Traffic State Estimation Using Probe Vehicle Data

by

Yan Zhao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in the University of Michigan
2020

Doctoral Committee:

Professor Henry X. Liu, Co-Chair
Professor Huei Peng, Co-Chair
Assistant Professor Neda Masoud
Associate Professor Gábor Orosz
Professor Yafeng Yin

Yan Zhao

zhaoyann@umich.edu

ORCID ID: 0000-0002-5246-059X

Dedication

To the shoulders of giants.

Acknowledgments

Pursuing the PhD degree at the University of Michigan has been an incredible journey for me. First of all, I am very grateful for the guidance, encouragement, and help from my advisor Prof. Henry Liu along the way. Prof. Liu has been a role model for me in both research and daily life. I am so fortunate to have the opportunity to work with him. During these years, not only have my technical and communication skills been improved significantly, but I also become more confident with doing research either independently or by collaborating with others. I could not have accomplished so much without the encouragement of Prof. Liu.

I am also thankful to Prof. Neda Masoud, Prof. Gábor Orosz, Prof. Huei Peng, and Prof. Yafeng Yin, who serve on my dissertation committee and provide me insightful feedback. I appreciate the opportunities to work with them in classes, seminars, and projects.

I would like to thank everyone in the Michigan Traffic Lab, particularly, Prof. Xuan Di, Dr. Yiheng Feng, Dr. Jianfeng Zheng, Prof. Xiaolei Guo, Dr. Weili Sun, Prof. Zhaosheng Zhang, Mr. Shengyin Shen, Mr. Shihong Huang, Dr. Wai Wong, Mr. Xingmin Wang, and Mr. Xintao Yan, who collaborated with me on different projects or co-authored papers with me in the past few years. I enjoyed studying with them, working with them, and playing with them. I also learned much from them.

I also would like to thank all my friends from classes, labs, organizations, and companies for helping me, inspiring me, and bringing happiness into my life.

Last but not least, I would like to express my deepest gratitude and love to my parents, my grandparents, my sister, and my girlfriend, unconditionally. I would not have gone so far without their boundless love.

Table of Contents

List of Figures

Figure

List of Tables

Table

List of Appendices

List of Abbreviations

AADT	Annual average daily traffic
AE	Approximate estimator
AVIS	Automatic vehicle identification system
BPCA	Bayesian principal component analysis
BSM	Basic safety message
CV	Connected vehicles
DGPS	Differential Global Positioning System
DP	Dynamic programming
EM	Expectation-Maximization
ETC	Electronic toll collection
FD	Fundamental diagram
GPS	Global Positioning System
HMM	Hidden Markov model
i.i.d.	Independent and identically distributed
IoT	Internet of Things
ITS	Intelligent transportation systems
KPPCA	Kernel probabilistic principal component analysis
LTE-V	Long-Term Evolution-Vehicle
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MLE	Maximum likelihood estimation
PCA	Principal component analysis
POI	Points of interest
PPCA	Probabilistic principal component analysis

PPCA-DF	PPCA-based data fusion
RMSE	Root mean square error
RSU	Road side unit
SAE	Society of Automobile Engineers
SVD	Singular value decomposition
SVD-DF	SVD-based data fusion
TOD	Time-of-day
V2I	Vehicle-to-infrastructure
V2V	Vehicle-to-vehicle
w.r.t.	With respect to

List of Symbols

Cycle-by-cycle queue length estimation

C	Total number of traffic signal cycles
l_i	Queue length in the ith cycle
$\hat{l}_i$	Estimated queue length in the ith cycle
l	Sequence of queue lengths
L_{max}	Upper bound of the queue length
p	Penetration rate of probe vehicles
q_i	Observed partial queue in the ith cycle
$\|q_i\|$	Length of the observed partial queue in the ith cycle
q	Sequence of observed partial queues
n_i	Number of probe vehicles in the ith cycle
π_j	(Initial) probability of the queue length being j
π	(Initial) queue length distribution
T_{jk}	Transition probability from queue length j to k
T	Transition matrix
E_{lq}	Probability of observing q from hidden state l
λ	Average arrival rate
a_{max}	Maximum number of vehicle arrivals in each cycle
s	Maximum number of served vehicles in each cycle
a_i^g	Number of vehicle arrivals in the green phase of the ith cycle
a_i^r	Number of vehicle arrivals in the red phase of the ith cycle

Approximate estimation (i.i.d. case)

s_i	Position of the first probe vehicle in the ith cycle
t_i	Position of the last probe vehicle in the ith cycle
X_i^j	A binary variable to indicate if the queue length in the ith cycle is j
C_j	Total number of queues of length j

Q^{obs}	Total length of all the (partially) observable queues
$\hat{Q}^{obs}$	Estimated total length of all the (partially) observable queues
Q^{hid}	Total length of the hidden queues
$\hat{Q}^{hid}$	Estimated total length of the hidden queues
Q^{probe}	Total number of probe vehicles in all the queues
$\hat{Q}^{all}$	Estimated total queue length
$\bar{c}_j$	Number of stopping probe vehicles at position j
$\hat{C}_j$	Estimation of $p\mathbb{E}(C_j)$
ω_k	Weight of the kth item in the least-square problem
H	Hellinger distance
V^{probe}	Traffic volume of probe vehicles
V^{all}	Traffic volume of all the vehicles

Maximum likelihood estimation (i.i.d. case)

$\mathcal{L}$	Likelihood function
$\mathcal{I}$	Fisher information
$\mathcal{N}$	Normal distribution
$\hat{p}$	Maximum likelihood estimator of p
p^*	True value of p
θ	Collection of parameters to be estimated
$\theta^{(0)}$	Initial guess of the parameters
$\theta^{(t)}$	Estimated parameters in the tth iteration
Q	Result of the E-step in the EM algorithm

Maximum likelihood estimation (non-i.i.d. case)

$\alpha_j(i)$	$P(q_1, q_2, \ldots, q_i, l_i = j; \theta)$
$\beta_j(i)$	$P(q_{i+1}, q_{i+2}, \ldots, q_C \mid l_i = j; \theta)$
$\alpha_j(d,i)$	$P(q_{d1}, q_{d2}, \ldots, q_{di}, l_{di} = j; \theta)$
$\beta_j(d,i)$	$P(q_{d,i+1}, q_{d,i+2}, \ldots, q_{dC} \mid l_{di} = j; \theta)$

Traffic volume estimation by data fusion

d	Number of locations
N	Number of days
X	Traffic volume matrix
$\hat{X}$	Estimated traffic volume matrix
x_n	Traffic volume vector for day n
x_{ij}	Traffic volume at location i on day j
Y	Traffic volume matrix of probe vehicles
y_n	Traffic volume vector of probe vehicles for day n
y_{ij}	Traffic volume of probe vehicles at location i on day j
σ_k	kth largest singular value of Y
u_k	kth left-singular vector of Y
r	Dimension of the low-rank subspace
$\hat{x}_n$	Estimation of x_n
α_n	Coordinates of $\hat{x}_n$ on the low-rank subspace
v_k	kth right-singular vector of Y
W	Matrix indicating if the entries in X are empty
w_n	nth column of W
v_{max}	Upper bound of traffic volumes
ν	Weight of the regularization term
$\mathcal{B}$	Binomial distribution
Λ	Projection matrix from the latent space to the data space
t_n	Latent vector in the PPCA model
μ_x	Mean traffic volume vector
ϵ_n	Gaussian error in the PPCA model
σ^2	Variance of the Gaussian error ϵ_n
I	Identity matrix
$\bar{x}$	Prior of x_n
η^2	Variance of the Gaussian approximation of probe vehicle sampling process
x_n^m	Missing data of the traffic volume vector x_n
x_n^o	Available data of the traffic volume vector x_n
q_n	Posterior distribution of the latent variables t_n and x_n^m

Abstract

Traffic problems are becoming a burden on cities across the world. To prevent traffic accidents, mitigate congestion, and reduce fuel consumption, a critical step is to have a good understanding of traffic. Traditionally, traffic conditions are monitored primarily by fixed-location sensors. However, fixed-location sensors only provide information about specific locations, and the installation and maintenance cost is very high. The advances in GPS-based technologies, such as connected vehicles and ride-hailing services, provide us an alternative approach to traffic monitoring. While these types of GPS-equipped probe vehicles travel on the road, a vast amount of trajectory data are being collected. As probe vehicle data contain rich information about traffic conditions, they have drawn much attention from both researchers and practitioners in the field of traffic management and control. Extensive literature has studied the estimation of traffic speeds and travel times using probe vehicle data. However, as for queue lengths and traffic volumes, which are critical for traffic signal control and performance measures, most of the existing estimation methods based on probe vehicles can hardly be implemented in practice. The main obstacle is the low market penetration of probe vehicles. Therefore, in this dissertation, we aim to develop probe vehicle based traffic state estimation methods that are suitable for the low penetration rate environment and can potentially be implemented in the real world.

First, we treat the traffic state in each location and each time point independently. We focus on estimating the queues forming at isolated intersections under light or moderate traffic. The existing methods often require prior knowledge of the queue length distribution or the probe vehicle penetration rate. However, these parameters are not available beforehand

in real life. Therefore, we propose a series of methods to estimate these parameters from historical probe vehicle data. Some of the methods have been validated using real-world probe vehicle data.

Second, we study traffic state estimation considering temporal correlations. The correlation of queue lengths in different traffic signal cycles is often ignored by the existing studies, although the phenomenon is commonly-observed in real life, such as the overflow queues induced by oversaturated traffic. To fill the gap, we model such queueing processes and observation processes using a hidden Markov model (HMM). Based on the HMM, we develop two cycle-by-cycle queue length estimation methods and an algorithm that can estimate the parameters of the HMM from historical probe vehicle data.

Lastly, we consider the spatiotemporal correlations of traffic states, with a focus on the estimation of traffic volumes. With limited probe vehicle data, it is difficult to estimate traffic volumes accurately if we treat each location and each time slot independently. Noticing that traffic volumes in different locations and different time slots are correlated, we propose to find the low-rank representation of traffic volumes and then reconstruct the unknown values by fusing probe vehicle data and fixed-location sensor data. Test results show that the proposed methods can reconstruct the traffic volumes accurately, and they have great potential for real-world applications.

In summary, this thesis systematically studies traffic state estimation based on probe vehicle data. Some of the proposed methods have been implemented in real life. We expect the methods to be implemented on an even larger scale and help transportation agencies solve more real-world traffic problems.

Chapter 1

Introduction

1.1 Background

The expansion of cities and the growth of urban populations impose a burden on the existing transportation infrastructure. Traffic congestion results in safety issues, long commutes, and a waste of energy. It is estimated that in 2017 alone, traffic congestion in the US caused 8.8 billion hours of total travel delay, wasted 3.3 billion gallons of fuel, and led to an economic loss of more than 305 billion dollars (Cookson, 2018; Schrank et al., 2019). Meanwhile, 37,133 people were killed in various motor vehicle traffic crashes in the US (National Highway Traffic Safety Administration, 2018).

Currently, many components of the transportation infrastructure are not well managed. For example, most transportation agencies in the US only retime their traffic signals every three to five years (Gordon, 2010; Lavrenz et al., 2016; Dunn et al., 2019). To prevent traffic accidents, mitigate traffic congestion, and reduce fuel consumption, intelligent transportation systems (ITS) are in great need. For better management of transportation infrastructure and improvement of traffic conditions, a critical step is to have a better understanding of traffic flows and transportation networks.

Traditionally, traffic conditions are monitored primarily by fixed-location sensors, such as loop detectors and cameras (Wang et al., 2019). Although fixed-location sensors are widely applied in current practice, there are a few drawbacks of these sensors, caused by the high installation and maintenance cost (Yoon et al., 2007; Work et al., 2008). First, the coverage of fixed-location sensors in transportation networks is usually low (Seo and Kusakabe, 2018;

Takenouchi et al., 2019). The sensors can only provide traffic information about locations where they are installed (Yoon et al., 2007; Seo et al., 2015; Guo et al., 2019). Second, although the sensors might measure traffic volumes directly, estimating other traffic states such as queue lengths is not straightforward if only fixed-location sensor data are available. Third, the lack of maintenance sometimes gives rise to missing data problems (Qu et al., 2009; Ran et al., 2016). The consequence is that a large number of intersections are still controlled by fixed-time signals and do not respond to short-term traffic fluctuations. Even for the intersections installed with adaptive traffic controllers, the performance often deteriorates due to sensor malfunction and missing data (Zheng and Liu, 2020).

Because of the drawbacks of fixed-location sensors, various alternative data sources have been proposed for traffic monitoring, including cellular signaling data (Tettamanti et al., 2012; Ran, 2013), probe vehicle data (Turner et al., 1998; Chen and Chien, 2001; Guo et al., 2019), and even satellite images (Seo and Kusakabe, 2018). Among all the alternatives, probe vehicle data have attracted the most attention from researchers and practitioners in recent years because of the advances in connected vehicle (CV) technologies, the prevalence of online navigation systems, and the emergence of ride-hailing services.

1.2 Traffic state estimation

1.2.1 Definition

Traffic states represent traffic conditions at a given location and time. Commonly used traffic state variables include traffic flow, traffic density, travel speed, travel time, and queue length. Some of the variables can be directly measured by sensors, although the measurement usually contains some errors. Some other variables, such as the queue length, often have to be inferred from the collected traffic data. Traffic state estimation refers to the process of estimating traffic state variables using the data collected by sensors such as loop detectors, cameras, and probe vehicles (Seo et al., 2017).

1.2.2 Importance

The estimation of traffic states is critical for many transportation-related applications. In order to respond to the change in traffic conditions, the adaptive control of traffic signals or ramp meters requires the information about traffic flows or queue lengths of different movements (Papageorgiou et al., 2003). Network-level traffic control strategies, such as perimeter control, rely heavily on the estimation of traffic states inside and outside the controlled region (Geroliminis et al., 2012). The performance measures of transportation systems also depend on the measurement or estimation of indices such as travel speed, travel time, and traffic flow (Cheng et al., 2012; Wang et al., 2019). The annual average daily traffic (AADT) on different roads are critical inputs to transportation planning and roadway design (Seo et al., 2015). Driving behavior advisory, which can potentially prevent traffic accidents and alleviate traffic congestion, is expected to utilize the information about traffic states as well (Zheng, 2016; Seo et al., 2017). Even for evacuation planning and management during natural disasters, monitoring the traffic states is also crucial (Wolshon et al., 2005a,b; Liu et al., 2007).

1.2.3 Challenges

There are a few challenges for traffic state estimation, especially in urban areas. First, real-world traffic is highly dynamic. Varying in space and time (Shahrbabaki et al., 2018), traffic conditions can also be influenced by other factors such as weather conditions, traffic incidents, and road construction.

Second, traffic models usually oversimplify real-world scenarios. Many traffic flow models consider traffic as a continuum. It might be a good approximation in highway-related scenarios. However, in urban areas, the existence of traffic signals and stop signs makes the approximation inappropriate (van Zuylen et al., 2010; Cheng et al., 2012; Shahrbabaki et al., 2018). At the microscopic level, car-following models often ignore the randomness in driving behaviors. In terms of travel behaviors, most classical models impose rationality assumptions

to travelers. Nevertheless, human behaviors are sometimes irrational and arbitrary.

Third, available traffic data are often limited and contain measurement errors (Seo and Bayen, 2017). The widely applied loop detectors can only record information about specific locations (Guo et al., 2019). Besides, due to the lack of maintenance, loop detector data often contain missing values (Qu et al., 2009; Kawasaki et al., 2019). Cameras used for traffic monitoring usually only cover intersections and arterials. The accuracy of vehicle automatic identification can be significantly influenced by weather and lighting conditions. As for probe vehicles, although their trajectories can cover a large area, the market penetration rate is still very low currently (Zheng and Liu, 2017; Wang et al., 2019; Zhao et al., 2019a). Therefore, probe vehicle data cannot directly provide us volume-related traffic state variables such as traffic flows and traffic densities (Seo et al., 2015).

1.3 Probe vehicle

1.3.1 Definition and types

Probe vehicles, sometimes called floating cars, refer to the vehicles that observe traffic conditions while floating in the traffic flow (Seo et al., 2015). Depending on the purposes why they are traveling on the road, probe vehicles can be divided into active probe vehicles and passive probe vehicles. Active probe vehicles refer to the vehicles that are intentionally sent into the traffic for data collection. On the contrary, passive probe vehicles collect traffic information while they travel for their own purposes (Turner et al., 1998). An example of passive probe vehicles is ride-hailing vehicles, of which the purpose of traveling is to pick up and deliver passengers, although meanwhile their locations are recorded and reported to the ride-hailing platforms. The platforms then use the data for applications such as vehicle routing and arrival time prediction (Li et al., 2018).

Probe vehicle technologies can be implemented in different ways. Early implementation includes mounting transmitters on signpost structures to track public transits, equipping vehicles with electronic tags for electronic toll collection (ETC), or identifying vehicle locations

using cellular geolocation (Turner et al., 1998). The most popular probe vehicles nowadays are enabled by the Global Positioning System (GPS).

GPS-based probe vehicles can be further classified into a few categories. Online navigation systems such as Google Maps offer navigation services to travelers who have the applications installed on their cellphones. In return, the locations of the travelers might be reported to the platform to improve the performance of the navigation systems. Connected vehicles can conduct vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications. When connected vehicles travel in transportation networks, the information about the vehicles is broadcast to the ambient environment. The broadcast information can help the surrounding vehicles make planning and control decisions. Traffic controllers can also adjust their control strategies and signal timing parameters accordingly, once the road side units (RSU) receive the information. The recently emerging ride-hailing platforms, such as Uber and DiDi, represent another type of GPS-based probe vehicle. In order to match drivers and riders and monitor the trips, the platforms collect the location data of the ride-hailing vehicles in real time. Other types of GPS-based probe vehicles include the vehicles equipped with the OnStar system (https://www.onstar.com), the buses or emergency vehicles equipped with GPS devices, and so on. The probe vehicles referred to in this thesis are mainly passive and GPS-based probe vehicles.

1.3.2 Probe vehicle data for ITS applications

In most cases, the information collected by a GPS device is the trajectories represented by a series of timestamps and locations (Zheng et al., 2009). Depending upon the communication protocol and the instrumentation of probe vehicles, we may also collect other information from probe vehicles. For instance, besides the vehicle location and status, the basic safety message (BSM) broadcast by connected vehicles may contain additional information such as the ambient air pressure and temperature (SAE J2735).

For traffic state estimation, the most commonly used probe vehicle data are the GPS

trajectories. However, due to random GPS errors, the longitude and latitude collected by the GPS devices usually do not perfectly match the roads on maps. Therefore, preprocessing is needed before using the trajectory data for traffic state estimation. The step that maps GPS points to the links of a transportation network is called map matching. Most popular map matching algorithms are graph-based algorithms (Newson and Krumm, 2009; Lou et al., 2009). Although GPS errors can be corrected to some extent during the map matching process, the accuracy of the matched trajectories is still hard to reach the lane level. To improve the quality of the GPS data, one may consider applying Differential GPS (DGPS), which reduces positioning errors using reference stations (Parkinson and Enge, 1996). Another factor influencing the quality of trajectory data is the sampling rate. The frequency for a connected vehicle to broadcast the BSM is 10 Hz. Ride-hailing vehicles mostly report their locations to central servers every few seconds (Zhao et al., 2019a). The trajectory data sampled at such rates are sufficient for most ITS applications.

From the perspective of information collection, one of the most remarkable characteristics of probe vehicle data is that each trajectory can cover different locations and different time slots. The three-dimensional time-space diagram in Figure 1.1 illustrates the information that can be extracted from several typical data sources. The straight lines in orange represent loop detectors that collect traffic information at specific locations throughout the entire time horizon. The horizontal plane in light blue stands for a snapshot of the transportation network taken by imaging satellites. The traffic information at the moment when the snapshot is taken can be extracted from the images by applying computer vision techniques. The blue curves represent the trajectories of probe vehicles, which extend in both spatial and temporal dimensions.

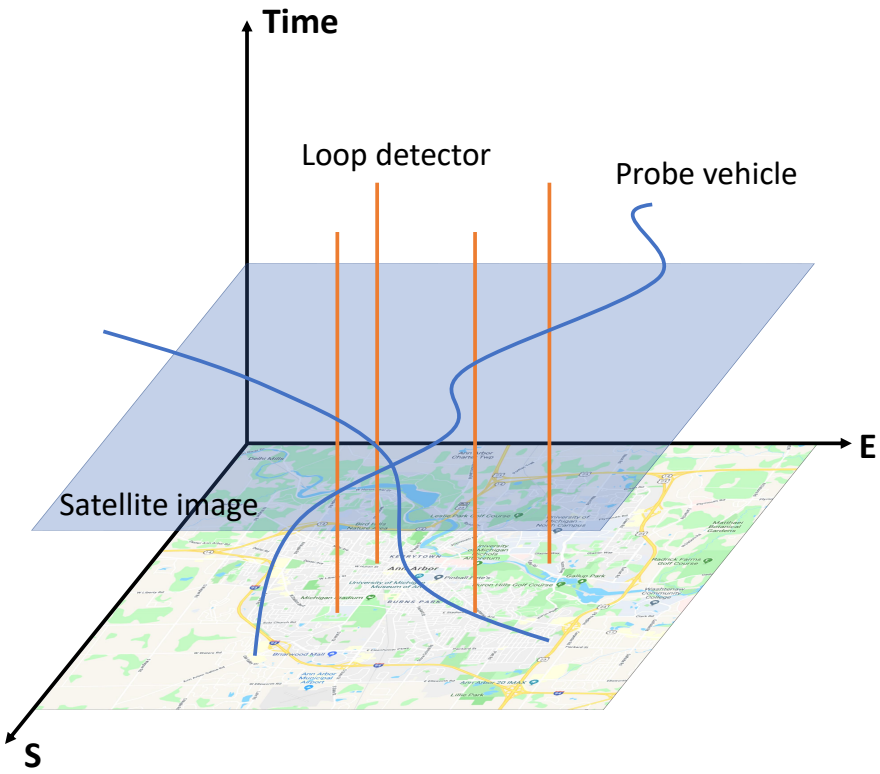


Figure 1.1: Traffic data collected by different sensors.

1.3.3 Pros and cons of using probe vehicle data

Using probe vehicle data as a data source for traffic state estimation has several advantages. First, the cost of collecting probe vehicle data is relatively low (Seo et al., 2015), especially when we consider the marginal cost of adding a new probe vehicle into an existing system. Also, compared to fixed-location sensors, maintaining a probe vehicle system is less expensive. Since the data collection process can be implemented in a crowd-sourcing manner, even if a few probe vehicles stop to work, the performance of the entire system is rarely influenced. Second, the coverage of probe vehicle data is broad. Probe vehicle trajectories cover times and locations whenever and wherever travelers drive the vehicles. Furthermore, at the places where the travel demand is higher or the traffic is more congested, usually more data can be collected, which gives us more insights into the traffic and helps us better solve the traffic problems. Third, with the rapid development of telecommunication technologies

(5G, LTE-V), connected vehicle technologies, and the Internet of Things (IoT), the market penetration of probe vehicles is likely to go up significantly. In the future, probe vehicle data may become more and more accessible (Zheng et al., 2018). Fourth, by adding other types of sensors, probe vehicles can collect abundant data other than just GPS trajectories, which would be beneficial to ITS applications. For example, if probe vehicles are equipped with spacing measurement devices, the spacing data can be easily used for estimating traffic densities (Seo et al., 2015).

On the other hand, however, a few concerns have been raised in terms of using probe vehicle data for traffic management and control. First, probe vehicle data may contain some privacy-sensitive information about personal mobility patterns. This kind of information can be potentially used to infer where the travelers live, when they go to work, and even who they are, which gives rise to privacy issues. Second, the current penetration rate of probe vehicles is still low in most areas. When using the data for traffic state estimation, the reliability has not yet been studied systematically, less some efforts on travel time estimation (Patire et al., 2015). Third, some probe vehicle data are very sparse due to low sampling rates, which can undermine the accuracy of map matching and lead to erroneous traffic state estimation results.

1.4 Literature overview

Since the early deployment of probe vehicles, the estimation of travel times and travel speeds has been studied extensively, mainly because it requires relatively low market penetration to achieve satisfactory accuracy (Srinivasan and Jovanis, 1996; Turner et al., 1998; Nakata and Takeuchi, 2004; Ramezani and Geroliminis, 2012; Zheng and Van Zuylen, 2013; Jenelius and Koutsopoulos, 2017). With the growth of probe vehicle market penetration in the past decade, researchers have started to pay more and more attention to the estimation of queue lengths and traffic volumes.

1.4.1 Queue length estimation

Queue length estimation is critical for traffic signal control and performance measures. Traditionally, queue lengths at signalized intersections are estimated by applying the shockwave theory to fixed-location sensor data (Skabardonis and Geroliminis, 2008; Liu et al., 2009; Lee et al., 2015; An et al., 2018). In recent years, a broad range of probe vehicle based methods was proposed. According to the methodologies applied, the existing studies on probe vehicle based queue length estimation can be classified into two main categories.

The first category of studies estimated queue lengths by applying the traffic flow theory to probe vehicle data. Ban et al. (2011) identified the break points of travel delays using probe vehicle data and estimated queue lengths in real time, with the assumption of uniform vehicle arrivals. Cetin (2012) focused on the oversaturated traffic conditions and estimated the back of the queue by determining the shockwave speed from probe vehicle trajectories. Li et al. (2013a) treated the cycle-by-cycle queue length evolution as a dynamic system and applied a Kalman filter to estimate the queue lengths. Hao et al. (2015) exploited the travel times of queueing vehicles to infer their positions and then estimated queue lengths using the inferred positions. Instead of estimating queue lengths, Ramezani and Geroliminis (2015) estimated the queue profiles using probe vehicle data, with a focus on the congested traffic conditions. Li et al. (2017) reconstructed the queue forming and discharging processes without additional signal timing information and found the maximum queue length accordingly. For most of the methods in this category, a sufficient high penetration rate is required to identify the shockwaves successfully.

The second category of studies tackled the problem from the perspective of the probability theory and statistics. Comert and Cetin (2009) showed that given the penetration rate of probe vehicles and the distribution of queue lengths, the position of the last probe vehicle in the queue alone would be sufficient for cycle-by-cycle queue length estimation. The authors also analyzed the relationship between the probe vehicle market penetration and estimation accuracy. Comert and Cetin (2011) extended the work by further considering the time when

the probe vehicles joined the queues. Following the early work, Comert (2013a,b) studied the effect of the data from stop line detection and also investigated the scenario when the penetration rate is not given. Instead of using the stopping positions of probe vehicles, Hao et al. (2014) proposed a Bayesian network to use the travel time information for queue length estimation and validated the proposed method using real-world data. By combining probe vehicle data and loop detector data, Shahrbabaki et al. (2018) developed a method that could estimate queue lengths and vehicle accumulations in links in real time. Mei et al. (2019) took a Bayesian approach and calculated the posterior distribution of queue lengths under the observations from probe vehicles.

From another point of view, almost all the existing probability theory based methods treated each intersection and each signal cycle independently, without considering the correlation of queue lengths in adjacent intersections and cycles. Among the methods based on the traffic flow theory, a few methods considered the temporal correlation of queue lengths in different cycles. Specifically, Li et al. (2013a) formulated the queueing dynamics into state-space equations; Cetin (2012) and Ramezani and Geroliminis (2015) considered the probe vehicle data in adjacent cycles when the traffic is oversaturated.

Table 1.1 summarizes the existing studies introduced above.

Table 1.1: Existing literature on probe vehicle based queue length estimation.

Literature	Methodology	Data source	Correlation
Ban et al. (2011)	Shockwave	PV	None
Cetin (2012)	Shockwave	PV	Temporal
Li et al. (2013a)	Shockwave, Kalman filter	PV, LD	Temporal
Hao et al. (2015)	Kinematic equation	PV	None
Ramezani and Geroliminis (2015)	Shockwave, Linear regression	PV	Temporal
Li et al. (2017)	Shockwave	PV	None
Comert and Cetin (2009)	Probability	PV	None
Comert and Cetin (2011)	Probability	PV	None
Comert (2013a)	Probability	PV, LD	None
Comert (2013b)	Probability	PV	None
Hao et al. (2014)	Bayesian network	PV	None
Shahrbabaki et al. (2018)	Probability	PV, LD	None
Mei et al. (2019)	Shockwave, Probability	PV	None

Note: PV - Probe vehicle, LD - Loop detector

1.4.2 Traffic volume estimation

As another important traffic state variable, the information of traffic volumes is of significance for traffic management and control as well. In current practice, traffic volume data are mainly collected by the costly fixed-location sensors. The growth of the market penetration of probe vehicles provides us a promising alternative data source. Over the past few years, diverse types of probe vehicle based methods have been proposed.

Some researchers estimated travel speeds from probe vehicle data for each road and then converted the intermediate results to traffic volumes by exploiting the relationship between travel speeds and traffic volumes (Shang et al., 2014; Lai and Huang, 2017). Some studies approached the problem from the perspective of the probability theory and statistics. Zheng and Liu (2017) proposed to estimate the vehicle arrival rate from the arrival times of probe

vehicles through maximum likelihood estimation (MLE), assuming vehicle arrivals at isolated intersections follow time-varying Poisson processes. Wang et al. (2019) combined shockwave analyses and Bayesian networks to estimate the average arrival rate of the assumed Poisson process. The estimated arrival rate was then used to calculate the posterior distribution of the cycle-based traffic volume. Similarly, Yao et al. (2019) integrated the shockwave theory and the probability theory and estimated cycle-based traffic volumes by using the trajectories of both the stopped and non-stopped vehicles. Almost all the methods of these types did not consider the correlation of traffic volumes either in nearby locations or in adjacent time slots, except the work by Luo et al. (2019), where the authors improved the estimation accuracy by considering the information from adjacent intersections.

A few Kalman filtering based methods took the temporal correlation of traffic volumes into account by formulating the estimation problems as dynamic systems (Aljamal et al., 2019; Chen and Levin, 2019). Besides, some recent studies attempted to estimate traffic volumes at a citywide level by considering the spatiotemporal correlation of traffic volumes in transportation networks. Cui et al. (2017) estimated the unknown traffic volumes by applying compressive sensing techniques. The correlation of traffic volumes in adjacent time slots was captured by a Toeplitz matrix; the correlation of traffic volumes in nearby locations was learned by fitting linear regression models to probe vehicle counts. Zhan et al. (2017) developed a hybrid framework that extracted some high-level features from calibrated fundamental diagrams and estimated traffic volumes by machine learning techniques. The model depends on various data sources, including probe vehicle data, partial loop detector data, points of interest (POI) data, and meteorology data. Using similar data sources, Meng et al. (2017a) modeled the spatiotemporal correlation of traffic volumes by a multi-layer affinity graph. Tang et al. (2019) focused on the fusion of probe vehicle data and surveillance camera data instead. The authors recovered vehicle trajectory data from the incomplete observations from cameras and captured multi-hop correlations between different roads and time slots by applying multi-view graph embedding.

Table 1.2 summarizes the existing studies in the literature.

Table 1.2: Existing literature on probe vehicle based traffic volume estimation.

Literature	Methodology	Data source	Correlation
Shang et al. (2014)	FD	PV	None
Lai and Huang (2017)	FD	PV	None
Zheng and Liu (2017)	MLE	PV	None
Wang et al. (2019)	Shockwave, MLE	PV	None
Yao et al. (2019)	Shockwave, MLE	PV	None
Luo et al. (2019)	Shockwave, MLE	PV	Spatial
Aljamal et al. (2019)	Kalman filter	PV	Temporal
Chen and Levin (2019)	Kalman filter	PV	Temporal
Cui et al. (2017)	Compressive sensing	PV, LD	Spatiotemporal
Zhan et al. (2017)	FD, Data-driven	PV, LD, POI, Weather	Spatiotemporal
Meng et al. (2017a)	Data-driven	PV, LD, POI, Weather	Spatiotemporal
Tang et al. (2019)	Data-driven	PV, LD, POI, Weather	Spatiotemporal

Note: PV - Probe vehicle, LD - Loop detector, FD - Fundamental diagram

1.5 Research scope

As discussed above, the vast amount of probe vehicle trajectory data is a promising substitute for the widely used fixed-location sensors. Extensive literature has shown that probe vehicle data can be used for estimating travel times or travel speeds. Nevertheless, although there are some existing methods for queue length and traffic volume estimation based on probe vehicle data, most of the methods impose strict assumptions and cannot be implemented on a large scale in real life. Therefore, the research scope of this thesis is to develop traffic state estimation methods that can avoid the restrictions and be potentially implemented on a large scale.

For queue length estimation using probe vehicle data, the traffic flow theory based meth-

ods usually assume the penetration rate is sufficiently high, and vehicle arrivals follow certain given processes such as uniform arrivals or Poisson arrivals (Ban et al., 2011; Cheng et al., 2012; Hao et al., 2014); the probability theory based methods not only assume the queue lengths in different traffic signal cycles are independent but also require the knowledge of the penetration rate and (or) queue length distribution. However, when implementing the methods in real life, prior information about the penetration rate and queue length distribution are not available beforehand. To fill the gap, one part of this thesis is dedicated to estimating the probe vehicle penetration rate and queue length distribution at the movement level for each signalized intersection. To this end, we propose a series of approximate estimators (AE) and a maximum likelihood estimator, which have been validated using both simulation and real-world datasets.

