower bound function, we have

$$\mathcal{Q}(\Theta^{(t+1)}, \Theta^{(t)}) \geq \mathcal{Q}(\Theta^{(t)}, \Theta^{(t)}) = \mathcal{LL}(\Theta^{(t)}) \tag{3.30}$$

Considering $\mathcal{Q}$ is the lower bound of $\mathcal{LL}$, the following relationship holds

$$\mathcal{LL}(\Theta^{(t+1)}) \geq \mathcal{Q}(\Theta^{(t+1)}, \Theta^{(t)}) \geq \mathcal{Q}(\Theta^{(t)}, \Theta^{(t)}) = \mathcal{LL}(\Theta^{(t)}) \tag{3.31}$$

Therefore, it indicates that the resulting total log likelihood would always non-decrease with iterations, and this EM algorithm can lead to a local maximum of log likelihood function $\mathcal{LL}$.

3.3 Confidence Interval Calculation Based on Profile Likelihood

In practice, it is also important to obtain the confidence intervals for estimated parameters, e.g., for the mean link travel times. The corresponding estimation can be approximated by the profile likelihood method, as briefly described below.

Let $\mu = (\mu_1, \ldots, \mu_n)$ denote the parameters of interest (mean link time in our context) and ϕ a vector of other parameters (i.e., nuisance parameters). Suppose we want to estimate the confidence interval for μ_1. We let $\mu_{-1} = (\mu_2, \ldots, \mu_n)$ and express the log-likelihood function as $\mathcal{LL}(\mu_1, \mu_{-1}, \phi)$. Then we may express the log-likelihood ratio statistic for parameter μ_1, denoted by $r(\mu_1)$, in terms of the profile likelihood function as

$$r(\mu_1) = 2\left\{\max_{\mu,\phi} \mathcal{LL}(\mu, \phi) - \max_{\mu_{-1},\phi} \mathcal{LL}(\mu_1, \mu_{-1}, \phi)\right\} \tag{3.32}$$

It can be shown that $r(\mu_1)$ is asymptotically distributed as χ_1^2 (chi-square distribution with one degree of freedom) when the sample size goes to infinity (see Bickel and Doksum [9]). Therefore, the 95% confidence interval for μ_1 can be approximated as

$$\left\{\mu_1 : r(\mu_1) \leq \chi_1^2(0.95)\right\} \tag{3.33}$$

The profile likelihood method may be computationally expensive for large-scale networks. An alternative approach to estimating confidence interval is through the observed Fisher information matrix (see Bickel and Doksum [9]). In the proposed framework, the observed information matrix can be computed on the last iteration of the EM procedure. We do not present details here, but interested readers can also refer to Louis [58].

3.4 Discussion on Correlation Between Link Travel Times

In the previous sections we consider the link travel times as independent of each other. Such assumption often leads to appropriate results for most application situations, though we do observe travel times on certain links show some degree of correlation. If we incorporate correlation into modeling, we need strong assumptions for the network travel times. The following provides a brief illustration of the proposed approach towards this direction.

We assume that the travel time on links follow the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ where μ denotes the vector of travel times on all n links, i.e., $\mu = (\mu_{a_1}, ..., \mu_{a_n})^T$, and Σ denotes the $n \times n$ covariance matrix, i.e., $\Sigma_{i,j} \triangleq \sigma^2_{i,j} = \text{Cov}(X_i, X_j)$ for travel time X_i and X_j on link i and j, respectively. Then the likelihood function (3.15) becomes

$$\begin{aligned}\mathcal{LL}(\mu, \Sigma, \pi \mid D) &= \sum_{i \in D^u} \log \left(\sum_{k \in K_i} \pi_k \mathcal{N}(x_i \mid \sum_a \delta_{k,a}\mu_a, \sum_a \delta_{k,a}\sigma^2_a + 2\sum_{a<b} \delta_{k,a}\delta_{k,b}\sigma^2_{a,b}) \right) \\ &+ \sum_{i \in D^l} \log \left(\mathcal{N}(x_i \mid \sum_a \delta_{i,a}\mu_a, \sum_a \delta_{i,a}\sigma^2_a + 2\sum_{a<b} \delta_{i,a}\delta_{i,b}\sigma^2_{a,b})) \right) \quad (3.34)\end{aligned}$$

Equations (3.18) and (3.19) for μ and Σ need to be adjusted accordingly. Therefore, the framework in Section 3.2.2 is still applicable in this case. We remark here that the likelihood function (3.34) involves more unknown parameters and hence has a

higher degree of freedom. In order to obtain reasonable results by the proposed framework in this case, some prior knowledge of link correlation may be needed.

4. METHOD II: TRIP SPLITTING APPROXIMATION*

In our earlier models, Gaussian distribution of link travel times gives rise to a trip time that follows a closed form distribution, which makes modeling technically tractable. However, the link time may follow other probability distributions than the Gaussian such as the log-normal, or a mixed distribution due to the recurrent traffic congestion, in which case, no closed form distribution for trip time is available. We propose to split the trip time among traversed links. Different approaches to splitting trip travel time would lead to different estimates. In this section, we propose a statistical method of trip splitting approximation and examine its properties, as in Yin et al. [99].*

We mainly focus on the case that the route of each trip observation is known and travel time on each link follows a certain general distribution. Then we also briefly discuss the case that the trip routes are unknown for some observations.

4.1 General Approach

We denote by D_p the set of trips traveling along path p, and the set P comprising of all paths of trips. In other words, the trips are grouped according to their paths. Let incidence indicator $\delta_{i,a}$ denote if trip i traverses link a. Trip time x_i, $i \in D_p$, actually comprises of unobserved $x_{i,a}$ on link a along path p, and hence $x_i = \sum_a \delta_{i,a} x_{i,a}$. We use $\xi_{i,a}$ to denote the corresponding random variable of travel time on link a for observed trip i, whose realized value being $x_{i,a}$, and denote by $f_a(\,\cdot\,;\Theta_a)$ its probability density function with the parameter vector Θ_a. Also assume that $\xi_{i,a}$ are independent for trip i. Since the link travel time is unobserved,

*Part of this section is reprinted with permission from "Link travel time inference using entry/exit information of trips on a network" by K. Yin, W. Wang, X.B. Wang, and T.M. Adams, 2015. *Transportation Research Part B: Methodological*, 80, 303-321, Copyright [2015] by Elsevier.

we have to maximize the following conditional expected log-likelihood function with respect to all parameters:

$$\mathcal{LL}(\Theta \mid D) = \sum_{p \in P} \sum_{i \in D_p} \mathbb{E}\left\{ \sum_a \delta_{i,a} \log f_a(\xi_{i,a}; \Theta_a) \,\middle|\, \sum_a \delta_{i,a} \xi_{i,a} = x_i \right\} \tag{4.1}$$

where Θ denotes the vector of all parameters in the above function. If we denote by $(x_{i,a})$ the row vector $(x_{i,a_1}, x_{i,a_2}, ..., x_{i,a_n})$ where n is number of links, and denote by $f(\cdot \mid \sum_a \delta_{i,a}\xi_{i,a} = x_i; \Theta) = \prod_a f_a(\cdot \mid \sum_a \delta_{i,a}\xi_{i,a} = x_i; \Theta_a)$ the conditional probability density function of $(\xi_{i,a})$ given the trip observation x_i, then we have

$$\mathcal{LL}(\Theta \mid D) = \sum_{p \in P} \sum_{i \in D_p} \int_{\mathbb{R}^n} f((x_{i,a}) \mid \textstyle\sum_a \delta_{i,a}\xi_{i,a} = x_i; \Theta) \sum_a \delta_{i,a} \log f_a(x_{i,a}; \Theta_a)\, d(x_{i,a}) \tag{4.2}$$

If $\delta_{i,a} = 0$ for some a in the integral in Equation (4.2), the corresponding $x_{i,a}$ will be automatically integrated out. Then the log-likelihood Equation (4.2) should be maximized according to the following

$$\mathcal{LL}(\Theta \mid D) = \sum_{p \in P} \sum_{i \in D_p} \int_{\sum_a \delta_{i,a} x_{i,a} = x_i} \frac{\prod_a f_a(x_{i,a}; \Theta_a)}{\mathbb{P}(\sum_a \delta_{i,a}\xi_{i,a} = x_i; \Theta)} \sum_a \delta_{i,a} \log f_a(x_{i,a}; \Theta_a)\, d(x_{i,a}) \tag{4.3}$$

where $\mathbb{P}(\sum_a \delta_{i,a}\xi_{i,a} = x_i; \Theta) = \int_{\sum_a \delta_{i,a} x_{i,a} = x_i} \prod_a f_a(x_{i,a}; \Theta_a)\, d(x_{i,a})$.

The difficulty of maximizing Equation (4.3) is the evaluation of the multi-dimensional integral. It may be possible to employ Monte Carlo techniques to evaluate the integral, especially when the probability density enjoys some special structures. However, it generally involves expensive computation even for a small-size network, therefore it is difficult to implement in practice.

One practical approach is to approximate the conditional probability density in Equation (4.2) by directly splitting path travel time onto links. We assume that trips

on the same path under similar traffic conditions have more or less a fixed fraction of the trip time for the same link. Let $w_{p,a}$ denote the proportion of travel time on link a among the total travel time on path p, and w be the vector of $w_{p,a}$. If the variation of the proportion $w_{p,a}$ is relatively small, the conditional probability density may be approximated by $\prod_a \delta(x_{i,a} - w_{p,a}x_i)$ where $\delta(\cdot)$ denotes the Dirac delta function (Gelfand and Shilov [32]). This Dirac delta function notation should not be confused with the incidence notation $\delta_{i,a}$. Then the problem of maximizing Equation (4.2) is approximated as [1]

$$\max_{\Theta,w} \mathcal{LL}(\Theta, w \mid D) = \sum_{p\in P} \sum_{i\in D_p} \sum_a \delta_{i,a} \log f_a(w_{p,a}x_i; \Theta_a), \tag{4.4}$$

subject to

$$\sum_a w_{p,a}\delta_{p,a} = 1, \text{ for any trip path } p, \tag{4.5}$$

$$0 \le w_{p,a} \le 1. \tag{4.6}$$

In Equation (4.5), $\delta_{p,a}$ is used for the convenience of notations, denoting if a trip along path p traverses link a. Note that we use $\delta_{p,a}$ instead of previous notation $\delta_{i,a}$ in order to go with the notation $w_{p,a}$, and we also enforce $w_{p,a} = 0$ if $\delta_{p,a} = 0$. Similar to the EM algorithm, an iterative approach to obtain the parameters Θ can be performed by repeating the following steps until convergence. Specifically, at k-th iteration, we have

Step 1 Estimate $w_{p,a}$ by using the estimates of Θ from Step 2 in the $(k-1)$-th iteration;

[1] This derivation applies the property of Dirac delta function: $\int \delta(x_{i,a} - w_{p,a}x_i)g(x_{i,a})\,dx_{i,a} = g(w_{p,a}x_i)$ for any function $g(\cdot)$.

Step 2 Estimate Θ by using $w_{p,a}$ obtained from Step 1.

We next consider a case with known paths of trips in which link times follow Gaussian distributions. This special case allows a comparison with Method I proposed earlier. Then we consider the case of link times following log-normal distributions.

4.2 Case of Gaussian Distribution

We assume that all link travel time variables $\xi_{i,a}$ are independent and follow Gaussian distributions, i.e., $\mathcal{N}(\cdot \mid \mu_a, {\sigma_a}^2)$. The objective (4.4) with constraints (4.5)–(4.6) leads to the following equations, by setting the partial derivative to zero for a specific link a with respective to its parameters μ_a and σ_a respectively,

$$\mu_a = \frac{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}w_{p,a}x_i}{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}}, \tag{4.7}$$

$$\sigma_a^2 = \frac{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}(w_{p,a}x_i-\mu_a)^2}{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}}, \tag{4.8}$$

where $\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}\neq 0$. Obviously, $\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}$ is the number of trip observations traversing link a. We maximize the objective (4.4) with respect to the ratio $w_{p,a}$ by considering Lagrange multipliers:

$$\sum_{p\in P}\sum_{i\in D_p}\sum_{a}\delta_{i,a}\cdot\log\left(\mathcal{N}(w_{p,a}x_i \mid \mu_a, {\sigma_a}^2)\right)+\sum_{p\in P}\lambda_p(\sum_{a}w_{p,a}\delta_{p,a}-1). \tag{4.9}$$

Taking the partial derivative with respective to $w_{p,a}$ and solving for λ_p and $w_{p,a}$, we obtain the following equations

$$\lambda_p=\frac{\sum_{i\in D_p}x_i(x_i-\sum_a\delta_{p,a}\mu_a)}{\sum_a\delta_{p,a}\sigma_a^2}, \text{for any trip path } p \text{ and } \sum_a\delta_{p,a}\sigma_a^2\neq 0, \tag{4.10}$$

and,

$$w_{p,a} = \mu_a \cdot \frac{\sum_{i \in D_p} x_i}{\sum_{i \in D_p} x_i^2} + \lambda_p \cdot \frac{\sigma_a^2}{\sum_{i \in D_p} x_i^2}, \text{for trip path } p, \text{ link } a \text{ and } \delta_{p,a} \neq 0 \quad (4.11)$$

There is a statistical interpretation for the Lagrange multiplier. Consider a simple case where all trips have the same path p. In this case, $\sum_{i \in D_p} \delta_{i,a}$ is the same for any link a along this fixed path, and is denoted by N_p. We also denote the sum of link travel time variance σ_a^2 along the path by σ_p^2. Then plugging Equation (4.7) into Equation (4.10), we have

$$\lambda_p = \frac{\sum_{i \in D_p} x_i^2 - \frac{(\sum_{i \in D_p} x_i)^2}{N_p}}{\sigma_p^2} = \frac{s_p^2}{\sigma_p^2}(N_p - 1), \quad (4.12)$$

where s_p^2 denotes the sample variance of trip travel time along path p. When a large number of trips are observed, $s_p^2 \approx \sigma_p^2$, which gives rise to $\lambda_p \approx N_p - 1$.

We also note that Equation (4.11) cannot guarantee positive $w_{p,a}$ in some extreme cases. If this situation happens, we may either directly solve the constrained optimization (4.4) –(4.6) with fixed μ_a, σ_a^2, or re-initialize $w_{p,a}$ and then perform the iterative algorithm.

To summarize, we apply the iterative algorithm here as:

Step 1: Initialize μ_a, σ_a for all links.

Step 2: Evaluate the Lagrange multipliers λ_p using the current parameter values based on Equation (4.10), and update the splitting ratios $w_{p,a}$ accordingly based on Equation (4.11).

Step 3: Re-estimate the parameters μ_a and σ_a using the current splitting ratios $w_{p,a}$ based on Equations (4.7) and (4.8).

Step 4: Evaluate the total log likelihood as Equation (4.4), and check for the

convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied, return to Step 1; otherwise, terminate.

4.3 Case of Log-Normal Distribution

We consider another case with known paths of trips, in which link travel time variables $\xi_{i,a}$ are independent and follow log-normal distributions for any link a of trip i. We have the probability density function $f_a(x_{i,a};\Theta) = \frac{1}{x_{i,a}}\mathcal{N}(\log x_{i,a} \mid \mu_a, {\sigma_a}^2)$. The objective (4.4) in this case becomes

$$\mathcal{LL}(\mu,\sigma,w \mid D) = \sum_{p\in P}\sum_{i\in D_p}\sum_{a}\delta_{i,a}\cdot\log\left(\frac{1}{w_{p,a}x_i}\mathcal{N}(\log(w_{p,a}x_i) \mid \mu_a, {\sigma_a}^2)\right). \quad (4.13)$$

Similarly, constraints (4.5)–(4.6) still hold.