When it comes to cycle-by-cycle queue length estimation, another aspect that is often ignored by the existing literature is the possible correlation of queue lengths in different traffic signal cycles. For example, in oversaturated traffic conditions, the queue length in the next cycle will be dependent on the queue length in the current cycle because of the overflow (residual) queue (Wang et al., 2019). Studying such scenarios is of both theoretical and practical significance. Therefore, another part of the thesis aims to develop queue length estimation methods that are suitable for this kind of scenario. We model the queue evolution and observation processes in a probe vehicle environment by using a hidden Markov model (HMM), where queue lengths are hidden states and probe vehicle data are observations. Accordingly, we propose two cycle-by-cycle queue length estimation methods, which take advantage of the observed information in multiple cycles and improve the estimation accuracy. We also develop an algorithm to estimate the key parameters of the HMM from historical probe vehicle data.

For traffic volume estimation, as mentioned in the previous subsections, the high cost of fixed-location sensors leads to the missing data problem and low coverage problem (Zhan et al., 2017; Cui et al., 2017). Nevertheless, when only low-market-penetration probe vehicle

data are available, it is difficult to estimate real-time traffic volume information if we consider each time slot and each road separately (Zheng and Liu, 2017). Noticing that traffic volumes in a transportation network are correlated spatially and temporally, we try to capture the correlation by fusing probe vehicle data and partial fixed-location sensor data, which are complementary to each other, as demonstrated by Figure 1.1. To approach the problem by data fusion, we apply both the singular value decomposition (SVD) and the probabilistic principal component analysis (PPCA) to estimate the unknown traffic volumes. These methods exploit the correlation of traffic volumes in different locations and different time slots and thereby achieve good estimation accuracy even if the probe vehicle data are sparse. Different from the existing methods, the spatiotemporal correlation is captured without using any other data sources such as POI data or weather data.

The three research topics introduced above are summarized in Figure 1.2. In the first topic, we treat different movements and different traffic signal cycles independently. The traffic state s_{ij} at time i and location j is estimated using the associated observation o_{ij}. In the second topic, we consider the correlation of queue lengths in different cycles. Specifically, the temporal correlation is captured by a Markov chain. The traffic state s_{ij} is estimated by also using the observations in the adjacent cycles. The correlation in the temporal dimension turns out to help us improve the queue length estimation accuracy. In the third topic, we consider the correlation not only in the temporal dimension but also in the spatial dimension. The spatiotemporal correlation is captured by low-rank representation methods, including the SVD and the PPCA. In this way, unknown traffic states are estimated by combing the data collected in different time slots and different roads.

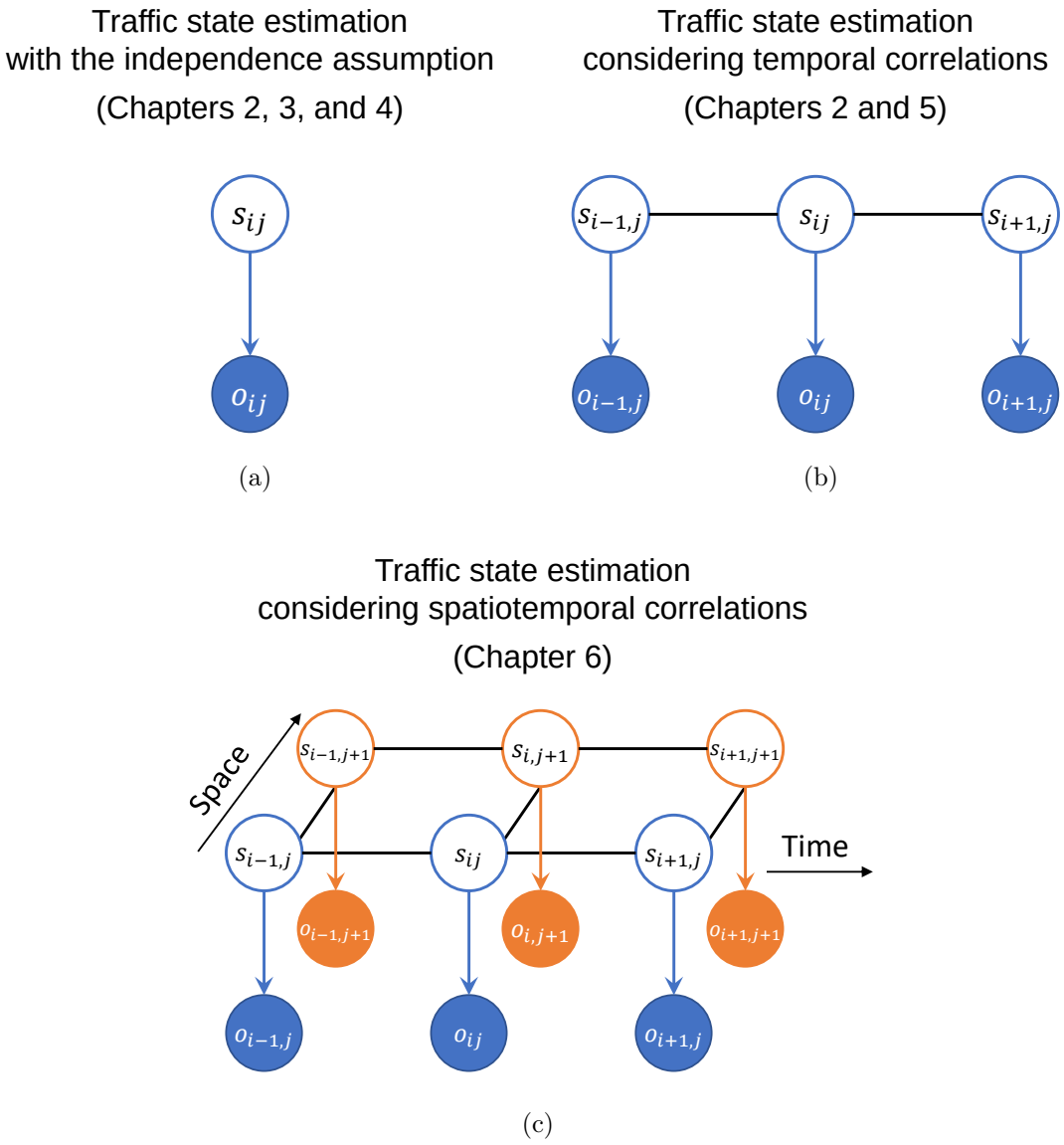


Figure 1.2: The framework of the three studied topics.

1.6 Contributions

This thesis systematically studies traffic state estimation in different scenarios and proposes a series of methodologies. For traffic state estimation with the independence assumption, we propose to estimate the queue length cycle by cycle using the observed stopping positions of probe vehicles. To obtain the penetration rate and the queue length distribution, which are required parameters, we aggregate historical probe vehicle data and develop a se-

ries of estimators. For traffic state estimation considering temporal correlations, we model the queueing process as a Markov chain and propose cycle-by-cycle queue length estimation methods that can be carried out by dynamic programming (DP). We also estimate the required parameters by aggregating the observation sequences over different days. For traffic state estimation considering spatiotemporal correlations, we achieve real-time traffic volume estimation by fusing probe vehicle data and partial fixed-location sensor data. Although the focus of this thesis is on the estimation of queue lengths and traffic volumes, the methodologies can be easily extended to the estimation of other traffic state variables. By aggregating historical data or exploiting spatiotemporal correlations, we overcome the obstacle to probe vehicle based traffic state estimation caused by the low penetration rate.

Besides the methodological contributions, this thesis also makes practical contributions. Developed with the aim of real-world implementation, all the methods in this thesis are suitable for the current low-penetration-rate probe vehicle environment. Some of the methods have already been implemented in real life. For example, the methods in Chapter 3 have been implemented by Didi Chuxing as part of their dynamic traffic control system. All the other methods have been validated using either real-world datasets or simulated datasets under reasonable assumptions. We expect the proposed methods in this thesis to be implemented on an even larger scale and help transportation agencies solve more real-world traffic problems.

1.7 Thesis overview

This thesis focuses on the application of probe vehicle data to traffic state estimation, particularly, the estimation of queue lengths and traffic volumes in different scenarios. The rest of this thesis is organized as follows.

In Chapter 2, we describe the queue length estimation problem and introduce several cycle-by-cycle queue length estimation methods. We first focus on the case where queue lengths in different cycles are assumed to be independent and identically distributed (i.i.d.). We introduce a few methods that can estimate queue lengths from probe vehicle observations

in this case. Then, we generalize the methods to the non-i.i.d. case, where we propose a hidden Markov model to solve the queue length estimation problem. Most of the content of this chapter can be found in Zhao and Liu (2020) and Zhao et al. (2020a).

In the i.i.d. case, the queue length distribution and probe vehicle penetration rate are required for queue length estimation. Chapters 3 and 4 present the approximate estimation and maximum likelihood estimation of these key parameters, respectively. Most of the work in these two chapters can be found in Zhao et al. (2019a,b) and Zhao and Liu (2020).

Chapter 5 focuses on the estimation of the parameters needed in the non-i.i.d. case. The parameters of the HMM, which include the penetration rate, initial probabilities, and transition probabilities, are estimated by applying the Expectation-Maximization (EM) algorithm and dynamic programming. The work can be found in Zhao et al. (2020a).

In Chapter 6, we present two data fusion based methods for traffic volume estimation. The first method captures the correlation of traffic volumes by applying the singular value decomposition to probe vehicle data. In the second method, we capture the correlation by generalizing the framework of the probabilistic principal component analysis. Most of the work can be found in Zhao et al. (2020b,c).

Finally, we provide concluding remarks and discuss some future research directions in Chapter 7.

Chapter 2

Cycle-by-cycle queue length estimation

2.1 Introduction

2.1.1 Background

Since traffic signals serve as critical components in urban traffic management systems, a better understanding of the performance of traffic signals can help transportation agencies better solve traffic problems. Queue length is one of the parameters that can be used for traffic signal control and performance measures. To estimate queue lengths, conventional approaches apply the shockwave theory to the data collected by fixed-location sensors, such as loop detectors (Skabardonis and Geroliminis, 2008; Liu et al., 2009; Lee et al., 2015; An et al., 2018). However, as the installation and maintenance cost of fixed-location sensors is very high, only a small portion of roadways are covered by the sensors.

With the emergence of connected vehicles and ride-hailing services, the deficiency of fixed-location sensors could potentially be overcome by the new data source: probe vehicle data. Although the current market penetration of probe vehicles is still low, probe vehicle data have a much broader coverage of roadways, and the cost is much lower compared to fixed-location sensors.

2.1.2 Related work

As introduced in Section 1.4, the existing methods for probe vehicle based queue length estimation can be classified into two categories. One category of literature estimated queue

lengths by applying the traffic flow theory to probe vehicle data (Ban et al., 2011; Cetin, 2012; Li et al., 2013a; Ramezani and Geroliminis, 2015; Li et al., 2017). Most of the literature in this branch requires a sufficient high penetration rate to identify the shockwaves. Another category of literature tackled the problem based on the probability theory and statistics (Comert and Cetin, 2009, 2011; Comert, 2013a,b; Hao et al., 2014; Comert, 2016; Shahrbabaki et al., 2018). The methods in this category usually require the knowledge of the probe vehicle penetration rate or the vehicle arrival process.

Most of the relevant studies summarized above focused on isolated intersections under moderate traffic conditions and treated the queues in different cycles independently. However, when there are overflow queues, or when the number of vehicle arrivals is correlated in different cycles, the queue lengths in different cycles are not independent anymore. These kinds of scenarios occur frequently in real life. In fact, although the estimation of queue lengths in such scenarios is not well studied in the context of probe vehicles, the modeling of dependent queues has drawn much attention from researchers since the 1950s. The modeling of overflow queues, an example of dependent queues, has been studied extensively. The ultimate goal for most literature on overflow queue modeling was to study the delay caused by traffic signals. Some early studies focused on the calculation of the mean overflow queue length, given the vehicle arrival process (Miller, 1963; Newell, 1965; McNeil, 1968). Some other studies attempted to obtain the distribution of the queue lengths (Haight, 1959; Darroch, 1964; Ohno, 1978; Heidemann, 1994; Mung et al., 1996; van Leeuwaarden, 2006), under the assumption of Poisson arrivals. In particular, a few studies modeled the cycle-to-cycle queue evolution using Markov chains (Newell, 1971; Olszewski, 1990, 1994; Viti and Van Zuylen, 2010; Igbinosun and Omosigho, 2016).

2.1.3 Contribution and organization of the chapter

As introduced above, most of the existing methods focused on the undersaturated traffic conditions and treated the queues in different cycles independently. There is litter literature

considering the correlation of queues when using probe vehicle data to estimate queue lengths. In fact, considering the correlation could potentially improve the estimation accuracy, as the observations in adjacent cycles contain additional information. Following the branch of literature based on the probability theory, this chapter systematically studies both the i.i.d. and non-i.i.d. cases. Specifically, in the i.i.d. case, we extend the work of Comert and Cetin (2009) by providing several queue length estimators; in the non-i.i.d. case, we propose a hidden Markov model to capture the correlation of queues in different cycles. Further, the hidden Markov model is also compatible with the i.i.d. case. When there is no correlation, the model will give the same results as the i.i.d. case. Such a unified framework for probe vehicle based queue length estimation is of both theoretical and practical importance.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the observations that can be used for queue length estimation in a probe vehicle environment. In Section 2.3, we present several queue length estimators under the i.i.d. assumption. In Section 2.4, we propose a hidden Markov model that relaxes the i.i.d. assumption and considers the correlation of different cycles. Based on the HMM, we propose two cycle-by-cycle queue length estimation methods. In Section 2.5, we validate the proposed methods by numerical experiments. Finally, Section 2.6 provides some concluding remarks.

2.2 Observation of queues in a probe vehicle environment

At a signalized intersection, a queue will form when the traffic signal turns red. Without loss of generality, we restrict our discussion to a specific time-of-day (TOD) and a specific single-lane movement controlled by fixed-time traffic signals. Here, the queue length refers to the number of vehicles in the queue at a given time point of the signal cycle, for instance, at the start of the green phase or the start of the red phase (overflow queue). Denote the queue length in the ith cycle by l_i. Denote the maximum possible queue length by L_{max}. Assume the probe vehicles are homogeneously mixed in the traffic flow. The trajectories of the probe vehicles can be recorded by GPS devices and stored in a database. From the trajectory

data, the stopping positions of the probe vehicles can be extracted. Denote the number of queueing probe vehicles in the ith cycle by n_i. By assuming a uniform space headway, we can infer the number of regular vehicles before the last probe vehicle in the queue. Denote the pattern of the observed partial queue by a tuple q_i, which consists of binary elements indicating the vehicle types. Denote the length of the tuple by $|q_i|$. $\forall k = 1, 2, \ldots, |q_i|$, if the kth vehicle in the queue is a probe vehicle, the kth element of q_i is set to 1; otherwise 0.

Figure 2.1 illustrates the observation process in the probe vehicle environment. The diagram on the left shows a queue formed by two probe vehicles (in yellow) and four regular vehicles (in white). The diagram on the right shows the pattern of the partial queue that can be observed or inferred directly. Specifically, the positions of the two probe vehicles can be easily extracted from the trajectory data; the three regular vehicles are inferred by dividing the distance between the two probe vehicles by the space headway. In this example, the pattern of the observed partial queue is represented by the tuple $(1, 0, 0, 0, 1)$.

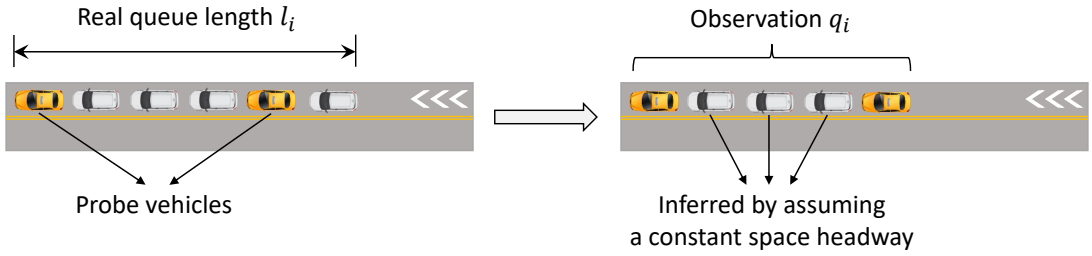


Figure 2.1: Queueing vehicles at a signalized intersection.

It is worth mentioning that in this thesis, the observation we use for queue length estimation only includes the stopping positions. Nevertheless, more information can be potentially used. For example, if the green phase starts immediately after a probe vehicle stops, it implies that the probe vehicle is the back of the queue.

2.3 Cycle-by-cycle queue length estimation in the i.i.d. case

Suppose the queue lengths in different cycles are independent and identically distributed, following a probability distribution denoted by a vector π. The jth element of π represents

$P(l_i = j), \forall j = 0, 1, \ldots, L_{max}$. Assume the queue length distribution π and probe vehicle penetration rate p remain roughly the same during the studied TOD. The patterns of the partially observed queues are governed by the queue length distribution and the penetration rate.

If the penetration rate p and the queue length distribution π are given, the queue lengths can be estimated cycle by cycle. From the perspective of the probability theory, we may apply the following two estimators. The first estimator is the maximum likelihood estimator, given by

$$\hat{l}_i = \underset{j}{\operatorname{argmax}} P\left(q_i \mid l_i = j\right) = \underset{j:|q_i| \le j \le L_{max}}{\operatorname{argmax}} \pi_j p^{n_i} \left(1-p\right)^{j-n_i} = \underset{j:|q_i| \le j \le L_{max}}{\operatorname{argmax}} \pi_j \left(1-p\right)^{j}. \quad (2.1)$$

The second estimator is the expected queue length conditional on the observed partial queue, given by

$$\hat{l}_i = \mathbb{E}\left(l_i \mid q_i\right) = \sum_{j=|q_i|}^{L_{max}} \frac{\pi_j p^{n_i} \left(1-p\right)^{j-n_i}}{\sum_{k=|q_i|}^{L_{max}} \pi_k p^{n_i} \left(1-p\right)^{k-n_i}} j = \sum_{j=|q_i|}^{L_{max}} \frac{\pi_j}{\sum_{k=|q_i|}^{L_{max}} \pi_k \left(1-p\right)^{k-j}} j. \quad (2.2)$$

Similar formulations of equation (2.2) were introduced in Comert and Cetin (2009). Given p and π, both of the estimators only depend on the position of the last probe vehicle $|q_i|$.

Figure 2.2 shows the mean absolute errors (MAE) when the two methods are used for queue length estimation under different penetration rates. The queue length distribution is chosen as $Poisson(\lambda = 10)$. The mean absolute errors of the three methods are calculated by

$$\mathbb{E}\left(\left|\hat{l}_i - l_i\right|\right) = \sum_{l_i=0}^{L_{max}} \sum_{|q_i|=0}^{l_i} P\left(l_i, |q_i|\right) \left|\hat{l}_i - l_i\right|, \quad (2.3)$$

where

$$P\left(|q_i|, l_i\right) = \begin{cases} \pi_{l_i} \left(1-p\right)^{l_i}, & |q_i| = 0 \\ \pi_{l_i} p \left(1-p\right)^{l_i - |q_i|}, & |q_i| \neq 0 \end{cases}. \quad (2.4)$$

The baseline method used for comparison is the naive estimation method which takes the

position of the last probe vehicle as an estimate of the queue length, that is, $\hat{l}_i = |q_i|$. Apparently, the naive estimation method usually leads to an underestimation. In general, the maximum likelihood estimator performs better when the penetration rate is high, because it gives the "most likely" estimate; the conditional expectation yields better estimation accuracy when the penetration rate is low, because it is essentially a weighted average and thus gives a more conservative estimate. Similar patterns are observed when setting λ to other values.

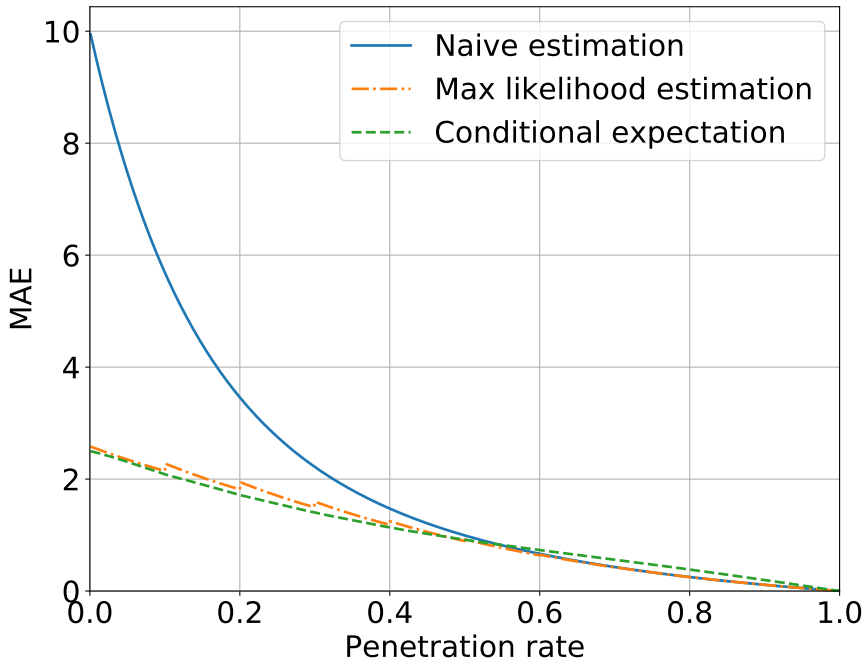


Figure 2.2: Estimation accuracy of the cycle-by-cycle estimation methods in the i.i.d. case.

Both of the estimators given by equation (2.1) and (2.2) require the knowledge of the penetration rate p and the queue length distribution π. In Chapters 3 and 4, we will discuss how to estimate the required parameters from historical probe vehicle data.

2.4 Cycle-by-cycle queue length estimation in the non-i.i.d. case

2.4.1 A hidden Markov model for queue evolution in the probe vehicle environment

Now we relax the i.i.d. assumption imposed in Section 2.3. To capture the correlation of different traffic signal cycles, we assume the stochastic process $\{l_i\}$ is a time-homogeneous Markov chain, that is

$$P(l_2 \mid l_1) = P(l_{i+1} \mid l_i) = P(l_{i+1} \mid l_1, l_2, \ldots, l_i), \forall i = 1, 2, \ldots, C-1. \tag{2.5}$$

An example that follows such properties is the overflow queue scenario described by Viti and Van Zuylen (2010), where the overflow queue at a signalized intersection is induced by an exogenous arrival process. Also, assume a vehicle will be served within two cycles, which is a sufficient condition for

$$P(q_i \mid l_i) = P(q_i \mid q_1, q_2, \ldots, q_{i-1}, l_1, l_2, \ldots, l_i), \forall i = 1, 2, \ldots, C. \tag{2.6}$$

After C cycles of observations, the hidden values of our interest are the sequence of queue lengths $l = \{l_1, l_2, \ldots, l_C\}$. The observations we have are the sequence of observed partial queues $q = \{q_1, q_2, \ldots, q_C\}$. Under the assumptions listed above, the queueing process and the observation process can be modeled by a hidden Markov model, where the hidden states are l and the observations are q, as illustrated in Figure 2.3.

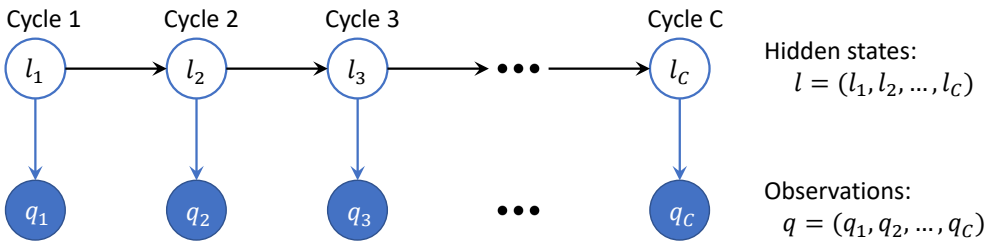


Figure 2.3: A hidden Markov model for the queueing process and observation process.

Denote the probabilities of the initial queue length by π, of which the jth element represents $P(l_1 = j), \forall j = 0, 1, \ldots, L_{max}$. Denote the transition probability matrix of the HMM by T. The element in the jth row and the kth column of T is

$$T_{jk} = P(l_{i+1} = k \mid l_i = j), \forall i = 1, 2, \ldots, C-1, \forall j, k = 0, 1, \ldots, L_{max}. \tag{2.7}$$

The probability of observing q_i from the hidden state l_i (emission probability) is

$$E_{l_i q_i} = P(q_i \mid l_i) = \begin{cases} p^{n_i}(1-p)^{l_i - n_i}, & l_i \geq |q_i| \\ 0, & l_i < |q_i| \end{cases}, \forall i = 1, 2, \ldots, C. \tag{2.8}$$

As equation (2.8) suggests, given the observations, the emission probabilities only depend on the penetration rate p. Therefore, the parameters of the HMM include π, T, and p. The hidden Markov model is also compatible with the i.i.d. case, where each row of T will be identical to the transpose of π, that is,

$$P(l_i = j \mid l_{i-1} = k) = P(l_i = j) = \pi_j, \forall i = 2, 3, \ldots, C, \forall j, k = 0, 1, \ldots, L_{max}. \tag{2.9}$$

2.4.2 Queue length estimation methods

When the parameters of the HMM are given, the observations from the probe vehicle data can be used to estimate the sequence of queue lengths. In this section, we propose two methods for cycle-by-cycle queue length estimation.

The first estimator is the maximum likelihood estimator, given by

$$\hat{l} = \underset{l}{\operatorname{argmax}} P(q_1, q_2, \ldots, q_C \mid l) = \underset{l}{\operatorname{argmax}} P(l_1) \prod_{i=2}^{C} P(l_i \mid l_{i-1}) \prod_{i=1}^{C} P(q_i \mid l_i). \tag{2.10}$$

The MLE can be obtained by applying the Viterbi algorithm (Viterbi, 1967; Forney, 1973). In the corresponding trellis diagram, as shown by Figure 2.4, each vertex represents a pos-

sible value of the hidden state, with which an emission probability is associated. Each arc represents a cycle-to-cycle state transition, with which a transition probability is associated. The Viterbi algorithm essentially finds the path that traverses all the stages (traffic signal cycles) on the graph with the largest product of transition and emission probabilities, which is represented by the red path in Figure 2.4. This process is also called HMM decoding in the literature.

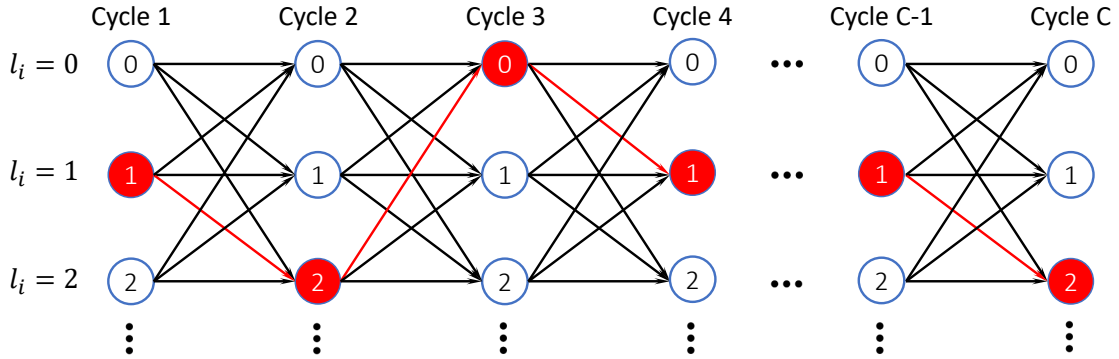


Figure 2.4: Graph representation of the maximum likelihood estimator in the non-i.i.d. case.

Another estimator is given by the expected queue length conditional on the sequential observations, that is,

$$\mathbb{E}\left(l_i \mid q_1, q_2, \ldots, q_i\right) = \sum_{j=1}^{L_{max}} \frac{P(q_1, q_2, \ldots, q_i, l_i = j)}{\sum_{k=0}^{L_{max}} P(q_1, q_2, \ldots, q_i, l_i = k)} j, \forall i = 1, 2, \ldots, C. \tag{2.11}$$

$\forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C$, the joint probability $P(q_1, q_2, \ldots, q_i, l_i = j)$ can be carried out recursively using the following equations (Baum et al., 1970).

$$P\left(q_1, l_1 = j\right) = \pi_j E_{jq_1}; \tag{2.12}$$

$$P(q_1, q_2, \ldots, q_i, l_i = j) = \sum_{k=0}^{L_{max}} P(q_1, q_2, \ldots, q_{i-1}, l_{i-1} = k) T_{kj} E_{jq_i}. \tag{2.13}$$

Both of the estimators given by equations (2.10) and (2.11) require the knowledge of the

penetration rate p, the initial queue length distribution π, and the transition matrix T. In Chapter 5, we will discuss how to estimate the required parameters from historical probe vehicle data.

2.5 Case studies

2.5.1 Simulation settings

To validate the proposed methods for queue length estimation, in the case study, we generate simulated data using a state-of-the-art overflow queue model proposed by Viti and Van Zuylen (2010). The model assumes that the number of vehicle arrivals in each traffic signal cycle is i.i.d., following a Poisson distribution. However, the number of vehicles that can be discharged in each cycle is limited. In some cycles, the number of vehicles arriving at the intersection might exceed the capacity, which leads to overflow queues. With the potential existence of overflow queues, the resultant cycle-to-cycle queueing process is a Markov chain, which makes it suitable for testing the proposed methods. The parameters of the queue model and their values are summarized in Table 2.1.

Table 2.1: The parameters of the Viti and Van Zuylen (2010) model and their values in the case study.

Parameters	Description	Value
λ	Average arrival rate	10
a_{max}	Maximum number of vehicle arrivals in each cycle	15
s	Maximum number of vehicles that can be served in each cycle	10
L_{max}	Maximum queue length	20

In the case study, we focus on the queues at the beginning of the (effective) red phases, namely, the overflow queues. The queues at the beginning of the green phases can be studied similarly. Figure 2.5 illustrates the queueing process characterized by the model. The colors of the vehicles indicate the traffic signal phases when the vehicles join the queue. At the

beginning of the ith cycle, there are five vehicles (l_i) in the overflow queue. During the red phase, six more vehicles (a_i^r) join the queue, resulting in a queue of size 11 when the traffic signal turns green. In the green phase, ten vehicles in the queue are discharged, while three more vehicles (a_i^g) join the queue. Consequently, the size of the overflow queue (l_{i+1}) is four at the beginning of cycle $i+1$. Obviously, the queue lengths in different cycles are correlated, and the stochastic process $\{l_i\}$ is a Markov chain.

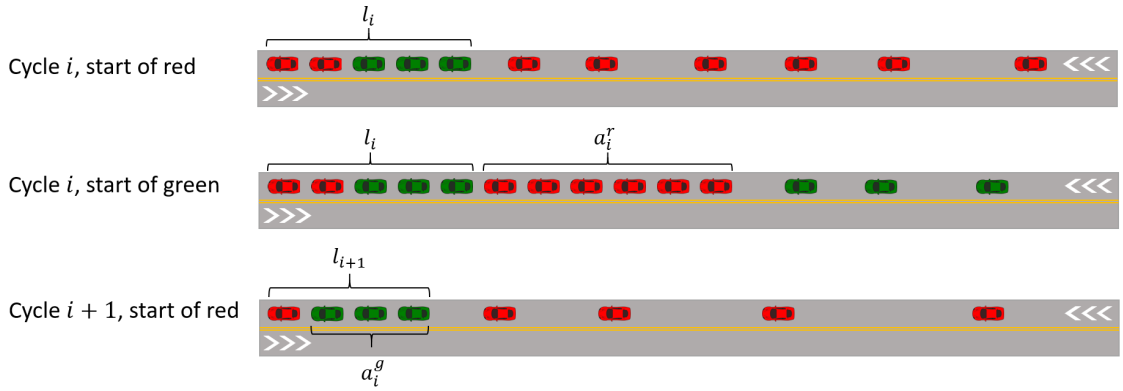


Figure 2.5: A typical scenario of overflow queues.

According to the derivations by Viti and Van Zuylen (2010), the transition probabilities of the Markov chain can be expressed as

$$T_{j0} = \begin{cases} \sum_{n=0}^{\min(s-j,a_{max})} P(n), & j \le s \\ 0, & j > s \end{cases}; \tag{2.14}$$

$$T_{jL_{max}} = \begin{cases} \sum_{n=L_{max}-j+s}^{a_{max}} P(n), & L_{max} - j + s \le a_{max} \\ 0, & \text{otherwise} \end{cases}; \tag{2.15}$$

$$T_{jk} = \begin{cases} P(s+k-j), & 0 \le s+k-j \le a_{max} \\ 0, & \text{otherwise} \end{cases}, \tag{2.16}$$

where $P(\cdot)$ denotes the probability mass function of the Poisson distribution, which is

$$P(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \forall n = 0, 1, \ldots \tag{2.17}$$

The initial distribution π is considered as the stationary distribution of the Markov chain, which can be obtained by solving the following linear equations

$$\pi_j = \sum_{k=0}^{L_{max}} \pi_k T_{kj}, \forall j = 0, 1, \ldots, L_{max}; \tag{2.18}$$

$$\sum_{j=0}^{L_{max}} \pi_j = 1. \tag{2.19}$$

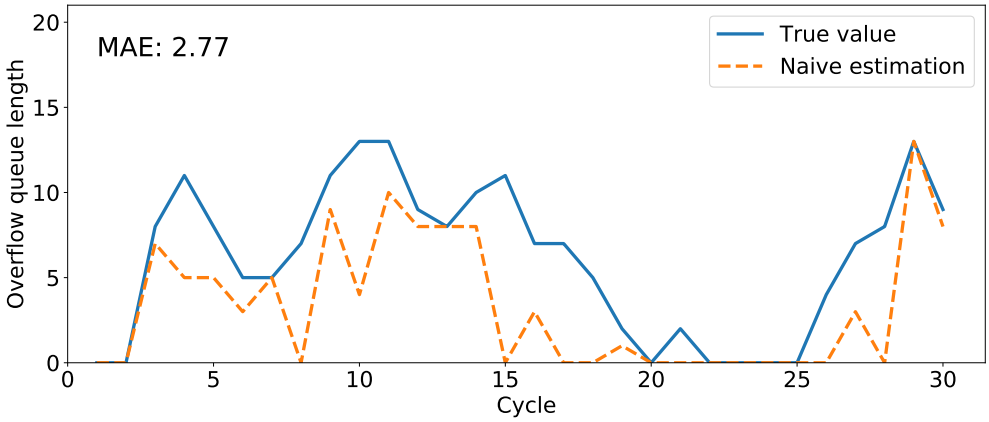
With the initial probabilities and transition probabilities specified above, we generate the ground-truth queues and their lengths $\{l_1, l_2, \ldots, l_C\}$. The penetration rate is set to 20%. With the penetration rate, we perform a Bernoulli trial for each vehicle in the queue to determine if it is a probe vehicle or a regular vehicle. Then, from the simulation data, we extract the observed partial queues $\{q_1, q_2, \ldots, q_C\}$, which are used as the input to the queue length estimation methods.

2.5.2 Results of cycle-by-cycle queue length estimation

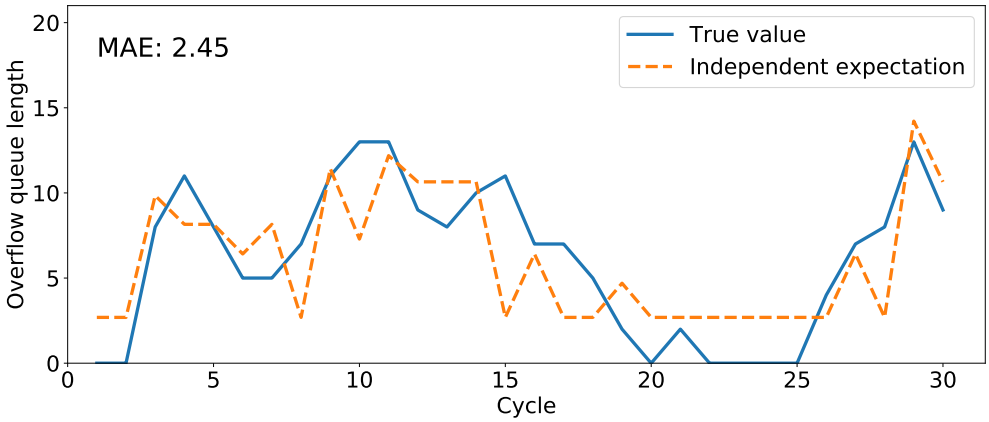
Figure 2.5 shows the cycle-by-cycle queue length estimation results when four different methods are applied to a 30-cycle observation sequence. The used measure of estimation accuracy is the mean absolute error. In Figure 2.5(a), the naive estimation refers to the estimator that takes the position of the last probe vehicle in the queue as an estimate of the queue length. Figure 2.5(b) shows the results when the correlation of different cycles is ignored. The queue length in each cycle is estimated independently by using the observation in the same cycle, which corresponds to equation (2.2) in the i.i.d. case. These two methods are considered as baseline methods.

Figures 2.5(c) and 2.5(d) show the estimation results of the methods proposed by equa-

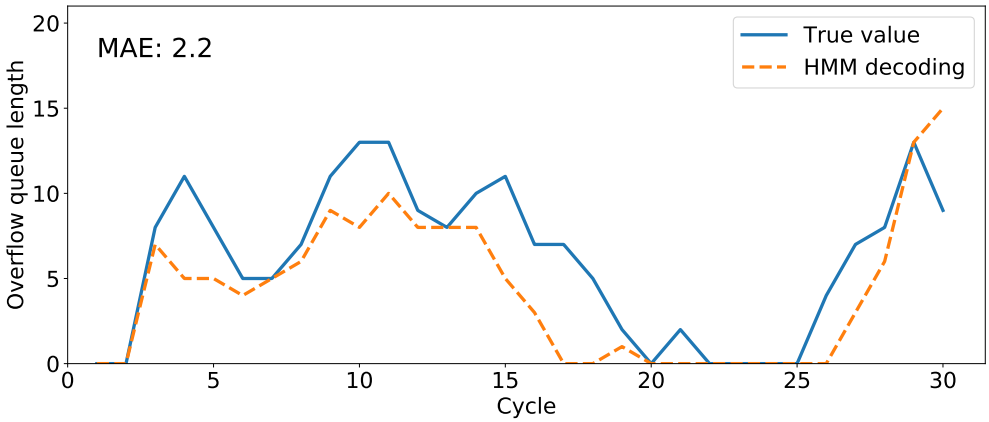
tions (2.10) and (2.11), respectively. The results indicate that the HMM-based methods outperform the two baseline methods. The reason for the improvement is that the HMM-based methods exploit the additional information in other cycles.



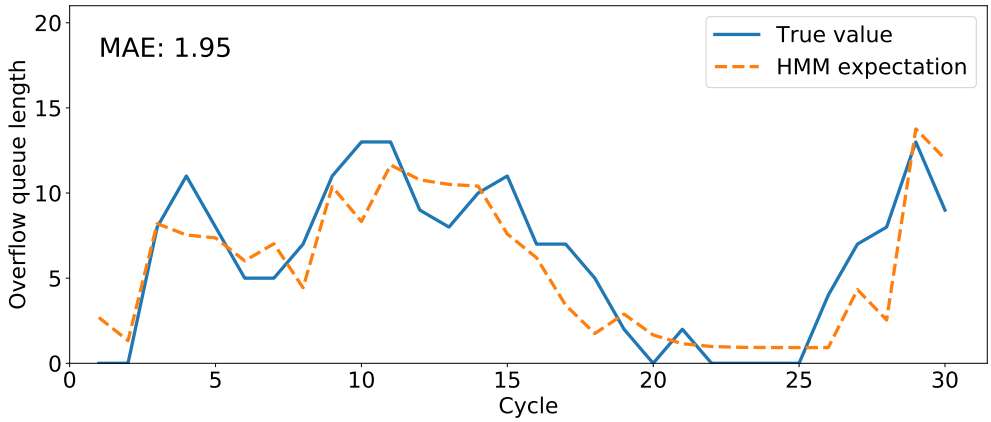
(a)



(b)



(c)



(d)

Figure 2.5: Cycle-by-cycle queue length estimation results with the given parameters using four different methods: (a) naive estimation; (b) expectation conditional on the observation in the current cycle; (c) maximum likelihood estimation (HMM decoding); and (d) expectation conditional on sequential observations.

The results above demonstrate how the algorithms perform when the penetration rate of probe vehicles is 20%. We run the experiments repeatedly to get the average estimation accuracy under different penetration rates. As Figure 2.6 shows, the HMM expectation method outperforms other methods when the penetration rate is low. When the penetration rate is high, both the HMM decoding method and the naive estimation method perform better than the other two methods. In general, the results indicate that considering the correlation of different cycles is beneficial to queue length estimation if the queueing process does not follow the i.i.d. assumption, especially when the penetration rate of probe vehicles is low.

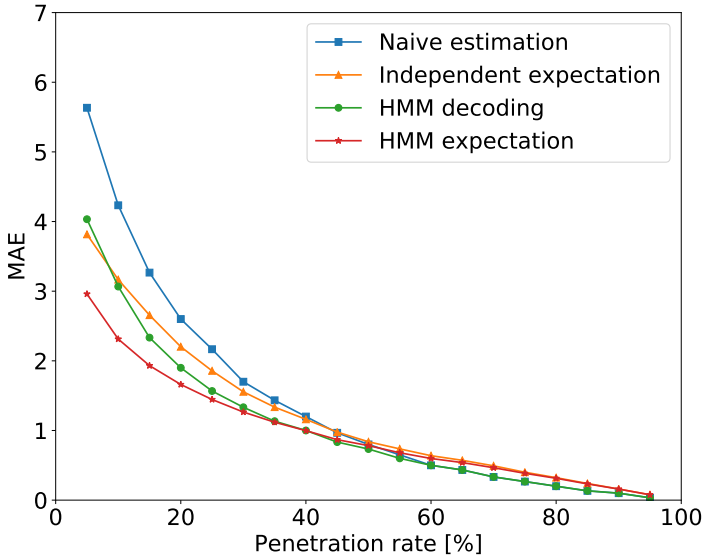


Figure 2.6: The comparison of the proposed methods and two baseline methods.

2.6 Conclusions

In this chapter, we systematically study the cycle-by-cycle queue length estimation using probe vehicle data. We first describe the queue length estimation problem in a probe vehicle environment and present two estimators in the case where queue lengths in different cycles are assumed to be independent and identically distributed. The i.i.d. assumption is appropriate for isolated intersections under light or moderate traffic conditions.

However, in some other scenarios, for example, when there are overflow queues, the i.i.d. assumption does not hold anymore. Such scenarios are often ignored by the existing literature in the context of probe vehicles. Therefore, the second part of this chapter focuses on the estimation of queue lengths in the non-i.i.d. case. We capture the correlation of the queues in different cycles by a hidden Markov model. Based on the HMM, we propose two cycle-by-cycle queue length estimation methods for the non-i.i.d. scenarios. We also compare their estimation accuracy with two baseline methods by numerical experiments. In the numerical experiments, we generate queue sequences using a state-of-the-art overflow queue model and obtain probe vehicle data by random sampling. The results of the experiments indicate that

considering the correlation of different cycles is beneficial for queue length estimation.

Throughout this chapter, we assume all the parameters, such as the penetration rate and queue length distribution, are given to us, as most of the relevant literature does. In the real world, when we implement the queue length estimation methods, the values of the parameters are not available beforehand. Instead, we need to estimate them from historical data. In the next few chapters, we will focus on the estimation of the required parameters.

Chapter 3

Parameter estimation for independent queues: approximate estimation

3.1 Introduction

3.1.1 Background

In chapter 2, we introduced the cycle-by-cycle queue length estimation methods enabled by probe vehicle data. When queue lengths in different cycles are assumed to be i.i.d., the probability theory based methods require the knowledge of the penetration rate of the probe vehicles and the distribution of queue lengths. Even for another category of methods in the existing literature, the traffic flow theory based methods, prior information about the penetration rate and queue length distribution is useful. However, the information is usually not available beforehand in real life.

Some recent studies attempted to estimate the probe vehicle penetration rate and queue length distribution. Comert (2016) derived several estimators of the penetration rate under the assumption of Poisson vehicle arrivals. Zheng and Liu (2017) proposed a maximum likelihood estimation method that can estimate the average arrival rate at signalized intersections. The authors assumed that the vehicle arrivals follow a time-varying Poisson process. However, the imposed Poisson arrival assumption limited its practical applications. Wong and Wong (2019) and Wong et al. (2019b) used the loop detector data and probe vehicle data collected in adjacent links to find the mean penetration rate, assuming there exists a probability distribution describing the penetration rates on adjacent links. Meng et al.

(2017b) also quantified the penetration rate variability based on land use variables using data collected in Hong Kong. The established model can be used to estimate penetration rates of links without detectors. However, the method might not be generalized to other locations. Wong et al. (2019a) proposed an unbiased estimator for the probe vehicle penetration rate. Nevertheless, the method cannot handle the cases when some of the queues are empty. In summary, most of the existing methods impose strong assumptions and can only be applied to a limited range of scenarios.

3.1.2 Contribution and organization of the chapter

In this chapter, we try to estimate the probe vehicle penetration rate and queue length distribution by making use of the stopping positions of probe vehicles. When the traffic is flowing, it is difficult to infer how many regular vehicles are around the probe vehicles. Consequently, it is almost impossible to estimate the penetration rate of the probe vehicles in the traffic. However, when the vehicles are stopping at the intersections, based on the empirical value of the space headway, we can roughly infer the number of vehicles in front of the last probe vehicle. Although the number of vehicles behind the last probe vehicle is still unknown, the incomplete information can still provide us an opportunity to estimate the required parameters. Since the proposed methods in this chapter have few external dependencies compared to the existing methods, they could overcome the limitations of the existing methods and be applied to a broader range of scenarios. The methods have been validated by both simulation and real-world data, showing good accuracy.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the general methodologies to estimate the required parameters. Specifically, we demonstrate how to obtain an approximated queue length distribution from the aggregated stopping positions of probe vehicles. We also classify queues into observable queues and hidden queues and show that the penetration rate can be estimated once we have an estimator of the total queue length. In Sections 3.3 and 3.4, we propose several estimators for observable queues and

hidden queues, respectively. In Section 3.5, we revisit the methodologies of estimating the penetration rate and queue length distribution and propose two general methods by making use of the results in Sections 3.3 and 3.4. We validate the proposed methods using simulation data and real-world probe vehicle data in Section 3.6. Finally, we summarize this chapter and discuss the limitations of the proposed methods in Section 3.7.

3.2 Methodology

As discussed in Chapter 2, if we assume the queue lengths in different cycles are i.i.d., the patterns of the observations from probe vehicles are governed by the penetration rate of probe vehicles and the queue length distribution. In this section, we present a general methodology for estimating the two governing parameters from historical data.

3.2.1 Approximate estimation of the queue length distribution

Suppose the trajectory data of the probe vehicles are collected for C traffic signal cycles. Denote the number of queues of length j in all the cycles by C_j. Then, according to the definition of the queue length distribution, we have

$$\pi_j = \frac{\mathbb{E}\left(C_j\right)}{C}, \forall j = 0, 1, \ldots, L_{max}. \tag{3.1}$$

Assuming the traffic signal timing is available, the value of C can be easily obtained. Therefore, if we can infer the value of $\mathbb{E}\left(C_j\right)$, then we can get an estimate of the queue length distribution.

One key insight is that the queue length distribution can be obtained from the aggregated data of vehicle queueing locations. Suppose the distribution of the stopping positions of all the vehicles (both probe vehicles and regular vehicles) is given. As illustrated by the left and middle diagrams in Figure 3.1, if the maximum queue length is 6, then the count of stopping vehicles at position 6 (the sixth stopping position behind the stop bar) is equal to the count

of queues of length 6, namely, C_6. Similarly, the count of stopping vehicles at position 5 is equal to the total count of queues of length 5 or 6, that is, $C_5 + C_6$; the count of stopping vehicles at position 4 is equal to the total count of queues of length 4, 5, or 6, so on and so forth. As a result, the distribution of the stopping positions has a decreasing trend with respect to the distance to the stop bar.

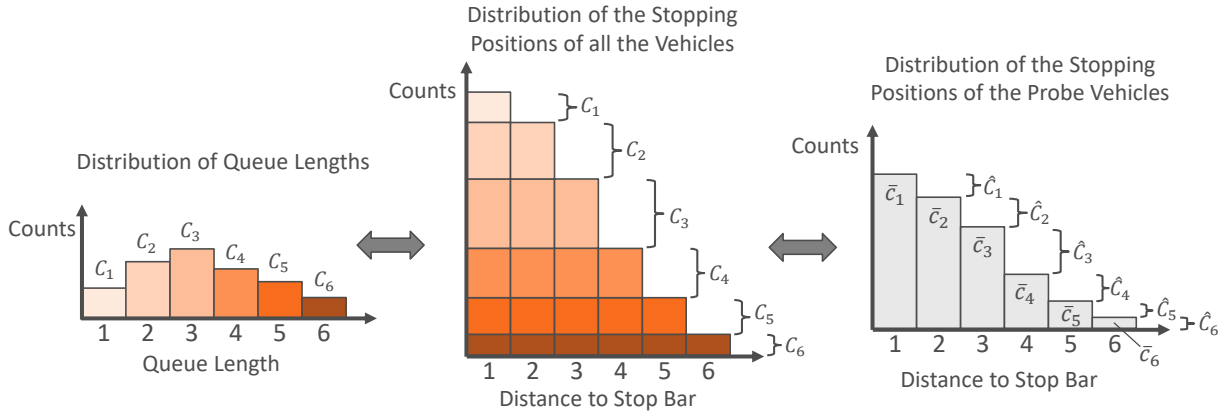


Figure 3.1: The relationship between the distributions of queue lengths and stopping positions.

Based on the reasoning above, if the distribution of the stopping positions of all the vehicles is given, $\mathbb{E}(C_j)$ can be approximated by the difference between the counts of stopping vehicles at position j and position $j+1$. Nevertheless, in reality, only the stopping positions of the probe vehicles are observable. Since the probe vehicles are assumed to be homogeneously mixed with other vehicles, the histogram of the stopping positions of the probe vehicles is a scaled-down version of the histogram of the stopping positions of all the vehicles. Therefore, as illustrated by the middle and right diagrams in Figure 3.1, $\hat{C}_j$, the difference between $\bar{c}_j$, the count of stopping probe vehicles at position j, and $\bar{c}_{j+1}$, the count of stopping probe vehicles at position $j+1$, can be used to approximate $p\mathbb{E}(C_j)$. Thus, we can estimate the queue length distribution by

$$\pi_j = \frac{p\mathbb{E}\left(C_j\right)}{pC} \approx \frac{\hat{C}_j}{pC}, \forall j = 1, 2, \ldots, L_{max}. \tag{3.2}$$

It is worth noting that due to randomness, the distribution of the stopping positions of the probe vehicles might not always follow the decreasing trend. When the decreasing trend is violated, the distribution cannot be directly used to calculate $\hat{C}_j$, because otherwise some of the values will be negative. In this case, $\hat{C}_j$ can be calculated by solving the following optimization problem.

$$\text{minimize} \quad \sum_{k=1}^{L_{max}} \omega_k \left(\sum_{j=k}^{L_{max}} \hat{C}_j - \bar{c}_k \right)^2 \tag{3.3}$$

$$\text{subject to} \quad \hat{C}_j \geq 0, \forall l = 1, 2, ...L_{max}. \tag{3.4}$$

The objective of the optimization problem is to minimize the difference between the ideal counts of stopping probe vehicles $\sum_{j=k}^{L_{max}} \hat{C}_j$ and the observed counts $\bar{c}_k$ for each location k, weighted by a coefficient ω_k. The constraints ensure the non-negativity of $\hat{C}_j$. From another perspective, the optimization problem tries to modify the distribution of the stopping positions of the probe vehicles to the least extent, so that non-negative $\hat{C}_j$ could be obtained.

3.2.2 Approximate estimation of the penetration rate

Denote the total number of probe vehicles in all the queues by Q^{probe}. Since the value of Q^{probe} can be easily obtained from the probe vehicle data, if we can get an estimate of the total queue length $\hat{Q}^{all}$ or express it as a function of p, the penetration rate of the probe vehicles can be easily obtained from the following equation

$$p = \frac{Q^{probe}}{\hat{Q}^{all}}. \tag{3.5}$$

To estimate the total queue length, we classify the queues into two categories: observable queues and hidden queues. Observable queues refer to the queues where there is at least one probe vehicle; hidden queues refer to the queues without any probe vehicles. Let s_i and t_i denote the positions of the first and the last probe vehicles in the ith cycle, respectively, if the queue is observable. Figure 3.2 shows an example of the queues in a probe vehicle

environment. In this example, l_1, l_2, l_4, l_5, l_7, and l_9 are (partially) observable because of the probe vehicles in the queues. l_3, l_6, and l_8 are hidden because there are no probe vehicles. Denote the total length of the observable queues and the total length of the hidden queues by Q^{obs} and Q^{hid}, respectively. In Figure 3.2, $Q^{obs} = l_1 + l_2 + l_4 + l_5 + l_7 + l_9 = 30$ and $Q^{hid} = l_3 + l_6 + l_8 = 7$.

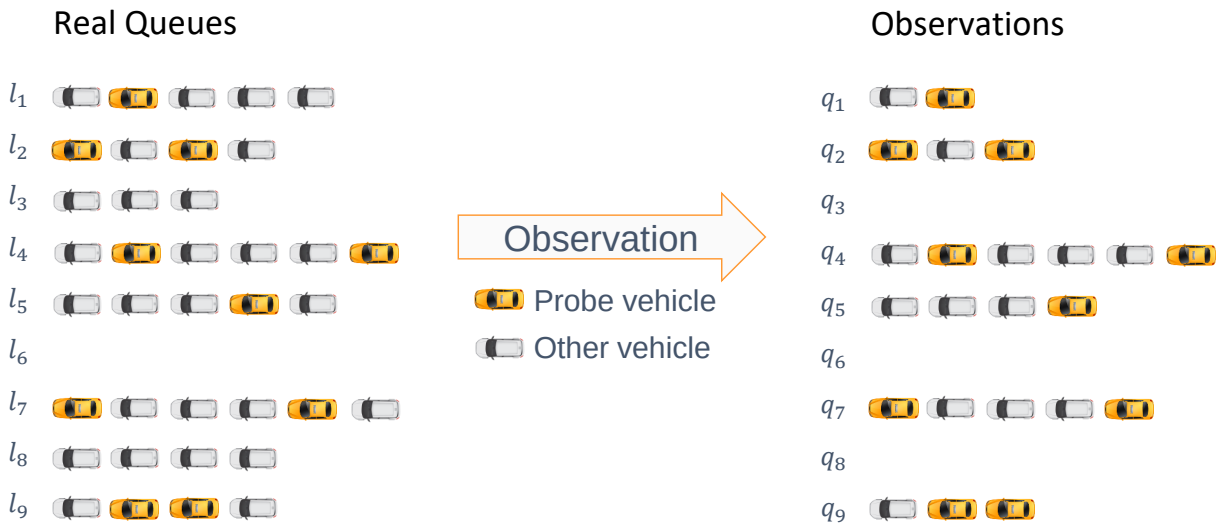


Figure 3.2: Observation process.

Estimating the total queue length is equivalent to estimating the sum of Q^{obs} and Q^{hid}. In the following two sections, we will discuss how to estimate the observable queues and hidden queues, respectively. Then, we will revisit the methodologies of estimating the penetration rate and queue length distribution in Section 3.5.

3.3 Estimation of observable queues

Q^{obs} can be estimated through two general approaches. Estimator 1, 2, and 3 are based on the fact that the queueing probe vehicles are expected to segregate the regular vehicles equally. These estimators only require the number of queueing probe vehicles and their stopping positions in each cycle, which can be easily obtained from the trajectory data. Therefore, the estimators give constant values. By contrast, estimator 4 is based on Bayes' theorem, which relies on the penetration rate p. Thus, estimator 4 is a function of p.

3.3.1 Estimator 1 using the first probe vehicles in the queues

Theorem 1.

Given that $n_i \geq 1$ in the ith cycle,

$$\mathbb{E}(l_i \mid n_i) = \mathbb{E}(s_i \mid n_i)(n_i + 1) - 1. \tag{3.6}$$

The proof is in Appendix A.

Theorem 1 states that given the number of probe vehicles in an observable queue, the expected queue length can be obtained from the expected stopping position of the first probe vehicle. Based on Theorem 1, given the number of probe vehicles in each cycle, the expected total length of the observable queues can be expressed as

$$\sum_{i:n_i \neq 0} \mathbb{E}(l_i \mid n_i) = \sum_{i:n_i \neq 0} \left(\mathbb{E}\left(s_i \mid n_i\right)(n_i + 1) - 1\right). \tag{3.7}$$

$$= \sum_{i:n_i \neq 0} \mathbb{E}\left(s_i \mid n_i\right)(n_i + 1) - \sum_{i:n_i \neq 0} 1 \tag{3.8}$$

$$= \sum_{j=1}^{L_{max}} \sum_{i:n_i = j} \mathbb{E}\left(s_i \mid n_i = j\right)(j + 1) - \sum_{i:n_i \neq 0} 1 \tag{3.9}$$

$$= \sum_{j=1}^{L_{max}} (j + 1) \sum_{i:n_i = j} \mathbb{E}\left(s_i \mid n_i = j\right) - \sum_{i:n_i \neq 0} 1. \tag{3.10}$$

Therefore, given the position of the first stopping probe vehicle s_i in the ith cycle, $\forall i \in \{1, 2, , \ldots, C\}$, by replacing the expected value $\mathbb{E}\left(s_i \mid n_i = j\right)$ by the sample mean $\frac{\sum_{i:n_i=j} s_i}{\sum_{i:n_i=j} 1}$, $\forall j \geq 1$, Q^{obs} can be estimated by

$$\hat{Q}_1^{obs} = \sum_{j=1}^{L_{max}} (j + 1) \sum_{i:n_i = j} s_i - \sum_{i:n_i \neq 0} 1 \tag{3.11}$$

$$= \sum_{j=1}^{L_{max}} \sum_{i:n_i = j} s_i\,(j + 1) - \sum_{i:n_i \neq 0} 1 \tag{3.12}$$

$$= \sum_{i:n_i \neq 0} s_i \left(n_i + 1\right) - \sum_{i:n_i \neq 0} 1 \tag{3.13}$$

$$= \sum_{i:n_i \neq 0} \left(s_i \left(n_i + 1\right) - 1\right). \tag{3.14}$$

3.3.2 Estimator 2 using the last probe vehicles in the queues

Theorem 2.

Given that $n_i \geq 1$ in the ith cycle,

$$\mathbb{E}(l_i \mid n_i) = \mathbb{E}(t_i \mid n_i)\frac{n_i + 1}{n_i} - 1. \tag{3.15}$$

The proof is in Appendix A.

Theorem 2 states that given the number of probe vehicles in an observable queue, the expected queue length can be obtained from the expected stopping position of the last probe vehicle. Based on Theorem 2, given the number of probe vehicles in each cycle, the expected total length of observable queues can be expressed as

$$\sum_{i:n_i \neq 0} \mathbb{E}(l_i \mid n_i) = \sum_{i:n_i \neq 0} \left(\mathbb{E}(t_i \mid n_i)\frac{n_i + 1}{n_i} - 1\right) \tag{3.16}$$

$$= \sum_{j=1}^{L_{max}} \left(\frac{j+1}{j}\right) \sum_{i:n_i = j} \mathbb{E}\left(t_i \mid n_i = j\right) - \sum_{i:n_i \neq 0} 1. \tag{3.17}$$

Following the similar derivations as estimator 1, given the position of the last stopping probe vehicle t_i in the ith cycle, $\forall i \in \{1, 2, \ldots, C\}$, by substituting the sample mean $\frac{\sum_{i:n_i=j} t_i}{\sum_{i:n_i=j} 1}$ for the expected value $\mathbb{E}\left(t_i \mid n_i = j\right), \forall j \geq 1$, Q^{obs} can be estimated by

$$\hat{Q}_2^{obs} = \sum_{i:n_i \neq 0} \left(t_i \frac{n_i + 1}{n_i} - 1\right). \tag{3.18}$$

3.3.3 Estimator 3 using the first and the last probe vehicles in the queues

Theorem 3.

Given that $n_i \geq 1$ in the ith cycle,

$$\mathbb{E}(l_i \mid n_i) = \mathbb{E}(s_i \mid n_i) + \mathbb{E}(t_i \mid n_i) - 1, \tag{3.19}$$

$$\mathbb{E}(l_i \mid n_i \geq 1) = \mathbb{E}(s_i \mid n_i \geq 1) + \mathbb{E}(t_i \mid n_i \geq 1) - 1. \tag{3.20}$$

The proof is in Appendix A.

Theorem 3 states that given the number of probe vehicles in an observable queue, the expected queue length can be obtained from the expected stopping positions of the first and the last probe vehicles. Based on Theorem 3, given the number of probe vehicles in each cycle, the expected total length of the observable queues can be expressed as

$$\sum_{i:n_i \neq 0} \mathbb{E}(l_i \mid n_i) = \sum_{i:n_i \neq 0} \left(\mathbb{E}(s_i \mid n_i) + \mathbb{E}(t_i \mid n_i) - 1\right). \tag{3.21}$$

$$= \sum_{j=1}^{L_{max}} \sum_{i:n_i=j} \left(\mathbb{E}\left(s_i \mid n_i = j\right) + \mathbb{E}\left(t_i \mid n_i = j\right) - 1\right) \tag{3.22}$$

Therefore, by substituting the sample means $\frac{\sum_{i:n_i=j} s_i}{\sum_{i:n_i=j} 1}$ and $\frac{\sum_{i:n_i=j} t_i}{\sum_{i:n_i=j} 1}$ for the expected values $\mathbb{E}(s_i \mid n_i = j)$ and $\mathbb{E}(t_i \mid n_i = j), \forall j \geq 1$, respectively, Q^{obs} can be estimated by

$$\hat{Q}_3^{obs} = \sum_{i:n_i \neq 0} \left(s_i + t_i - 1\right). \tag{3.23}$$

The mechanism behind $\hat{Q}_3^{obs}$ is intuitive. Figure 3.3 shows an example, where the queue in the kth cycle is the reverse of the queue in the jth cycle. It implies that the number of vehicles behind the last probe vehicle in the jth cycle is equal to the number of vehicles in front of the first probe vehicle in the kth cycle. Because of the symmetry, these two queues have the same probability of occurring. Therefore, even though the number of vehicles

behind the last probe vehicle in a cycle is unknown, as long as the sample size is sufficient, the missing number could be compensated by the number of vehicles in front of the first probe vehicle in another cycle. Essentially, $\hat{Q}_3^{obs}$ is obtained by summing up the position of the last probe vehicle t_i and the number of vehicles in front of the first probe vehicle $s_i - 1$, which can be regarded as a compensation of the missing vehicles in the rear.

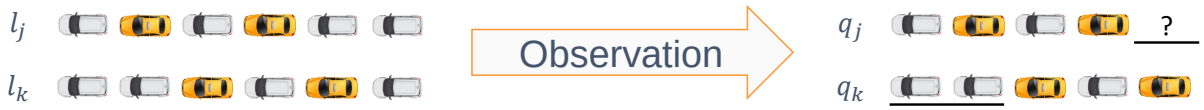


Figure 3.3: The missing part compensated by another queue.

3.3.4 Estimator 4 based on Bayes' theorem

Given all the observed partial queues, according to equation (2.2), the conditional expectation of the total length of the observable queues can be expressed as

$$\sum_{i:n_i\neq 0} \mathbb{E}(l_i \mid q_i) = \sum_{i:n_i\neq 0} \sum_{j=|q_i|}^{L_{max}} \frac{\pi_j}{\sum_{k=|q_i|}^{L_{max}} \pi_k \left(1-p\right)^{k-j}} j. \tag{3.24}$$

Substituting equation (3.2) into equation (3.24) gives an estimate of Q^{obs}

$$\hat{Q}_4^{obs}(p) = \sum_{i:n_i\neq 0} \sum_{j=|q_i|}^{L_{max}} \frac{\hat{C}_j}{\sum_{k=|q_i|}^{L_{max}} \hat{C}_k \left(1-p\right)^{k-j}} j, \tag{3.25}$$

which is a function of the penetration rate p.

3.4 Estimation of hidden queues

After estimating Q^{obs}, the following question is how to estimate Q^{hid}, as there is no probe vehicle in the corresponding cycles. Fortunately, the fact that no probe vehicle is in the queues also contains information. In this section, two estimators of Q^{hid} will be presented. Similar to $\hat{Q}_4^{obs}(p)$, estimator 1 of Q^{hid} applies Bayes' theorem to the hidden queues directly.

Estimator 2 utilizes the ratio between the probability of being observable and the probability of being hidden for each queue, to estimate the total length of the hidden queues.

3.4.1 Estimator 1 based on Bayes' theorem

Similar to equation (3.24), given the fact that no probe vehicle is observed in the hidden queues, the expected total length of the hidden queues can be expressed as

$$\sum_{i:n_i=0} \mathbb{E}(l_i \mid q_i) = \sum_{i:n_i=0} \sum_{j=0}^{L_{max}} \frac{\pi_j}{\sum_{k=0}^{L_{max}} \pi_k (1-p)^{k-j}} j. \tag{3.26}$$

Therefore, an estimator of Q^{hid} can be given by

$$\hat{Q}_1^{hid}(p) = \sum_{i:n_i=0} \sum_{j=0}^{L_{max}} \frac{\hat{C}_j}{\sum_{k=0}^{L_{max}} \hat{C}_k (1-p)^{k-j}} j. \tag{3.27}$$

Please note that different from equation (3.25), the summation over j in equation (3.27) starts from 0. Here shows how to find $\hat{C}_0$, an estimate of $p\mathbb{E}(C_0)$.