For the estimates of parameters in Step 2 of the proposed iterative approach in Section 4.1, we have the following equations by fixing $w_{p,a}$:

$$\mu_a = \frac{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}\log(w_{p,a}x_i)}{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}}, \quad (4.14)$$

$$\sigma_a^2 = \frac{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}(\log(w_{p,a}x_i)-\mu_a)^2}{\sum_{p\in P}\sum_{i\in D_p}\delta_{i,a}}. \quad (4.15)$$

For the estimates of $w_{p,a}$ in Step 1 in Section 4.1, there is no closed-form expression. Therefore, we can estimate $w_{p,a}$ through the nonlinear optimization (4.13) with constraints (4.5)–(4.6) by fixing μ_a and σ_a at each iteration.

4.4 Case with Unknown Route Trips

We briefly discuss the case that the routes of some trip observations are unknown in this section. Similar to the previous solution framework, we can apply the EM steps at iterations to infer the unknown routes as well as travel time estimates. If link travel time follows a Gaussian distribution, the derivation in both the E-step

and the M-step may lead to a closed form. While for other general distributions such as log-normal distribution, it may not be possible to obtain closed form solutions in some steps. Taking log-normal distributed link travel time as an example, the splitting ratios may be derived after applying some approximations to the sum of log-normal random variables. One can also apply some optimization methods to solve for the parameters in M-step, though the computational cost would become prohibitive. In general, the EM algorithm is desirable as long as either the E-step or M-step can be solved easily. However, we leave the detailed discussions for the future work.

5. EXPERIMENTAL RESULTS*

This section numerically tests the proposed models and procedures on networks, as in Yin et al. [99].*

First we test the EM algorithm of Method I to see individual algorithm efficiency, followed by testing the trip splitting method for the log-normal distributed link travel times. We also compare the estimates from using both Method I and Method II with link times following Gaussian distributions. All the numerical tests in this section are conducted on a Windows 7 x64 Workstation with two 2.70 GHz CPUs and 4 GB RAM. We code the algorithms in MATLAB, and the convergence criterion is set that the gap of objective value of total likelihood from two consecutive iterations is no larger than 1e-4.

5.1 Test EM Algorithm for the Case with Unknown Route Trips

5.1.1 Test Method I on a Simple Network with 9 Directional Links

Figure 5.1 shows a simple test network consisting of 9 directional links. Given that all the link travel times are independently and normally distributed, trips are generated/observed to guarantee that the rank of the link-path incidence matrix has a rank equal to the number of links to estimate (i.e., a full rank system). In addition, trips on two OD pairs with 'unknown' routes are also generated: from A to F and from C to D, respectively, which means actual routes traversed by those trips are kept from the observed trips and are inferred instead by the proposed procedure earlier. The path information for trips of 'unknown' paths is kept as the ground truth for assessing the estimated paths.

*Part of this section is reprinted with permission from "Link travel time inference using entry/exit information of trips on a network" by K. Yin, W. Wang, X.B. Wang, and T.M. Adams, 2015. *Transportation Research Part B: Methodological*, 80, 303-321, Copyright [2015] by Elsevier.

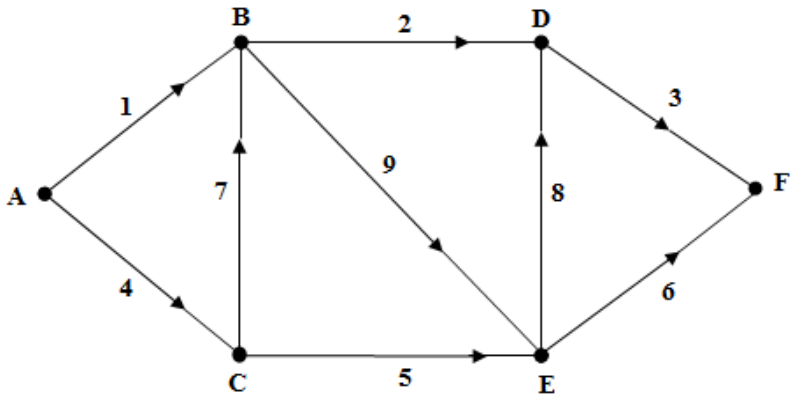

Figure 5.1: A simple test network.

We randomly generate link travel times with an arbitrary mean between 40 and 80 and a standard deviation between 6 and 20 for each link. Different times on the same link are experienced by trips, all following a normal distribution of the same mean and variance. OD trips are generated whose total travel time is the sum of the link times traversed. The generated link times are used as ground truth to assess the link estimates from the proposed methods. As for the test sample size, we generate 50 trips along each link, 50 trips covering multiple links, and also 200 and 50 trips for the two unknown-route OD pairs respectively.

The candidate paths of the two OD pairs with unknown routes are enumerated as: From A to F: [1, 2, 3]; [4, 5, 6]; From C to D: [5, 8]; [7, 2]; [7, 9, 8]. The numbers in each bracket represent the traversed links sequentially.

We then compute the estimates of link means by following the proposed EM algorithm in Method I, and solve for variance estimates by maximizing the total likelihood function. The variances are solved by coding the interior point method in MATLAB for the constrained nonlinear program. The total computational time to obtain estimates is about 2 minutes.

Table 5.1: Estimated and Ground Truth Values of Parameters for Each Link

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
1	72.6	68.4	5.78%	18.7	20.1	7.35%
2	59.6	58.1	2.39%	10.7	12.0	11.70%
3	68.9	68.8	0.09%	8.1	7.6	5.63%
4	53.6	58.4	8.84%	14.5	13.5	7.22%
5	63.0	60.6	3.80%	17.8	18.9	6.25%
6	66.7	65.9	1.19%	14.4	14.9	3.07%
7	57.0	61.2	7.50%	12.0	12.5	4.36%
8	63.3	63.7	0.58%	13.7	14.3	4.24%
9	68.9	69.9	1.50%	18.3	15.6	14.95%
MAPE	-	-	3.52%	-	-	7.20%

Table 5.1 summarizes the resulting estimates and errors, where the Mean Absolute Percentage Error (MAPE) is recorded for each estimate. To illustrate, MAPE for the link mean estimate is calculated as

$$MAPE = \frac{1}{n} \sum_{a=1}^{n} \frac{\mid \mu_a - \hat{\mu}_a \mid}{\mu_a} \tag{5.1}$$

where n denotes the total number of links, $\hat{\mu}_a$ denotes the estimated mean travel time on link a, and μ_a denotes its ground truth mean value.

We also obtain the resulting mixing coefficients for each Gaussian mixture model from optimization of the total likelihood function, and their estimates from the iterative EM algorithm serve as initial guess for the nonlinear optimization. As for trips with unknown routes, the truth is that trips from A to F are equally split between the two alternative paths, and trips from C to D all traverse the path [7, 9, 8]. Table 5.2 shows the estimated mixing coefficients for trips on the two OD pairs, which demonstrates close proximity to the true path choices.

Table 5.2: Estimated Mixing Coefficients for Unlabeled Trips

OD Pair	Candidate Paths	Mixing Coefficients	
AF	$[1, 2, 3]$	π_1^{AF}	0.5765
	$[4, 5, 6]$	π_2^{AF}	0.4235
CD	$[5, 8]$	π_1^{CD}	0
	$[7, 2]$	π_2^{CD}	0
	$[7, 9, 8]$	π_3^{CD}	1

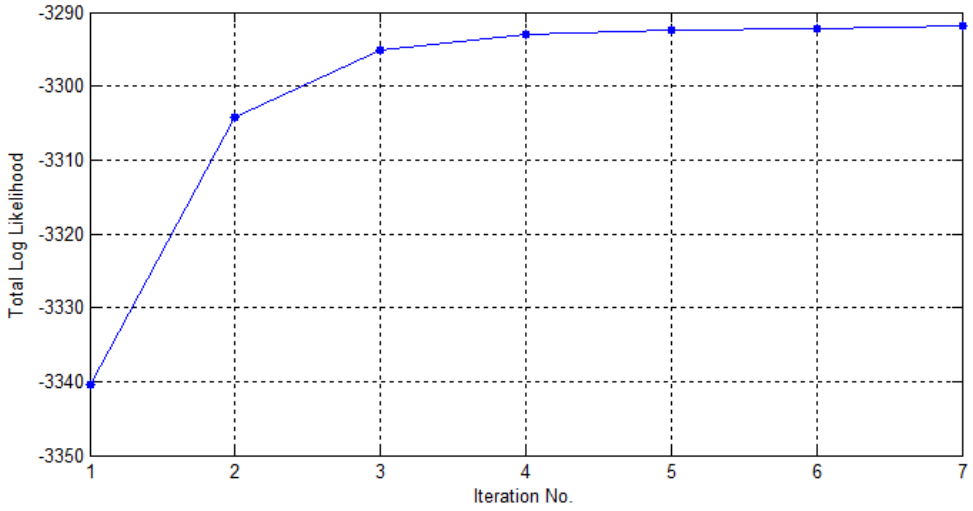

Figure 5.2: The objective value of total log likelihood with iterations for EM method on the 9-link network.

Besides, Figure 5.2 shows that the proposed EM algorithm results in fast improvement of total likelihood to its convergence at iterations.

In addition, we also test the effect of trip observations along a single link and the unknown-route trip observations on estimation accuracy. As shown in Table 5.3, the input setting is the same, except the varying number of trips along each link. We compare the resulting mean estimates using only single-link trips, trips without unlabeled ones, and all the trips respectively. The results demonstrate that the

Table 5.3: Comparison of Mean Estimates with Basic Setting

Trips along Each Link	**Percent**	**MAPE for Mean Estimates**		
		Use Single-link Trips	**Use Labeled Trips**	**Use All Trips**
10	23.08%	9.11%	8.94%	6.86%
20	37.50%	8.89%	8.43%	7.18%
40	54.55%	6.64%	5.97%	5.93%
60	64.29%	4.53%	4.32%	3.44%
80	70.59%	2.43%	2.40%	2.33%

Table 5.4: Comparison of Mean Estimates with Modified Setting

Trips along Each Link	**Percent**	**MAPE for Mean Estimates**		
		Use Single-link Trips	**Use Labeled Trips**	**Use All Trips**
10	13.04%	16.14%	13.93%	10.14%
20	23.08%	13.69%	12.77%	8.33%
40	37.50%	13.63%	12.46%	8.19%
60	47.37%	10.44%	10.15%	7.57%
80	54.55%	6.27%	6.19%	5.06%

proposed method incorporating unknown-route trips can generally lead to more accurate estimates, especially when the single-link observations are insufficient. Then, we modify the input setting to make the sampling of single-link observations more biased, and also double the number of multiple-link trip observations. The resulting estimates are compared in Table 5.4, which indicates that if single-link observations are biased and insufficient, the proposed method can fully utilize all trip observations and make more effective improvement for the mean estimates, while the simple estimation using only single-link observations incurs larger errors.

Table 5.5: Basic Input Information to Generate Test Sample for the Case with Unknown Route Trips

Type of Generated Trips		**Number of Trips**
Trips along each link		10
Trips covering multiple links		550
Trips with unlabeled paths		300
Setting of Randomly Generated Parameters		**Value of Bounds**
Mean	Upper Bound	70
	Lower Bound	40
Standard Deviation	Upper Bound	20
	Lower Bound	6

Table 5.6: Estimate Errors for All Links

Estimated Parameters	**MAPE**
Mean	4.68%
Standard Deviation	12.16%

5.1.2 Test Method I on Sioux Falls Network

The Sioux Falls network in Figure 5.3 consists of 76 links and 24 nodes. The corresponding link number is marked along each link. Based on the predetermined Gaussian distribution on each link, we randomly generate the sample of trip travel times. Trips with 'unknown' routes are also generated similarly as for the 9-link network earlier. The trips generated again guarantees a full rank system. The input information for the estimation analysis is summarized in Table 5.5. We code the iterative procedure in MATLAB, and also use the optimization toolbox in MATLAB to solve the constrained nonlinear problem.

Table 5.6 illustrates estimation errors using the Sioux Falls network, which ap-

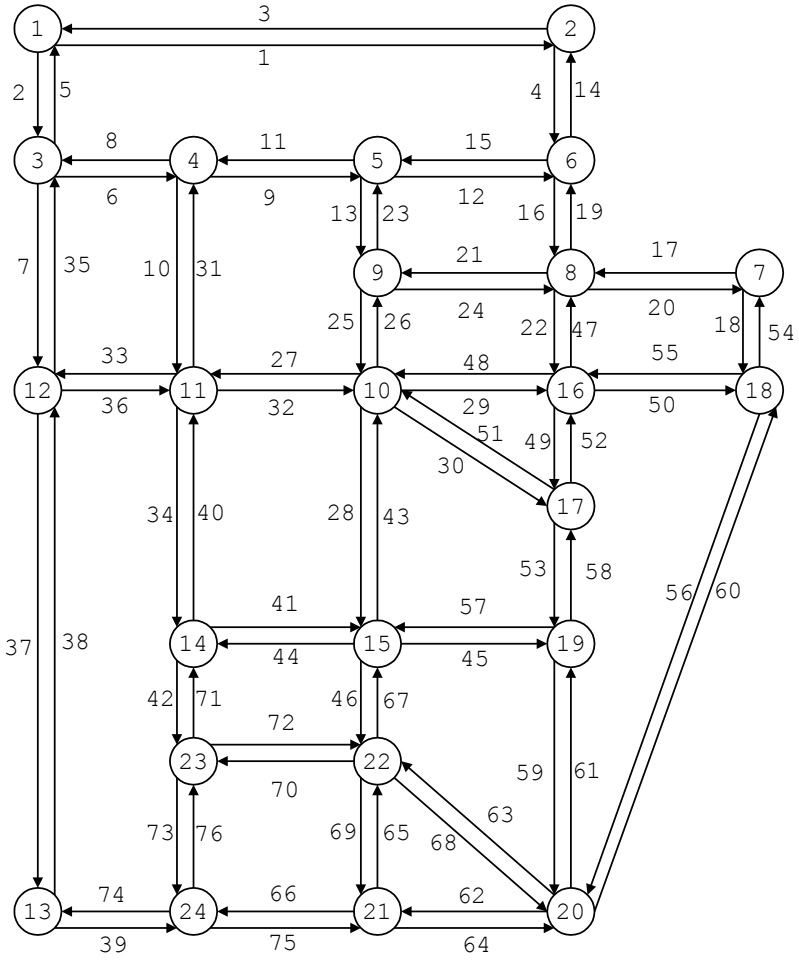

Figure 5.3: Sioux Falls test network.