In all the queues, the expected number of queues of length 0 is

$$\mathbb{E}(C_0) = C - \sum_{j=1}^{L_{max}} \mathbb{E}(C_j). \tag{3.28}$$

Therefore, multiplying p on the two sides of the equation gives

$$p\mathbb{E}(C_0) = pC - \sum_{j=1}^{L_{max}} p\mathbb{E}(C_j). \tag{3.29}$$

$\hat{C}_0$, an estimate of $p\mathbb{E}(C_0)$, can be easily given by

$$\hat{C}_0 = pC - \sum_{j=1}^{L_{max}} \hat{C}_j. \tag{3.30}$$

Now, all the parameters except p on the right-hand side of equation (3.27) can be calculated.

Therefore, $\hat{Q}_1^{hid}(p)$ is a function of only p.

3.4.2 Estimator 2 using the probabilities of being observed and being hidden

Define a binary variable X_i^j to indicate if the queue length in the ith cycle is j, that is,

$$X_i^j = \begin{cases} 1, & l_i = j \\ 0, & l_i \neq j \end{cases}, \tag{3.31}$$

where $j = 0, 1, \ldots, L_{max}$. Obviously, $C_j = \sum_{i=1}^{C} X_i^j$.

Among the observable queues, $\forall j = 1, 2, \ldots, L_{max}$, the expected number of queues of length j can be expressed as

$$\sum_{i:n_i \neq 0} \mathbb{E}\left(X_i^j \mid q_i\right) = \sum_{i:n_i \neq 0} \left(P(X_i^j = 1 \mid q_i) \cdot 1 + P(X_i^j = 0 \mid q_i) \cdot 0\right) \tag{3.32}$$

$$= \sum_{i:n_i \neq 0} \left(P(l_i = j \mid q_i) \cdot 1 + P(l_i \neq j \mid q_i) \cdot 0\right) \tag{3.33}$$

$$= \sum_{i:n_i \neq 0} P(l_i = j \mid q_i). \tag{3.34}$$

For a queue of length j, the probability of being hidden (without any probe vehicle) is $(1-p)^j$; the probability of being observable (with at least one probe vehicle) is $1-(1-p)^j$. Therefore, the expected total length of the hidden queues can be estimated by

$$\sum_{j=1}^{L_{max}} \left(\frac{(1-p)^j}{1-(1-p)^j} \sum_{i:n_i \neq 0} \mathbb{E}\left(X_i^j \mid q_i\right) \right) j = \sum_{j=1}^{L_{max}} \frac{(1-p)^j}{1-(1-p)^j} \sum_{i:n_i \neq 0} P(l_i = j \mid q_i) j \tag{3.35}$$

$$= \sum_{i:n_i \neq 0} \sum_{j=|q_i|}^{L_{max}} \frac{(1-p)^j}{1-(1-p)^j} \frac{\pi_j}{\sum_{k=|q_i|}^{L_{max}} \pi_k (1-p)^{k-j}} j \tag{3.36}$$

Then, an estimator of Q^{hid}, the total length of the hidden queues, can be defined as

$$\hat{Q}_2^{hid}(p) = \sum_{i:n_i \neq 0} \sum_{j=|q_i|}^{L_{max}} \frac{(1-p)^j}{1-(1-p)^j} \frac{\hat{C}_j}{\sum_{k=|q_i|}^{L_{max}} \hat{C}_k (1-p)^{k-j}} j. \tag{3.37}$$

3.5 Estimation of the parameters

In this section, we propose two different methods for penetration rate estimation. Extending the methodology presented in Section 3.2.2, we estimate p by establishing and solving a single-variable equation. Method 1 is based upon the equivalence between the different estimators. Method 2 exploits the fact that the proportion of probe vehicles in the queues is approximately equal to the penetration rate. Once p is estimated, we can easily estimate the queue length distribution using equation (3.2).

3.5.1 Method 1

When estimating Q^{obs}, estimator 1, 2, and 3 can generate constant results, whereas estimator 4 is a function of p. Since the four estimators are of the same variable Q^{obs}, it is intuitive to establish the following single-variable equation

$$\hat{Q}_i^{obs} = \hat{Q}_4^{obs}(p), \forall i = 1, 2, 3. \tag{3.38}$$

Solving the equation will yield an estimate of the penetration rate p. Similarly, when estimating Q_{hid}, both estimator 1 and estimator 2 are functions of p. Therefore, another single-variable equation can be given by

$$\hat{Q}_1^{hid}(p) = \hat{Q}_2^{hid}(p). \tag{3.39}$$

A more general formulation of this method can be expressed as follows.

$$\hat{Q}_i^{obs}(p) + \hat{Q}_j^{hid}(p) = \hat{Q}_m^{obs}(p) + \hat{Q}_n^{hid}(p). \tag{3.40}$$

As long as it is an equation with a single unknown variable p, solving it will give an estimate of the penetration rate. Both the left-hand side and the right-hand side of equation (3.40) can be regarded as estimators of the total queue length.

3.5.2 Method 2

Another way to establish a single-variable equation for p is shown by equation (3.41).

$$\frac{Q^{probe}}{\hat{Q}_i^{obs}(p) + \hat{Q}_j^{hid}(p)} = p, \forall i = 1, 2, 3, 4, \forall j = 1, 2, \tag{3.41}$$

The left-hand side of equation (3.41) can be interpreted as an estimate of the frequency of observing a probe vehicle in the queues. The right-hand side is the probability of observing a probe vehicle. When the sample size is large enough, the two sides will be very close to each other. Similarly, solving the equation yields an estimate of p.

In practice, it is usually hard to find the analytical solutions of equations (3.38), (3.39), (3.40), or (3.41). Instead, an iterative algorithm should be applied. One may search p from an upper bound to 0 with a small step size until the difference between the left-hand side and the right-hand side reaches certain stopping criteria. The upper bound can be taken as $\frac{Q^{probe}}{\sum_{i=1}^{L_{max}} |q_i|}$ since it is an overestimate of the penetration rate p.

3.6 Validation and evaluation

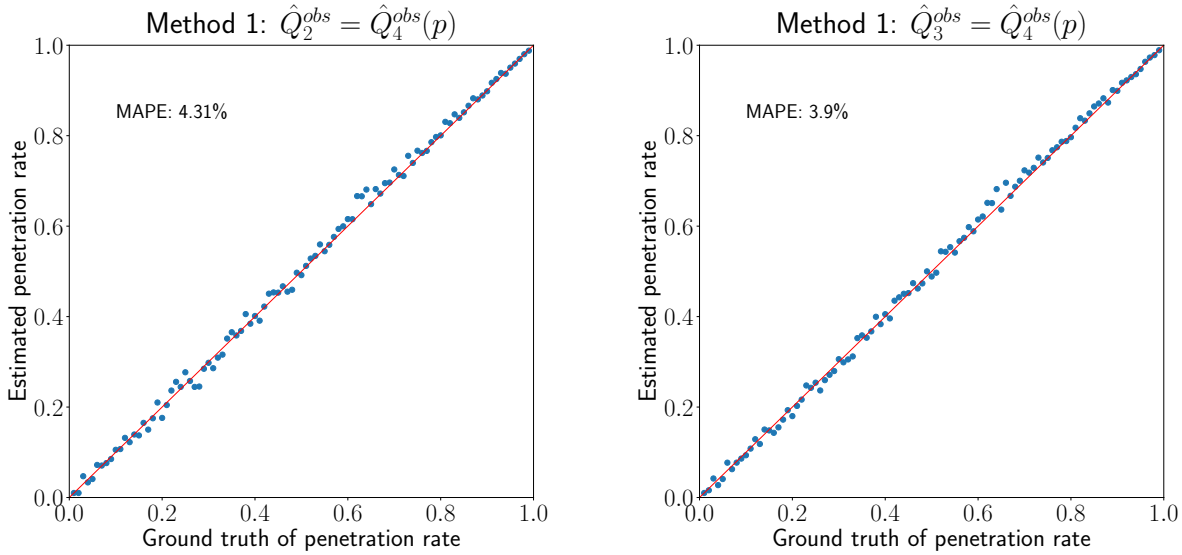
3.6.1 Simulation

This test mainly focuses on the estimation of the penetration rate. For demonstration purposes, the testing dataset is generated by a simulation of Poisson processes, although any other stochastic process can also be applied. In order to study the robustness of the proposed

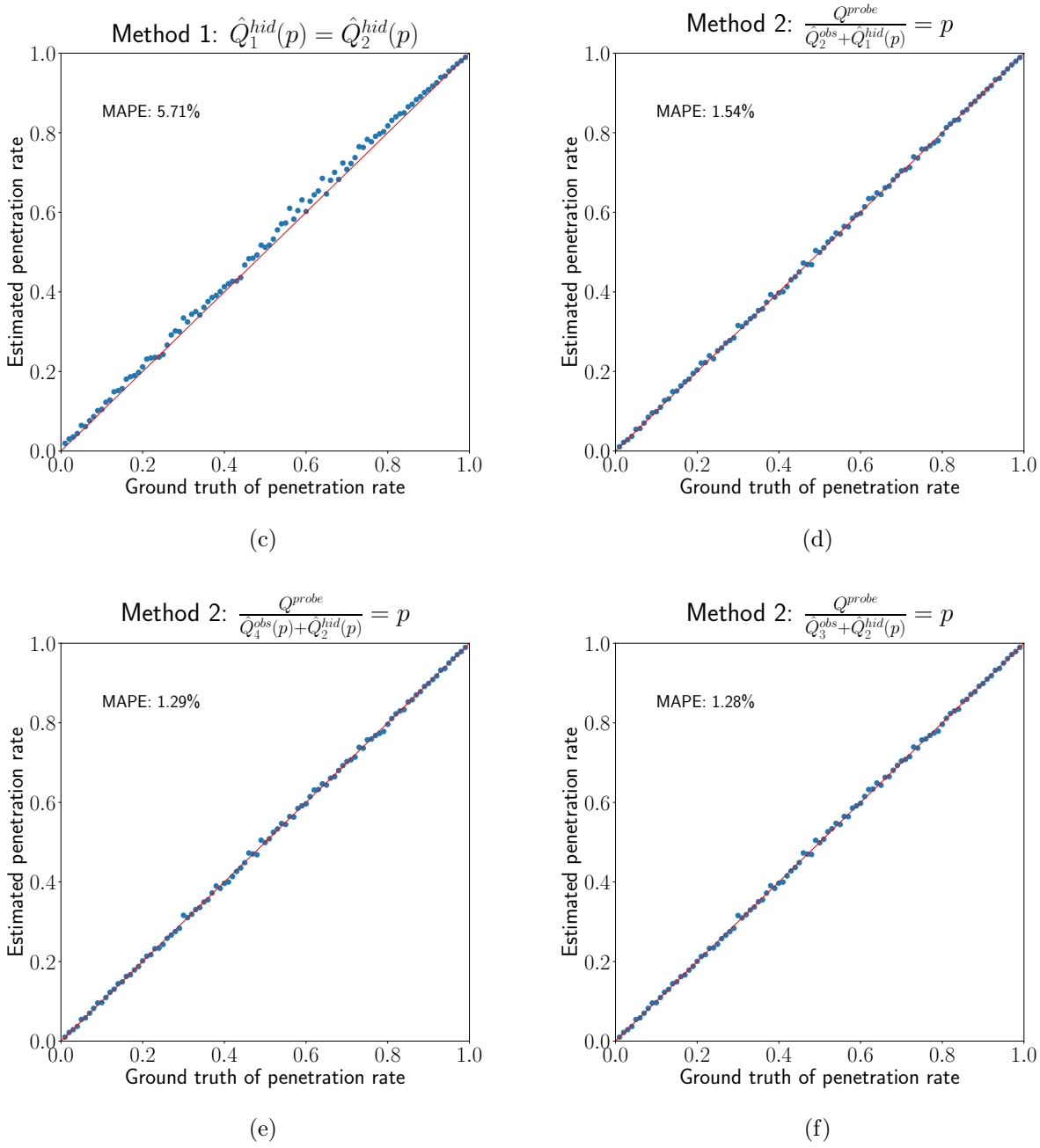
methods, in each test, we test the methods under different penetration rates, ranging from 1% to 99%.

The comparison of different methods

Figure 3.4 shows the results of penetration rate estimation using six different submethods introduced in Section 3.5. We generate the 1000-cycle simulation data by a Poisson process with an average arrival rate of 10 during the red phase. The horizontal axes represent the ground truth of the penetration rates. The vertical axes represent the estimated values. The used measure of the estimation accuracy is the mean absolute percentage error (MAPE). As Figure 3.4 shows, the dots in blue are very close to the diagonals, which implies that the methods can estimate the penetration rate very accurately. The results show that the higher the penetration rate is, the better the estimation results tend to be. It is intuitive because when the penetration rate is very low, only a tiny portion of vehicles can be observed. By contrast, if the penetration rate is very close to 100%, there will be little missing information, and the estimation results would be more accurate.



(a) (b)



(c) (d)

(e) (f)

Figure 3.4: The results of penetration rate estimation using different methods.

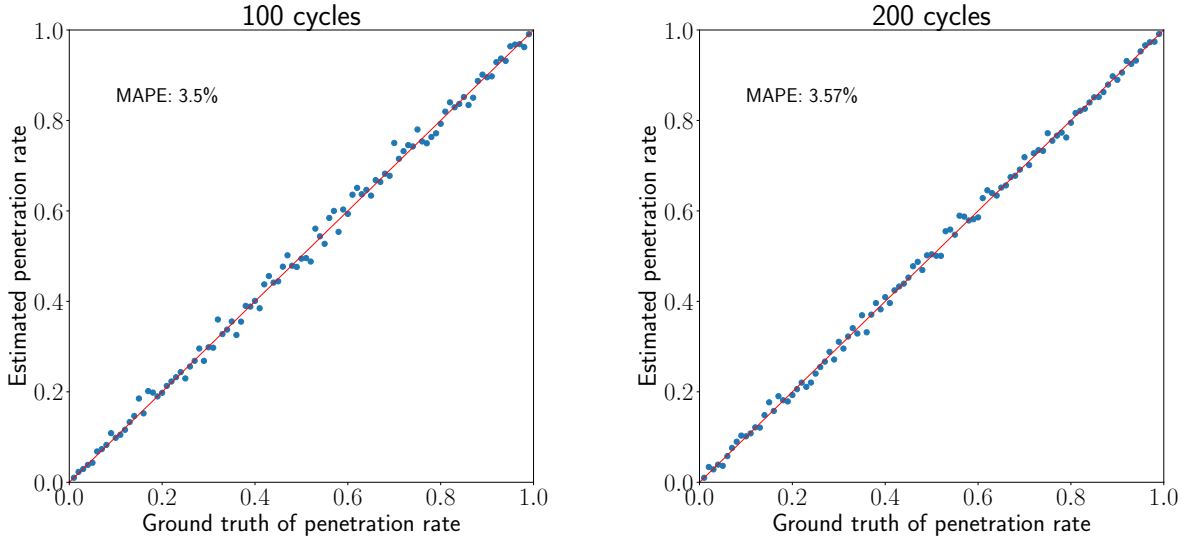
In general, method 2 outperforms method 1. To better understand the mechanism behind method 2, we define an inverse proportional function

$$f(x) = \frac{M}{x}, \tag{3.42}$$

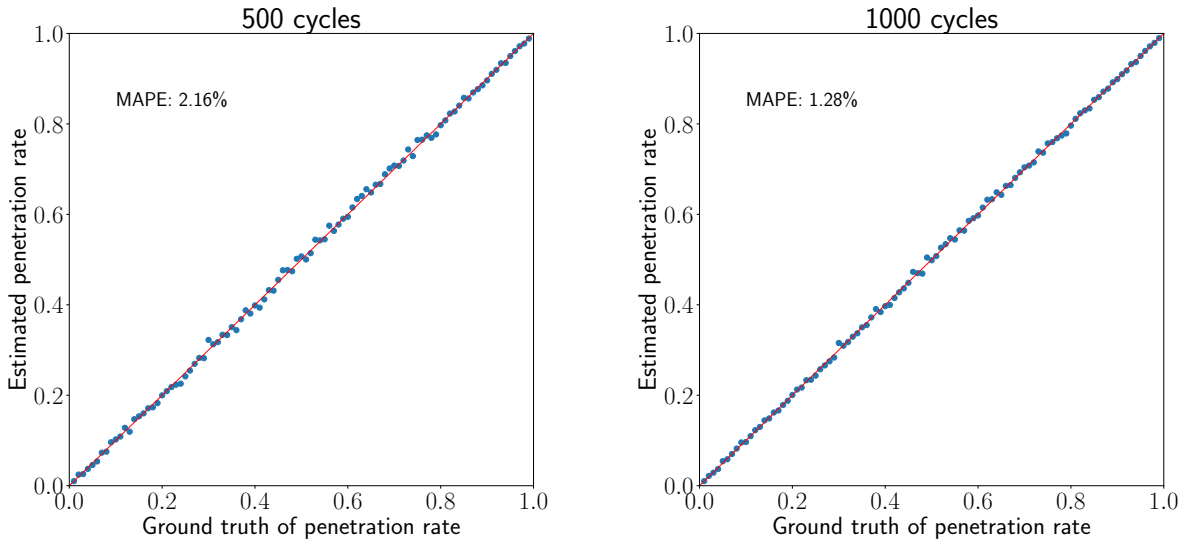
where M is a positive constant. When $x \gg \sqrt{M}$, the absolute value of the derivative is $|f'(x)| = \frac{M}{x^2} \ll 1$. In method 2, as equation (3.41) shows, the denominator of the left-hand side is $\hat{Q}_i^{obs}(p) + \hat{Q}_j^{hid}(p)$, which is much larger than $\sqrt{Q^{probe}}$. Therefore, due to the property of the inverse proportional function, the error in $\hat{Q}_i^{obs}(p) + \hat{Q}_j^{hid}(p)$ only results in an error of p that is orders of magnitude smaller.

The effect of sample size

In order to demonstrate the impact of sample size on the estimation accuracy, the data of 100 cycles, 200 cycles, 500 cycles, and 1000 cycles are used in four rounds of tests, respectively. The submethod $\frac{Q^{probe}}{\hat{Q}_3^{obs}(p)+\hat{Q}_2^{hid}(p)} = p$ is applied. The results in Figure 3.5 show that better results can be obtained when the sample size is larger.



(a) (b)



(c) (d)

Figure 3.5: The results of penetration rate estimation with different sample sizes.

The effect of the arrival rate during the red phase

To study the impact of the arrival rate on the estimation accuracy, we apply the same submethod to four different Poisson processes, of which the average arrival rates in the red phase are 3, 5, 10, and 15, respectively. In each test, 1000 cycles of data are used. The results in Figure 3.6 show that the larger the arrival rate is, the more accurate the estimation tends to be. The reason is that a higher arrival rate implies more observations of the probe vehicles, which can generally improve the estimation accuracy.

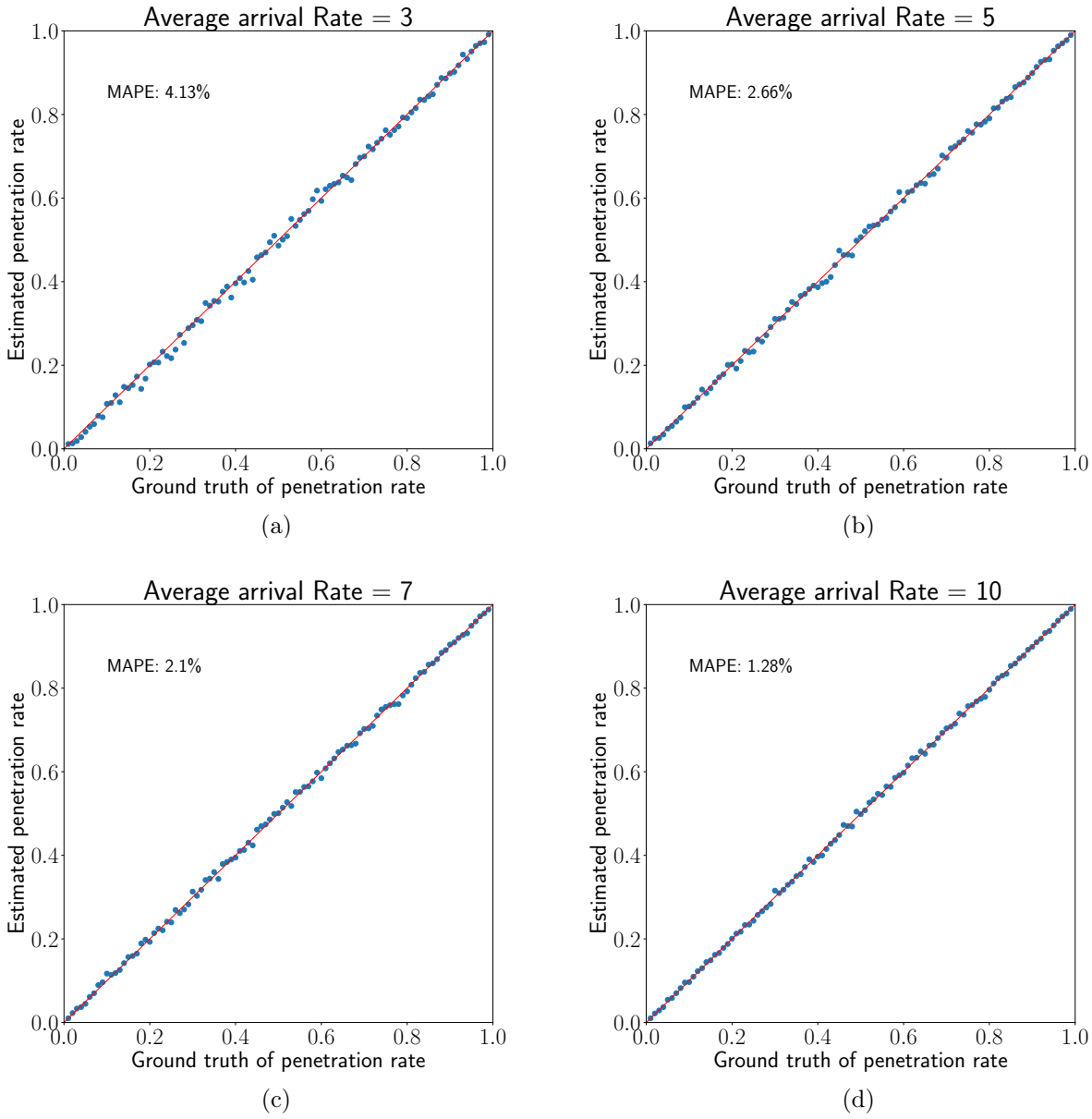


Figure 3.6: The results of penetration rate estimation with different arrival rates.

Results of queue length distribution

To investigate the estimation accuracy of the queue length distribution, we fix the arrival rate to 5 and use 1000 cycles of data. We use the Hellinger distance to measure the dissimilarity between the estimated distribution and the true distribution. For discrete probability distributions $F = (f_1, f_2, \dots f_k)$ and $G = (g_1, g_2, \dots, g_k)$ defined on the same probability

space, the Hellinger distance between F and G is defined as

$$H(F,G) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{f_i}-\sqrt{g_i}\right)^2}. \tag{3.43}$$

Figure 3.7 shows the comparison of the true distribution and estimated distribution under four different penetration rates. In general, with a higher penetration rate, we can get a more accurate estimate of the queue length distribution. More results will be presented in Section 4.4 when we compare the approximate estimator and the maximum likelihood estimator.

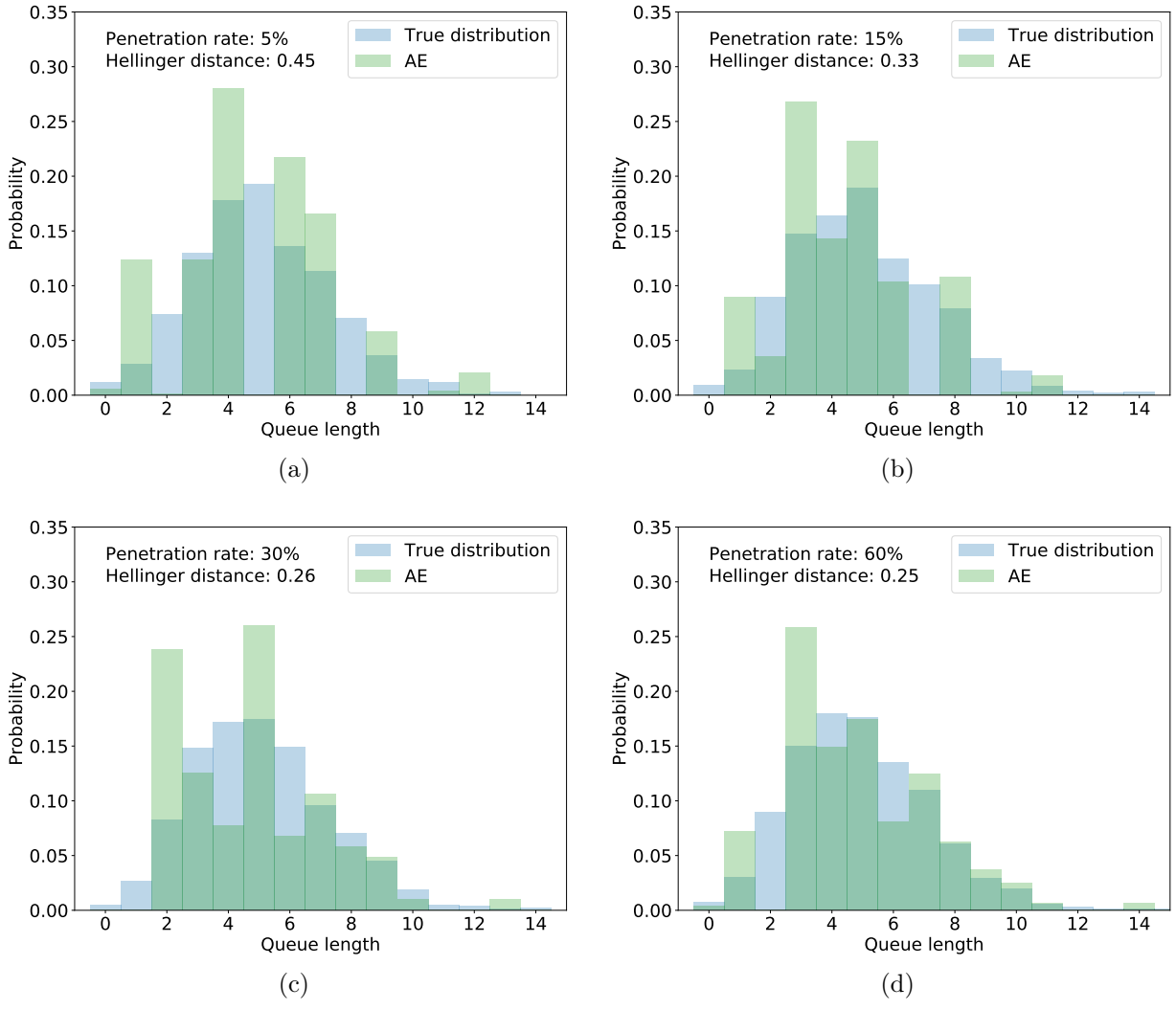


Figure 3.7: Estimation results of queue length distributions under different probe vehicle penetration rates: (a) 5%; (b) 15%; (c) 30%; and (d) 60%.

3.6.2 Real-world data

We also test the proposed methods using real-world probe vehicle data. The probe vehicle data were collected by DiDi Chuxing from the vehicles offering its ride-hailing services in an area in Suzhou, Jiangsu Province, China, shown in Figure 3.8. The data of the 15 workdays from May 8, 2018, to May 28, 2018, are used for validation. The GPS trajectories of the DiDi vehicles in the selected area are mapped onto the transportation network by a map matching algorithm (Newson and Krumm, 2009). For each movement and each one-hour time slot, the "snapshots" of the trajectory data are taken to extract the observed partial queues. Due to the accuracy of the trajectory data, the average space headway for the queueing vehicles could not be easily estimated. Therefore, its value is empirically set to 7.5 m/veh. For the movements with multiple lanes, since the accuracy of the trajectory data cannot reach the lane level, the stopping vehicles are randomly assigned to the different lanes. The random assignment process is repeated 50 times to get an average estimate.

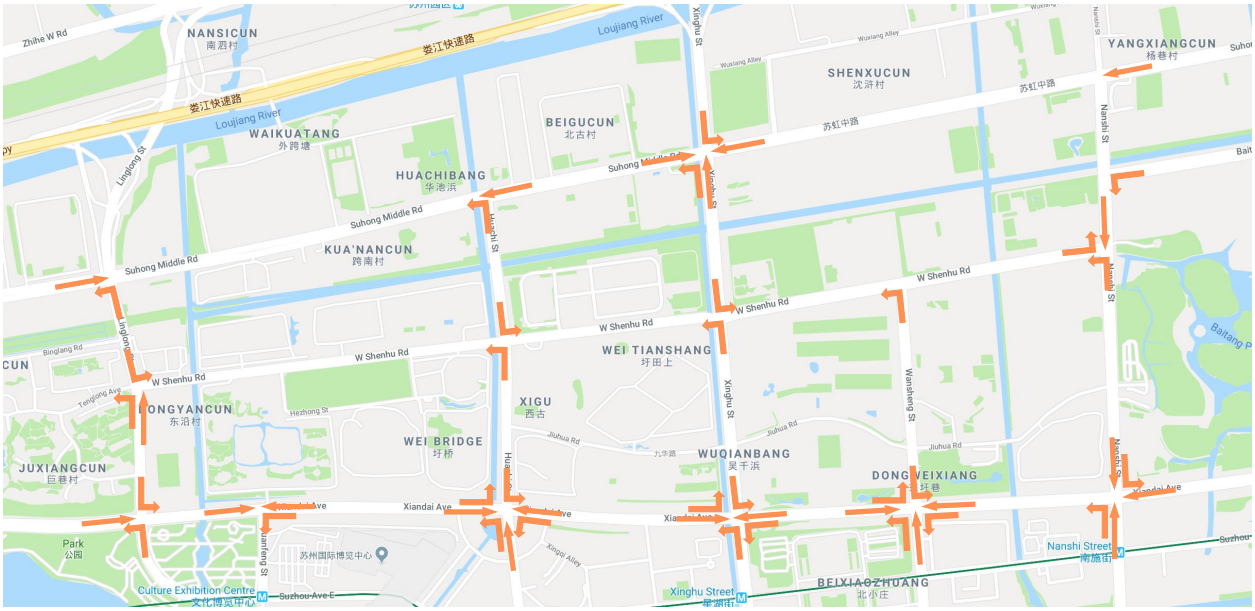


Figure 3.8: The studied movements in Suzhou.

The methodology in this chapter does not apply to the right-turn movements, as there might not be queues. Also, due to the accuracy of the data, it is almost impossible to deal

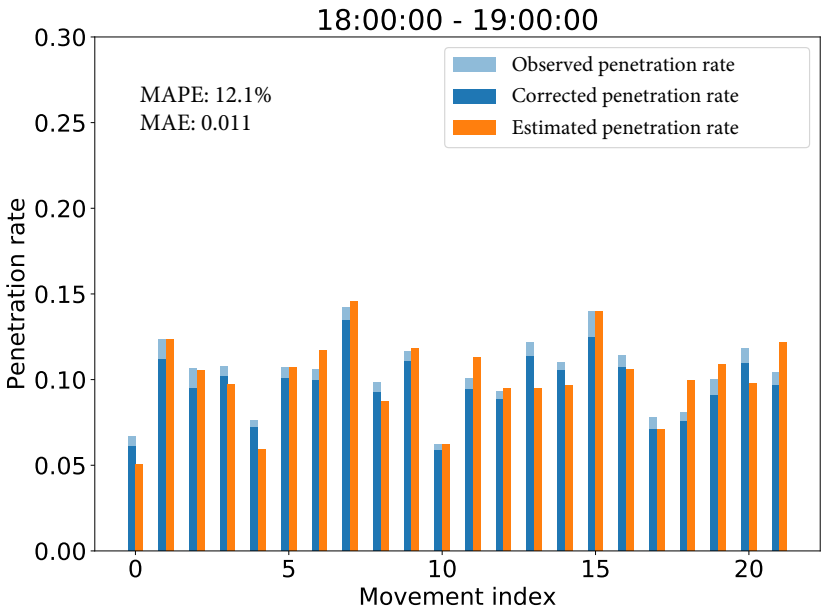
with the lanes with mixed movements. The studied movements are represented by the arrows in Figure 3.8. In total, 22 through movements and 31 left-turn movements are studied.