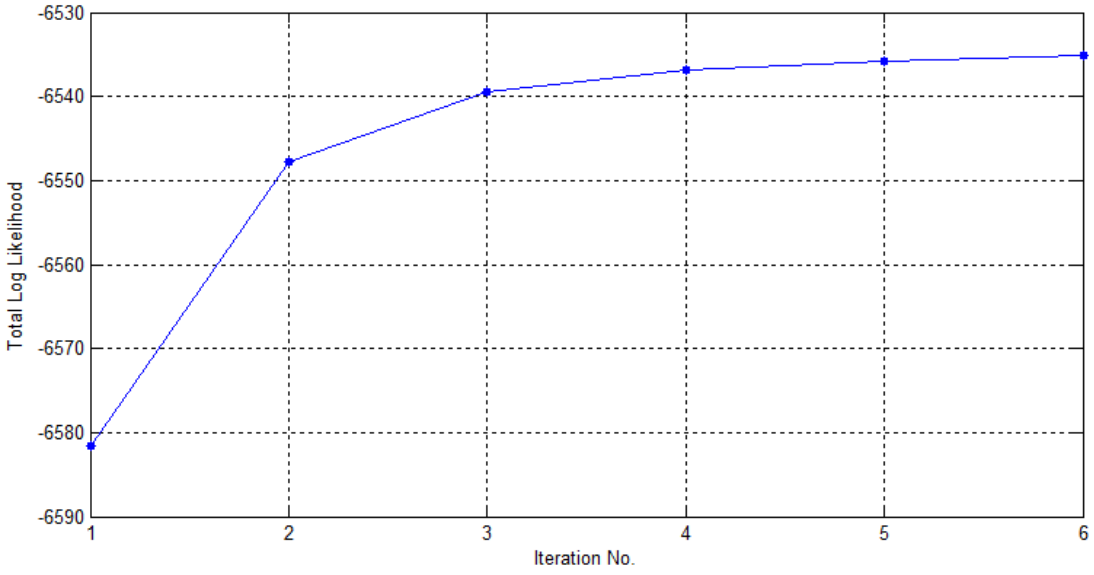

Figure 5.4: The objective value of total log likelihood with iterations for EM method on the Sioux Falls network.

pear to be within range of general acceptance. Figure 5.4 shows that for the Sioux Falls network, the proposed EM algorithm can still lead to convergence of the total likelihood within 11 iterations.

As noted earlier, the application of nonlinear solver in MATLAB may experience computational issues due to the large number of variables. The total computational time here is nearly 20 minutes. Besides, its ability to search for good solutions appears challenged. Therefore, design of heuristic algorithms for this particular constrained nonlinear problem is meaningful in future studies.

To further examine the computational performance of proposed EM algorithm, we test the same Sioux Falls network but with varying number of unknown-route trips as illustrated in Table 5.7. Figure 5.5 displays that the total computational time increases fast with larger sample of unknown-route trips.

Table 5.7: Computational Time of EM Method for Varying Number of Unknown-route Trips

Input for Unknown-route Trips		**Computational Time(min)**
OD Pairs / Trips Per Pair	**Total Trips**	
5 / 15	75	13.26
5 / 30	150	15.00
5 / 45	225	18.15
5 / 60	300	24.49
5 / 75	375	34.69
5 / 90	450	56.05

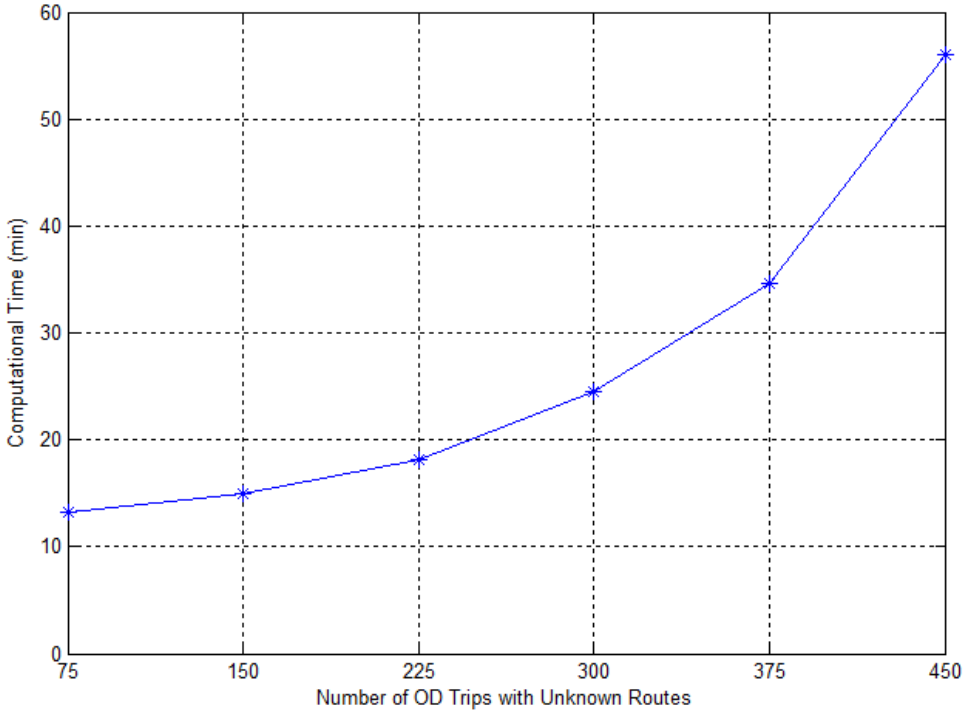

Figure 5.5: Computational time with varying number of unknown-route trips.

Table 5.8: Illustration of Generated Trip Itineraries

Path No.	1	2	3	4	5	6
Link Sequence	[1,2,3]	[1,9,8,3]	[4,5,6]	[5,6]	[7,9,8]	[4,7,9]

5.2 Test Trip Splitting Method for the Case of Log-Normal Distribution

5.2.1 Test Method II on a Simple Network with 9 Directional Links

In this case, all the link travel times are log-normally distributed. We generate link times with randomly selected mean value between 40 and 80, and standard deviation between 6 and 20, similarly as before. The ground truth of parameters for each link are the same as listed in Table 5.1. Specifically, we sample both the trips along a single link and trips covering multiple links as illustrated in Table 5.8, where the numbers in each bracket represent the traversed links sequentially, and we randomly generate 50 trips on each path.

We solve for the parameter estimates for link travel times using trip splitting method. It takes less than 1 minute to obtain the estimates. The errors of resulting estimates are displayed in Table 5.9, where Mean and SD denote the mean and standard deviation of travel times on each link following log-normal distribution, while Mu and Sigma denote the parameters of corresponding normal distribution.

We also find that the splitting ratios between the same pair of links for different trips are consistent. The results are illustrated in Table 5.10, where $w_{p,a}$ denotes the resulting proportion of travel time on link a among those trips along path p.

Besides, Figure 5.6 shows that the trip splitting method results in fast improvement of total likelihood with iterations to convergence.

Then, we test the effect of standard deviations of link travel times on the estima-

Table 5.9: Estimate Errors for Each Link

Link No.	**Estimate Errors**			
	Mean	**SD**	**Mu**	**Sigma**
1	0.60%	13.01%	0.33%	13.19%
2	1.61%	1.94%	0.40%	0.34%
3	1.84%	8.03%	0.41%	6.03%
4	3.22%	23.68%	1.20%	25.47%
5	5.88%	25.53%	1.13%	20.31%
6	2.44%	12.07%	0.49%	9.68%
7	4.56%	2.68%	1.18%	6.79%
8	0.12%	16.67%	0.20%	16.48%
9	4.40%	24.66%	1.41%	27.25%
MAPE	**2.74%**	**14.25%**	**0.75%**	**13.95%**

Table 5.10: Comparison of Splitting Ratios between a Same Link Pair along Various Paths

Link Pairs		**Notations**	**Results**
Links 1 and 3	Path 1	$w_{1,1}/w_{1,3}$	1.0543
	Path 2	$w_{2,1}/w_{2,3}$	1.0559
Links 5 and 6	Path 3	$w_{3,5}/w_{3,6}$	0.9018
	Path 4	$w_{4,5}/w_{4,6}$	0.9036
Links 7 and 9	Path 5	$w_{5,7}/w_{5,9}$	0.8223
	Path 6	$w_{6,7}/w_{6,9}$	0.8157
Links 8 and 9	Path 2	$w_{2,8}/w_{2,9}$	0.8724
	Path 5	$w_{5,8}/w_{5,9}$	0.8793

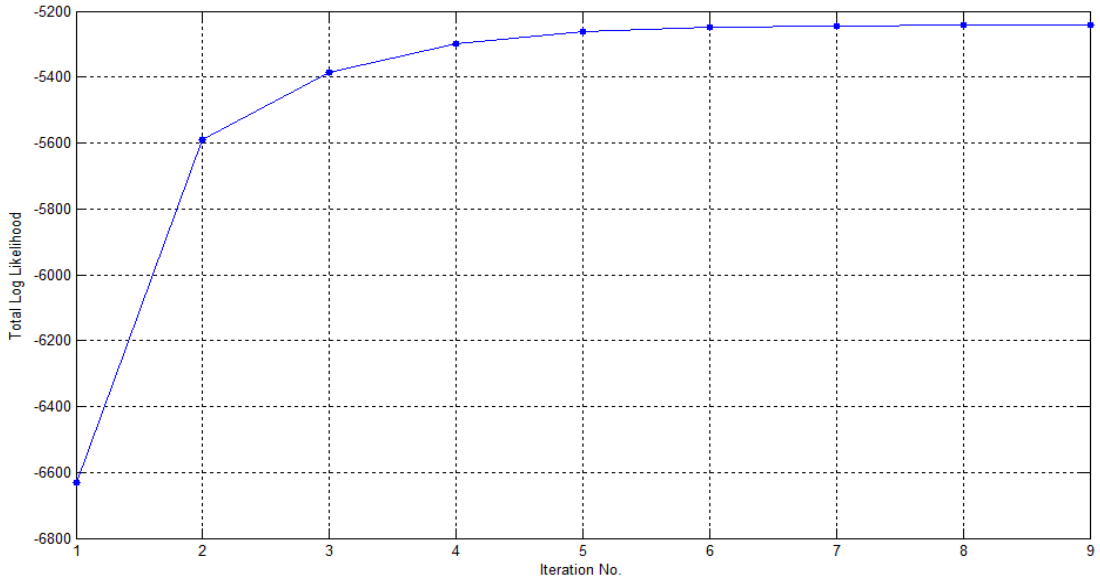

Figure 5.6: The objective value of total log likelihood with iterations for trip splitting method on the 9-link network.

tion accuracy using trip splitting method. Table 5.11 shows that the errors of mean estimates increase with the larger standard deviations of link times, where Mean, SD and Mu are defined the same as for Table 5.9. It is worth noting that the application of trip splitting method would have an issue as the link travel times become more unstable (i.e., with particularly large standard deviation). This can be explained by its underlying assumption of relatively stable traffic conditions on the network. For the practical applications, those individual links with heavy congestion or unexpected incidents need to be identified and carefully examined, which is beyond the scope of this study.

5.2.2 Test Method II on Sioux Falls Network

We additionally test log-normal distributions for links on the Sioux Falls network. The basic input information is summarized in Table 5.12. Note that all trips are with

Table 5.11: Comparison of Mean Estimates with Varying Standard Deviations

Interval to Generate Random Numbers for SD	**MAPE for Estimates**	
	Mean	**Mu**
[3, 6]	0.91%	0.23%
[6, 20]	2.74%	0.75%
[20, 30]	4.97%	1.40%
[30, 40]	7.25%	2.06%

Table 5.12: Basic Input Information to Generate Test Sample for the Case of Log-Normal Distribution

Type of Generated Trips		**Number of Trips**
Trips along each arc		10
Trips covering multiple links		810
Setting of Randomly Generated Parameters		**Value of Bounds**
Mean	Upper Bound	70
	Lower Bound	40
Standard Deviation	Upper Bound	20
	Lower Bound	6

labeled paths. Table 5.13 illustrates the estimation errors from the trip splitting approximation, where Mean, SD, Mu, and Sigma are defined the same as in Table 5.9. Figure 5.7 indicates a fast convergence of the total likelihood with iterations.

In the case of log-normal distribution, solving for splitting ratios in the nonlinear optimization can be computationally expensive at iterations. The total computational time is about 15 minutes in this example. A good initial point of splitting ratios is important. In practice we can always refer to the travel speeds or distances along consecutive links to provide a good starting point of splitting ratios. Besides, those trips traversing a single link are important to trip splitting method. A sample

Table 5.13: Estimate Errors for All Links

Estimated Parameters	**MAPE**
Mean	5.48%
SD	15.73%
Mu	1.40%
Sigma	14.98%

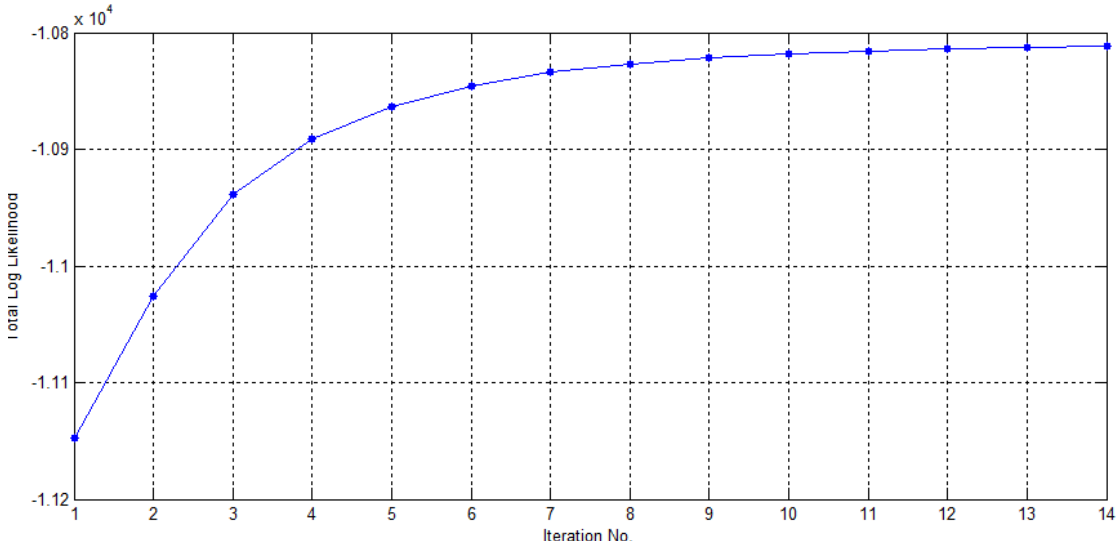

Figure 5.7: The objective value of total log likelihood with iterations for trip splitting method on the Sioux Falls network.

Table 5.14: Comparison of Estimate Errors Using Both Methods on the 9-link Network

Link No.	**Estimate Errors**			
	Mean		**Standard Deviation**	
	Method I	**Method II**	**Method I**	**Method II**
1	0.27%	0.68%	12.01%	14.44%
2	1.85%	1.85%	6.24%	2.21%
3	0.65%	1.50%	3.35%	1.05%
4	4.57%	4.16%	11.70%	25.66%
5	5.00%	4.93%	4.77%	20.77%
6	3.45%	3.30%	2.05%	10.60%
7	3.73%	4.22%	10.31%	7.42%
8	0.40%	1.51%	0.85%	17.12%
9	4.63%	4.09%	7.30%	28.84%
MAPE	**2.73%**	**2.92%**	**6.51%**	**14.23%**

with sufficient single-link trips can help get accurate estimates.

5.3 Compare the Estimates Using Two Methods for the Case of Gaussian Distribution

In the case of Gaussian distributions for link travel times, we compare the estimates from using both Method I and Method II (i.e., trip splitting method) as they apply to the simple 9-link network and the Sioux Falls network respectively. The test sample size is the same as used in Section 5.2. Tables 5.14 and 5.15 summarize the estimates from both methods. Besides, we also calculate the 95% confidence interval of the mean estimates for several links on the 9-link network, as illustrated in Table 5.16. The resulting confidence intervals for mean estimates appear very close under two model approaches.

The trip splitting method of Method II generally runs very fast at iterations compared with Method I in the case of Gaussian distribution. For example, it takes

Table 5.15: Comparison of Estimate Errors Using Both Methods on the Sioux Falls Network

Estimated Parameters	**MAPE**	
	Method I	**Method II**
Mean	4.98%	5.38%
Standard Deviation	9.00%	15.59%

Table 5.16: Illustration of 95% Confidence Interval Calculation for 9-link Network

Link No.	**Method I**		**Method II**	
	Mean Estimate	**CI**	**Mean Estimate**	**CI**
1	72.79	[69.20, 76.99]	73.08	[70.30, 76.49]
4	56.10	[53.38, 58.61]	55.88	[53.73, 58.12]
8	63.59	[60.75, 66.43]	63.67	[61.26, 65.91]

only a few seconds for Sioux Falls network using trip splitting method, as compared with nearly 3 minutes using Method I. This is likely because Method I takes relatively long time to solve for variance estimates in the nonlinear optimization. However, accuracy of the trip splitting method may be in the check, especially that of the variance estimates. To summarize the trade off again, the trip splitting method may incur larger errors with estimating the standard deviation of link travel time in trade for a much faster time compared with Method I. In terms of the mean estimates, both methods are comparably competitive.

6. DISCUSSION OF THE TWO METHODS*

This section discusses the advantages and disadvantages of both methods, as in Yin et al. [99].*

The two proposed methods in this dissertation are based on the additive property of link time distributions, i.e., whether the summation of link travel times has a closed-form distribution. Starting from the likelihood principle, both methods, whenever necessary, decompose the link travel time inference into structural steps that share the same spirit of the EM machinery. The key strategy is the introduction of the augmented data (or complete data), namely augmenting the observed data with hidden (unobserved) variables that represent the problem structure.

In Method I, the unobserved variables represent the path choices for individual travelers with unknown routes. While the proposed method involve path inference, it mainly focuses on the estimation of model parameters (so as to approximate the real values) and the stable solution, rather than the accuracy of individual path inference. The investigation of the case with all trips of known routes reveals its connection to a least squares solution. And the analysis on the property of mean estimates when trip variance estimates are within reasonable errors demonstrates the validity of our iterative calculation of mean and variance. The hard-assignment algorithm that addresses the case with some trips of unknown routes usually provides the initial solution to the soft-assignment algorithm. Because of easy computation, solution from hard assignment can also serve as a crude approximation to the real values. When dealing with the uncertainty of path choices, applying the mixture

*Part of this section is reprinted with permission from "Link travel time inference using entry/exit information of trips on a network" by K. Yin, W. Wang, X.B. Wang, and T.M. Adams, 2015. *Transportation Research Part B: Methodological*, 80, 303-321, Copyright [2015] by Elsevier.

model in the soft assignment is more appropriate.

While similar methods to Method I have somewhat been studied in literature, the proposed Method II appears new. The method of splitting trip travel time is straightforward, but directly applying it cannot guarantee certain property of results. We show that this method can be viewed from the statistical perspective, and redesign the method through maximum conditional likelihood function. The trip splitting method is fast in computation compared to Method I, and can apply to various link time distributions. But it requires many parameters. Some variable selection techniques (e.g., Fan et al. [30]) can be used to overcome the proliferation of parameters. Moreover, since the E-step involves a probability inference for the augmented variables based on the observed data, properly defining the augmented variables can help improve the convergence of the algorithm (Meng and Van Dyk [62]). The proposed framework of trip splitting is built mainly from the statistical perspective, which can further combine the results from conventional traffic flow theory in order to obtain more reliable estimates for practical applications. For example, one may incorporate the empirical speed-volume relationship into the iterative procedure to generate reasonable splitting ratios.

The maximum-likelihood model framework can also be extended to deal with the correlation among different links. Our modeling approaches will lay a methodological foundation that we can use to extend to a dynamic network with time-dependent link travel times (e.g., Xing et al. [97]), to the OD flow estimation (e.g., Parry and Hazelton [71]), to the travel time reliability (e.g., Ng et al. [67]), or to the day-to-day dynamic travel pattern inference (e.g., Parry and Hazelton [72]), as future study efforts. As different sources of traffic data become available (see Zheng et al. [103]; Mori et al. [64]), we will find the proposed statistical framework useful.

7. CONCLUSIONS

Link travel time estimation on a roadway network is essential for performance assessment in order to improve traffic mobility and network efficiency. It is made possible now by widely available traffic data. This dissertation develops model framework based on statistical inference methods for link travel time estimation using entry/exit information of trips on a network.

First, we propose a method considering that the trip time has a closed-form distribution, using independent Gaussian distribution for link travel time as an example. We particularly analyze the property of mean estimates and investigate the uniqueness of solutions in the derived EM algorithm. To overcome the modeling challenge that random link times do not typically add up to a trip time with close-form distribution, we develop a trip splitting method assuming a relatively reliable way to partition the trip time between links. The proposed trip splitting method applies to the general case with arbitrary link travel time distribution, although with varying complexity in computation, making it potentially applicable to many traffic situations. And it is also statistically justified for the network estimation problem.

The proposed methods are tested and compared numerically on two networks, a simple 9-link network and the Sioux Falls network. The experimental results indicate that both methods perform well and generate quality estimates, and that the trip splitting method generally runs much faster. A trade off is that the trip splitting method incurs larger errors for standard deviation estimates than the first method.

Worthy of a special mention is that link travel time inference on a network is complicated and much still remains to be addressed. For example, can we obtain reliable link travel time estimates if the mapping relationship between trip itineraries

and link travel times has the rank deficiency issue? How do we improve the computational performance when solving for large number of estimates for a large-size network? Besides, further extension to the Bayesian approach with more traffic data available as prior information is also worth our examinations, and the application to various realistic networks with empirical data would be desirable in the future work.

REFERENCES

[1] B. Abdulhai and S. M. Tabib. Spatio-temporal inductance-pattern recognition for vehicle re-identification. *Transportation Research Part C: Emerging Technologies*, 11(3):223–239, 2003.

[2] H. Al-Deek and E. B. Emam. New methodology for estimating reliability in transportation networks with degraded link capacities. *Journal of Intelligent Transportation Systems*, 10(3):117–129, 2006.

[3] H. Al-Deek, A. Mohamed, and A. Radwan. Operational benefits of electronic toll collection: case study. *Journal of Transportation Engineering*, 123(6):467–477, 1997.

[4] S. Arroyo and A. L. Kornhauser. Modeling travel time distributions on a road network. In *Proceedings of the 2005 TRB Annual Conference*, 2005.

[5] H. Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, 2007.

[6] J. Barceló, L. Montero, L. Marqués, and C. Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, 2175(1):19–27, 2010.

[7] R. L. Bertini and S. Tantiyanugulchai. Transit buses as traffic probes: empirical evaluation using geo-location data. *Transportation Research Record: Journal of the Transportation Research Board*, 1870:35–45, 2004.

[8] A. Bhaskar, E. Chung, and A.-G. Dumont. Estimation of travel time on urban networks with midlink sources and sinks. *Transportation Research Record: Journal of the Transportation Research Board*, 2121(1):41–54, 2009.

[9] P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics*. Prentice Hall, 2000.

[10] C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[11] K. S. Chan, W. H. K. Lam, and M. L. Tam. Real-time estimation of arterial travel times with spatial travel time covariance relationships. *Transportation Research Record: Journal of the Transportation Research Board*, 2121(1):102–109, 2009.

[12] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *Intelligent Transportation Systems, IEEE Transactions on*, 5(1):42–53, 2004.

[13] A. Chen, Z. Ji, and W. Recker. Travel time reliability with risk-sensitive travelers. *Transportation Research Record: Journal of the Transportation Research Board*, (1783):27–33, 2002.

[14] C. Chen, A. Skabardonis, and P. Varaiya. Travel-time reliability as a measure of service. *Transportation Research Record: Journal of the Transportation Research Board*, (1855):74–79, 2003.

[15] M. Chen and S. I. Chien. Determining the number of probe vehicles for freeway travel time estimation by microscopic simulation. *Transportation Research Record: Journal of the Transportation Research Board*, 1719(1):61–68, 2000.

[16] S. Clark and D. Watling. Modelling network travel time reliability under stochastic demand. *Transportation Research Part B: Methodological*, 39(2):119–140, 2005.

[17] B. Coifman. Improved velocity estimation using single loop detectors. *Transportation Research Part A: Policy and Practice*, 35(10):863–880, 2001.

[18] B. Coifman. Estimating travel times and vehicle trajectories on freeways using dual loop detectors. *Transportation Research Part A: Policy and Practice*, 36(4):351–364, 2002.

[19] B. Coifman and M. Cassidy. Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research Part A: Policy and Practice*, 36(10):899–917, 2002.

[20] B. Coifman and E. Ergueta. Improved vehicle reidentification and travel time measurement on congested freeways. *Journal of Transportation Engineering*, 129(5):475–483, 2003.

[21] B. Coifman and S. Krishnamurthy. Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C: Emerging Technologies*, 15(3):135–153, 2007.

[22] D. J. Dailey. Travel-time estimation using cross-correlation techniques. *Transportation Research Part B: Methodological*, 27(2):97–107, 1993.

[23] D. J. Dailey. A statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B: Methodological*, 33(5):313–322, 1999.

[24] K. Davidson. A flow travel time relationship for use in transportation planning. In *Australian Road Research Board (ARRB) Conference, 3rd, 1966, Sydney*, volume 3, 1966.

[25] P. Davies, C. Hill, and N. Emmott. Automatic vehicle identification to support driver information systems. In *Vehicle Navigation and Information Systems Conference, 1989. Conference Record*, pages A31–A35. IEEE, 1989.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[27] F. Dion and H. Rakha. Estimating spatial travel times using automatic vehicle identification data. In *82nd Annual Meeting Preprint CD-ROM, Transportation Research Board, Washington, DC*, pages 12–16, 2003.

[28] F. Dion and H. Rakha. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transportation Research Part B: Methodological*, 40(9):745–766, 2006.

[29] E. Emam and H. Ai-Deek. Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, (1959):140–150, 2006.

[30] J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National Science Review*, 1:293–314, 2014.

[31] Y. Feng, J. Hourdos, and G. A. Davis. Probe vehicle based real-time traffic monitoring on urban roadways. *Transportation Research Part C: Emerging Technologies*, 40:160–178, 2014.

[32] I. M. Gelfand and G. Shilov. *Generalized functions. Vol. I: Properties and operations, Translated by Eugene Saletan.* Academic Press, New York, 1964.

[33] F. Guo, H. Rakha, and S. Park. Multistate model for travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, (2188):46–54, 2010.

[34] A. Haghani, M. Hamedi, K. F. Sadabadi, S. Young, and P. Tarnoff. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160(1):60–68, 2010.

[35] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

[36] B. Hellinga, P. Izadpanah, H. Takada, and L. Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C: Emerging Technologies*, 16(6):768–782, 2008.

[37] B. R. Hellinga and L. Fu. Reducing bias in probe-based arterial link travel time estimates. *Transportation Research Part C: Emerging Technologies*, 10(4):257–273, 2002.

[38] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1679–1693, 2012.

[39] A. Hofleitner, R. Herring, and A. Bayen. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological*, 46(9):1097–1122, 2012.

[40] J. Hopkin, D. Crawford, and I. Catling. Travel time estimation. In *Summary of the European Workshop Organized by the SERTI Project, Avignon*, 2001.

[41] HoustonTranStar. Transtar description. (http://traffic.houstontranstar.org). 2001.

[42] S.-R. Hu, S. Peeta, and C.-H. Chu. Identification of vehicle sensor locations for link-based network traffic applications. *Transportation Research Part B: Methodological*, 43(8):873–894, 2009.

[43] T. Hunter, T. Das, M. Zaharia, P. Abbeel, and A. M. Bayen. Large-scale estimation in cyberphysical systems using streaming data: a case study with arterial traffic estimation. *Automation Science and Engineering, IEEE Transactions on*, 10(4):884–898, 2013.

[44] T. Hunter, R. Herring, P. Abbeel, and A. Bayen. Path and travel time inference from gps probe vehicle data. *NIPS Analyzing Networks and Learning with Graphs*, 2009.

[45] E. Jenelius and H. N. Koutsopoulos. Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53:64–81, 2013.

[46] K. Jintanakul, L. Chu, and R. Jayakrishnan. Bayesian mixture model for estimating freeway travel time distributions from small probe samples from multiple days. *Transportation Research Record: Journal of the Transportation Research Board*, (2136):37–44, 2009.

[47] H. Jula, M. Dessouky, and P. A. Ioannou. Real-time estimation of travel times along the arcs and arrival times at the nodes of dynamic stochastic networks. *IEEE Transactions on Intelligent Transportation Systems*, 9(1):97–110, 2008.

[48] E. Kazagli and H. N. Koutsopoulos. Arterial travel time estimation from automatic number plate recognition data. In *Proceedings of the 92nd Annual TRB Meeting, Washington, DC*, 2013.

[49] J. Kim and H. S. Mahmassani. A finite mixture model of vehicle-to-vehicle and day-to-day variability of traffic network travel times. *Transportation Research Part C: Emerging Technologies*, 46:83–97, 2014.

[50] G. Leduc. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, 1:55, 2008.

[51] R. Li, G. Rose, and M. Sarvi. Evaluation of speed-based travel time estimation models. *Journal of transportation engineering*, 132(7):540–547, 2006.

[52] X. Li and Y. Ouyang. Reliable sensor deployment for network traffic surveillance. *Transportation research part B: methodological*, 45(1):218–231, 2011.

[53] Z. Li, D. A. Hensher, and J. M. Rose. Willingness to pay for travel time reliability in passenger transport: a review and some new empirical evidence. *Transportation Research Part E: Logistics and Transportation Review*, 46(3):384–403, 2010.

[54] W.-H. Lin, A. Kulkarni, and P. Mirchandani. Short-term arterial travel time prediction for advanced traveler information systems. In *Intelligent Transportation Systems*, volume 8, pages 143–154. Taylor & Francis, 2004.

[55] H. X. Liu, W. Recker, and A. Chen. Uncovering the contribution of travel time reliability to dynamic route choice using real-time loop data. *Transportation Research Part A: Policy and Practice*, 38(6):435–453, 2004.

[56] J. Long, Z. Gao, and W. Szeto. Discretised link travel time models based on cumulative flows: formulations and properties. *Transportation Research Part*

B: Methodological, 45(1):232–254, 2011.

[57] S. Lorkowski, P. Mieth, and R. Schäfer. New its applications for metropolitan areas based on floating car data. In *ECTRI Young Researcher Seminar, Den Haag*, 2005.

[58] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[59] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. Mining smart card data for transit riders travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[60] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[61] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[62] X.-L. Meng and D. Van Dyk. The em algorithman old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.

[63] E. F. Morgul, K. Ozbay, S. Iyer, and J. Holguin-Veras. Commercial vehicle travel time estimation in urban networks using gps data from multiple sources. In *Transportation Research Board 92nd Annual Meeting*, number 13-4439, 2013.

[64] U. Mori, A. Mendiburu, M. Álvarez, and J. A. Lozano. A review of travel time estimation and forecasting for advanced traveller information systems.

Transportmetrica A: Transport Science, (ahead-of-print):1–39, 2014.

[65] K. C. Mouskos, E. Niver, L. J. Pignataro, S. Lee, N. Antonion, and L. Papadopoulos. Transmit system evaluation. *Final Report, Newark: New Jersey Institute of Technology*, 1998.

[66] D. H. Nam and D. R. Drew. Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. *Journal of Transportation Engineering*, 122(3):185–191, 1996.

[67] M. Ng, W. Szeto, and S. Travis Waller. Distribution-free travel time reliability assessment with probability inequalities. *Transportation Research Part B: Methodological*, 45(6):852–866, 2011.

[68] R. B. Noland and J. W. Polak. Travel time variability: a review of theoretical and empirical issues. *Transport Reviews*, 22(1):39–54, 2002.

[69] Y. Ohba, H. Ueno, and M. Kuwahara. Travel time calculation method for expressway using toll collection system data. In *Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEJ/JSAI International Conference on*, pages 471–475. IEEE, 1999.

[70] P. V. Palacharla and P. C. Nelson. Application of fuzzy logic and neural networks for dynamic travel time estimation. *International Transactions in Operational Research*, 6(1):145–160, 1999.