Most of the signalized intersections in the selected area are monitored by the camera-based automatic vehicle identification systems (AVIS) that can record the timestamps when vehicles go through the intersections. Nevertheless, not all the vehicles could be successfully identified by the cameras, and thus the vehicle counts often underestimate the actual traffic volumes. Therefore, for each camera, its identification rate is estimated by the ratio of the number of identified DiDi Vehicles and the total number of DiDi vehicles passing the camera. Then, the real "ground truth" of the traffic volumes are projected by dividing the vehicle counts by the estimated identification rates. Then, with the ground-truth traffic volume V^{all} and the traffic volume of probe vehicles V^{probe}, we obtained the ground-truth penetration rate using the following equation

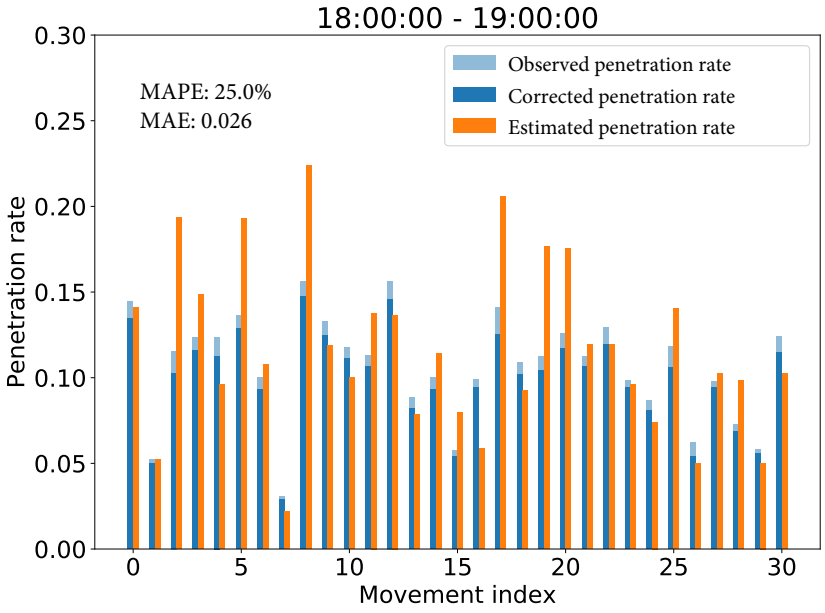
$$p = \frac{V^{probe}}{V^{all}}. \tag{3.44}$$

Results

Figure 3.9 shows the results of penetration rate estimation for the studied movements for the TOD 18:00-19:00. The estimation results show that the applied method $\frac{Q^{probe}}{\hat{Q}_4^{obs}+\hat{Q}_2^{hid}(p)} = p$ can estimate penetration rates accurately. Figures 3.9(a) and 3.9(b) show the results for the through and left-turn movements, respectively. Since the traffic volumes of the through movements are much larger compared to the left-turn movements, more probe vehicle samples can be used for estimation, and the corresponding performance is better.



(a)



(b)

Figure 3.9: The penetration rates of the probe vehicles in: (a) through movements; and (b) left-turn movements.

Compared to the results of the simulation data, the estimation accuracy is undermined when the method is applied to the real-world data, due to multiple reasons. First, although the map matching algorithm can mitigate the effect of GPS errors at the data preprocessing

stage, the errors in the real-world trajectory data could still influence the estimation accuracy. Second, in the real world, for each movement and each one-hour time slot, the penetration rate and the queueing pattern might vary slightly during the studied 15 workdays. Third, the average space headway for the queueing vehicles is set empirically, which might introduce some biases into the results. If the data with higher accuracy are available, the value of the average space headway should be estimated independently for each movement and each time slot.

3.7 Conclusions

In this chapter, we propose a general framework and a series of methods for estimating the parameters needed for queue length estimation in the i.i.d. case. For each specific movement and each specific time slot, the penetration rate of the probe vehicles and the queue length distribution are estimated by using the aggregated historical trajectory data of the probe vehicles.

The proposed methods do not assume the type of vehicle arrival process or the queueing process. The proposed methods do not require high penetration rates and would be feasible for use in reality nowadays. Therefore, compared to the existing methods in the literature, they can be applied to a broader range of scenarios. The tests by both the simulation and the real-world data show good estimation accuracy, indicating that the proposed methods could be used for traffic signal control and performance measures at signalized intersections.

A few issues should be considered when implementing the method in real life. Although the GPS errors can be partially corrected by map matching algorithms (Newson and Krumm, 2009; Lou et al., 2009), the error in the longitudinal direction might still undermine the estimation accuracy. For movements with multiple lanes, since the accuracy of the current GPS data usually cannot reach the lane level, it is difficult to know which vehicle is stopping on which lane. To walk around the issue, one may randomly assign the observed probe vehicles into different lanes, by assuming that the queue lengths in different lanes of the

same movement share similar patterns. For a specific TOD and movement, the data in one day are usually not sufficient to give accurate estimation results. Since the penetration rate and queue length distribution in different days is often similar, aggregating the data over several days can augment the dataset.

Chapter 4

Parameter estimation for independent queues: maximum likelihood estimation

4.1 Introduction

4.1.1 Background

In Chapter 2, we systematically introduced cycle-by-cycle queue length estimation methods using probe vehicle data. It was shown that the penetration rate of probe vehicles and the queue length distribution are the key parameters required by the estimation methods in the i.i.d. case. However, the values of these critical parameters are usually not given. Therefore, in Chapter 3, we tried to estimate the parameters from historical probe vehicle data and proposed a series of approximate estimators. We first obtained the approximated distribution of queue lengths from the aggregated data of stopping positions and then constructed the single-variable equation to solve for the penetration rate. The penetration rate could not be used to readjust the shape of the queue length distribution, as the two parameters were almost estimated independently.

4.1.2 Contribution and organization of the chapter

In this chapter, we propose to estimate the penetration rate and the queue length distribution simultaneously using maximum likelihood estimation. Similar to the approximate estimators presented in Chapter 3, the proposed method in this chapter does not impose any assumptions on the magnitude of the penetration rate or the form of the queue length distri-

bution. Therefore, the method can be applied to a wider range of scenarios compared to the existing methods in the literature. Based on the properties of the MLE, we also analyze the asymptotic standard error of the estimator, which can be regarded as the theoretical limit of its accuracy. Validation results from numerical experiments show that the proposed MLE improves the overall estimation accuracy compared to the approximate estimators. We also investigate the impact of the traffic intensity and sample size on the estimation accuracy. The estimated penetration rate and queue length estimation could enable many existing methods to estimate queue lengths cycle by cycle.

The rest of this chapter is organized as follows. In Section 4.2, we present the formulation of the maximum likelihood estimation problem. In Section 4.3, we elaborate on how to solve the MLE problem by applying the EM algorithm. In Section 4.4, we systematically test the performance of the proposed estimator in different scenarios and compare it with the approximate estimators introduced in Chapter 3. We provide discussion and concluding remarks in Section 4.5.

4.2 Maximum likelihood estimation of the penetration rate and queue length distribution

The objective of this section is to formulate the maximum likelihood estimation of the penetration rate of probe vehicles p and the queue length distribution π. For compactness, we denote the collection of parameters by θ, namely, $\theta = (p, \pi)$. Suppose we are given C cycles of historical probe vehicle data, from which we can extract the observed partial queues $q = \{q_1, q_2, \ldots, q_C\}$. The latent variables are the actual queue lengths $l = \{l_1, l_2, \ldots, l_C\}$. Based on the assumption that queue lengths in different traffic signal cycles are i.i.d., the likelihood function with respect to (w.r.t.) θ is

$$\mathcal{L}(\theta; q) = \prod_{i=1}^{C} P\left(q_i; \theta\right) = \prod_{i=1}^{C} \sum_{l_i=0}^{L_{max}} P\left(l_i, q_i; \theta\right), \tag{4.1}$$

where the joint probability of the queue length l_i and the observed partial queue q_i is

$$P(l_i, q_i; \theta) = \begin{cases} \pi_{l_i} p^{n_i} (1-p)^{l_i - n_i}, & l_i \geq |q_i| \\ 0, & l_i < |q_i| \end{cases}. \tag{4.2}$$

Therefore, the maximum likelihood estimation of θ can be obtained by maximizing the log-likelihood function $\log \mathcal{L}(\theta; q)$, that is,

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{C} \log \left(\sum_{j=|q_i|}^{L_{max}} \pi_j (1-p)^{j-n_i} p^{n_i} \right) && (4.3) \\ \text{subject to} \quad & \pi_j \geq 0, \forall j = 0, 1, \ldots, L_{max} && (4.4) \\ & \sum_{j=0}^{L_{max}} \pi_j = 1 && (4.5) \\ & 0 < p < 1. && (4.6) \end{aligned}$$

Constraints (4.4) and (4.5) ensure the validity of the queue length distribution. Constraint (4.6) guarantees that the penetration rate is in the range of $(0, 1)$.

The asymptotic normality of the maximum likelihood estimator can be used to analyze the ideal performance of the estimator. For the parameter p, given the log-likelihood function $\log \mathcal{L}(\theta; q)$, the corresponding Fisher information is

$$\begin{aligned} \mathcal{I}(p) &= -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial p^2} \log \mathcal{L}(\theta; q) \right) && (4.7) \\ &= -C \mathbb{E}_\theta \left(\frac{\mathcal{L}'' \mathcal{L} - \mathcal{L}'^2}{\mathcal{L}^2} \right) && (4.8) \\ &= -C \sum_{|q_i|=0}^{L_{max}} \sum_{n_i=0}^{|q_i|} \sum_{l_i=|q_i|}^{L_{max}} P(|q_i|, n_i, l_i) \frac{\mathcal{L}'' \mathcal{L} - \mathcal{L}'^2}{\mathcal{L}^2}, && (4.9) \end{aligned}$$

where

$$P(|q_i|, n_i, l_i) = \begin{cases} \pi_{l_i} C_{|q_i|-1}^{n_i - 1} p^{n_i} (1-p)^{l_i - n_i} & |q_i| > 0 \\ \pi_{l_i} (1-p)^{l_i} & |q_i| = 0 \end{cases}. \tag{4.10}$$

The asymptotic normality property states that the maximum likelihood estimator $\hat{p}$ converges in distribution to a normal distribution, that is,

$$\hat{p} \to \mathcal{N}\left(p^*, \frac{1}{\mathcal{I}(p^*)C}\right), \tag{4.11}$$

where p^* represents the true value of the penetration rate.

The objective of the optimization problem is to maximize the log-likelihood function w.r.t. θ. The log-likelihood function is concave w.r.t. π, but not concave w.r.t. p. Therefore, convex optimization methods cannot be directly applied here. To solve the problem, we resort to the EM algorithm.

4.3 The EM algorithm

As an iterative algorithm for solving maximum likelihood estimation problems, the EM algorithm is commonly used when there are latent variables in the likelihood function (Dempster et al., 1977). Instead of solving the original optimization problem, the EM algorithm converts it into a series of problems that can be more easily solved.

4.3.1 E-step

Given a feasible solution $\theta^{(t)}$, we evaluate the expectation of the complete-data log-likelihood function $\log P\left(q_i, l_i; \theta\right)$ under the posterior distribution of the latent variable l_i.

$$\begin{aligned} Q\left(\theta; \theta^{(t)}\right) &= \sum_{i=1}^{C} \mathbb{E}_{l_i|q_i,\theta^{(t)}} \left(\log P\left(q_i, l_i; \theta\right)\right) && (4.12) \\ &= \sum_{i=1}^{C} \sum_{l_i=0}^{L_{max}} P\left(l_i \mid q_i; \theta^{(t)}\right) \log P\left(q_i, l_i; \theta\right) && (4.13) \\ &= \sum_{i=1}^{C} \sum_{l_i=|q_i|}^{L_{max}} \frac{P\left(q_i, l_i; \theta^{(t)}\right)}{\sum_{k=|q_i|}^{L_{max}} P\left(q_i, k; \theta^{(t)}\right)} \log P\left(q_i, l_i; \theta\right). && (4.14) \end{aligned}$$

This step is equivalent to constructing a lower bound of the original log-likelihood function (Jain et al., 2017; Balakrishnan et al., 2017). Substituting equation (4.2) into (4.14) gives

$$Q\left(\theta;\theta^{(t)}\right)=\sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_j^{(t)}\left(n_i\log p+(j-n_i)\log(1-p)+\log\pi_j\right)}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}, \tag{4.15}$$

which is a concave function of θ.

4.3.2 M-step

The M-step updates the estimate of θ by maximizing $Q\left(\theta;\theta^{(t)}\right)$, that is,

$$\begin{aligned}
\text{maximize}\quad & \sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_j^{(t)}\left(n_i\log p+(j-n_i)\log(1-p)+\log\pi_j\right)}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}} && (4.16)\\
\text{subject to}\quad & \pi_j\geq 0,\forall j=0,1,\ldots,L_{max} && (4.17)\\
& \sum_{j=0}^{L_{max}}\pi_j=1 && (4.18)\\
& 0<p<1. && (4.19)
\end{aligned}$$

As long as $\theta^{(t)}$ is in the feasible region, the optimization problem will be feasible. The analytical solutions can be expressed as

$$p^{(t+1)}=\frac{\sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_j^{(t)}n_i}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}}{\sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_j^{(t)}j}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}}, \tag{4.20}$$

$$\pi_j^{(t+1)}=\frac{\sum_{i:|q_i|\leq j}^{C}\frac{\pi_j^{(t)}}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}}{\sum_{m=0}^{L_{max}}\sum_{i:|q_i|\leq m}^{C}\frac{\pi_m^{(t)}}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-m}\pi_k^{(t)}}},\forall j=0,1,\ldots,L_{max}. \tag{4.21}$$

The detailed solving process can be found in Appendix B.

The EM algorithm guarantees $\log\mathcal{L}\left(\theta^{(t+1)};q\right)\geq\log\mathcal{L}\left(\theta^{(t)};q\right)$. In other words, the updated estimate will be no worse than the previous estimate in the sense of likelihood (Demp-

ster et al., 1977). Since the log-likelihood function is bounded from above, according to the monotone convergence theorem, the solution will always converge after iterating the E-steps and M-steps (Wu et al., 1983).

4.3.3 Initial point

The previous subsections have elaborated the E-step and the M-step. This subsection describes how to select an initial guess of the parameters. In Chapter 3, we proposed a series of approximate estimators for the probe vehicle penetration rate. Applying any of the methods will give an initial estimate of the penetration rate $p^{(0)}$. As explained in the previous sections, when p is fixed to $p^{(0)}$, $\log \mathcal{L}\left(\pi; q, p^{(0)}\right)$ will be a concave function w.r.t. π. Therefore, solving the following optimization problem gives an initial guess of the queue length distribution, namely, $\pi^{(0)} = \text{argmax}_{\pi} \log \mathcal{L}\left(\pi; q, p^{(0)}\right)$.

$$\text{maximize} \quad \sum_{i=1}^{C} \log \left(\sum_{j=|q_i|}^{L_{max}} \pi_j \left(1 - p^{(0)}\right)^{j-n_i} \left(p^{(0)}\right)^{n_i} \right) \tag{4.22}$$

$$\text{subject to} \quad \pi_j \geq 0, \forall j = 0, 1, \ldots, L_{max} \tag{4.23}$$

$$\sum_{j=0}^{L_{max}} \pi_j = 1. \tag{4.24}$$

Then, $\theta^{(0)} = \left(p^{(0)}, \pi^{(0)}\right)$ can be used as an initial estimate of θ to start the EM algorithm. The steps of the EM algorithm is summarized in Algorithm 4.1.

Algorithm 4.1 The EM algorithm for estimating the penetration rate and queue length distribution.

Calculate $p^{(0)}$
$\pi^{(0)} \leftarrow \mathbf{argmax}_{\pi} \log \mathcal{L}(\pi; q, p^{(0)})$
$\theta^{(0)} \leftarrow \left(p^{(0)}, \pi^{(0)}\right)$
$t \leftarrow 0$
Set a stopping threshold ϵ
while True **do**
  E-step: $Q\left(\theta; \theta^{(t)}\right) \leftarrow \sum_{i=1}^{C} \mathbb{E}_{l_i|q_i,\theta^{(t)}} \log P\left(q_i, l_i; \theta\right)$
  M-step: $\theta^{(t+1)} \leftarrow \mathbf{argmax}_{\theta} Q\left(\theta; \theta^{(t)}\right)$
  if $|\log \mathcal{L}(\theta^{(t+1)}; q) - \log \mathcal{L}(\theta^{(t)}; q)|/|\log \mathcal{L}(\theta^{(t)}; q)| < \epsilon$ **then**
    break
  end if
  $t \leftarrow t + 1$
end while

4.4 Numerical experiments

4.4.1 Simulation environment and performance measures

The proposed method is validated using numerical experiments. Although the proposed method is not restricted in terms of the form of the queue length distribution, we focus on Poisson distributions, which is commonly used in the relevant literature. Similar to the settings introduced in Section 3.6, we draw queue length samples from the pre-determined distribution. For each vehicle in the queue drawn from the distribution, its vehicle type is determined by a Bernoulli trial based on the penetration rate. Then, from each queue sample, the pattern in front of (including) the last probe vehicle is extracted as the observed partial queue. Finally, such partial queues are considered as the input to the proposed estimator.

We measure the accuracy of the estimated penetration rate using the mean absolute percentage error. For queue length distributions, we use the Hellinger distance, which has been introduced in Section 3.6.

4.4.2 Results

We generate 1000 cycles of queues from the distribution $Poisson(\lambda = 5)$, which corresponds to the case where the average arrival rate in the red phase is five. The box plot in Figure 4.1 shows the estimation results when the numerical experiments are repeated 500 times. The horizontal axis represents the ground-truth penetration rates used for generating the simulation data. The vertical axis represents the estimated penetration rates. The ends of the boxes indicate the first and third quartiles. The ends of whiskers are the min and max. The dashed line in red is a reference line. Figure 4.1 shows that the proposed method could estimate penetration rates accurately.

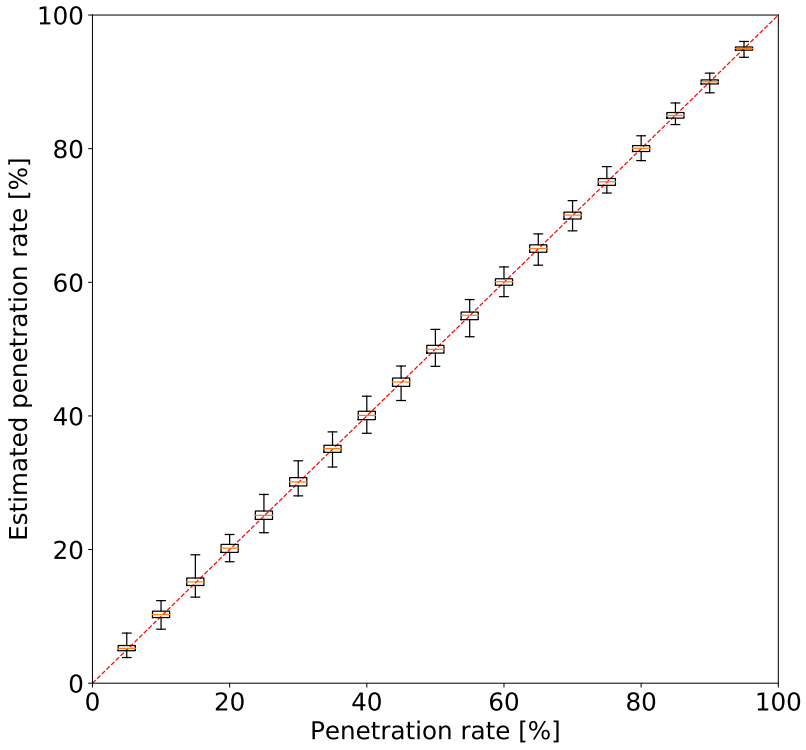


Figure 4.1: Estimation results for penetration rates.

By applying the property of asymptotic normality, we can obtain the asymptotic standard error of the MLE. Figure 4.2 shows the comparison between the calculated asymptotic standard errors and the actual errors generated by the EM algorithm. The asymptotic values represent what the MLE should asymptotically achieve when the sample size goes to

infinity and the problem is perfectly solved. One reason why there exist gaps between the two curves is that the likelihood function is non-concave, and the EM algorithm does not guarantee global optimum. Figure 4.2 also shows that when the penetration rate is higher, the gaps between the asymptotic values and the actual values tend to be smaller.

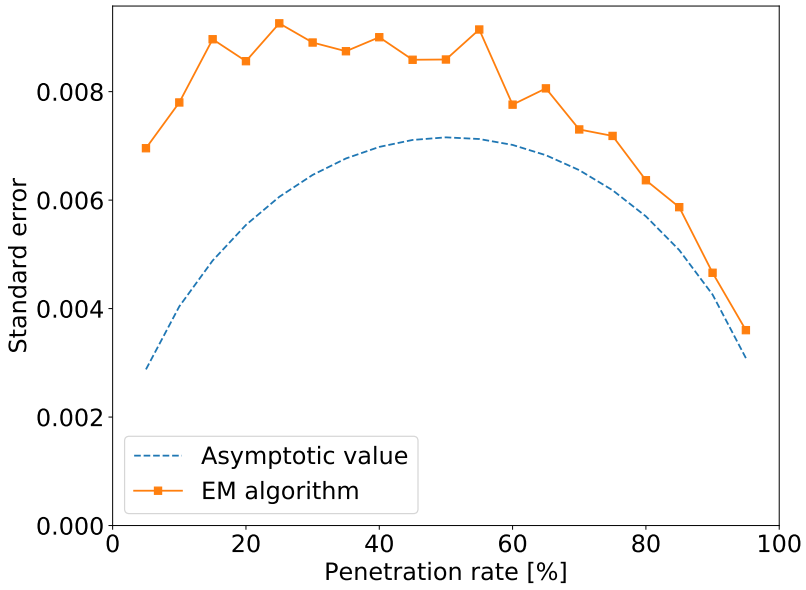


Figure 4.2: The comparison of asymptotic standard errors and the actual errors given by the EM algorithm.

Figure 4.3 shows the estimation results of queue length distributions in four representative scenarios. In each plot, the distribution in blue stands for the ground-truth queue length distribution. The distribution in orange represents the estimated queue length distribution. When the penetration rate is low, the information contained in the observations is limited, which results in a rough estimate of the queue length distribution. When the penetration rate gets higher, the proposed method could reconstruct the distribution more accurately.

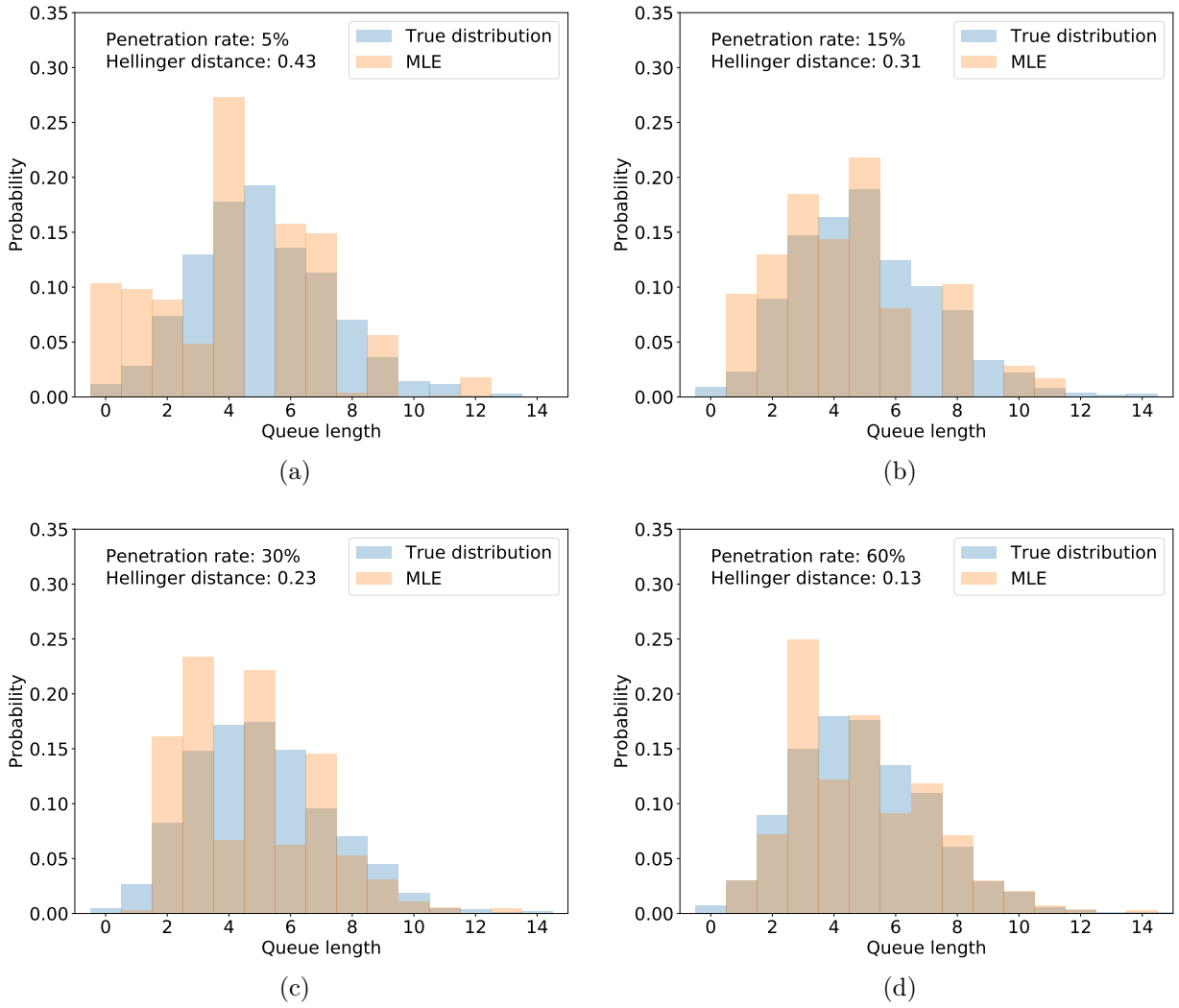
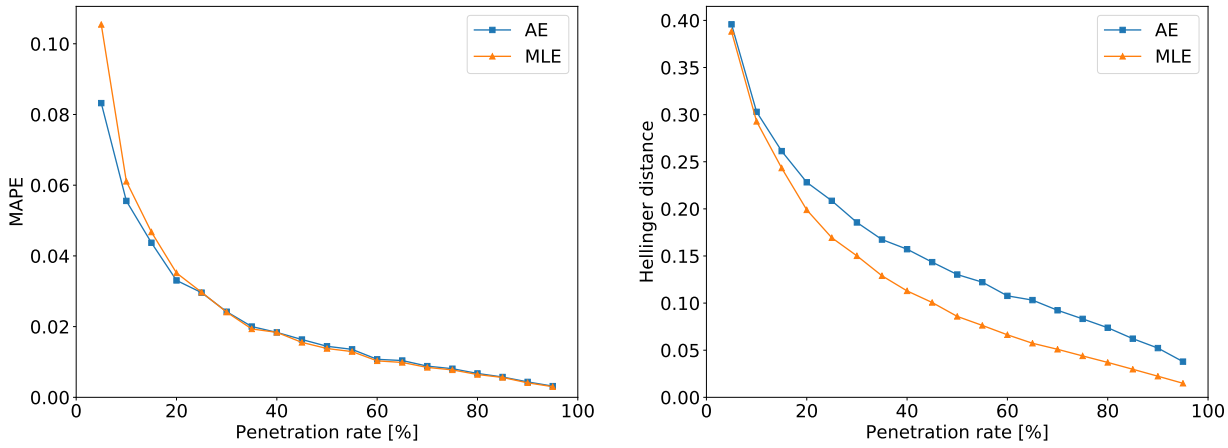


Figure 4.3: Estimation results of queue length distributions under different probe vehicle penetration rates: (a) 5%; (b) 15%; (c) 30%; and (d) 60%.

We compare the performance of the MLE with the approximate estimators proposed in Chapter 3. Figures 4.4(a) and 4.4(b) show the estimation results of penetration rates and queue length distributions based on the two methods under different true penetration rates, respectively. In general, for penetration rate estimation, the approximate estimators perform slightly better when the penetration rate is low, whereas the proposed MLE has the edge over the AE when the penetration rate is higher than 25%. When it comes to the estimation of queue length distributions, the MLE outperforms the approximate estimators significantly. The results also indicate that increasing the penetration rate will improve

estimation accuracy.



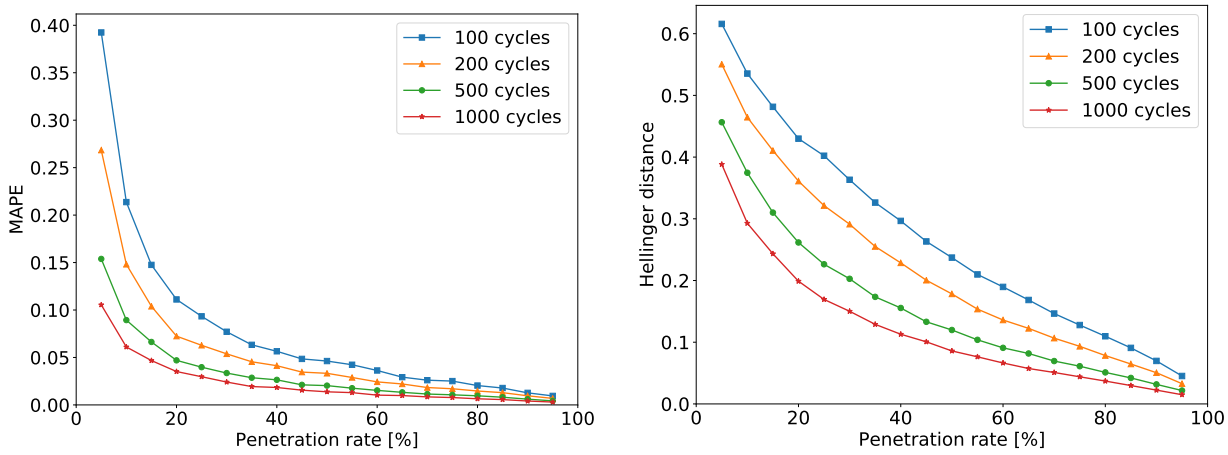
(a) (b)

Figure 4.4: The comparison of the AE and the MLE: (a) estimation of penetration rates and (b) estimation of queue length distributions.

4.4.3 Sensitivity Analysis

The impact of sample size

To investigate the impact of sample size on estimation accuracy, we fix the true queue length distribution to $Poisson(\lambda = 5)$ and use 100, 200, 500, and 1000 cycles of simulated probe vehicle data to estimate the parameters, respectively. Figure 4.5 shows the comparison of the results generated using different sample sizes. For both parameters, better estimation accuracy is achieved when a larger sample size is used.

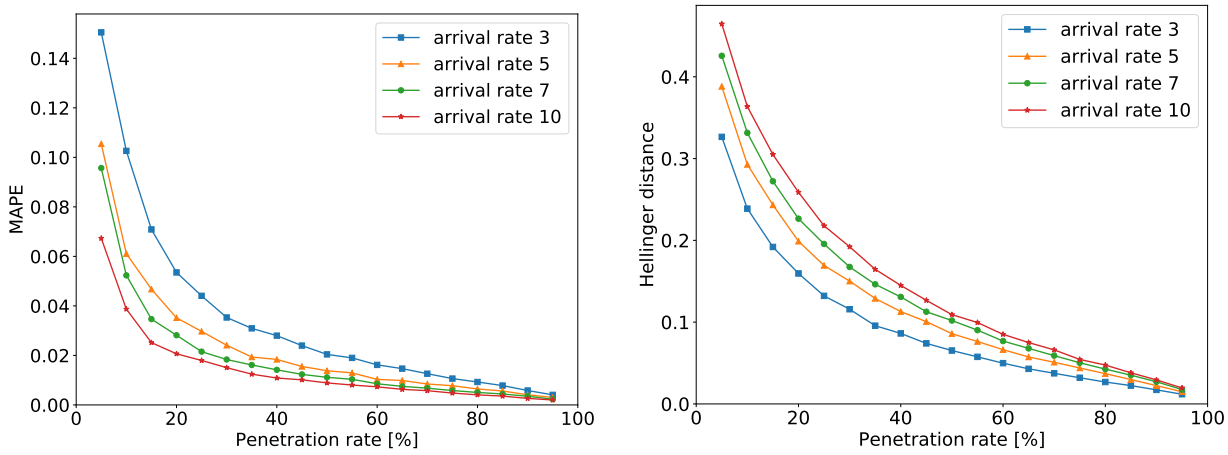


(a) (b)

Figure 4.5: The impact of sample size on the estimation results for: (a) penetration rates and (b) queue length distributions.

The impact of the average arrival rate in the red phase

To investigate how the average arrival rate in the red phase influences the estimation accuracy, we set the average arrival rate to 3, 5, 7, and 10 in four experiments, respectively. In each experiment, 1000 cycles of probe vehicle data are used. Figure 4.6 shows a comparison of the results under different arrival rates. For penetration rates, better estimation accuracy is achieved when the average arrival rate is higher. The reason is that a higher arrival rate implies more probe vehicle samples. However, for queue length distributions, the results show the opposite when the accuracy is measured by the Hellinger distance. It is because when the arrival rate is higher, L_{max} tends to be larger, and therefore, the number of decision variables in the optimization problem also grows, which makes it harder to estimate π.



(a) (b)

Figure 4.6: The impact of arrival rates on the estimation results for: (a) penetration rates and (b) queue length distributions.

4.5 Conclusions

Many probe vehicle based cycle-by-cycle queue length estimation methods require the knowledge of the probe vehicle penetration rate and queue length distribution at the studied intersection. However, the estimation of the two parameters has not been extensively studied. This chapter proposes a maximum likelihood estimation method to estimate the parameters using historical probe vehicle data. Due to the non-concavity of the log-likelihood function, we apply the EM algorithm to solve the problem iteratively. We validate the proposed method and study the impacts of the sample size and the average arrival rate by conducting numerical experiments.

There are certain limitations of the proposed method. For instance, the estimation methods proposed in this chapter and Chapter 3 take the stopping positions of the probe vehicles as the features to infer the penetration rate and the queue length distribution. However, there might not be queues forming at non-signalized intersections or in right-turn movements. Also, the queueing patterns in the shared left-through (right-through) lanes could be different from other left-turn (right-turn) lanes or through lanes.

There are also a few directions one may explore to further improve the estimation accuracy. In this chapter, only the stopping locations of the queueing probe vehicles are used for estimation. The estimation accuracy might be improved by taking into account the time when the vehicles join the queue. In general, this chapter takes a frequentist approach by applying maximum likelihood estimation to give point estimates of the parameters. One may introduce prior distributions of the parameters and approach the problem from a Bayesian perspective. Finally, the focus of this chapter is on queues that are independent and identically distributed. In the next chapter, we will deal with the case where queues in different cycles are correlated.

Chapter 5

Parameter estimation for dependent queues: maximum likelihood estimation

5.1 Introduction

5.1.1 Background

Understanding the queueing processes at signalized intersections can help with traffic management and control. Most of the existing studies introduced in Section 1.4.1 focused on isolated intersections under light or moderate traffic conditions and treated the queues in different cycles independently. However, there are many common scenarios where the queue lengths in different cycles are dependent, for example, overflow queues. In these scenarios, considering the correlations can potentially improve the queue length estimation accuracy, as the information in adjacent cycles can also be used for estimation. In Chapter 2, we proposed a hidden Markov model to deal with such scenarios. With only probe vehicle data, the real queue lengths might not be directly observable. Therefore, the queue lengths are considered as hidden states of the HMM. For each queue, the pattern of the stopping positions of probe vehicles can be observed. The observed patterns are considered as the observations of the HMM. The model is well suited for modeling dependent queues in probe vehicle environments. Based on the hidden Markov model, we provide two cycle-by-cycle queue length estimation methods, which require the knowledge of the HMM parameters, including transition probabilities, initial probabilities, and the probe vehicle penetration rate. In the real world, however, when the methods are implemented, the parameters of

the HMM are not given. Instead, the parameters should be estimated from historical probe vehicle data.

5.1.2 Contribution and organization of the chapter

In this chapter, we focus on estimating the parameters of the HMM from historical probe vehicle data. Similar to the i.i.d. case, we formulate the problem as a maximum likelihood estimation problem and obtain the solutions by applying the EM algorithm. We also validate the cycle-by-cycle estimation methods proposed in Chapter 2 using the parameters estimated from historical probe vehicle data.

The rest of this chapter is organized as follows. In Section 5.2, we formulate the maximum likelihood estimation problem, of which the objective function is non-concave because of the hidden variables. In Section 5.3, we solve the problem by combining the EM algorithm and dynamic programming. Considering that the data collected in real life usually span multiple days, we also extend the solutions to the multi-day case. In Section 5.4, we validate the proposed algorithm by a case study, of which the simulation settings are consistent with the one in Chapter 2. We also analyze the impact of the sample size on estimation accuracy. Finally, in Section 5.5, we provide some concluding remarks.

5.2 Maximum likelihood estimation of the HMM parameters

In order to carry out the cycle-by-cycle estimation methods introduced in Section 2.4, we need to estimate the parameters of the hidden Markov model from historical probe vehicle data. This section elaborates on the maximum likelihood estimation of the parameters $\theta = (\pi, T, p)$.

Given the collection of observations $q = \{q_1, q_2, \ldots, q_C\}$, the likelihood function of θ is

$$\mathcal{L}(\theta; q) = \sum_{l} P\left(q, l; \theta\right) = \sum_{l} P(l_1) \prod_{i=2}^{C} P(l_i \mid l_{i-1}) \prod_{i=1}^{C} P(q_i \mid l_i). \tag{5.1}$$

The likelihood function is marginalized due to the existence of the hidden variables, namely, the queue length sequence $l = \{l_1, l_2, \ldots, l_C\}$. After substituting the parameters into equation (5.1), the log-likelihood function can be expressed as

$$\log \mathcal{L}(\theta; q) = \log \left(\sum_{l} \pi_{l_1} \prod_{i=2}^{C} T_{l_{i-1} l_i} \prod_{i=1}^{C} p^{n_i} (1-p)^{l_i - n_i} \right) \tag{5.2}$$

The maximum likelihood estimation of θ can be formulated as the following optimization problem.

$$\begin{aligned}
\text{maximize} \quad & \log \left(\sum_{l} \pi_{l_1} \prod_{i=2}^{C} T_{l_{i-1} l_i} \prod_{i=1}^{C} p^{n_i} (1-p)^{l_i - n_i} \right) && (5.3) \\
\text{subject to} \quad & \pi_j \geq 0, \forall j = 0, 1, \ldots, L_{max} && (5.4) \\
& \sum_{j=0}^{L_{max}} \pi_j = 1 && (5.5) \\
& T_{jk} > 0, \forall j, k = 0, 1, \ldots, L_{max} && (5.6) \\
& \sum_{k=0}^{L_{max}} T_{jk} = 1, \forall j = 0, 1, \ldots, L_{max} && (5.7) \\
& 0 < p < 1. && (5.8)
\end{aligned}$$

The objective is to maximize the log-likelihood function $\log \mathcal{L}(\theta; q)$. The decision variable is $\theta = (\pi, T, p)$. Constraints (5.4) and (5.5) ensure the validity of the initial probabilities. Constraints (5.6) and (5.7) guarantee the validity of the transition probabilities. Constraint (5.8) states that the penetration rate should be between 0 and 1.

However, due to the existence of hidden states, the objective function is not concave w.r.t. θ. Therefore, we resort to the EM algorithm (Dempster et al., 1977) to solve the maximum likelihood estimation problem.

5.3 The EM algorithm

5.3.1 E-step

Following the standard EM algorithm, in the E-step, we evaluate the expectation of the complete data log-likelihood function, under the posterior distribution of hidden states based on the current estimate of the parameters $\theta^{(t)}$, which is

$$Q\left(\theta;\theta^{(t)}\right) = \mathbb{E}_{l|q;\theta^{(t)}} \log P\left(q,l;\theta\right) \tag{5.9}$$

$$= \sum_{l} P\left(l \mid q;\theta^{(t)}\right)\left(\log \pi_{l_1} + \sum_{i=2}^{C} \log T_{l_{i-1}l_i} + \sum_{i=1}^{C}\left(n_i \log p + (l_i - n_i)\log\left(1-p\right)\right)\right). \tag{5.10}$$

5.3.2 M-step

In the M step, we obtain a new estimate of the parameters, $\theta^{(t+1)}$, by maximizing $Q\left(\theta;\theta^{(t)}\right)$ subject to the constraints, which is

$$\begin{aligned}
\text{maximize} \quad & Q\left(\theta;\theta^{(t)}\right) && (5.11)\\
\text{subject to} \quad & \pi_j \geq 0, \forall j = 0,1,\ldots,L_{max} && (5.12)\\
& \sum_{j=0}^{L_{max}} \pi_j = 1 && (5.13)\\
& T_{jk} > 0, \forall j,k = 0,1,\ldots,L_{max} && (5.14)\\
& \sum_{k=0}^{L_{max}} T_{jk} = 1, \forall j = 0,1,\ldots,L_{max} && (5.15)\\
& 0 < p < 1. && (5.16)
\end{aligned}$$

The objective function of the M-step is concave, and all the constraints are convex. The analytical solutions to the problem can be obtained by setting the derivatives of the corresponding Lagrangian to zero. Specifically, the solutions are

$$\pi_j^{(t+1)} = \frac{\sum_{l:l_1=j} P\left(l \mid q; \theta^{(t)}\right)}{\sum_l P\left(l \mid q; \theta^{(t)}\right)}, \forall j = 0, 1, \ldots, L_{max}; \tag{5.17}$$

$$T_{jk}^{(t+1)} = \frac{\sum_l P\left(l \mid q; \theta^{(t)}\right) \sum_{i:2\leq i\leq C, l_{i-1}=j, l_i=k} 1}{\sum_{m=0}^{L_{max}} \sum_l P\left(l \mid q; \theta^{(t)}\right) \sum_{i:2\leq i\leq C, l_{i-1}=j, l_i=m} 1}, \forall j, k = 0, 1, \ldots, L_{max}; \tag{5.18}$$

$$p^{(t+1)} = \frac{\sum_{i=1}^{C} n_i}{\sum_l P\left(l \mid q; \theta^{(t)}\right) \sum_{i=1}^{C} l_i}. \tag{5.19}$$

The detailed solving process can be found in Appendix C. Nevertheless, the calculation of the solutions given by equations (5.17)-(5.19) requires enumerating all the possible sequences of the hidden states l, which is intractable. Thus, we resort to dynamic programming to carry out the solutions (Baum et al., 1970).

5.3.3 The forward-backward algorithm

Define two sets of auxiliary variables

$$\alpha_j^{(t)}(i) = P\left(q_1, q_2, \ldots, q_i, l_i = j; \theta^{(t)}\right), \forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C, \tag{5.20}$$

$$\beta_j^{(t)}(i) = P\left(q_{i+1}, q_{i+2}, \ldots, q_C \mid l_i = j; \theta^{(t)}\right), \forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C. \tag{5.21}$$

The solutions in equations (5.17)-(5.19) can be reformulated as

$$\pi_j^{(t+1)} = \frac{\alpha_j^{(t)}(1)\beta_j^{(t)}(1)}{\sum_{k=0}^{L_{max}} \alpha_k^{(t)}(1)\beta_k^{(t)}(1)}, \forall j = 0, 1, \ldots, L_{max}; \tag{5.22}$$

$$T_{jk}^{(t+1)} = \frac{\sum_{i=1}^{C-1} \alpha_j^{(t)}(i) T_{jk}^{(t)} E_{kq_{i+1}}^{(t)} \beta_k^{(t)}(i+1)}{\sum_{m=0}^{L_{max}} \sum_{i=1}^{C-1} \alpha_j^{(t)}(i) T_{jm}^{(t)} E_{mq_{i+1}}^{(t)} \beta_m^{(t)}(i+1)}, \forall j, k = 0, 1, \ldots, L_{max}; \tag{5.23}$$

$$p^{(t+1)} = \frac{\sum_{j=0}^{L_{max}} \sum_{i=1}^{C} \alpha_j^{(t)}(i)\beta_j^{(t)}(i) n_i}{\sum_{j=0}^{L_{max}} \sum_{i=1}^{C} \alpha_j^{(t)}(i)\beta_j^{(t)}(i) j}. \tag{5.24}$$

After the reformulation, the intractable enumeration of all possible l is avoided. The update rules for (π, T, p) now become tractable. The auxiliary variables defined in equations (5.22)-(5.24) can be carried out by dynamic programming using the following equations.

$$\alpha_j^{(t)}(1) = \pi_j^{(t)} E_{jq_1}^{(t)}, \forall j = 0, 1, \ldots, L_{max}; \tag{5.25}$$

$$\alpha_j^{(t)}(i) = \sum_{k=0}^{L_{max}} \alpha_k^{(t)}(i-1) T_{kj}^{(t)} E_{jq_i}^{(t)}, \forall j = 0, 1, \ldots, L_{max}, \forall i = 2, 3, \ldots, C. \tag{5.26}$$

$$\beta_j^{(t)}(C) = 1, \forall j = 0, 1, \ldots, L_{max}, \tag{5.27}$$

$$\beta_j^{(t)}(i) = \sum_{k=0}^{L_{max}} T_{jk}^{(t)} E_{kq_{i+1}}^{(t)} \beta_k^{(t)}(i+1), \forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C-1. \tag{5.28}$$

5.3.4 Considering the data of different days

The results above are for the case where the observations are collected in consecutive traffic signal cycles. In this subsection, we generalize the results to the case where probe vehicle data of the same TOD are collected for D days with C cycles per day, which is a common scenario in the context of probe vehicle based parameter estimation.

Given the collection of observations $\{q_{di}, \forall d = 1, 2, \ldots, D, \forall i = 1, 2, \ldots, C\}$, the update rules of the parameters can be derived following the same procedures as the previous subsection.

$$\pi_j^{(t+1)} = \frac{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \alpha_j^{(t)}(d,1) \beta_j^{(t)}(d,1)}{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \sum_{j=0}^{L_{max}} \alpha_j^{(t)}(d,1) \beta_j^{(t)}(d,1)}, \forall j = 0, 1, \ldots, L_{max}; \tag{5.29}$$

$$T_{jk}^{(t+1)} = \frac{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \sum_{i=1}^{C-1} \alpha_j^{(t)}(d,i) T_{jk}^{(t)} E_{kq_{d,i+1}}^{(t)} \beta_k^{(t)}(d,i+1)}{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \sum_{k=0}^{L_{max}} \sum_{i=1}^{C-1} \alpha_j^{(t)}(d,i) T_{jk}^{(t)} E_{kq_{d,i+1}}^{(t)} \beta_k^{(t)}(d,i+1)}, \tag{5.30}$$

$$\forall j, k = 0, 1, \ldots, L_{max};$$

$$p^{(t+1)} = \frac{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \sum_{j=0}^{L_{max}} \sum_{i=1}^{C} \alpha_j^{(t)}(d,1)\beta_j^{(t)}(d,1) n_{di}}{\sum_{d=1}^{D} \frac{1}{\sum_{m=0}^{L_{max}} \alpha_m^{(t)}(d,C)} \sum_{j=0}^{L_{max}} \sum_{i=1}^{C} \alpha_j^{(t)}(d,i)\beta_j^{(t)}(d,i) j}. \tag{5.31}$$

$\forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C_d, \forall d = 1, 2, \ldots, D$, the auxiliary variables are defined as

$$\alpha_j^{(t)}(d,i) = P\left(q_{d1}, q_{d2}, \ldots, q_{di}, l_{di} = j; \theta^{(t)}\right), \tag{5.32}$$

$$\beta_j^{(t)}(d,i) = P\left(q_{d,i+1}, q_{d,i+2}, \ldots, q_{dC} \mid l_{di} = j; \theta^{(t)}\right). \tag{5.33}$$

$\forall d = 1, 2, \ldots, D$, the auxiliary variables can be carried out recursively as follows.

$$\alpha_j^{(t)}(d,1) = \pi_j^{(t)} E_{jq_{d1}}^{(t)}, \forall j = 0, 1, \ldots, L_{max}; \tag{5.34}$$

$$\alpha_j^{(t)}(d,i) = \sum_{k=0}^{L_{max}} \alpha_k^{(t)}(d,i-1) T_{kj}^{(t)} E_{jq_{di}}^{(t)}, \forall j = 0, 1, \ldots, L_{max}, \forall i = 2, 3, \ldots, C; \tag{5.35}$$

$$\beta_j^{(t)}(d,C) = 1, \forall j = 0, 1, \ldots, L_{max}, \tag{5.36}$$

$$\beta_j^{(t)}(d,i) = \sum_{k=0}^{L_{max}} T_{jk}^{(t)} E_{kq_{d,i+1}}^{(t)} \beta_k^{(t)}(d,i+1), \forall j = 0, 1, \ldots, L_{max}, \forall i = 1, 2, \ldots, C-1. \tag{5.37}$$

The results in this subsection will degenerate to the results in Section 5.3.3 if we set the number of days D to 1.

5.4 Case studies

5.4.1 Simulation settings

The focus of this case study is to validate that the maximum likelihood estimator proposed in this chapter can successfully estimate the parameters of the HMM from historical probe vehicle data, and the estimated parameters can be used for cycle-by-cycle queue length estimation. The simulation settings are the same as the case study in Section 2.5, where we demonstrated how to estimate queue lengths cycle by cycle if the parameters of the HMM are

given. Specifically, we generate queue sequences using the overflow queue model proposed by Viti and Van Zuylen (2010). The penetration rate of probe vehicles is set to be 20%.

According to the overflow queue model and the simulation settings, we generate a dataset containing the queues of 30 days. To simulate the amount of data that can be collected for a one-hour TOD, we generate 30 cycles of data for each day, considering the length of a traffic signal cycle is roughly two minutes. Then, from the generated queue data, we perform Bernoulli trials to determine if each vehicle in the queues is a probe vehicle or a regular vehicle. The data of the probe vehicles are used as the input to the estimation algorithm.

5.4.2 Parameter estimation

We apply the proposed method to the simulated probe vehicle data of 30 days. We generate the initial guess of the parameters randomly and then update the parameters iteratively through equations (5.29)-(5.31) until the estimated values converge or the number of iterations reaches a preset threshold. Figure 5.1(a) is a visualization of the transition matrix estimated from the historical data, which captures the key features of the ground-truth transition matrix shown in Figure 5.1(b). Figure 5.1(c) illustrates the convergence of the penetration rate during the estimation process. The figure indicates that the parameter estimation algorithm converges to the estimated value very quickly. Figure 5.1(d) shows the comparison of the estimated initial probabilities and their true values. In general, the estimated distribution is close to the true distribution.

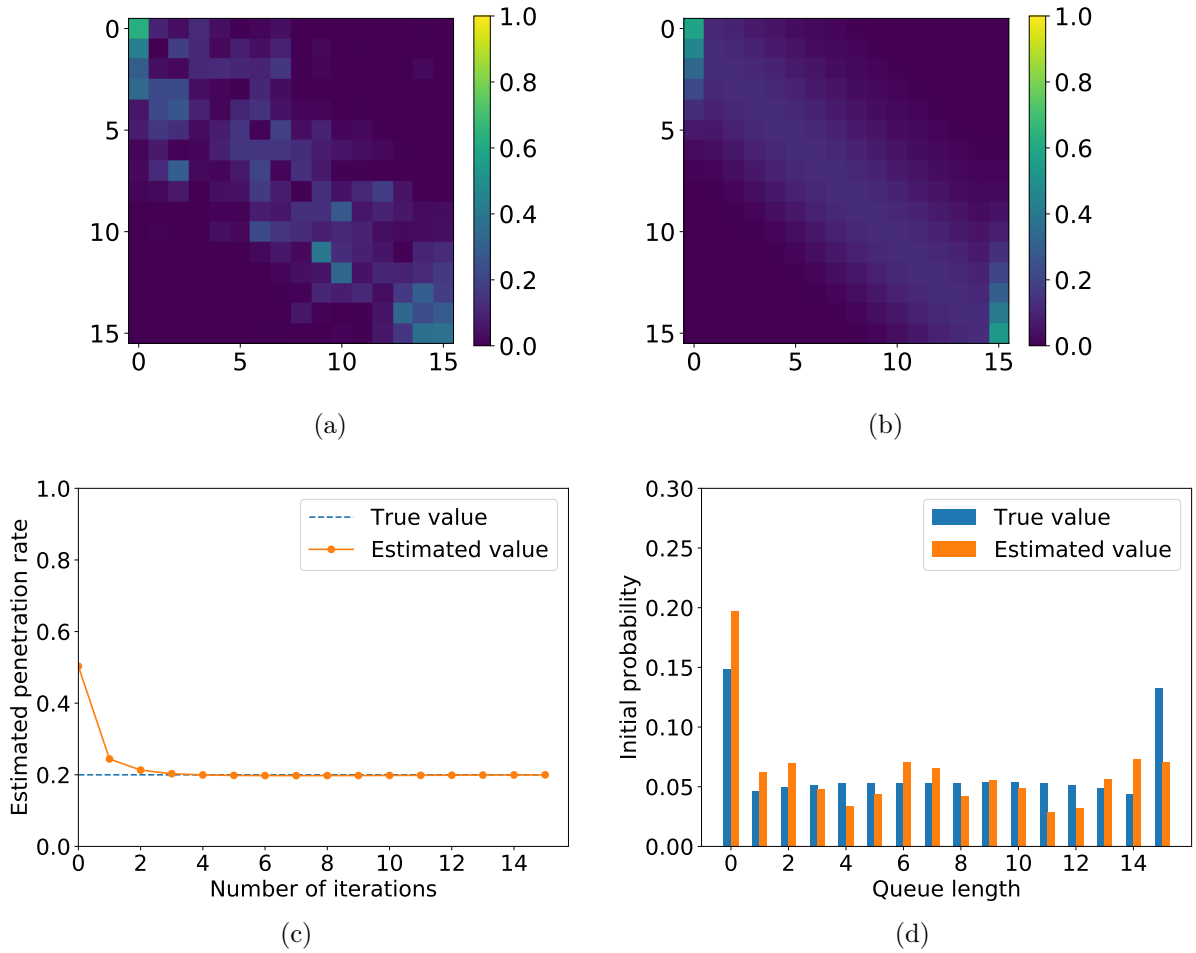
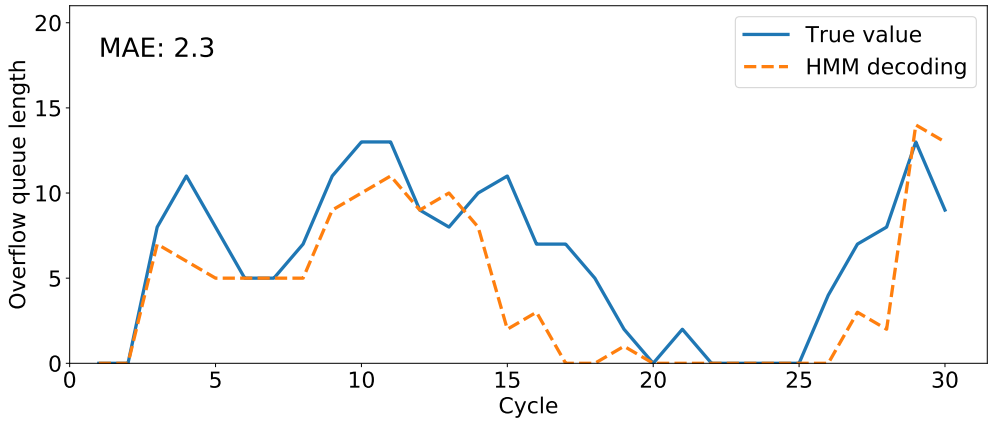
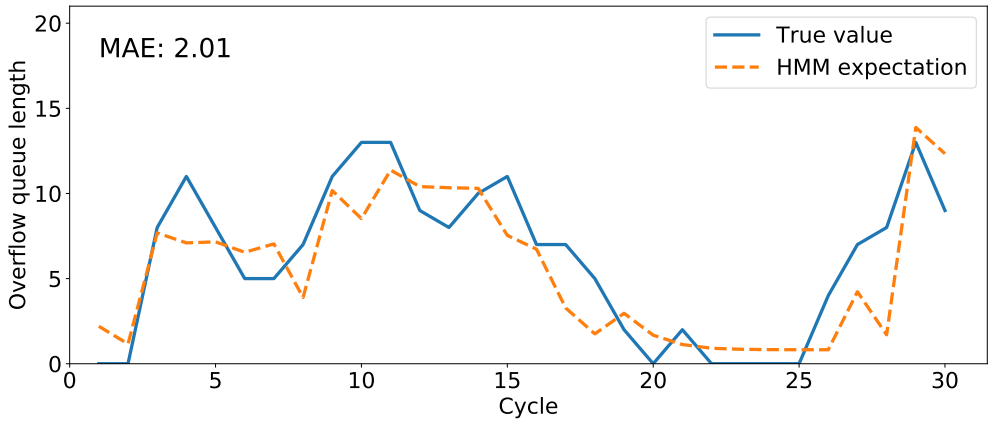


Figure 5.1: The estimation results of the parameters: (a) the estimated transition matrix; (b) the true transition matrix; (c) the estimation process of the penetration rate; and (d) the estimated initial probabilities compared to the true values.

With the estimated parameters of the hidden Markov model, we again apply the cycle-by-cycle queue length estimation methods to a 30-cycle observation sequence. Figure 5.2 shows the corresponding estimation results. As introduced in Chapter 2, the HMM decoding method represents the maximum likelihood estimator corresponding to equation (2.10), and the HMM expectation method corresponds to equation (2.11). Compared with the results shown in Figure 2.5, the proposed methods still outperform the baseline methods. It is concrete evidence of the effectiveness of the two cycle-by-cycle queue length estimation methods and the parameter estimation algorithm proposed in this chapter.



(a)



(b)

Figure 5.2: Cycle-by-cycle queue length estimation results using the learned parameters by: (a) maximum likelihood estimation (decoding); and (b) expectation conditional on sequential observations.

5.4.3 The impact of penetration rates

The results above illustrate the performance of the algorithms in one example. We run the experiments repeatedly to get the average estimation accuracy under different probe vehicle penetration rates. Figure 5.3 shows the results. Obviously, a larger penetration rate implies a higher estimation accuracy. When the penetration rate is very low, for example, 5%, it is hard for the algorithm to estimate the parameters accurately. Therefore, the queue length estimation accuracy of the HMM-based methods is undermined. Nevertheless, the HMM expectation method still outperforms the baseline methods. It indicates that the

estimated parameters can be used for queue length estimation under different penetration rates. Compared to the performance when the parameters are given, which is shown in Figure 2.6, the HMM-based methods here achieve almost the same accuracy when the penetration rate is higher than 20%.

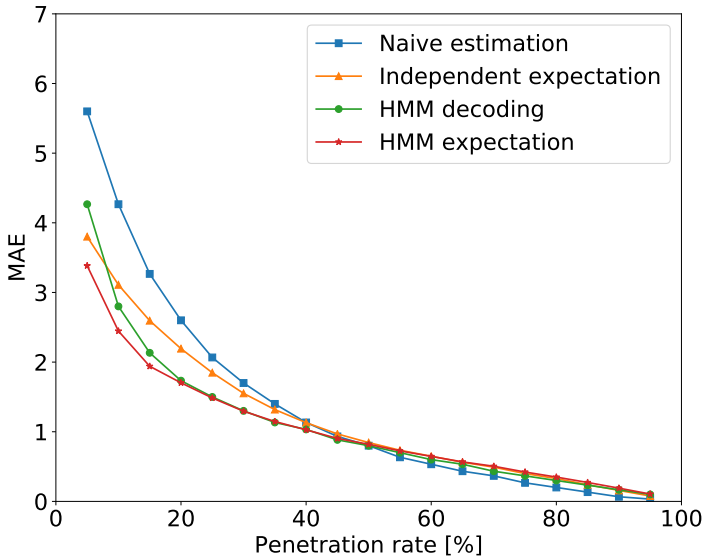


Figure 5.3: The comparison of the proposed methods and two baseline methods when the parameters of the HMM are estimated from historical data.

5.4.4 The impact of sample size

To investigate the impact of sample size on the estimation results, we use 5, 15, 30, and 60 days of historical probe vehicle data to estimate the parameters of the HMM and then estimate the queue length using the estimated parameters. Figures 5.4(a) and 5.4(b) show the performance of the two HMM-based queue length estimation methods, respectively. The results suggest that the more historical data we use, the better the estimation results will be. However, it is worth noting that the effect of the sample size would be marginal if more than 30 days of data are used.

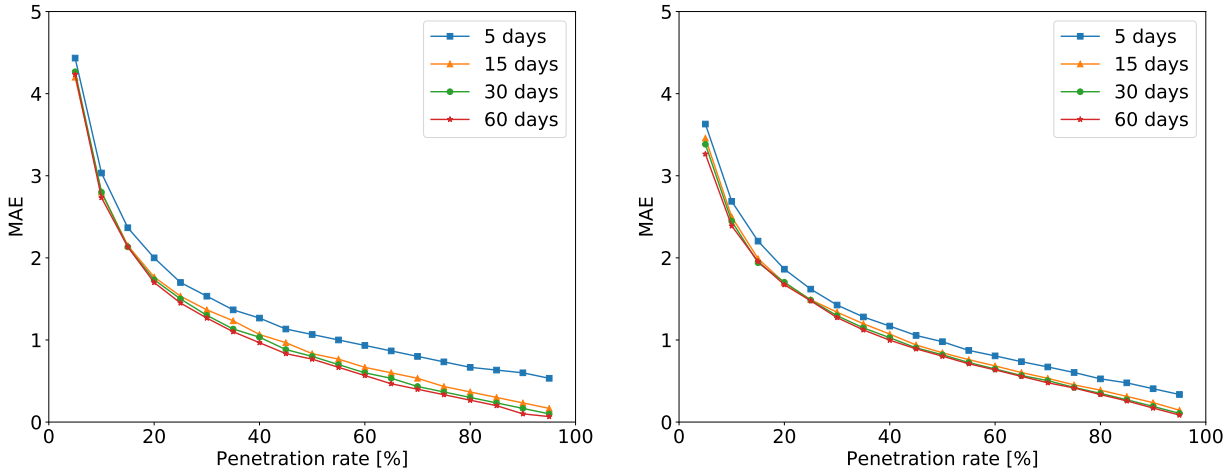


Figure 5.4: The impact of sample size when the parameters of the HMM are estimated from historical data: (a) maximum likelihood estimation (decoding); and (b) expectation conditional on sequential observations.

5.5 Conclusions

The dependent queue scenario is common in the real world. The correlation of the queues in different traffic signal cycles may come from two sources: the existence of overflow queues or the dependence of the numbers of arrivals in different cycles. However, this kind of scenario has not been well studied in the context of probe vehicles by the existing literature. In Chapter 2, we proposed a hidden Markov model to model the queueing process and the observation process in probe vehicle environments. The hidden states of the HMM are the queue lengths, and the observations are the stopping positions of probe vehicles. The HMM is governed by the transition probabilities, initial probabilities, and the penetration rate of probe vehicles. Based on the HMM, we proposed two cycle-by-cycle queue length estimation methods, which require the knowledge of the parameters mentioned above.

Since the parameters are not available beforehand in real life, in this chapter, we focus on estimating the required parameters using historical probe vehicle data. We estimate the parameters by applying maximum likelihood estimation to the queue sequences collected in multiple cycles. Because the corresponding log-likelihood function is non-concave, we

resort to the EM algorithm. The update rules generated by the EM algorithm are not tractable mathematically. Therefore, we convert the update rules by applying dynamic programming techniques. We validate the proposed estimation method using a case study of overflow queues. The validation results show that the proposed parameter estimation algorithm can adequately learn the parameters of the HMM from historical probe vehicle data, and the proposed cycle-by-cycle queue length estimation methods still outperform the baseline methods even with the estimated parameters. Finally, we analyze the effects of the penetration rate and sample size on estimation accuracy.

Chapter 6

Traffic volume estimation by data fusion

6.1 Introduction

6.1.1 Background

Traffic volume information plays a critical role in transportation planning, roadway design, and traffic signal control. In conventional transportation systems, traffic volumes are primarily measured by fixed-location sensors, such as loop detectors. Although widely applied, loop detectors have the following two drawbacks. The first drawback is that the collected data often contain missing values, which might be caused by hardware malfunction. Another drawback of loop detectors is that they usually only cover a small subset of links in a transportation network, due to the high installation and maintenance costs (Zhan et al., 2016). Therefore, loop detectors usually measure very limited traffic volume information, which restrains our understanding of the traffic at the network level.

To tackle the first problem, the missing data problem, abundant literature applied data imputation methods to loop detector data. The key idea of these methods is to exploit the spatiotemporal correlation of the traffic volume data. The methods can be roughly divided into three categories. The first category is based on the variants of principal component analysis (PCA), which includes probabilistic PCA, kernel probabilistic PCA (KPPCA), and Bayesian PCA (BPCA) (Qu et al., 2008, 2009; Ilin and Raiko, 2010; Li et al., 2013b; Asif et al., 2016). The second category is based on the matrix (tensor) completion. The methods in this category usually represent traffic volume data as a matrix (tensor) and impute the

traffic data by matrix (tensor) decomposition (Tan et al., 2013; Asif et al., 2016; Ran et al., 2016; Goulart et al., 2017; Chen et al., 2019a,b). The third category mainly contains data-driven machine learning methods, including neural networks (Duan et al., 2016; Zhuang et al., 2018; Chen and Levin, 2019), k-nearest neighbors (Tak et al., 2016), and CoKriging methods (Bae et al., 2018).