[71] K. Parry and M. L. Hazelton. Estimation of origin–destination matrices from link counts and sporadic routing data. *Transportation Research Part B: Methodological*, 46(1):175–188, 2012.

[72] K. Parry and M. L. Hazelton. Bayesian inference for day-to-day dynamic traffic models. *Transportation Research Part B: Methodological*, 50:104–115, 2013.

[73] K. F. Petty, P. Bickel, M. Ostland, J. Rice, F. Schoenberg, J. Jiang, and Y. Ritov. Accurate estimation of travel times from single-loop detectors. *Transportation Research Part A: Policy and Practice*, 32(1):1–17, 1998.

[74] A. Pushkar, F. L. Hall, and J. A. Acha-Daza. Estimation of speeds from single-loop freeway flow and occupancy data using cusp catastrophe theory model. *Transportation Research Record*, (1457), 1994.

[75] M. Rahmani and H. N. Koutsopoulos. Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, 30:41–54, 2013.

[76] H. Rakha, I. El-Shawarby, M. Arafeh, and F. Dion. Estimating path travel-time reliability. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, pages 236–241. IEEE, 2006.

[77] S. Samaranayake, S. Blandin, and A. Bayen. A tractable class of algorithms for reliable routing in stochastic networks. *Transportation Research Part C: Emerging Technologies*, 20(1):199–217, 2012.

[78] S. Sananmongkhonchai, P. Tangamchit, and P. Pongpaibool. Road traffic estimation from multiple gps data using incremental weighted update. In *ITS Telecommunications, 2008. ITST 2008. 8th International Conference on*, pages 62–66. IEEE, 2008.

[79] K. K. Sanwal and J. Walrand. Vehicles as probes. *California Partners for Advanced Transit and Highways (PATH)*, 1995.

[80] T. Siripirote, A. Sumalee, D. P. Watling, and H. Shao. Updating of travel behavior model parameters and estimation of vehicle trip chain based on plate scanning. *Journal of Intelligent Transportation Systems*, 18:393–409, 2013.

[81] V. P. Sisiopiku and N. M. Rouphail. Toward the use of detector output for arterial link travel time estimation: A literature review. *Transportation Research Record*, (1457), 1994.

[82] A. Skabardonis and R. Dowling. Improved speed-flow relationships for planning applications. *Transportation Research Record: Journal of the Transportation Research Board*, (1572):18–23, 1997.

[83] F. Soriguera, D. Rosas, and F. Robusté. Travel time measurement in closed toll highways. *Transportation Research Part B: Methodological*, 44(10):1242–1267, 2010.

[84] H. Spiess. Technical noteconical volume-delay functions. *Transportation Science*, 24(2):153–158, 1990.

[85] K. K. Srinivasan and P. P. Jovanis. Determination of number of probe vehicles required for reliable travel time measurement in urban network. *Transportation Research Record: Journal of the Transportation Research Board*, 1537(1):15–22, 1996.

[86] C. Sun, S. G. Ritchie, K. Tsai, and R. Jayakrishnan. Use of vehicle signature analysis and lexicographic optimization for vehicle reidentification on freeways. *Transportation Research Part C: Emerging Technologies*, 7(4):167–185, 1999.

[87] SwRI. Automatic vehicle identification model deployment initiative - system design document. *Report Prepared for TransGuide Texas Department of Transportation, Southwest Research Institute, San Antonio, TX*, 1998.

[88] M. L. Tam and W. H. Lam. Application of automatic vehicle identification technology for real-time journey time estimation. *Information Fusion*, 12(1):11–19, 2011.

[89] S. M. Turner, W. L. Eisele, R. J. Benz, and D. J. Holdener. Travel time data collection handbook. Technical report, 1998.

[90] S. M. Turner and D. J. Holdener. Probe vehicle sample sizes for real-time information: the houston experience. In *Vehicle Navigation and Information Systems Conference, 1995. Proceedings. In conjunction with the Pacific Rim TransTech Conference. 6th International VNIS.'A Ride into the Future'*, pages 3–10. IEEE, 1995.

[91] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *Proceeding of the 20th SIGKDD conference on Knowledge Discovery and Data Mining*, 2014.

[92] T. Wen, C. C. Li, C. J. Che, L. D. Zhong, and X. Xin. Research on link travel time reliability model based on massive expressway toll data. *Applied Mechanics and Materials*, 505:719–726, 2014.

[93] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Travel time estimation for ambulances using bayesian data augmentation. *The Annals of Applied Statistics*, 7(2):1139–1161, 2013.

[94] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Large-network travel time distribution estimation for ambulances. *Submitted to European Journal of Operational Research*, 2014.

[95] C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[96] C. Xie, R. L. Cheu, and D.-H. Lee. Calibration-free arterial link speed estimation model using loop data. *Journal of Transportation Engineering*, 127(6):507–514, 2001.

[97] T. Xing, X. Zhou, and J. Taylor. Designing heterogeneous sensor networks for estimating and predicting path travel time dynamics: An information-theoretic modeling approach. *Transportation Research Part B: Methodological*, 57:66–90, 2013.

[98] J.-L. Ygnace, C. Drane, Y. B. Yim, and R. De Lacvivier. Travel time estimation on the san francisco bay area network using cellular phones as probes. *California Partners for Advanced Transit and Highways (PATH) UCB-ITS-PWP-2000-18*, September 2000.

[99] K. Yin, W. Wang, X. B. Wang, and T. M. Adams. Link travel time inference using entry/exit information of trips on a network. *Transportation Research Part B: Methodological*, 80:303–321, 2015.

[100] T. Yokota and D. Tamagawa. Route identification of freight vehicle's tour using gps probe data and its application to evaluation of on and off ramp usage of expressways. *Procedia-Social and Behavioral Sciences*, 39:255–266, 2012.

[101] X. Zhang and J. A. Rice. Short-term travel time prediction. *Transportation Research Part C: Emerging Technologies*, 11(3):187–210, 2003.

[102] F. Zheng and H. Van Zuylen. Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 31:145–157, 2013.

[103] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology (ACM TIST)*, 2014.

[104] L. Zou, J.-M. Xu, and L.-X. Zhu. Arterial speed studies with taxi equipped with global positioning receivers as probe vehicle. In *Wireless Communica-*

tions, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on, volume 2, pages 1343–1347. IEEE, 2005.

APPENDIX A

SUPPLEMENT TO SECTION 3

A.1 Alternative Representation of Equations (3.5) and (3.6)

We present another way to format Equations (3.5) and (3.6) by the path-link relationship. The same approach is applied to other equations in the dissertation. Let the path-link matrix be $\Delta_P = [\delta_{p_i,a}]_{v\times n}$, where v is the number of all paths, p_i the i-th path connecting an origin and a destination, and n the total number of links. We also denote by n_{p_i} the number of observations and by $\bar{x}_{p_i}$ the average of the travel time along path p_i. Then Equations (3.5) and (3.6) are read as

$$0 = \sum_{p_i} \frac{\delta_{p_i,a} n_{p_i} (\sum_b \delta_{p_i,b}\mu_b - \bar{x}_{p_i})}{\sum_b \delta_{p_i,b}\sigma_b^2}, \tag{A.1}$$

$$0 = \sum_{p_i} \frac{\delta_{p_i,a} n_{p_i}}{(\sum_b \delta_{p_i,b}\sigma_b^2)^2} \left(\sum_b \delta_{p_i,b}\sigma_b^2 - \frac{1}{n_{p_i}} \sum_{j\in p_i} (x_j - \sum_b \delta_{p_i,b}\mu_b)^2 \right). \tag{A.2}$$

where $j \in p_i$ means that the observation j associates with route p_i. Let $\tilde{\Delta}_P$ be the matrix Δ_P scaled by $n_{p_i}(\sum_b \delta_{p_i,b}\sigma_b^2)^{-1}$ for all $\delta_{p_i,a}$ in the row p_i, and let $\bar{X}$ be the vector of $\bar{x}_{p_i}$. Then Equation (A.1) can be also written as

$$\tilde{\Delta}_P^T \Delta_P \cdot \mu = \tilde{\Delta}_P^T \cdot \bar{X}. \tag{A.3}$$

Comparing with the matrix representation in Section 3.1.1, the above equation is more compact and saves memory for numerical computation. However, it seems that Equation (A.2) does not have a more compact representation than the one in Section 3.1.1.

A.2 Introduction of K-means Clustering Algorithm

K-means clustering is a method often used to identify groups or clusters of data points. The term of K-means was first proposed by MacQueen [60] in 1967. It aims to partition N observations into K clusters in a way that each observation belongs to the cluster with the nearest mean value, serving as the prototype of the cluster. The problem is defined as follows.

Given a data set $\{x_1, ..., x_N\}$ consisting of N observations of a random D-dimensional variable x, the objective is to find an assignment of these observations into K clusters, as well as a set of D-dimensional vectors $\{\mu_k\}$ where μ_k denotes the prototype associated with the k^{th} cluster, such that the sum of squared distances of each data point to its closest vector μ_k is minimized. The objective is formulated as below.

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\|x_i - \mu_k\|^2. \tag{A.4}$$

where $r_{ik} \in \{0, 1\}$ is binary indicator variables denoting if data point x_i is assigned to cluster k.

The basic idea of K-means algorithm is the successive optimization with respect to r_{ik} and μ_k: given the initial values for the $\{\mu_k\}$, in the first stage minimize the objective with respect to r_{ik} with fixed values of μ_k; then in the second stage minimize the objective with respect to μ_k with fixed values of r_{ik}. Repeat this two-stage optimization until convergence. These two stages of updating r_{ik} and μ_k essentially corresponds to the E(expectation) and M(maximization) steps respectively in the EM algorithm.

Therefore, an iterative procedure to solve this problem involves two successive steps as

Step 1: Assign each data point to the cluster whose mean yields the least within-cluster sum of squares:

$$r_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in K} \|x_i - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

Step 2: Update the mean (i.e. centroid) of all the data points assigned to each cluster:

$$\mu_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}}, \tag{A.6}$$

which is the result of minimizing the objective of within-cluster sum of squares, with the r_{ik} held fixed.

Repeat these two steps until there is no further change in the assignments.

This procedure is known as K-means algorithm. The convergence of this algorithm is assured since each step reduces the value of the objective function. However, it may converge to a local optimum rather than global optimum. The convergence properties of this K-means algorithm have been studied by MacQueen [60]. A commonly used initialization method is to randomly choose a subset of K data points to get the cluster centers μ_k and use them as the initial means. Typically one can use multiple runs from random starting guesses, and chooses the solution with the smallest within-cluster sum of squares. This K-means algorithm is essentially a variant of the generalized EM algorithm.

APPENDIX B

SUPPLEMENT TO SECTION 5

As for testing EM algorithm for the case with unknown-route trips, Table B.1 summarizes the resulting estimates and errors for each link on the Sioux Falls network.

Table B.1: Detailed Estimates for Testing EM Algorithm on Sioux Falls Network

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
1	64.44	65.75	2.03	18.68	21.46	14.87
2	68.72	71.87	4.59	12.80	11.77	8.05
3	60.36	58.43	3.19	16.61	18.85	13.48
4	60.84	62.88	3.35	10.44	8.36	19.95
5	61.28	58.62	4.34	16.57	14.37	13.25
6	47.65	48.23	1.20	13.08	15.75	20.36
7	47.31	49.96	5.62	19.01	19.06	0.26
8	48.58	51.12	5.24	16.60	18.58	11.90
9	54.08	53.23	1.57	6.17	5.53	10.25
10	62.44	59.76	4.30	12.31	14.39	16.96
11	68.86	70.19	1.94	6.06	6.73	11.01
12	47.91	47.67	0.50	8.04	6.37	20.74
13	42.28	41.26	2.41	9.36	8.23	12.11
14	67.00	65.23	2.65	11.17	12.63	13.08

Table B.1: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
15	41.79	43.09	3.10	9.29	9.41	1.30
16	48.89	50.75	3.80	16.43	13.05	20.53
17	53.08	54.63	2.94	12.25	10.42	14.93
18	68.17	69.77	2.34	18.26	16.10	11.87
19	45.12	45.67	1.22	9.19	9.51	3.53
20	52.26	49.21	5.84	14.33	13.37	6.71
21	47.87	50.16	4.77	17.21	14.68	14.72
22	46.95	48.06	2.38	12.84	14.10	9.81
23	47.86	45.94	4.01	10.69	10.79	0.93
24	60.96	62.95	3.26	8.77	7.29	16.90
25	45.49	45.83	0.76	9.36	10.78	15.19
26	41.27	41.01	0.63	7.00	6.47	7.58
27	64.01	61.63	3.72	12.35	11.51	6.81
28	55.81	51.74	7.29	11.84	10.44	11.81
29	54.69	56.02	2.42	10.75	11.53	7.25
30	49.04	51.08	4.14	15.82	15.16	4.15
31	60.08	62.52	4.08	8.67	8.19	5.54
32	43.62	44.21	1.35	14.25	14.30	0.31
33	50.32	53.08	5.48	14.18	13.83	2.43
34	45.36	44.23	2.50	11.92	9.42	20.95
35	59.63	60.61	1.63	11.71	9.29	20.62
36	47.99	49.90	3.97	8.15	6.37	21.80
37	60.28	60.04	0.40	10.05	10.73	6.77

Table B.1: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
38	40.20	41.78	3.93	14.43	14.53	0.71
39	54.20	52.95	2.32	8.14	8.66	6.40
40	42.73	51.04	19.43	14.07	16.86	19.89
41	47.09	48.25	2.47	7.67	6.06	21.04
42	47.69	52.69	10.49	14.59	14.62	0.20
43	56.34	58.74	4.26	15.06	17.27	14.63
44	50.97	48.69	4.48	16.69	18.14	8.72
45	65.83	63.43	3.66	12.79	12.47	2.51
46	47.28	49.48	4.66	12.19	9.25	24.14
47	48.11	46.11	4.15	8.76	7.79	11.07
48	63.71	65.65	3.04	19.29	17.13	11.20
49	57.64	54.71	5.08	8.17	6.99	14.38
50	60.36	65.37	8.31	12.93	9.82	24.10
51	66.37	88.18	32.86	19.84	15.76	20.57
52	62.16	67.47	8.54	14.20	16.87	18.74
53	67.85	65.61	3.30	14.12	14.05	0.53
54	50.87	44.04	13.43	6.69	5.54	17.22
55	60.85	62.38	2.51	12.99	13.37	2.95
56	52.51	54.44	3.67	8.88	11.03	24.14
57	62.14	64.50	3.81	6.89	6.81	1.20
58	68.17	66.23	2.85	10.22	8.51	16.73
59	41.63	43.92	5.51	8.48	6.36	25.04
60	64.55	62.86	2.61	7.40	6.37	13.91

Table B.1: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
61	67.36	65.59	2.63	7.46	7.17	3.90
62	47.27	43.34	8.32	6.75	7.68	13.75
63	41.86	40.43	3.42	10.18	10.79	6.03
64	64.77	66.08	2.01	10.73	10.62	1.09
65	48.64	45.99	5.44	15.70	11.26	28.23
66	66.71	64.95	2.65	17.19	18.86	9.72
67	43.52	50.34	15.65	14.97	13.60	9.15
68	50.99	47.93	6.01	11.17	8.47	24.12
69	46.87	40.26	14.11	14.99	12.66	15.54
70	43.37	46.08	6.25	16.98	15.39	9.39
71	45.62	44.14	3.26	9.73	11.25	15.71
72	41.78	40.28	3.59	10.42	10.31	1.08
73	43.79	43.01	1.78	7.88	7.48	5.13
74	56.67	57.49	1.44	8.58	10.36	20.73
75	47.73	45.36	4.98	11.56	13.51	16.92
76	66.13	63.69	3.69	10.91	8.22	24.68
MAPE	-	-	4.68	-	-	12.16

As for testing trip splitting method for the case of log-normal distributed link travel times, Table B.2 summarizes the resulting estimates and errors for each link on the Sioux Falls network.