When it comes to the second drawback of loop detectors, the low coverage problem, using solely loop detector data is not sufficient to solve the problem. If loop detectors are not installed at the location where the traffic volume information is of our interest, the data imputation methods introduced above could not be applied, because all of the methods require at least one observed data point for each location. We introduced some probe vehicle based traffic volume methods in Section 1.4.2. For the methods without considering the spatiotemporal correlations in traffic volumes, it is difficult to achieve good accuracy with low penetration rate probe vehicle data alone. For the methods which take the correlation into account, usually, multiple data sources, such as POI data and meteorology data, are required to construct the similarity between different roads and time slots.

6.1.2 Contribution and organization of the chapter

In this chapter, we propose to simultaneously address the two challenges, namely, the missing data problem and the low coverage problem, by combining loop detector data and probe vehicle data**.** On the one hand, despite the low coverage, when loop detectors function well, they could give the complete vehicle counts at specific locations. On the other hand, although the penetration rate of probe vehicles is low currently, probe vehicles usually have broad coverage and do not suffer from the maintenance issues. Therefore, the fusion of the two data sources makes their advantages complementary to each other.

Noticing that probe vehicles can be considered as samples of the entire traffic, we first apply singular value decomposition to probe vehicle data and obtain an approximated low-rank structure of traffic volumes, which captures the spatiotemporal correlation. The low-rank

structure is then used for estimating the unknown traffic volumes. To further improve the accuracy, we propose a second data fusion method based on the framework of the probabilistic principal component analysis. The PPCA-based model finds the low-rank structure by using both probe vehicle data and loop detector data, which turns out to be more robust.

The rest of the chapter is organized as follows. In Section 6.2, we introduce the matrix representation of loop detector data and probe vehicle data. In Section 6.3, we present an SVD-based data fusion method that captures the correlation of traffic volumes by matrix factorization and reconstructs the unknown traffic volumes by minimizing the reconstruction error of the loop detector measurement. In Section 6.4, we describe the distributions of loop detector data and probe vehicle data and extend the classical PPCA framework by considering both data sources. We also elaborate on how to find the low-rank structure behind traffic data based on the model and how to reconstruct the unknown traffic volumes using the estimated parameters. In Section 6.5, we validate the proposed methods using a real-world loop detector dataset and a probe vehicle dataset generated by simulation in both the missing data scenario and low coverage scenario. Finally, we summarize this chapter in Section 6.6.

6.2 Matrix representation of loop detector data and probe vehicle data

For a specific time-of-day, suppose we are interested in the traffic volume information at d locations during N days. Denote the traffic volumes by a matrix $X \in \mathbb{R}^{d \times N}$, of which the element x_{ij} in the ith row and jth column represents the traffic volume at location i on day j. Due to the low coverage of loop detectors, the values in some rows of X may not be available. Similarly, due to the malfunction of loop detectors, some entries of X may be missing as well. Define an indicator matrix W, such that $w_{ij} = 1$ if the traffic volume data x_{ij} is not available, otherwise 0.

Suppose at the d locations of our interest, the penetration rate of probe vehicles is p. In

other words, when a vehicle is arbitrarily selected, its probability of being a probe vehicle is p. From the trajectory data of probe vehicles, we can extract the number of probe vehicles passing by each location on each day. Denote the traffic volume of probe vehicles by a matrix Y, which has the same size as X. For location i and day j, the probe vehicle volume y_{ij} is a fraction of the entire traffic volume x_{ij}. In particular, y_{ij} follows a binomial distribution $\mathcal{B}(x_{ij}, p)$.

In this chapter, we assume the penetration rate of probe vehicles in all the d locations are the same. In the real world, as Figure 3.9 suggests, the penetration rate in different places can be different. To satisfy the assumption, we can classify the locations we are interested in into several groups according to the estimated penetration rate, so that the penetration rate within each group is more or less the same.

6.3 Data fusion by singular value decomposition

Since traffic volumes are correlated spatially and temporally, the traffic volume matrix X introduced above can be approximated by a low-rank matrix (Qu et al., 2009; Tan et al., 2013; Coogan et al., 2017; Feng et al., 2018). If we know the low-rank structure, for instance, the principal components, then we can easily reconstruct the missing entries in X. However, in real life, obtaining the low-rank structure of X can be difficult, due to the missing data problem and the low coverage problem. To obtain the low-rank structure, we resort to the probe vehicle data. First, the traffic volume matrix Y of probe vehicles is complete. Second, since probe vehicles are samples of the entire traffic, Y should share a similar low-rank structure with X. Especially when the penetration rate is high or when the traffic volume is large, Y is almost a linearly scaled-down version of X.

We apply the SVD to find the low-rank representation of the probe vehicle traffic volume matrix Y. The SVD of Y is

$$Y = U\Sigma V^T = \sum_{k=1}^{d} \sigma_k u_k v_k^T. \tag{6.1}$$

The column vectors in $U = [u_1, u_2, \ldots, u_d]$ form a orthogonal basis of $\mathbb{R}^d$. $V = [v_1, v_2, \ldots, v_N]$ is also an orthogonal matrix. Σ is a diagonal matrix with nonnegative singular values σ_1, $\sigma_2, \ldots, \sigma_d$ of Y sorted in descending order on the diagonal. Because of the correlation in traffic volumes, the traffic volume vector x_n or y_n lie very close to the span of the first r orthogonal column vectors $U_r = [u_1, u_2, \ldots, u_r]$, where $r < d$. In other words, the probe vehicle traffic volume matrix Y can be approximated by

$$Y \approx \sum_{k=1}^{r} \sigma_k u_k v_k^T. \tag{6.2}$$

The approximation not only captures the main features in the original data but also discards noises hidden in the traffic volume fluctuation.

We now use the low-rank structure of Y to reconstruct the missing values in X. For each column x_n of X, we hope to find its projection $\hat{x}_n = U_r \alpha_n$ on the subspace $\mathbf{span}\{u_1, u_2, \ldots, u_r\}$ by solving the following optimization problem.

$$\text{minimize} \quad \|w_n * (x_n - U_r \alpha_n)\|_2 \tag{6.3}$$

$$\text{subject to} \quad \mathbf{0} \preceq U_r \alpha_n \preceq v_{max} \mathbf{1}, \tag{6.4}$$

where the operator "$*$" represents the Hadamard product (entry-wise product). The decision variable is α_n, which can be regarded as the coordinates of $\hat{x}_n$ in the subspace. v_{max} denotes the upper bound of traffic volumes. When the number of non-missing entries in x_n is smaller than r, we add a regularization term $\nu \|\alpha_n\|_2$ to the objective function to avoid overfitting. The optimization problem basically finds the coordinates of x_n in the subspace by minimizing the reconstruction error of the non-missing entries.

When each column of X is projected to the subspace, the estimated traffic volume matrix is then $\hat{X} = U_r[\alpha_1, \alpha_2, \ldots, \alpha_N]$. We then take the corresponding entries in $\hat{X}$ as the estimation of the missing values in X. For convenience, we abbreviate the proposed SVD-based data fusion method as SVD-DF.

6.4 Data fusion by probabilistic principal component analysis

6.4.1 PPCA fundamentals

Besides the SVD-DF model, we propose another method based on the probabilistic principal component analysis. The PPCA model (Tipping and Bishop, 1999) assumes that each d-dimensional vector x_n, which is the nth column of X, depends on a r-dimensional latent vector t_n through the following linear-Gaussian model

$$x_n = \Lambda t_n + \mu_x + \epsilon_n, \tag{6.5}$$

where the latent vector t_n follows an isotropic multivariate Gaussian distribution $\mathcal{N}(0, I)$, and Λ is a $d \times r$ projection matrix. μ_x represents the mean of all columns, and ϵ_n is a d-dimensional isotropic Gaussian noise following $\mathcal{N}(0, \sigma^2 I)$, where σ^2 indicates the magnitude of the noise. The intuition behind the formulation is that with $r < d$, the original d-dimensional sample data can be represented in a sparse way by mapping a r-dimensional vector in the latent variable space to the sample data space using the projection matrix Λ. According to the model, given the latent variable t_n, the conditional distribution of x_n is

$$x_n | t_n \sim \mathcal{N}\left(\mu_x + \Lambda t_n, \sigma^2 I\right). \tag{6.6}$$

The distribution of x_n is

$$x_n \sim \mathcal{N}\left(\mu_x, \Lambda\Lambda^T + \sigma^2 I\right). \tag{6.7}$$

6.4.2 Distribution of probe vehicle traffic volumes

The traffic volume of probe vehicles represents the number of probe vehicles passing by a location during a specific period. As mentioned in the previous sections, given the penetration rate p, the traffic volume of probe vehicles at location i on day j follows the binomial distribution $y_{ij} \sim \mathcal{B}(x_{ij}, p)$. The binomial distribution can be approximated by

a Gaussian distribution with the same mean and variance (Shiryayev, 1984). Therefore, the probe vehicle traffic volume vector y_n, the nth column of Y, approximately follows the Gaussian distribution

$$y_n \sim \mathcal{N}\left(x_n p, \operatorname{diag}\left(x_n p\left(1-p\right)\right)\right). \tag{6.8}$$

One reason why we approximate the binomial distribution using a Gaussian is that the PPCA framework applies to continuous random variables, whereas real-world traffic volumes can only take integer values. The Gaussian approximation makes it easy to consider the loop detector data and probe vehicle data together.

We further approximate the distribution of y_n for mathematical simplification. First, we substitute a prior, $\bar{x}$, for traffic volume vector x_n in the covariance of the distribution. For example, the prior can be the average traffic volume $\sum_{i=1}^{N} x_n/N$. Second, we decouple the mean and covariance by replacing $p\left(1-p\right)$ with η^2. Consequently, the approximated probability distribution of the probe vehicle traffic volume becomes

$$y_n \sim \mathcal{N}\left(x_n p, \operatorname{diag}\left(\bar{x}\eta^2\right)\right). \tag{6.9}$$

The approximation significantly reduces the mathematical complexity of the model and leads to efficient solving processes, as will be shown later.

6.4.3 Traffic volume reconstruction by data fusion

In the missing data scenario, some of the entries in X might be missing due to loop detector malfunction. The missing entries mostly show a random pattern. In the low coverage scenario, loop detectors are not installed in some locations of our interest, and the entries of the corresponding rows in X will also be empty. Our goal is to reconstruct the missing values of X in both scenarios by fusing the non-missing values of X and the probe vehicle data Y.

For each column x_n of X, we divide x_n into two parts x_n^m and x_n^o, following the notation

in Marlin (2008), where x_n^m refers to the missing part, and x_n^o represents the non-missing part. Then, the latent variables of the PPCA model are x_n^m and t_n. Given the observed data x_n^o, the parameters Λ, μ_x, σ^2 of the PPCA model can be estimated by maximum likelihood estimation. The objective of the PPCA-based method (Marlin, 2008) is to maximize the log-likelihood function

$$\log \mathcal{L}\left(\Lambda, \mu_x, \sigma^2; x_n^o\right) = \sum_{n=1}^{N} \log P_{\Lambda,\mu_x,\sigma^2}\left(x_n^o\right). \tag{6.10}$$

In this study, we incorporate probe vehicle data into the PPCA framework and propose a PPCA-based data fusion (PPCA-DF) model. The observed data include not only the non-missing loop detector data x_n^o but also the probe vehicle data y_n, $\forall n = 1, 2, \ldots N$. The latent variables remain the same, i.e., x_n^m and t_n, but the parameters of the model become $\Lambda, \mu_x, \sigma^2, p,$ and η^2. For conciseness, we denote the collection of parameters by θ. The PPCA-DF model estimates the parameters by maximizing the log-likelihood function

$$\log \mathcal{L}\left(\theta; x_n^o, y_n\right) = \sum_{n=1}^{N} \log P_\theta\left(x_n^o, y_n\right) \tag{6.11}$$

$$= \sum_{n=1}^{N} \log \left(\int \int P_\theta\left(x_n^m, x_n^o, y_n, t_n\right) dx_n^m dt_n \right). \tag{6.12}$$

The complete-data likelihood function in the marginal log-likelihood function (6.12) can be expressed as

$$P_\theta\left(x_n^m, x_n^o, y_n, t_n\right) = P_\theta\left(x_n, y_n, t_n\right) \tag{6.13}$$

$$= P_\theta\left(t_n\right) P_\theta\left(x_n | t_n\right) P_\theta\left(y_n | x_n, t_n\right) \tag{6.14}$$

$$= P_\theta\left(t_n\right) P_\theta\left(x_n | t_n\right) P_\theta\left(y_n | x_n\right). \tag{6.15}$$

Equation (6.14) is converted to equation (6.15) because y_n is independent of t_n given x_n.

The probability density functions of t_n, $x_n|t_n$, and $y_n|x_n$ under parameter θ are

$$P_\theta(t_n) = (2\pi)^{-\frac{r}{2}} e^{-\frac{1}{2}t_n^T t_n}, \tag{6.16}$$

$$P_\theta(x_n|t_n) = \left(2\pi\sigma^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2\sigma^2}(x_n-\Lambda t_n-\mu_x)^T(x_n-\Lambda t_n-\mu_x)}, \tag{6.17}$$

$$P_\theta(y_n|x_n) = \frac{1}{\sqrt{\Pi_{i=1}^d \bar{x}_i}} \left(2\pi\eta^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2}(y_n-px_n)^T\left[\mathrm{diag}\left(\bar{x}\eta^2\right)\right]^{-1}(y_n-px_n)}. \tag{6.18}$$

Note that in the missing data scenario, $\bar{x}$ in equation (6.18) can be obtained by averaging the non-missing values in each row of X. In the low coverage scenario, when there is no loop detector installed at location i, we approximate $\bar{x}_i$ by scaling up the average probe vehicle traffic volume using the penetration rate.

By substituting the probability density functions into equation (6.15), the complete-data log-likelihood function can be expressed as

$$\begin{aligned}\log P_\theta(x_n, y_n, t_n) = &-\frac{(r+2d)}{2}\log(2\pi) - \frac{1}{2}t_n^T t_n - \frac{d}{2}\log\left(\sigma^2\right) \\ &- \frac{1}{2\sigma^2}(x_n-\Lambda t_n-\mu_x)^T(x_n-\Lambda t_n-\mu_x) - \frac{1}{2}\sum_{i=1}^{d}\log\left(\bar{x}_i\eta^2\right) \\ &- \frac{1}{2}(y_n-px_n)^T\left[\mathrm{diag}\left(\bar{x}\eta^2\right)\right]^{-1}(y_n-px_n)\end{aligned} \tag{6.19}$$

Substituting equation (6.19) into equation (6.12) gives a non-concave objective function of the maximum likelihood estimation problem. Therefore, we apply the EM algorithm to solve it (Dempster et al., 1977).

6.4.4 The EM algorithm

In the E-step, we evaluate the expectation of the complete-data log-likelihood function under the posterior distribution of the latent variables given the current estimate $\theta^{(k)}$. Mathematically, the expectation can be expressed as $\mathbb{E}_{t_n,x_n^m|x_n^o,y_n;\theta^{(k)}}\left[\log\mathcal{L}(\theta; x_n, y_n, t_n)\right]$, where $\log\mathcal{L}(\theta; x_n, y_n, t_n) = \log P_\theta(x_n, y_n, t_n)$.

To get the probability density function of the posterior distribution, we first derive the

joint distribution of x_n, y_n, and t_n, which is

$$(x_n, y_n, t_n)\,;\theta^{(k)} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x^{(k)} \\ \mu_y^{(k)} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{x_n x_n}^{(k)} & \Sigma_{x_n y_n}^{(k)} & \Sigma_{x_n t_n}^{(k)} \\ \Sigma_{y_n x_n}^{(k)} & \Sigma_{y_n y_n}^{(k)} & \Sigma_{y_n t_n}^{(k)} \\ \Sigma_{t_n x_n}^{(k)} & \Sigma_{t_n y_n}^{(k)} & \Sigma_{t_n t_n}^{(k)} \end{bmatrix}\right), \tag{6.20}$$

where the covariance matrix can be expressed as

$$\begin{bmatrix} (\sigma^2)^{(k)} I + \Lambda^{(k)} \left(\Lambda^{(k)}\right)^T & p^{(k)} \left((\sigma^2)^{(k)} I + \Lambda^{(k)} \left(\Lambda^{(k)}\right)^T\right) & \Lambda^{(k)} \\ p^{(k)} \left((\sigma^2)^{(k)} I + \Lambda^{(k)} \left(\Lambda^{(k)}\right)^T\right)^T & \operatorname{diag}\left(\bar{x}(\eta^2)^{(k)}\right) + \left(p^{(k)}\right)^2 \left((\sigma^2)^{(k)} I + \Lambda^{(k)} \left(\Lambda^{(k)}\right)^T\right) & p^{(k)}\Lambda^{(k)} \\ \left(\Lambda^{(k)}\right)^T & p^{(k)} \left(\Lambda^{(k)}\right)^T & I \end{bmatrix}. \tag{6.21}$$

Then, according to the Gaussian conditional distribution formula, the conditional distribution of the latent variables x_n^m and t_n given the observed data x_n^o and y_n is still Gaussian. For conciseness, we denote the distribution of $\left(t_n, x_n^m | x_n^o, y_n; \theta^{(k)}\right)$ by $q_n^{(k)}\left(t_n, x_n^m\right)$, which is

$$q_n^{(k)}\left(t_n, x_n^m\right) : t_n, x_n^m | x_n^o, y_n; \theta^{(k)} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{t_n | x_n^o, y_n}^{(k)} \\ \mu_{x_n^m | x_n^o, y_n}^{(k)} \end{bmatrix}, \begin{bmatrix} \Sigma_{t_n | x_n^o, y_n}^{(k)} & \Sigma_{t_n x_n^m | x_n^o, y_n}^{(k)} \\ \left(\Sigma_{t_n x_n^m | x_n^o, y_n}^{(k)}\right)^T & \Sigma_{x_n^m | x_n^o, y_n}^{(k)} \end{bmatrix}\right), \tag{6.22}$$

where

$$\begin{bmatrix} \mu_{t_n | x_n^o, y_n}^{(k)} \\ \mu_{x_n^m | x_n^o, y_n}^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu_{x_n^m}^{(k)} \end{bmatrix} + \begin{bmatrix} \Sigma_{t_n x_n^o}^{(k)} & \Sigma_{t_n y_n}^{(k)} \\ \Sigma_{x_n^m x_n^o}^{(k)} & \Sigma_{x_n^m y_n}^{(k)} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o}^{(k)} & \Sigma_{x_n^o y_n}^{(k)} \\ \Sigma_{y_n x_n^o}^{(k)} & \Sigma_{y_n y_n} \end{bmatrix}^{-1} \left(\begin{bmatrix} x_n^o \\ y_n \end{bmatrix} - \begin{bmatrix} \mu_{x_n^o}^{(k)} \\ \mu_y^{(k)} \end{bmatrix}\right), \tag{6.23}$$

$$\Sigma_{t_n|x_n^o,y_n}^{(k)} = \Sigma_{t_n t_n}^{(k)} - \begin{bmatrix} \Sigma_{t_n x_n^o}^{(k)} & \Sigma_{t_n y_n}^{(k)} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o}^{(k)} & \Sigma_{x_n^o y_n}^{(k)} \\ \Sigma_{y_n x_n^o}^{(k)} & \Sigma_{y_n y_n}^{(k)} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{t_n x_n^o}^{(k)} & \Sigma_{t_n y_n}^{(k)} \end{bmatrix}^T, \tag{6.24}$$

$$\Sigma_{t_n x_n^m|x_n^o,y_n}^{(k)} = \Sigma_{t_n x_n^m}^{(k)} - \begin{bmatrix} \Sigma_{t_n x_n^o}^{(k)} & \Sigma_{t_n y_n}^{(k)} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o}^{(k)} & \Sigma_{x_n^o y_n}^{(k)} \\ \Sigma_{y_n x_n^o}^{(k)} & \Sigma_{y_n y_n}^{(k)} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{x_n^m x_n^o}^{(k)} & \Sigma_{x_n^m y_n}^{(k)} \end{bmatrix}^T, \tag{6.25}$$

$$\Sigma_{x_n^m|x_n^o,y_n}^{(k)} = \Sigma_{x_n^m x_n^m}^{(k)} - \begin{bmatrix} \Sigma_{x_n^m x_n^o}^{(k)} & \Sigma_{x_n^m y_n}^{(k)} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o}^{(k)} & \Sigma_{x_n^o y_n}^{(k)} \\ \Sigma_{y_n x_n^o}^{(k)} & \Sigma_{y_n y_n}^{(k)} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{x_n^m x_n^o}^{(k)} & \Sigma_{x_n^m y_n} \end{bmatrix}^T. \tag{6.26}$$

Finally, we evaluate the expected complete-data log-likelihood function under the posterior distribution $q_n^{(k)}(t_n, x_n^m)$, which is

$$\begin{aligned}
&\mathbb{E}_{q_n^{(k)}}\left[\log \mathcal{L}\left(\theta; x_n, y_n, t_n\right)\right] \\
=&\int\int q_n^{(k)}\left(t_n, x_n^m\right)\left(\log P\left(t_n\right) + \log P\left(x_n|t_n\right) + \log P\left(y_n|x_n\right)\right) dx_n^m dt_n \\
=&-\frac{(r+2d)}{2}\log\left(2\pi\right) - \frac{1}{2}\mathbb{E}_{q_n^{(k)}}\left[t_n^T t_n\right] - \frac{d}{2}\log\left(\sigma^2\right) \\
&-\frac{1}{2\sigma^2}\mathbb{E}_{q_n^{(k)}}\left[\left(x_n - \Lambda t_n - \mu_x\right)^T\left(x_n - \Lambda t_n - \mu_x\right)\right] - \frac{1}{2}\sum_{i=1}^{d}\log\left(\bar{x}_i\eta^2\right) \\
&-\frac{1}{2}\mathbb{E}_{q_n^{(k)}}\left[\left(y_n - p x_n\right)^T\left[\mathrm{diag}\left(\bar{x}\eta^2\right)\right]^{-1}\left(y_n - p x_n\right)\right].
\end{aligned} \tag{6.27}$$

In the M-step, taking into account all the available loop detector data and probe vehicle data, we maximize the sum of the expectation in terms of the parameters θ, which is

$$Q\left(\theta; \theta^{(k)}\right) = \sum_{n=1}^{N}\mathbb{E}_{q_n^{(k)}}\left[\log \mathcal{L}\left(\theta; x_n, y_n, t_n\right)\right]. \tag{6.28}$$

The solutions to the optimization problem are

$$\mu_x^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n\right] - \Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]\right), \tag{6.29}$$

$$\Lambda^{(k+1)} = \left(\sum_{n=1}^{N}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n t_n^T\right] - \mu_x^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]^T\right)\right)\left(\sum_{n=1}^{N}\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right]\right)^{-1}, \tag{6.30}$$

$$\begin{aligned}\left(\sigma^2\right)^{(k+1)} = \frac{1}{Nd}\sum_{n=1}^{N}\Big(&\mathrm{tr}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]\right) + \left(\mu_x^{(k)}\right)^T\mu_x^{(k)} + \mathrm{tr}\left(\left(\Lambda^{(k)}\right)^T\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right]\right)\\ &-2\left(\mu_x^{(k)}\right)^T\mathbb{E}_{q_n^{(k)}}\left[x_n\right] - 2\mathrm{tr}\left(\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[x_n t_n^T\right]^T\right) + 2\left(\mu_x^{(k)}\right)^T\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]\Big),\end{aligned} \tag{6.31}$$

$$p^{(k+1)} = \left(\sum_{n=1}^{N} y_n^T \mathrm{diag}\left(\bar{x}\left(\eta^2\right)^{(k)}\right)^{-1}\mathbb{E}_{q_n^{(k)}}\left[x_n\right]\right)\left(\sum_{n=1}^{N}\mathrm{tr}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]\mathrm{diag}\left(\bar{x}\left(\eta^2\right)^{(k)}\right)^{-1}\right)\right)^{-1}, \tag{6.32}$$

$$\left(\eta^2\right)^{(k+1)} = \frac{1}{Nd}\sum_{n=1}^{N}\sum_{i=1}^{d}\frac{1}{\bar{x}_i}\left(y_{ni}^2 - 2p^{(k)}y_{ni}\mathbb{E}_{q_n^{(k)}}\left[x_n\right]_i + \left(p^{(k)}\right)^2\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]_{ii}\right). \tag{6.33}$$

The solutions are concisely expressed in terms of expectations derived from equation (6.22). The detailed solving process and the expressions of the expectations are in Appendix D.

6.4.5 Estimating the unknown traffic volumes

Performing the update rules given above iteratively leads to the convergence of the estimated θ. With the estimated parameters of the PPCA-DF model, the posterior predictive distribution of the missing data is a Gaussian distribution given by

$$x_n^m | x_n^o, y_n \sim \mathcal{N}\left(\mu_{x_n^m|x_n^o,y_n}, \Sigma_{x_n^m|x_n^o,y_n}\right), \tag{6.34}$$

where

$$\mu_{x_n^m|x_n^o,y_n} = \mu_{x_n^m} + \begin{bmatrix} \Sigma_{x_n^m x_n^o} & \Sigma_{x_n^m y} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o} & \Sigma_{x_n^o y_n} \\ \Sigma_{y_n x_n^o} & \Sigma_{y_n y_n} \end{bmatrix}^{-1} \left(\begin{bmatrix} x_n^o \\ y_n \end{bmatrix} - \begin{bmatrix} \mu_{x_n^o} \\ \mu_y \end{bmatrix} \right), \quad (6.35)$$

$$\Sigma_{x_n^m|x_n^o,y_n} = \Sigma_{x_n^m x_n^m} - \begin{bmatrix} \Sigma_{x_n^m x_n^o} & \Sigma_{x_n^m y_n} \end{bmatrix} \begin{bmatrix} \Sigma_{x_n^o x_n^o} & \Sigma_{x_n^o y_n} \\ \Sigma_{y_n x_n^o} & \Sigma_{y_n y_n} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{x_n^m x_n^o} & \Sigma_{x_n^m y_n} \end{bmatrix}^T. \quad (6.36)$$

We can estimate the unknown traffic volumes in the nth column of X by the mean $\mu_{x_n^m|x_n^o,y_n}$.

6.5 Case studies

In this section, we first introduce the ground-truth dataset we use for validation and how we generate the input to the models from the ground truth. Then, we validate the proposed methods using the generated input in different scenarios and compare their performance with the baseline methods.

6.5.1 Ground-truth dataset

To evaluate the performance of the proposed traffic volume estimation methods, we need a ground-truth traffic volume dataset. The ground-truth dataset used here is the PORTAL Arterial Data (https://portal.its.pdx.edu/fhwa) collected from the loop detectors on 82nd Ave in Portland, Oregon. We aggregate the data to 15-min intervals in the preprocessing stage. The specific 15 loop detectors we use are of IDs 253, 254, 255, 256, 409, 410, 411, 412, 414, 415, 416, 712, 713, 714, and 715. We use the data of 15 workdays spanning from October 21 to November 10, 2011. Figure 6.1 shows the average traffic volumes over the 15 workdays collected by the 15 loop detectors. In general, the traffic volumes at different locations fluctuate in a similar trend, which implies strong correlations.

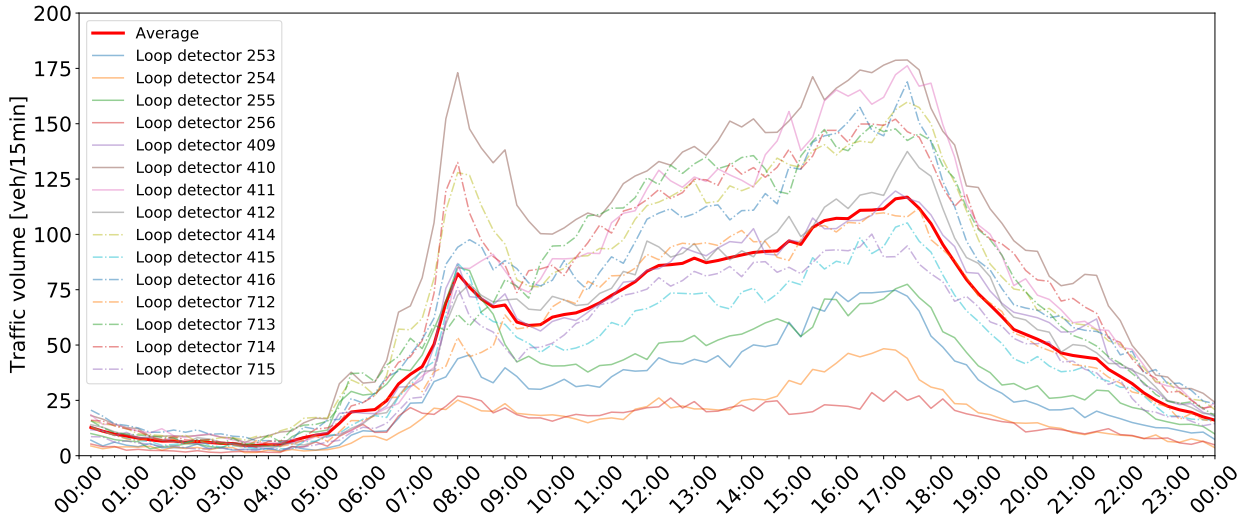


Figure 6.1: The average traffic volumes at the 15 locations.

6.5.2 Experimental settings

Probe vehicle data

Since we do not have access to the real-world probe vehicle data collected from the studied locations, for validation purposes, we generate the probe vehicle data through sampling. For a given penetration rate p and a specific TOD, the probe vehicle traffic volume y_{ij} at location i on day j is sampled from the ground-truth whole-population traffic volume. Sampling from the binomial distribution is equivalent to performing Bernoulli trials for all the vehicles to determine if they are probe vehicles or regular vehicles. After this step, we obtain the simulated probe vehicle traffic volume matrix Y.

Loop detector data with missing entries

We simulate two missing data patterns to characterize the two different scenarios of our interest, namely, the missing data scenario and the low coverage scenario. For the missing data scenario, given a missing ratio, we perform a Bernoulli trial to decide if each entry in the ground-truth traffic volume matrix is missing or not. The process simulates missing data caused by the occasional loop detector malfunction. For the low coverage scenario,

we randomly remove several rows of the ground-truth traffic volume matrix. This process simulates the situation where loop detectors only cover a subset of locations. After this step, we obtain the simulated loop detector traffic volume matrix X. In this case study, for each TOD, the size of X and Y is 15×15, as there are 15 loop detectors and 15 days.

Measure of accuracy

We evaluate the performance of the proposed methods by the root mean square error (RMSE). Only the missing entries are taken into account when calculating the RMSE. For both methods, we reconstruct the traffic volumes for each TOD separately and then combine all the results to calculate the overall performance measure.

6.5.3 Results of the missing data scenario

Figure 6.2 illustrates the estimation process in the missing data scenario. The input data include the loop detector data X with randomly missing entries and the probe vehicle traffic volume matrix Y. Using the SVD-DF model or the PPCA-DF model, we can reconstruct the missing traffic volumes. Since this case study is concerned with traffic volume estimation for 15-min intervals, we set $r = 1$ for both methods. For a larger interval such as 60 min, increasing the rank might give rise to better performance.

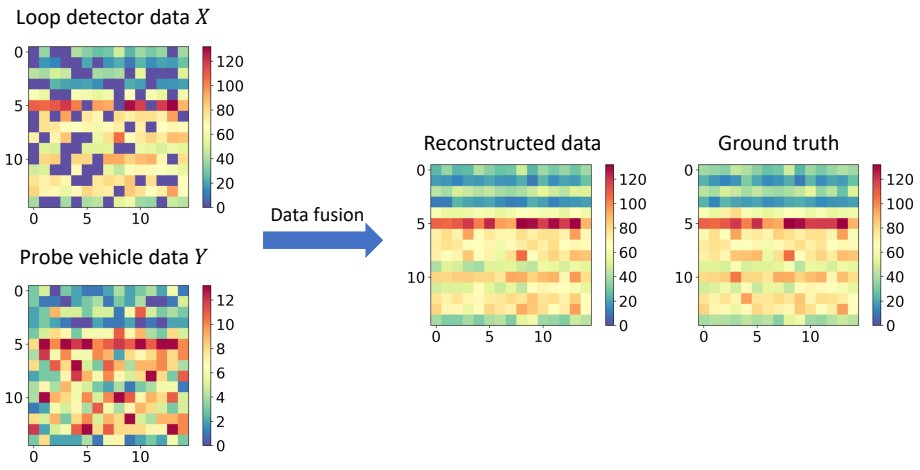
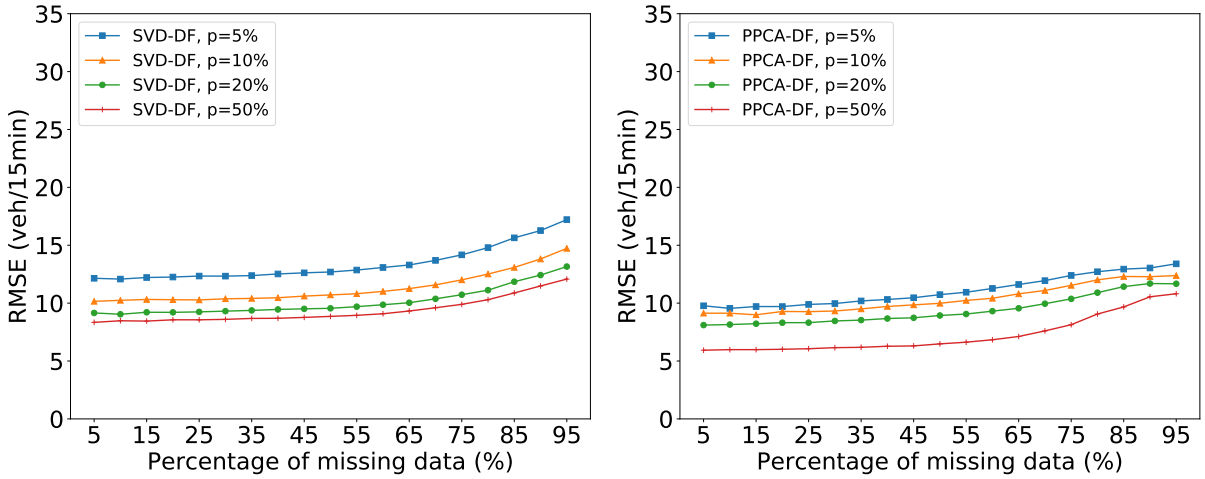


Figure 6.2: Traffic volume reconstruction for the missing data scenario.