Table B.2: Detailed Estimates for Testing Trip Splitting on Sioux Falls Network

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
1	64.44	77.61	20.43	18.68	22.77	21.91
2	68.72	81.36	18.40	12.80	9.99	21.90
3	60.36	55.52	8.02	16.61	18.61	12.05
4	60.84	60.01	1.36	10.44	8.10	22.41
5	61.28	58.17	5.08	16.57	18.23	10.04
6	47.65	50.34	5.65	13.08	9.65	26.22
7	47.31	44.73	5.44	19.01	16.13	15.13
8	48.58	49.43	1.77	16.60	14.28	13.95
9	54.08	54.87	1.46	6.17	5.36	13.08
10	62.44	58.87	5.72	12.31	14.45	17.43
11	68.86	73.30	6.46	6.06	8.07	33.01
12	47.91	46.08	3.83	8.04	6.67	16.99
13	42.28	38.86	8.08	9.36	7.91	15.44
14	67.00	69.66	3.96	11.17	13.57	21.50
15	41.79	38.85	7.05	9.29	8.37	9.89
16	48.89	50.41	3.11	16.43	18.49	12.59
17	53.08	55.00	3.62	12.25	6.98	43.05
18	68.17	65.94	3.28	18.26	19.61	7.40
19	45.12	45.56	0.97	9.19	6.25	31.98
20	52.26	44.07	15.67	14.33	11.51	19.64
21	47.87	47.83	0.09	17.21	14.45	16.04
22	46.95	50.60	7.77	12.84	8.96	30.25

Table B.2: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
23	47.86	48.12	0.55	10.69	9.46	11.54
24	60.96	64.06	5.08	8.77	6.46	26.38
25	45.49	46.44	2.10	9.36	8.43	9.94
26	41.27	42.50	2.98	7.00	6.18	11.66
27	64.01	58.98	7.86	12.35	11.16	9.65
28	55.81	54.72	1.95	11.84	8.99	24.04
29	54.69	55.06	0.67	10.75	9.18	14.63
30	49.04	50.86	3.71	15.82	15.69	0.80
31	60.08	61.90	3.04	8.67	8.18	5.56
32	43.62	44.17	1.26	14.25	10.08	29.28
33	50.32	49.45	1.71	14.18	12.99	8.34
34	45.36	42.05	7.30	11.92	9.35	21.57
35	59.63	60.07	0.73	11.71	9.48	19.00
36	47.99	51.20	6.69	8.15	6.69	17.90
37	60.28	63.48	5.31	10.05	11.16	11.08
38	40.20	36.09	10.23	14.43	12.79	11.34
39	54.20	54.51	0.57	8.14	8.89	9.20
40	42.73	46.79	9.50	14.07	11.17	20.59
41	47.09	50.48	7.20	7.67	6.17	19.56
42	47.69	50.37	5.62	14.59	14.89	2.05
43	56.34	53.78	4.54	15.06	17.71	17.59
44	50.97	45.29	11.14	16.69	15.30	8.33
45	65.83	64.53	1.98	12.79	8.34	34.79

Table B.2: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
46	47.28	49.11	3.87	12.19	9.26	24.10
47	48.11	50.08	4.09	8.76	5.53	36.81
48	63.71	70.66	10.91	19.29	15.81	18.03
49	57.64	56.52	1.95	8.17	7.68	6.00
50	60.36	59.47	1.48	12.93	9.65	25.35
51	66.37	76.42	15.14	19.84	16.25	18.10
52	62.16	67.36	8.36	14.20	13.75	3.19
53	67.85	65.15	3.98	14.12	16.89	19.63
54	50.87	48.97	3.75	6.69	6.75	0.87
55	60.85	63.31	4.04	12.99	11.13	14.27
56	52.51	53.82	2.49	8.88	7.79	12.27
57	62.14	68.14	9.66	6.89	8.03	16.54
58	68.17	67.08	1.61	10.22	8.30	18.77
59	41.63	43.87	5.39	8.48	6.65	21.58
60	64.55	60.80	5.80	7.40	5.41	26.94
61	67.36	65.90	2.17	7.46	5.19	30.44
62	47.27	43.33	8.34	6.75	7.58	12.21
63	41.86	41.02	2.01	10.18	9.73	4.37
64	64.77	65.30	0.82	10.73	10.50	2.13
65	48.64	42.43	12.77	15.70	18.69	19.10
66	66.71	70.97	6.38	17.19	15.28	11.10
67	43.52	48.17	10.67	14.97	16.33	9.08
68	50.99	49.06	3.79	11.17	9.88	11.56

Table B.2: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
69	46.87	38.31	18.26	14.99	15.63	4.30
70	43.37	44.90	3.54	16.98	15.31	9.86
71	45.62	47.74	4.64	9.73	9.04	7.03
72	41.78	39.95	4.38	10.42	10.19	2.23
73	43.79	40.78	6.88	7.88	7.61	3.48
74	56.67	55.13	2.71	8.58	8.11	5.45
75	47.73	45.52	4.64	11.56	10.88	5.87
76	66.13	64.22	2.90	10.91	8.96	17.92
MAPE	-	-	5.48	-	-	15.73

As for comparing the estimates using both methods for the case of Gaussian distributed link travel times, Table B.3 and Table B.4 summarize the resulting estimates and errors using Method I and Method II respectively for each link on the Sioux Falls network.

Table B.3: Detailed Estimates Using Method I on Sioux Falls Network

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
1	64.44	72.32	12.22	18.68	21.46	14.88
2	68.72	77.88	13.34	12.80	10.92	14.66

Table B.3: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
3	60.36	54.20	10.22	16.61	17.06	2.72
4	60.84	60.34	0.83	10.44	9.23	11.54
5	61.28	58.39	4.72	16.57	14.47	12.66
6	47.65	50.05	5.03	13.08	14.28	9.17
7	47.31	47.19	0.25	19.01	19.13	0.62
8	48.58	51.84	6.72	16.60	18.03	8.59
9	54.08	53.17	1.68	6.17	5.88	4.65
10	62.44	58.09	6.98	12.31	13.25	7.66
11	68.86	69.95	1.59	6.06	6.87	13.35
12	47.91	47.33	1.22	8.04	6.35	20.97
13	42.28	39.39	6.84	9.36	9.23	1.35
14	67.00	69.29	3.42	11.17	13.02	16.58
15	41.79	38.76	7.26	9.29	9.01	2.98
16	48.89	51.68	5.71	16.43	14.89	9.36
17	53.08	55.41	4.40	12.25	11.76	4.01
18	68.17	65.26	4.27	18.26	16.14	11.62
19	45.12	45.69	1.25	9.19	8.22	10.58
20	52.26	43.36	17.03	14.33	13.68	4.55
21	47.87	48.67	1.67	17.21	14.59	15.23
22	46.95	50.12	6.75	12.84	13.39	4.22
23	47.86	48.39	1.11	10.69	9.43	11.78
24	60.96	64.32	5.50	8.77	7.16	18.35
25	45.49	44.83	1.45	9.36	11.43	22.13

Table B.3: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
26	41.27	42.88	3.90	7.00	6.08	13.12
27	64.01	58.91	7.96	12.35	11.72	5.14
28	55.81	52.67	5.63	11.84	12.63	6.70
29	54.69	55.24	1.01	10.75	10.84	0.83
30	49.04	51.08	4.14	15.82	15.16	4.15
31	60.08	61.92	3.07	8.67	9.72	12.13
32	43.62	43.98	0.84	14.25	16.31	14.42
33	50.32	49.68	1.27	14.18	13.25	6.52
34	45.36	44.66	1.55	11.92	10.87	8.78
35	59.63	59.87	0.39	11.71	9.58	18.19
36	47.99	50.97	6.20	8.15	6.66	18.30
37	60.28	60.69	0.67	10.05	11.82	17.60
38	40.20	38.82	3.44	14.43	15.06	4.34
39	54.20	53.12	2.00	8.14	8.15	0.11
40	42.73	47.89	12.06	14.07	13.61	3.26
41	47.09	49.56	5.26	7.67	6.30	17.84
42	47.69	50.50	5.88	14.59	14.12	3.24
43	56.34	52.70	6.46	15.06	16.40	8.85
44	50.97	48.26	5.33	16.69	18.72	12.16
45	65.83	64.53	1.98	12.79	11.63	9.04
46	47.28	51.77	9.49	12.19	10.92	10.45
47	48.11	50.07	4.08	8.76	7.69	12.22
48	63.71	70.60	10.81	19.29	19.99	3.61

Table B.3: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
49	57.64	53.91	6.47	8.17	7.18	12.03
50	60.36	59.76	1.00	12.93	11.74	9.24
51	66.37	77.34	16.52	19.84	17.26	13.00
52	62.16	67.53	8.64	14.20	12.72	10.43
53	67.85	66.38	2.16	14.12	13.73	2.80
54	50.87	48.87	3.94	6.69	7.81	16.61
55	60.85	64.19	5.48	12.99	14.05	8.19
56	52.51	52.50	0.01	8.88	9.33	4.98
57	62.14	64.50	3.80	6.89	7.36	6.83
58	68.17	67.52	0.95	10.22	9.42	7.76
59	41.63	44.21	6.21	8.48	8.34	1.67
60	64.55	60.85	5.73	7.40	6.75	8.84
61	67.36	66.09	1.90	7.46	7.11	4.66
62	47.27	42.90	9.25	6.75	7.25	7.30
63	41.86	41.96	0.24	10.18	11.06	8.69
64	64.77	65.35	0.90	10.73	10.42	2.89
65	48.64	43.27	11.04	15.70	15.11	3.71
66	66.71	69.42	4.06	17.19	17.00	1.09
67	43.52	47.22	8.50	14.97	14.30	4.48
68	50.99	49.55	2.82	11.17	10.23	8.39
69	46.87	40.95	12.63	14.99	14.55	2.91
70	43.37	46.51	7.25	16.98	16.00	5.77
71	45.62	46.21	1.29	9.73	10.77	10.74

Table B.3: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
72	41.78	39.83	4.68	10.42	10.65	2.18
73	43.79	41.09	6.18	7.88	6.35	19.48
74	56.67	54.89	3.14	8.58	9.45	10.14
75	47.73	45.22	5.27	11.56	12.43	7.54
76	66.13	63.76	3.59	10.91	9.09	16.66
MAPE	-	-	4.98	-	-	9.00

Table B.4: Detailed Estimates Using Method II on Sioux Falls Network

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
1	64.44	74.57	15.71	18.68	19.34	3.51
2	68.72	77.74	13.13	12.80	8.25	35.54
3	60.36	54.20	10.22	16.61	20.06	20.78
4	60.84	60.34	0.83	10.44	8.23	21.12
5	61.28	57.69	5.85	16.57	15.73	5.05
6	47.65	51.25	7.54	13.08	9.36	28.49
7	47.31	48.37	2.25	19.01	13.15	30.80
8	48.58	52.66	8.40	16.60	10.76	35.15
9	54.08	53.15	1.73	6.17	5.10	17.25
10	62.44	58.09	6.98	12.31	15.25	23.91

Table B.4: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
11	68.86	69.88	1.48	6.06	6.91	13.98
12	47.91	45.90	4.20	8.04	7.64	4.89
13	42.28	38.93	7.91	9.36	7.95	15.05
14	67.00	69.29	3.42	11.17	13.02	16.58
15	41.79	38.76	7.26	9.29	9.01	2.98
16	48.89	51.74	5.84	16.43	18.37	11.81
17	53.08	55.39	4.36	12.25	10.93	10.79
18	68.17	65.85	3.41	18.26	16.02	12.28
19	45.12	45.95	1.84	9.19	8.05	12.38
20	52.26	43.36	17.03	14.33	13.68	4.55
21	47.87	48.67	1.67	17.21	14.59	15.23
22	46.95	51.10	8.85	12.84	11.55	10.06
23	47.86	48.39	1.11	10.69	9.43	11.78
24	60.96	64.32	5.50	8.77	6.16	29.75
25	45.49	46.49	2.21	9.36	8.19	12.44
26	41.27	42.88	3.90	7.00	6.08	13.12
27	64.01	57.83	9.66	12.35	10.80	12.60
28	55.81	53.53	4.08	11.84	8.55	27.79
29	54.69	54.44	0.46	10.75	9.30	13.56
30	49.04	51.08	4.14	15.82	15.16	4.15
31	60.08	62.49	4.03	8.67	8.09	6.67
32	43.62	45.63	4.60	14.25	9.70	31.96
33	50.32	49.68	1.27	14.18	13.25	6.52

Table B.4: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
34	45.36	43.47	4.16	11.92	12.50	4.87
35	59.63	59.12	0.86	11.71	7.69	34.29
36	47.99	50.61	5.44	8.15	6.17	24.31
37	60.28	60.41	0.20	10.05	9.86	1.90
38	40.20	39.27	2.31	14.43	8.70	39.73
39	54.20	52.26	3.59	8.14	7.99	1.79
40	42.73	46.39	8.56	14.07	12.14	13.73
41	47.09	49.27	4.64	7.67	5.70	25.74
42	47.69	50.50	5.88	14.59	14.12	3.24
43	56.34	52.70	6.46	15.06	18.40	22.13
44	50.97	48.47	4.92	16.69	14.42	13.62
45	65.83	65.81	0.03	12.79	8.27	35.36
46	47.28	49.51	4.70	12.19	11.99	1.64
47	48.11	49.80	3.52	8.76	7.15	18.33
48	63.71	71.69	12.52	19.29	14.83	23.12
49	57.64	55.24	4.16	8.17	7.49	8.34
50	60.36	59.20	1.93	12.93	11.82	8.59
51	66.37	77.34	16.52	19.84	14.26	28.12
52	62.16	67.53	8.64	14.20	12.72	10.43
53	67.85	63.88	5.85	14.12	16.85	19.34
54	50.87	48.88	3.91	6.69	6.92	3.43
55	60.85	65.73	8.01	12.99	9.44	27.32
56	52.51	52.69	0.33	8.88	7.82	11.94

Table B.4: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error (%)**	**Truth**	**Estimate**	**Error (%)**
57	62.14	64.64	4.03	6.89	7.30	5.91
58	68.17	67.52	0.95	10.22	8.42	17.55
59	41.63	44.21	6.21	8.48	6.34	25.26
60	64.55	60.85	5.73	7.40	5.75	22.34
61	67.36	66.09	1.90	7.46	5.31	28.80
62	47.27	42.90	9.25	6.75	8.25	22.11
63	41.86	41.51	0.83	10.18	8.91	12.42
64	64.77	65.35	0.90	10.73	10.42	2.89
65	48.64	43.27	11.04	15.70	13.11	16.45
66	66.71	72.07	8.03	17.19	15.93	7.28
67	43.52	47.22	8.50	14.97	17.30	15.56
68	50.99	49.55	2.82	11.17	8.23	26.30
69	46.87	39.95	14.76	14.99	16.19	8.03
70	43.37	46.51	7.25	16.98	15.10	11.07
71	45.62	46.19	1.24	9.73	8.99	7.62
72	41.78	39.83	4.68	10.42	10.65	2.18
73	43.79	40.36	7.84	7.88	6.89	12.63
74	56.67	54.14	4.47	8.58	9.39	9.39
75	47.73	45.22	5.27	11.56	12.43	7.54
76	66.13	62.93	4.84	10.91	8.75	19.81
MAPE	-	-	5.38	-	-	15.59

We also conduct the numerical tests with more sufficient trip observations along single links on the Sioux Falls network. As for testing EM algorithm for the case with unknown-route trips, we modify the input setting as shown in Table B.5.

Table B.6 illustrates the average estimation errors for all links, and Table B.7 summarizes the resulting estimates and errors for each link on the Sioux Falls network.

Table B.5: Modified Input Information to Generate Test Sample for the Case with Unknown Route Trips

Type of Generated Trips		**Number of Trips**
Trips along each link		50
Trips covering multiple links		110
Trips with unlabeled paths		300
Setting of Randomly Generated Parameters		**Value of Bounds**
Mean	Upper Bound	70
	Lower Bound	40
Standard Deviation	Upper Bound	20
	Lower Bound	6

Table B.6: Estimate Errors of All Links with Modified Setting

Estimated Parameters	**MAPE**
Mean	2.49%
Standard Deviation	7.82%

Table B.7: Detailed Estimates for Testing EM Algorithm on Sioux Falls Network with Modified Setting

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
1	64.44	69.56	7.95%	18.68	23.56	26.11%
2	44.88	45.32	0.98%	7.67	7.83	2.10%
3	58.06	57.22	1.44%	9.68	10.52	8.66%
4	43.34	43.40	0.14%	16.92	15.29	9.68%
5	49.04	47.47	3.20%	12.59	12.91	2.50%
6	63.89	63.67	0.34%	7.38	7.24	1.97%
7	52.97	50.54	4.60%	17.55	17.47	0.48%
8	65.67	64.54	1.71%	15.03	13.44	10.59%
9	52.70	52.35	0.68%	7.27	7.04	3.23%
10	60.50	62.35	3.05%	13.65	11.70	14.33%
11	45.76	46.38	1.35%	7.94	7.18	9.57%
12	55.43	57.51	3.75%	18.38	15.55	15.40%
13	40.51	39.37	2.82%	7.69	8.28	7.66%
14	63.57	61.17	3.77%	13.19	10.39	21.24%
15	66.85	65.46	2.08%	7.00	7.30	4.22%
16	62.03	61.51	0.83%	6.72	6.37	5.14%
17	60.84	55.29	9.12%	16.61	17.47	5.16%
18	69.64	68.76	1.26%	8.39	7.74	7.70%
19	66.83	71.20	6.53%	13.23	13.84	4.56%
20	44.27	45.29	2.32%	6.35	7.10	11.77%
21	58.78	59.85	1.82%	7.93	7.10	10.44%
22	47.09	47.75	1.40%	8.48	8.44	0.51%

Table B.7: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
23	52.42	52.26	0.30%	12.89	14.43	11.95%
24	49.95	51.14	2.38%	8.13	8.14	0.09%
25	64.02	72.50	13.23%	18.55	17.99	3.00%
26	48.26	48.17	0.19%	16.03	14.28	10.93%
27	54.75	55.64	1.63%	6.99	5.92	15.40%
28	40.47	42.99	6.23%	18.09	18.68	3.26%
29	52.20	49.47	5.24%	7.58	7.46	1.53%
30	63.17	64.03	1.37%	9.19	9.71	5.59%
31	44.44	42.97	3.29%	8.77	8.00	8.78%
32	65.11	67.35	3.44%	19.60	20.41	4.18%
33	57.54	57.98	0.76%	9.99	10.19	2.03%
34	52.97	53.53	1.05%	16.49	15.67	4.98%
35	63.55	62.78	1.21%	12.51	12.03	3.84%
36	51.79	52.44	1.25%	8.51	9.22	8.39%
37	57.48	56.36	1.96%	17.97	15.78	12.17%
38	46.58	46.65	0.14%	10.56	9.97	5.57%
39	62.90	66.73	6.09%	17.57	17.99	2.38%
40	49.12	52.85	7.61%	12.77	13.49	5.69%
41	41.98	42.46	1.13%	9.86	9.46	4.04%
42	42.83	44.31	3.45%	18.29	16.56	9.42%
43	60.26	59.75	0.85%	18.65	20.83	11.71%
44	42.11	42.09	0.03%	6.97	6.15	11.72%
45	67.62	66.93	1.03%	12.98	12.23	5.76%

Table B.7: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
46	64.51	64.56	0.08%	8.65	8.13	6.04%
47	50.74	50.92	0.35%	12.85	11.35	11.65%
48	51.45	53.03	3.08%	13.95	12.82	8.07%
49	69.42	69.21	0.30%	15.03	16.94	12.72%
50	59.60	58.39	2.04%	12.86	14.31	11.33%
51	60.38	65.31	8.17%	19.43	19.27	0.81%
52	65.11	66.15	1.60%	17.67	19.41	9.87%
53	42.05	41.61	1.04%	7.19	5.60	22.09%
54	42.69	44.74	4.80%	17.56	17.55	0.07%
55	54.65	56.27	2.95%	8.24	7.58	7.94%
56	67.32	68.61	1.91%	18.39	21.04	14.40%
57	42.75	43.72	2.27%	18.72	19.79	5.75%
58	66.29	65.10	1.78%	14.54	14.69	1.01%
59	51.92	50.35	3.02%	12.71	13.67	7.51%
60	43.40	43.90	1.17%	10.96	11.64	6.19%
61	66.58	67.98	2.10%	8.99	8.13	9.59%
62	57.23	58.19	1.68%	11.78	10.88	7.60%
63	69.06	70.70	2.38%	7.38	6.65	9.91%
64	54.61	55.16	1.01%	9.67	8.50	12.08%
65	65.79	65.00	1.20%	15.49	14.87	3.98%
66	56.20	56.32	0.22%	9.09	9.05	0.47%
67	64.28	66.29	3.12%	14.52	13.83	4.75%
68	47.01	47.34	0.71%	7.35	6.19	15.79%

Table B.7: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
69	58.57	58.10	0.80%	8.31	7.12	14.34%
70	42.44	44.74	5.43%	12.76	12.07	5.42%
71	68.85	67.92	1.35%	13.99	12.70	9.22%
72	57.56	58.51	1.64%	7.67	7.46	2.72%
73	55.52	55.46	0.11%	18.64	18.81	0.91%
74	44.99	45.60	1.36%	8.09	9.45	16.76%
75	50.81	49.04	3.50%	17.60	17.95	1.99%
76	41.19	42.34	2.79%	12.57	14.79	17.68%
MAPE	-	-	2.49%	-	-	7.82%

As for testing trip splitting method for the case of log-normal distributed link travel times, we also modify the input setting as shown in Table B.8.

Table B.9 illustrates the average estimation errors for all links, and Table B.10 summarizes the resulting estimates and errors for each link on the Sioux Falls network.

Table B.8: Modified Input Information to Generate Test Sample for the Case of Log-Normal Distribution

Type of Generated Trips		**Number of Trips**
Trips along each link		50
Trips covering multiple links		810
Setting of Randomly Generated Parameters		**Value of Bounds**
Mean	Upper Bound	70
	Lower Bound	40
Standard Deviation	Upper Bound	20
	Lower Bound	6

Table B.9: Estimate Errors of All Links with Modified Setting for the Case of Log-Normal Distribution

Estimated Parameters	**MAPE**
Mean	2.47%
SD	13.36%
Mu	0.60%
Sigma	13.03%

Table B.10: Detailed Estimates for Testing Trip Splitting on Sioux Falls Network with Modified Setting

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
1	64.44	69.31	7.55%	18.68	21.52	15.22%
2	44.88	45.01	0.29%	7.67	7.80	1.77%
3	58.06	56.98	1.86%	9.68	10.27	6.07%
4	43.34	41.79	3.56%	16.92	11.56	31.72%
5	49.04	47.55	3.04%	12.59	10.51	16.56%
6	63.89	64.20	0.49%	7.38	6.93	6.12%
7	52.97	50.41	4.84%	17.55	16.77	4.47%
8	65.67	66.53	1.32%	15.03	13.00	13.49%
9	52.70	52.87	0.31%	7.27	6.02	17.25%
10	60.50	61.99	2.46%	13.65	11.98	12.26%
11	45.76	46.55	1.73%	7.94	7.27	8.46%
12	55.43	56.39	1.72%	18.38	10.75	41.50%
13	40.51	38.91	3.94%	7.69	6.43	16.40%
14	63.57	60.04	5.55%	13.19	8.09	38.68%
15	66.85	65.54	1.96%	7.00	7.15	2.10%
16	62.03	61.52	0.82%	6.72	5.27	21.58%
17	60.84	55.94	8.06%	16.61	16.13	2.88%
18	69.64	67.96	2.41%	8.39	6.58	21.55%
19	66.83	70.08	4.86%	13.23	11.67	11.84%
20	44.27	45.22	2.15%	6.35	5.39	15.10%
21	58.78	59.66	1.49%	7.93	7.24	8.66%
22	47.09	47.58	1.03%	8.48	8.55	0.83%

Table B.10: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
23	52.42	52.08	0.65%	12.89	12.63	2.06%
24	49.95	50.66	1.41%	8.13	8.21	0.94%
25	64.02	73.28	14.46%	18.55	17.67	4.74%
26	48.26	47.03	2.54%	16.03	10.62	33.79%
27	54.75	55.29	0.98%	6.99	5.86	16.17%
28	40.47	40.96	1.22%	18.09	18.69	3.31%
29	52.20	49.51	5.15%	7.58	7.59	0.19%
30	63.17	64.11	1.49%	9.19	9.86	7.23%
31	44.44	42.88	3.51%	8.77	6.36	27.48%
32	65.11	67.59	3.81%	19.60	21.23	8.35%
33	57.54	57.24	0.52%	9.99	8.20	17.95%
34	52.97	52.66	0.58%	16.49	15.67	4.95%
35	63.55	63.71	0.26%	12.51	11.25	10.06%
36	51.79	52.55	1.46%	8.51	9.43	10.82%
37	57.48	55.11	4.12%	17.97	11.29	37.17%
38	46.58	46.26	0.69%	10.56	9.90	6.29%
39	62.90	68.89	9.53%	17.57	16.87	4.00%
40	49.12	51.28	4.40%	12.77	9.91	22.39%
41	41.98	42.02	0.09%	9.86	9.52	3.38%
42	42.83	42.70	0.32%	18.29	16.46	9.99%
43	60.26	58.59	2.77%	18.65	17.43	6.57%
44	42.11	41.94	0.40%	6.97	6.15	11.80%
45	67.62	66.21	2.09%	12.98	9.91	23.66%

Table B.10: Continued

Link No.	Mean			Standard Deviation		
	Truth	Estimate	Error	Truth	Estimate	Error
46	64.51	64.73	0.35%	8.65	8.67	0.20%
47	50.74	49.51	2.43%	12.85	8.94	30.44%
48	51.45	49.34	4.10%	13.95	8.21	41.12%
49	69.42	70.55	1.63%	15.03	17.76	18.19%
50	59.60	60.53	1.55%	12.86	12.93	0.60%
51	60.38	65.34	8.21%	19.43	20.85	7.29%
52	65.11	65.53	0.64%	17.67	13.21	25.24%
53	42.05	41.65	0.97%	7.19	5.56	22.61%
54	42.69	42.43	0.61%	17.56	11.92	32.15%
55	54.65	56.17	2.77%	8.24	7.83	5.01%
56	67.32	69.82	3.71%	18.39	16.53	10.12%
57	42.75	43.92	2.75%	18.72	22.87	22.18%
58	66.29	65.55	1.11%	14.54	10.74	26.14%
59	51.92	50.67	2.41%	12.71	13.49	6.09%
60	43.40	43.38	0.05%	10.96	10.01	8.74%
61	66.58	67.34	1.15%	8.99	8.24	8.43%
62	57.23	58.02	1.39%	11.78	11.03	6.34%
63	69.06	71.14	3.02%	7.38	7.27	1.58%
64	54.61	55.37	1.40%	9.67	8.54	11.65%
65	65.79	66.09	0.46%	15.49	13.04	15.81%
66	56.20	57.14	1.67%	9.09	9.10	0.02%
67	64.28	66.14	2.89%	14.52	14.32	1.43%
68	47.01	47.10	0.20%	7.35	6.24	15.09%

Table B.10: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
69	58.57	58.87	0.51%	8.31	7.20	13.35%
70	42.44	43.76	3.12%	12.76	10.64	16.61%
71	68.85	67.02	2.66%	13.99	10.57	24.44%
72	57.56	58.14	1.01%	7.67	7.67	0.00%
73	55.52	53.67	3.34%	18.64	17.09	8.33%
74	44.99	45.88	1.98%	8.09	9.77	20.75%
75	50.81	49.77	2.05%	17.60	13.02	26.03%
76	41.19	42.69	3.65%	12.57	12.80	1.83%
MAPE	-	-	2.47%	-	-	13.36%

As for comparing the estimates using both methods for the case of Gaussian distributed link travel times, Table B.11 and Table B.12 summarize the resulting estimates and errors using Method I and Method II respectively for each link on the Sioux Falls network. Table B.13 illustrates the average estimation errors for all links from both methods.

Table B.11: Detailed Estimates Using Method I on Sioux Falls Network with Modified Setting

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
1	64.44	67.79	5.19%	18.68	22.17	18.68%
2	44.88	44.99	0.25%	7.67	7.78	1.46%
3	58.06	56.83	2.12%	9.68	10.45	7.96%
4	43.34	43.41	0.17%	16.92	14.89	12.04%
5	49.04	47.94	2.24%	12.59	12.76	1.33%
6	63.89	63.65	0.38%	7.38	7.16	3.05%
7	52.97	50.25	5.14%	17.55	17.62	0.37%
8	65.67	66.60	1.42%	15.03	14.38	4.31%
9	52.70	52.07	1.20%	7.27	6.99	3.91%
10	60.50	62.39	3.12%	13.65	11.73	14.08%
11	45.76	46.68	2.01%	7.94	7.16	9.89%
12	55.43	57.80	4.27%	18.38	17.58	4.34%
13	40.51	39.32	2.95%	7.69	7.96	3.50%
14	63.57	60.25	5.22%	13.19	10.23	22.44%
15	66.85	65.49	2.03%	7.00	7.29	4.11%
16	62.03	61.37	1.06%	6.72	6.31	6.08%
17	60.84	55.39	8.97%	16.61	17.51	5.42%
18	69.64	67.77	2.68%	8.39	7.47	10.91%
19	66.83	69.92	4.62%	13.23	13.45	1.65%
20	44.27	45.38	2.52%	6.35	7.02	10.56%
21	58.78	59.75	1.66%	7.93	7.14	9.93%
22	47.09	47.59	1.04%	8.48	8.46	0.20%

Table B.11: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
23	52.42	51.90	0.99%	12.89	14.30	10.89%
24	49.95	50.66	1.42%	8.13	8.09	0.46%
25	64.02	72.44	13.14%	18.55	18.70	0.84%
26	48.26	47.80	0.94%	16.03	14.11	12.00%
27	54.75	55.23	0.87%	6.99	5.98	14.44%
28	40.47	40.84	0.91%	18.09	18.43	1.88%
29	52.20	49.37	5.43%	7.58	7.53	0.57%
30	63.17	64.03	1.37%	9.19	9.71	5.59%
31	44.44	43.26	2.65%	8.77	8.19	6.70%
32	65.11	67.35	3.44%	19.60	20.41	4.18%
33	57.54	56.87	1.16%	9.99	10.09	1.02%
34	52.97	52.86	0.21%	16.49	15.79	4.21%
35	63.55	63.46	0.14%	12.51	11.99	4.16%
36	51.79	52.41	1.20%	8.51	9.28	9.09%
37	57.48	55.96	2.65%	17.97	16.00	10.95%
38	46.58	46.38	0.43%	10.56	9.98	5.50%
39	62.90	67.95	8.03%	17.57	17.87	1.72%
40	49.12	51.51	4.87%	12.77	13.37	4.71%
41	41.98	42.10	0.27%	9.86	9.52	3.37%
42	42.83	43.34	1.18%	18.29	16.63	9.03%
43	60.26	57.84	4.02%	18.65	21.54	15.51%
44	42.11	42.06	0.11%	6.97	6.18	11.31%
45	67.62	66.23	2.07%	12.98	12.16	6.32%

Table B.11: Continued

Link No.	Mean			Standard Deviation		
	Truth	Estimate	Error	Truth	Estimate	Error
46	64.51	64.32	0.29%	8.65	8.31	3.95%
47	50.74	50.07	1.32%	12.85	11.22	12.68%
48	51.45	50.12	2.59%	13.95	12.34	11.52%
49	69.42	70.00	0.84%	15.03	17.41	15.84%
50	59.60	58.99	1.02%	12.86	14.31	11.33%
51	60.38	65.31	8.17%	19.43	19.27	0.81%
52	65.11	64.97	0.22%	17.67	17.59	0.43%
53	42.05	41.88	0.42%	7.19	5.63	21.64%
54	42.69	44.52	4.30%	17.56	17.66	0.55%
55	54.65	56.25	2.92%	8.24	7.62	7.50%
56	67.32	69.31	2.95%	18.39	19.82	7.77%
57	42.75	42.52	0.54%	18.72	21.89	16.95%
58	66.29	65.04	1.87%	14.54	14.21	2.25%
59	51.92	50.36	2.99%	12.71	13.78	8.42%
60	43.40	43.25	0.35%	10.96	11.59	5.67%
61	66.58	67.45	1.31%	8.99	8.15	9.39%
62	57.23	58.20	1.70%	11.78	10.90	7.48%
63	69.06	71.06	2.90%	7.38	6.61	10.50%
64	54.61	55.06	0.83%	9.67	8.37	13.48%
65	65.79	65.93	0.22%	15.49	14.77	4.61%
66	56.20	56.77	1.01%	9.09	9.14	0.51%
67	64.28	66.27	3.09%	14.52	13.93	4.11%
68	47.01	47.26	0.54%	7.35	6.24	15.12%

Table B.11: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
69	58.57	59.02	0.77%	8.31	7.18	13.68%
70	42.44	44.19	4.13%	12.76	12.85	0.72%
71	68.85	67.06	2.60%	13.99	12.69	9.27%
72	57.56	58.15	1.02%	7.67	7.59	0.99%
73	55.52	53.88	2.96%	18.64	17.73	4.92%
74	44.99	45.60	1.34%	8.09	9.55	18.04%
75	50.81	49.87	1.86%	17.60	18.03	2.42%
76	41.19	42.66	3.57%	12.57	14.07	11.90%
MAPE	-	-	2.35%	-	-	7.30%

Table B.12: Detailed Estimates Using Method II on Sioux Falls Network with Modified Setting

Link No.	**Mean**			**Standard Deviation**		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
1	64.44	68.14	5.74%	18.68	20.09	7.52%
2	44.88	44.99	0.25%	7.67	7.78	1.46%
3	58.06	56.83	2.12%	9.68	10.45	7.96%
4	43.34	43.49	0.35%	16.92	12.21	27.84%
5	49.04	47.66	2.80%	12.59	10.77	14.51%
6	63.89	63.45	0.68%	7.38	6.71	9.13%
7	52.97	50.25	5.14%	17.55	17.62	0.37%
8	65.67	66.93	1.92%	15.03	12.99	13.53%
9	52.70	52.15	1.04%	7.27	5.90	18.93%
10	60.50	62.39	3.12%	13.65	11.73	14.08%
11	45.76	46.68	2.01%	7.94	7.16	9.89%
12	55.43	58.15	4.91%	18.38	10.84	41.04%
13	40.51	39.27	3.05%	7.69	6.70	12.89%
14	63.57	60.31	5.12%	13.19	8.61	34.75%
15	66.85	65.49	2.03%	7.00	7.29	4.11%
16	62.03	61.23	1.29%	6.72	5.40	19.65%
17	60.84	55.39	8.97%	16.61	17.51	5.42%
18	69.64	67.54	3.01%	8.39	6.65	20.76%
19	66.83	69.75	4.36%	13.23	11.25	15.01%
20	44.27	45.25	2.23%	6.35	5.36	15.54%
21	58.78	59.75	1.66%	7.93	7.14	9.93%
22	47.09	47.59	1.04%	8.48	8.46	0.20%

Table B.12: Continued

Link No.	Mean			Standard Deviation		
	Truth	Estimate	Error	Truth	Estimate	Error
23	52.42	51.47	1.80%	12.89	12.89	0.06%
24	49.95	50.66	1.42%	8.13	8.09	0.46%
25	64.02	72.69	13.54%	18.55	15.76	15.02%
26	48.26	47.90	0.75%	16.03	11.18	30.27%
27	54.75	54.86	0.19%	6.99	5.97	14.61%
28	40.47	40.84	0.91%	18.09	18.43	1.88%
29	52.20	48.89	6.34%	7.58	7.66	1.09%
30	63.17	64.03	1.37%	9.19	9.71	5.59%
31	44.44	42.92	3.41%	8.77	6.50	25.94%
32	65.11	67.35	3.44%	19.60	20.41	4.18%
33	57.54	56.74	1.38%	9.99	8.24	17.54%
34	52.97	52.86	0.21%	16.49	15.79	4.20%
35	63.55	63.48	0.10%	12.51	11.27	9.94%
36	51.79	52.41	1.20%	8.51	9.28	9.09%
37	57.48	56.03	2.52%	17.97	11.78	34.45%
38	46.58	46.38	0.43%	10.56	9.98	5.50%
39	62.90	68.14	8.34%	17.57	15.31	12.89%
40	49.12	51.57	5.00%	12.77	9.64	24.50%
41	41.98	42.10	0.27%	9.86	9.52	3.37%
42	42.83	43.34	1.18%	18.29	16.63	9.03%
43	60.26	58.17	3.47%	18.65	18.03	3.34%
44	42.11	42.06	0.11%	6.97	6.18	11.31%
45	67.62	66.14	2.20%	12.98	10.18	21.54%

Table B.12: Continued

Link No.	Mean			Standard Deviation		
	Truth	Estimate	Error	Truth	Estimate	Error
46	64.51	63.98	0.81%	8.65	8.89	2.71%
47	50.74	50.09	1.28%	12.85	9.24	28.08%
48	51.45	50.17	2.49%	13.95	8.66	37.93%
49	69.42	70.00	0.84%	15.03	17.41	15.84%
50	59.60	59.10	0.85%	12.86	12.25	4.68%
51	60.38	65.31	8.17%	19.43	19.27	0.81%
52	65.11	65.20	0.13%	17.67	13.22	25.20%
53	42.05	41.88	0.42%	7.19	5.63	21.64%
54	42.69	44.70	4.71%	17.56	12.16	30.78%
55	54.65	56.25	2.92%	8.24	7.62	7.50%
56	67.32	69.77	3.64%	18.39	15.96	13.20%
57	42.75	42.52	0.54%	18.72	21.89	16.95%
58	66.29	65.08	1.82%	14.54	10.90	25.04%
59	51.92	50.36	2.99%	12.71	13.78	8.42%
60	43.40	43.10	0.68%	10.96	10.03	8.54%
61	66.58	67.45	1.31%	8.99	8.15	9.39%
62	57.23	58.20	1.70%	11.78	10.90	7.48%
63	69.06	70.81	2.53%	7.38	6.95	5.81%
64	54.61	54.63	0.04%	9.67	8.47	12.43%
65	65.79	65.87	0.12%	15.49	12.81	17.28%
66	56.20	56.51	0.55%	9.09	8.86	2.61%
67	64.28	66.27	3.09%	14.52	13.93	4.11%
68	47.01	47.26	0.54%	7.35	6.24	15.12%

Table B.12: Continued

Link No.	Mean			Standard Deviation		
	Truth	**Estimate**	**Error**	**Truth**	**Estimate**	**Error**
69	58.57	59.02	0.77%	8.31	7.18	13.68%
70	42.44	44.53	4.92%	12.76	10.55	17.28%
71	68.85	67.08	2.57%	13.99	10.84	22.52%
72	57.56	58.15	1.02%	7.67	7.59	0.99%
73	55.52	53.88	2.96%	18.64	17.73	4.92%
74	44.99	45.60	1.34%	8.09	9.55	18.04%
75	50.81	50.19	1.23%	17.60	13.33	24.30%
76	41.19	42.92	4.20%	12.57	12.40	1.38%
MAPE	-	-	2.42%	-	-	13.09%

Table B.13: Comparison of Estimate Errors Using Both Methods on Sioux Falls Network with Modified Setting

Estimated Parameters	MAPE	
	Method I	**Method II**
Mean	2.35%	2.42%
Standard Deviation	7.30%	13.09%

To provide supplemental information of estimate accuracy with respect to the sample size of single-link observations, we test the sample sizes with different numbers of single-link observations along each link, and summarize the resulting estimate

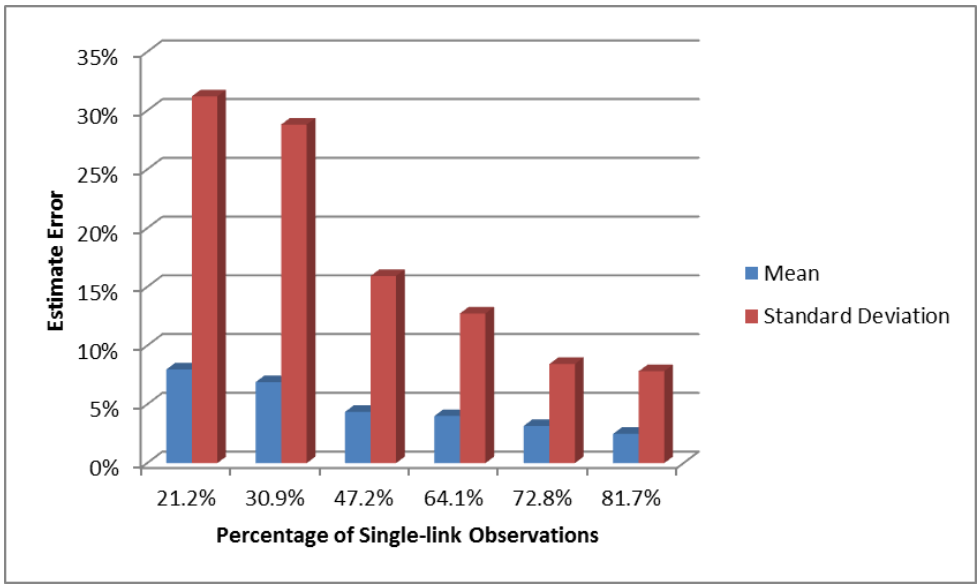

Figure B.1: Estimate errors with various sample sizes of single-link observations on Sioux Falls network for the case of unknown route trips.

errors accordingly. Figure B.1 displays the average errors on Sioux Falls Network for the case of unknown-route trips, and Figure B.2 displays the average estimate errors on Sioux Falls Network for the case of log-normal distribution.

For the case of log-normal distributed link travel times with unknown-route trips, we do some extra tests using the Gaussian distribution as an approximation, such that the Gaussian mixture model and EM algorithm can be applied as introduced in the framework of Method I. The resulting parameter estimates are compared with the ground true of mean and standard deviation for each link on Sioux Falls network. Figure B.3 displays the estimate errors using this approximation, with different numbers of single-link observations in the test sample. It is shown that when the percentage of total single-link observations accounts for about 50%, the estimate error under this approximation appears acceptable, with the resulting average error

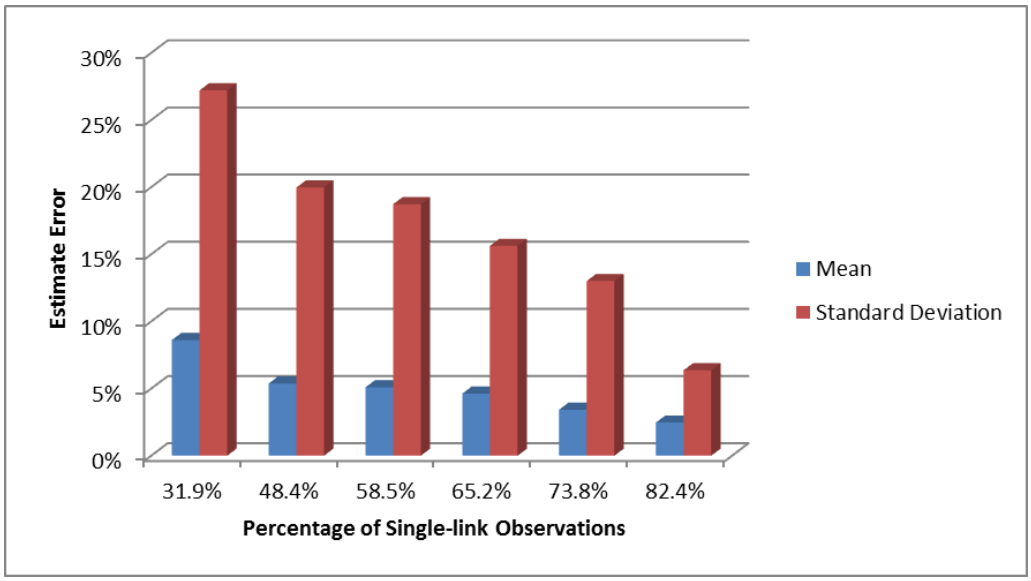

Figure B.2: Estimate errors with various sample sizes of single-link observations on Sioux Falls network for the case of log-normal distribution.

of mean estimates less than 5% and the average error of standard deviation estimates less than 20%.

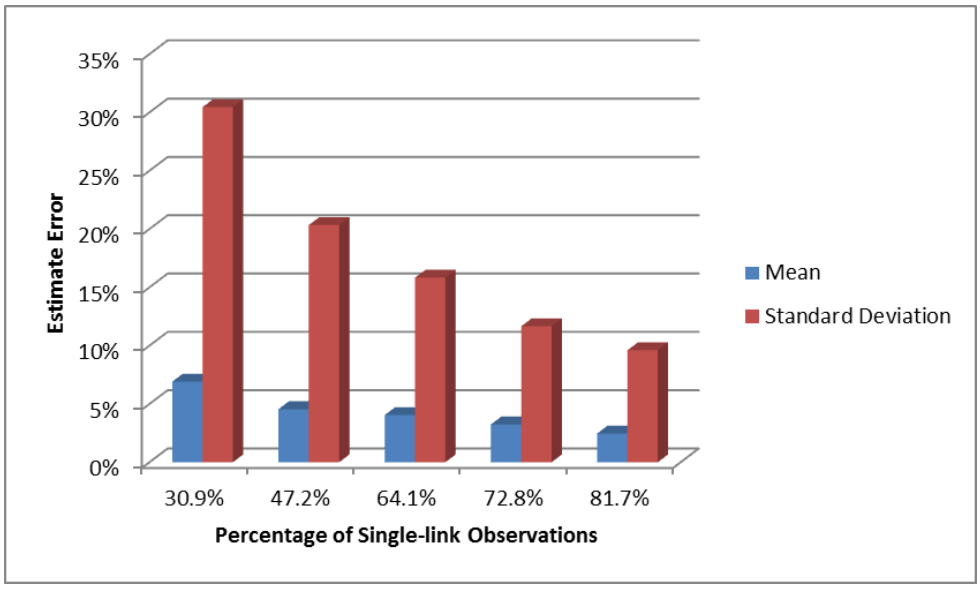

Figure B.3: Estimate errors with various sample sizes of single-link observations on Sioux Falls network using Gaussian approximation.