The impact of missing ratios and penetration rates

We enumerate the missing ratio from 5% to 95%, with a step size of 5%. At the same time, we test the methods under different penetration rates, including 1%, 5%, 20%, and 50%. Figures 6.3(a) and 6.3(b) show the estimation results of the two proposed methods under different conditions, respectively. In general, for both models, the estimation accuracy decreases as the missing ratio increases. It is because when the missing ratio is low, non-missing entries can provide sufficient information for us. By contrast, when the missing ratio is high, the number of remaining entries is very limited, which results in inaccuracy when estimating the unknown traffic volumes.

The probe vehicle penetration rate is another critical parameter that influences the performance of the methods. The results indicate that the proposed methods can already reconstruct traffic volumes accurately when the penetration rate is only 10%. However, with a higher penetration rate, the spatiotemporal correlation of the traffic volume data can be better retained in the probe vehicle data; therefore, the performance is even better. Considering the practical applicability, we will use the 10% penetration rate in the following experiments.



(a) (b)

Figure 6.3: Performance of the two models under different penetration rates and different missing ratios in the missing data scenario: (a) SVD-DF and (b) PPCA-DF.

Comparison with the existing methods

We compare the proposed methods with two baseline methods. The first baseline method is the direct scaling method, which reconstructs the unknown traffic volumes by scaling up the traffic volumes of the probe vehicles using the penetration rate directly. The second method is the probabilistic principal component analysis used by Qu et al. (2009) and Li et al. (2013b), which captures the low-rank structure by solving a maximum likelihood estimation problem. Figure 6.4 shows the comparison results.

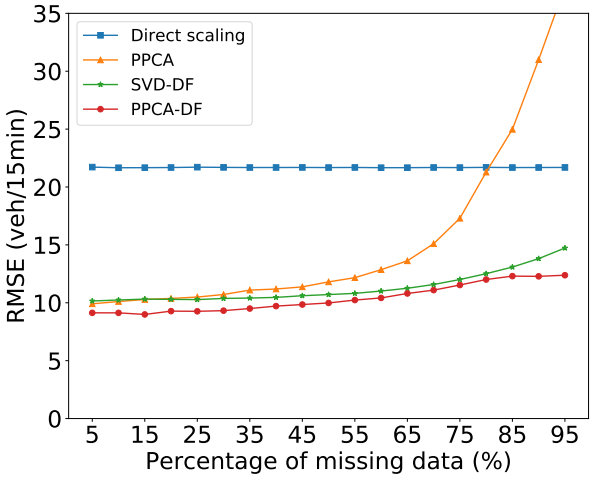


Figure 6.4: Comparison of different methods in the missing data scenario.

As the results suggest, the proposed methods consistently outperform the direct scaling method. The reason is that scaling up the probe vehicle traffic volume directly will amplify its variance, especially when the penetration rate is not high enough. However, the direct scaling method does not use any information from other locations and time slots to reduce the variance. The proposed methods also yield better performance than the PPCA method for most missing ratios. Especially when the missing ratio is high, the PPCA baseline method cannot reconstruct the missing values accurately with limited information. It implies that the probe vehicle data is an appropriate data source for finding the embedded spatiotemporal correlations. The results validate the idea that incorporating probe vehicle data can provide a robust approach to the reconstruction of traffic volumes.

Figure 6.5 shows the errors of the methods in different TODs. The figure corresponds to the scenario when the percentage of missing data is 50%. The proposed methods outperform the baseline methods in almost all the TODs. The performance of the PPCA-DF model slightly outperforms the SVD-DF model. The RMSE is smaller in the night time compared to the day time. It is because the ground-truth traffic volumes are much smaller in the night time, as shown in Figure 6.1. The ratio between the error and the ground truth is actually larger in the night time, due to the smaller sample size of probe vehicle data.

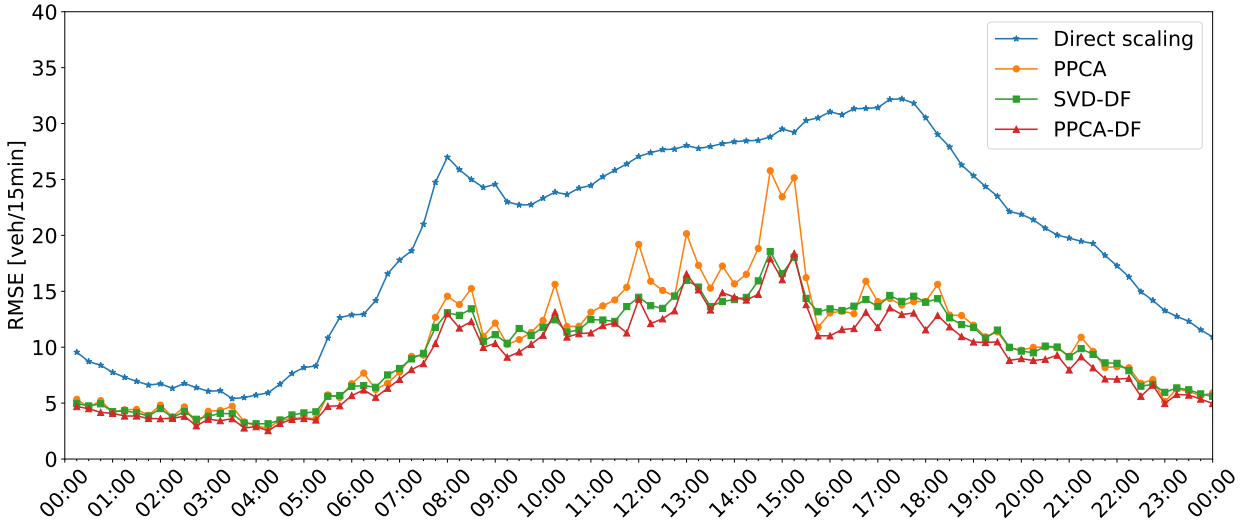


Figure 6.5: The accuracy of different methods in different TODs.

6.5.4 Results of the low coverage scenario

The low coverage scenario is more challenging. In this case, not all the locations we study are covered by loop detectors. In other words, an entire row of the traffic volume matrix X can be missing. Figure 6.6 illustrates the whole process of traffic volume reconstruction in the low coverage scenario. Similar to the missing data scenario, the input data include the loop detector data X with missing rows and the probe vehicle data Y. The proposed methods estimate the unknown traffic volumes by fusing the two data sources.

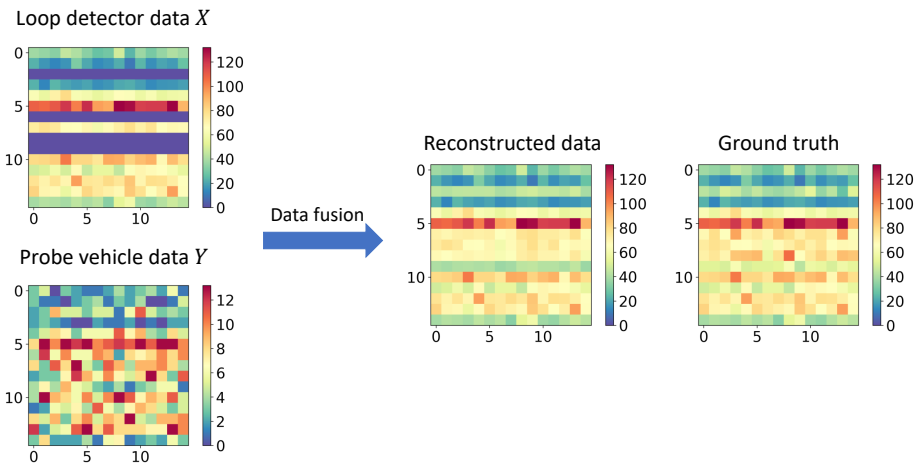
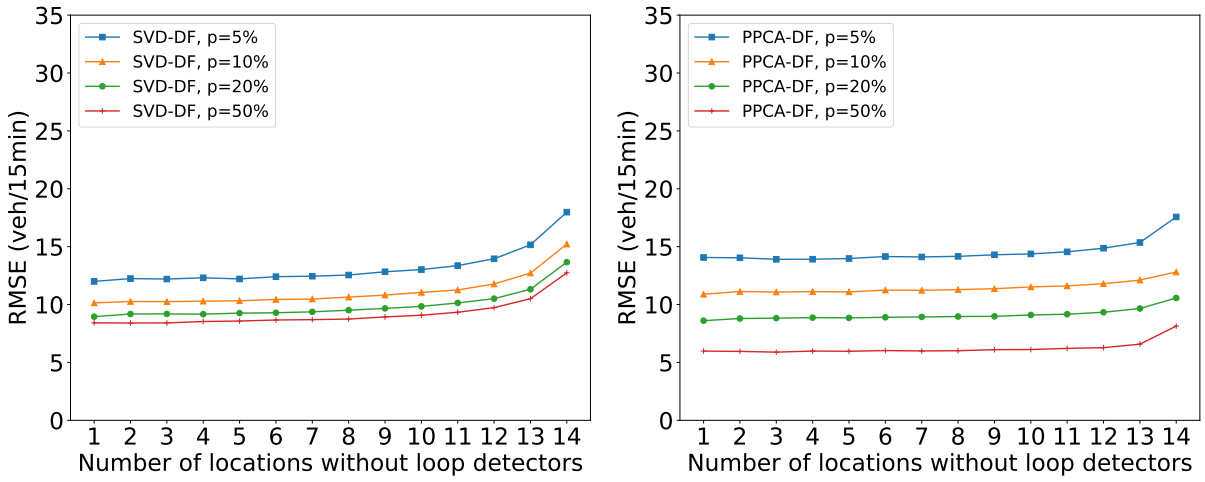


Figure 6.6: Traffic volume reconstruction for the low coverage scenario.

The impact of missing ratios and penetration rates

Figures 6.8(a) and 6.8(b) show the estimation results of the two proposed methods under different missing ratios and penetration rates. Similar to the missing data scenario, for both methods, a lower missing ratio or a higher penetration rate leads to better estimation accuracy. Even in the scenario where multiple locations are not covered with loop detectors, a 10% penetration rate can still enable the proposed methods to reconstruct traffic volumes accurately. Again, considering practical conditions, we use the 10% penetration rate in the following experiments to compare our methods with the benchmark.



(a) (b)

Figure 6.7: Performance of the two models under different penetration rates and different numbers of missing rows in the low coverage scenario: (a) SVD-DF and (b) PPCA-DF.

Comparison with the existing method

Since the PPCA method cannot deal with the low coverage scenario, we only use the direct scaling method as the baseline method. The comparison results are shown in Figure 6.8. From the results, we can see that the performance of the SVD-DF method and the PPCA-DF method is very close. When the number of locations without loop detectors is small, the SVD-DF model outperforms PPCA-DF slightly; when the number of missing rows is large, PPCA-DF performs better instead. Compared to the direct scaling method, both of the proposed methods perform better significantly. It is because the proposed methods consider the spatiotemporal correlation in the traffic volume data, whereas the direct scaling method considers each location and each time slot independently.

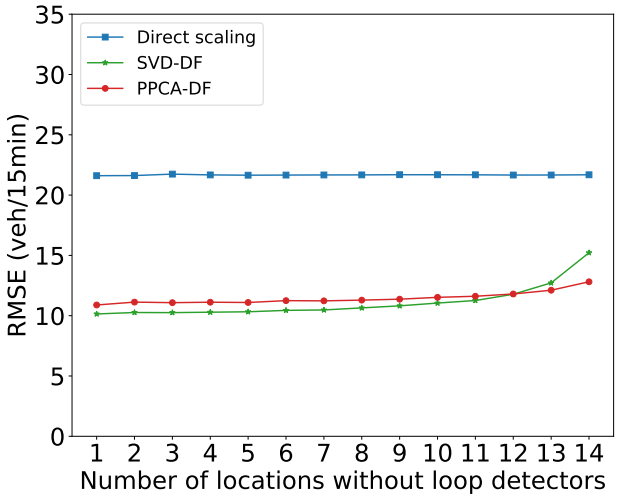


Figure 6.8: Comparison of different methods in the low coverage scenario.

Figure 6.9 shows the errors of the methods in different TODs. The figure corresponds to the scenario where loop detectors only cover seven out of the 15 locations. The proposed methods outperform the baseline methods in almost all the TODs. In general, the SVD-DF model slightly outperforms the PPCA-DF model. The trend of the error in a day is similar to the missing data scenario, which is shown in Figure 6.5.

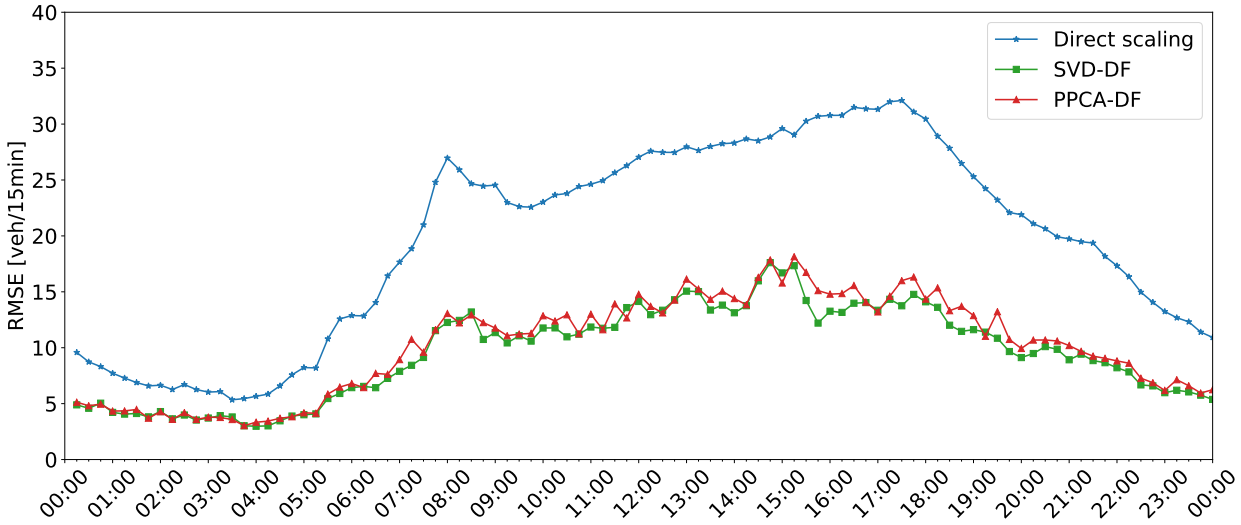


Figure 6.9: The accuracy of different methods in different TODs.

6.6 Conclusions

In this chapter, we propose two data fusion methods for traffic volume reconstruction by exploiting the low-rank structures contained in traffic data. The first method is based on the singular value decomposition. It first utilizes the probe vehicle data to approximate the low-rank structure of the loop detector data and then leverage the non-missing loop detector data to carry out the traffic volume reconstruction. The second method is a probabilistic model, which extends the framework of the probabilistic principal component analysis to include both loop detector data and probe vehicle data in the formulation. The proposed methods offer a unified framework to deal with the two challenges of loop detectors, namely, the missing data problem and the low coverage problem.

We examine the performance of the SVD-DF and PPCA-DF methods using a real-world loop detector dataset. The results show that both methods can achieve excellent performance when dealing with the two problems. The validation results also demonstrate that the proposed methods outperform the baseline methods, even when the penetration rate of probe vehicles is only 10%. Therefore, the proposed methods have enormous potentials for practical applications.

Chapter 7

Summary and future directions

7.1 Summary of the thesis

The tremendous amount of probe vehicle data collected from connected vehicles, ride-hailing vehicles, and vehicles using online navigation systems provide us a new perspective on traffic state estimation. Although there has been extensive literature on probe vehicle based travel time and travel speed estimation, the estimation of queue lengths and traffic volumes has not been well studied. This thesis aims to develop innovative traffic state estimation methodologies that can be implemented in real life, with a focus on queue length and traffic volume estimation. Specifically, this thesis has mainly covered the following topics.

First, we studied the estimation of queue lengths, under the assumption that queues in different traffic signal cycles are independent and identically distributed. Some existing studies already pointed out a few methods for estimating queue lengths cycle by cycle under this assumption. However, the parameters needed by the methods, including the penetration rate of probe vehicles and the queue length distribution, are not available beforehand in real life, which makes the existing methods hard to be implemented on a large scale. To overcome the limitations, we proposed a series of approximate estimators and a maximum likelihood estimator to estimate the required parameters. We validated the proposed methods using both simulation and real-world datasets, which showed that the proposed methods could achieve good estimation accuracy even with low-penetration-rate probe vehicle data.

Second, by relaxing the i.i.d. assumption usually imposed by relevant literature, we systematically studied the estimation of queue lengths when the queues in different cycles are

correlated. The correlation of queue lengths is a common phenomenon in the real world; however, it was ignored by most literature studying probe vehicle based traffic state estimation. Studying the estimation of queue lengths in such scenarios is of both theoretical and practical significance. We approached the problem by modeling the queueing process and observation process using a hidden Markov model. Based on the hidden Markov model, we were able to estimate queue lengths cycle by cycle in the non-i.i.d. case and estimated all the required parameters of the HMM from historical probe vehicle data. It turned out that considering the correlation of different cycles could improve the queue length estimation accuracy.

In the third part of the thesis, we focused on the estimation of traffic volumes, which are critical for ITS applications as well. Noticing that traffic volumes are correlated not only in the temporal dimension but also in the spatial dimension, we proposed to extract the hidden structure behind traffic volumes by applying low-rank representation techniques. The proposed SVD-DF and PPCA-DF models both enabled us to capture the correlation and estimate the unknown traffic volumes. Validation results showed that the proposed methods are promising for real-world applications.

In summary, this thesis presented a series of methods for probe vehicle based traffic state estimation. The proposed methods paved the way for some critical ITS applications based on probe vehicle data, which can help us better understand real-world urban traffic and solve traffic problems.

7.2 Future directions

This thesis not only provides a series of methodologies and solutions to critical traffic state estimation problems but also points out several research questions that should be addressed in the future.

First, most existing research on probe vehicle data based traffic state estimation usually only provide estimation methods; however, the theoretical limit of the probe vehicle data for

ITS applications has not been studied systematically. Considering that the penetration rate of probe vehicles in most places is still low currently, if we replace the fixed-location sensors with probe vehicle data, the reliability of the new system should be evaluated.

Second, besides probe vehicle data and fixed-location sensor data studied in this thesis, there are diverse data sources that can be used for traffic monitoring. For example, a traveler may report an accident and complain about the traffic congestion on social media; a sudden change of weather may imply a change in traffic demand. Transportation management agencies may own different types of data as well, such as traffic counts, image data, trajectory data, and incident reports. Although there is abundant literature on traffic condition inference from a single data source, how to take advantage of different data sources and improve the estimation quality has not been well studied. Another related research question is about the optimal deployment of traffic sensors. As different data sources have different strengths and weaknesses, given a limited budget, a transportation agency may need to deploy a combination of sensors and collect traffic data from multiple sources. A thorough understanding of data fusion would allow transportation agencies to make better decisions on investment.

Third, the use of probe vehicle data sometimes raises privacy concerns, as the data may contain some privacy-sensitive information. In nature, ITS applications only require the knowledge of the overall traffic, regardless of personal mobility patterns. Ideally, if we preprocess probe vehicle data appropriately, the processed data should still pertain useful traffic information while getting ride of the privacy-sensitive information. A systematic study of privacy-aware methods for processing probe vehicle data is needed.

Fourth, traffic state estimation is just one step in the process of solving traffic problems. The development of probe vehicle technologies also brings tremendous opportunities to other ITS applications. There have been some efforts on other probe vehicle based applications, such as map generation (Ahmed et al., 2015), curbside management (He et al., 2018), and adaptive traffic signal control (Feng et al., 2015; Zheng and Liu, 2020). Nevertheless, the potential of probe vehicle data has not been fully explored. With the growth of probe vehicle

market penetration, we expect to see probe vehicle data play an even more important role in solving real-world traffic problems.

Appendix A

Proof of the observable queue theorems

Definitions

For $k, n \in \mathbb{N}$ and $n \geq k$,

$$C_n^k = \frac{n!}{k!(n-k)!}, \tag{A.1}$$

$$A_n^k = \frac{n!}{(n-k)!}. \tag{A.2}$$

Theorem 1

For conciseness, l_i, n_i, s_i, t_i are represented by l, n, s, t, respectively.

$$\mathbb{E}(s \mid n, l) = \frac{l+1}{n+1}, \tag{A.3}$$

$$\mathbb{E}(l \mid n) = \mathbb{E}(s \mid n)(n+1) - 1, \tag{A.4}$$

where $n \geq 1$.

Proof:

$$\mathbb{E}(s \mid n, l) = \sum_{s=1}^{l-n+1} P(s \mid n, l)s \tag{A.5}$$

$$= \sum_{s=1}^{l-n+1} \frac{nC_{l-n}^{s-1} A_{s-1}^{s-1} A_{l-s}^{l-s}}{A_l^l} s \tag{A.6}$$

$$= \sum_{s=1}^{l-n+1} \frac{nA_{l-s}^{n-1}}{A_l^n} s \tag{A.7}$$

$$= \frac{n}{A_l^n} \sum_{s=1}^{l-n+1} A_{l-s}^{n-1} s \tag{A.8}$$

$$= \frac{n}{A_l^n} \sum_{k=0}^{l-n} A_{n+k-1}^{n-1} (l-n+1-k) \tag{A.9}$$

$$= \frac{n}{A_l^n} \sum_{k=0}^{l-n} A_{n+k-1}^{n-1} (l+1) - \frac{n}{A_l^n} \sum_{k=0}^{l-n} A_{n+k-1}^{n-1} (n+k) \tag{A.10}$$

$$= (l+1) \sum_{k=0}^{l-n} \frac{(n+k-1)!(l-n)!n!}{k!l!(n-1)!} - \frac{n}{A_l^n} \sum_{k=0}^{l-n} A_{n+k}^n \tag{A.11}$$

$$= \frac{l+1}{C_l^n} \sum_{k=0}^{l-n} C_{n+k-1}^{n-1} - \frac{n}{C_l^n} \sum_{k=0}^{l-n} C_{n+k}^n \tag{A.12}$$

$$= (l+1) \frac{C_l^n}{C_l^n} - n \frac{C_{l+1}^{n+1}}{C_l^n} \tag{A.13}$$

$$= (l+1) - n \frac{l+1}{n+1} \tag{A.14}$$

$$= \frac{l+1}{n+1}. \tag{A.15}$$

Chu's theorem (Merris, 2003) is applied when converting equation (A.12) to equation (A.13).

Then, based on the results above,

$$\mathbb{E}(s \mid n) = \sum_{s=1}^{L_{max}} P(s \mid n) s \tag{A.16}$$

$$= \sum_{s=1}^{L_{max}} \sum_{l=s+n-1}^{L_{max}} P(s \mid n, l) P(l \mid n) s \tag{A.17}$$

$$= \sum_{l=n}^{L_{max}} \sum_{s=1}^{l-n+1} P(s \mid n, l) P(l \mid n) s \tag{A.18}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n) \sum_{s=1}^{l-n+1} P(s \mid n, l) s \tag{A.19}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n) \mathbb{E}(s \mid n, l) \tag{A.20}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n) \frac{l+1}{n+1} \tag{A.21}$$

$$= \frac{1}{n+1} \sum_{l=n}^{L_{max}} P(l \mid n)(l+1) \tag{A.22}$$

$$= \frac{1}{n+1} \left(\mathbb{E}(l \mid n) + 1 \right). \tag{A.23}$$

This is equivalent to

$$\mathbb{E}(l \mid n) = \mathbb{E}(s \mid n)(n+1) - 1. \tag{A.24}$$

Theorem 2

For conciseness, l_i, n_i, s_i, t_i are represented by l, n, s, t, respectively.

$$\mathbb{E}(t \mid n, l) = n \frac{l+1}{n+1}, \tag{A.25}$$

$$\mathbb{E}(l \mid n) = \mathbb{E}(l \mid n) \frac{n+1}{n} - 1, \tag{A.26}$$

where $n \geq 1$.

Proof:

$$\mathbb{E}(t \mid n, l) = \sum_{t=n}^{l} P(t \mid n, l) t \tag{A.27}$$

$$= \sum_{t=n}^{l} \frac{n C_{l-n}^{l-t} A_{t-1}^{t-1} A_{l-t}^{l-t}}{A_l^l} t \tag{A.28}$$

$$= \sum_{t=n}^{l} \frac{n A_{t-1}^{n-1}}{A_l^n} t \tag{A.29}$$

$$= n \sum_{t=n}^{l} \frac{A_t^n}{A_l^n} \tag{A.30}$$

$$= n \sum_{t=n}^{l} \frac{C_t^n}{C_l^n} \tag{A.31}$$

$$= \frac{n}{C_l^n} \sum_{k=0}^{l-n} C_{n+k}^n \tag{A.32}$$

$$= \frac{n C_{l+1}^{n+1}}{C_l^n} \tag{A.33}$$

$$= n\frac{l+1}{n+1}. \tag{A.34}$$

Then, based on the results above,

$$\mathbb{E}(t \mid n) = \sum_{t=n}^{L_{max}} P(t \mid n)t \tag{A.35}$$

$$= \sum_{t=n}^{L_{max}} \sum_{l=t}^{L_{max}} P(t \mid n, l)P(l \mid n)t \tag{A.36}$$

$$= \sum_{l=n}^{L_{max}} \sum_{t=n}^{l} P(t \mid n, l)P(l \mid n)t \tag{A.37}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n) \sum_{t=n}^{l} P(t \mid n, l)t \tag{A.38}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n)\mathbb{E}(t \mid n, l) \tag{A.39}$$

$$= \sum_{l=n}^{L_{max}} P(l \mid n)n\frac{l+1}{n+1} \tag{A.40}$$

$$= \frac{n}{n+1} \sum_{l=n}^{L_{max}} P(l \mid n)(l+1) \tag{A.41}$$

$$= \frac{n}{n+1} \left(\mathbb{E}(l \mid n) + 1\right). \tag{A.42}$$

This is equivalent to

$$\mathbb{E}(l \mid n) = \mathbb{E}(t \mid n)\frac{n+1}{n} - 1. \tag{A.43}$$

Theorem 3

For conciseness, l_i, n_i, s_i, t_i are represented by l, n, s, t, respectively.

$$\mathbb{E}(l \mid n \geq 1) = \mathbb{E}(s \mid n \geq 1) + \mathbb{E}(t \mid n \geq 1) - 1. \tag{A.44}$$

Proof:

First of all,

$$P(t = l - s + 1 \mid n \geq 1, l) = p(1-p)^{l-(l-s+1)} \tag{A.45}$$

$$= p(1-p)^{s-1} \tag{A.46}$$

$$= P(s \mid n \geq 1, l), \text{ if } 1 \leq s \leq l. \tag{A.47}$$

Then,

$$\mathbb{E}(s \mid n \geq 1) = \sum_{s=1}^{L_{max}} P(s \mid n \geq 1)s \tag{A.48}$$

$$= \sum_{s=1}^{L_{max}} \sum_{l=s}^{L_{max}} P(s \mid n \geq 1, l)P(l \mid n \geq 1)s \tag{A.49}$$

$$= \sum_{l=1}^{L_{max}} \sum_{s=1}^{l} P(s \mid n \geq 1, l)P(l \mid n \geq 1)s \tag{A.50}$$

$$= \sum_{l=1}^{L_{max}} \sum_{s=1}^{l} P(t = l - s + 1 \mid n \geq 1, l)P(l \mid n \geq 1)s \tag{A.51}$$

$$= \sum_{l=1}^{L_{max}} \sum_{t=1}^{l} P(t \mid n \geq 1, l)P(l \mid n \geq 1)(l - t + 1), \tag{A.52}$$

$$\mathbb{E}(t \mid n \geq 1) = \sum_{t=1}^{L_{max}} P(t \mid n \geq 1)t \tag{A.53}$$

$$= \sum_{t=1}^{L_{max}} \sum_{l=t}^{L_{max}} P(t \mid n \geq 1, l)P(l \mid n \geq 1)t \tag{A.54}$$

$$= \sum_{l=1}^{L_{max}} \sum_{t=1}^{l} P(t \mid n \geq 1, l)P(l \mid n \geq 1)t. \tag{A.55}$$

Therefore,

$$\mathbb{E}(s \mid n \geq 1) + \mathbb{E}(t \mid n \geq 1) - 1 = \sum_{l=1}^{L_{max}} \sum_{t=1}^{l} P(t \mid n \geq 1, l)P(l \mid n \geq 1)(l-1) + 1 \tag{A.56}$$

$$= \sum_{l=1}^{L_{max}} P(l \mid n \geq 1)(l-1) + 1 \tag{A.57}$$

$$= \sum_{l=1}^{L_{max}} P(l \mid n \geq 1)l \tag{A.58}$$

$$= \mathbb{E}(l \mid n \geq 1). \tag{A.59}$$

Alternatively, Theorem 3 can also be proved by combining Theorem 1 and Theorem 2.

Appendix B

Analytical solution of the EM algorithm in Chapter 4

The solutions can be calculated by constructing the Lagrangian. Alternatively, we can eliminate the equality constraint by substituting $\pi_{L_{max}} = 1 - \sum_{k=0}^{L_{max}-1} \pi_k$ into the objective function, that is,

$$Q\left(\theta;\theta^{(t)}\right) = \sum_{i=1}^{C} \sum_{j=|q_i|}^{L_{max}-1} \frac{\pi_j^{(t)} \left(n_i \log p + (j-n_i)\log(1-p) + \log \pi_j\right)}{\sum_{k=|q_i|}^{L_{max}} \left(1-p^{(t)}\right)^{k-j} \pi_k^{(t)}} \tag{B.1}$$

$$- \sum_{i=1}^{C} \frac{\pi_{L_{max}}^{(t)} \left(n_i \log p + (L_{max}-n_i)\log(1-p) + \log\left(1 - \sum_{k=0}^{L_{max}-1} \pi_k\right)\right)}{\sum_{k=|q_i|}^{L_{max}} \left(1-p^{(t)}\right)^{k-L_{max}} \pi_k^{(t)}}. \tag{B.2}$$

The first order derivatives are

$$\frac{\partial Q\left(\theta;\theta^{(t)}\right)}{\partial p} = \sum_{i=1}^{C} \sum_{j=|q_i|}^{L_{max}} \frac{\pi_j^{(t)}}{\sum_{k=|q_i|}^{L_{max}} \left(1-p^{(t)}\right)^{k-j} \pi_k^{(t)}} \left(n_i \frac{1}{p} - (j-n_i)\frac{1}{1-p}\right); \tag{B.3}$$

$$\forall j = 0, 1, 2, ..., L_{max} - 1,$$

$$\frac{\partial Q\left(\theta;\theta^{(t)}\right)}{\partial \pi_j} = \sum_{i:|q_i|\le j}^{C} \frac{\pi_j^{(t)}}{\left(\sum_{k=|q_i|}^{L_{max}} \left(1-p^{(t)}\right)^{k-j} \pi_k^{(t)}\right) \pi_j} - \sum_{i=1}^{C} \frac{\pi_{L_{max}}^{(t)}}{\left(\sum_{k=|q_i|}^{L_{max}} \left(1-p^{(t)}\right)^{k-L_{max}} \pi_k^{(t)}\right) \pi_{L_{max}}}. \tag{B.4}$$

Therefore, setting the derivatives to zero gives

$$\left.\frac{\partial Q\left(\theta;\theta^{(t)}\right)}{\partial p}\right|_{p=p^{(t+1)}}=0\Leftrightarrow p^{(t+1)}=\frac{\sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_j^{(t)}n_i}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}}{\sum_{i=1}^{C}\sum_{j=|q_i|}^{L_{max}}\frac{\pi_l^{(t)}l}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}}}, \tag{B.5}$$

$$\left.\frac{\partial Q\left(\theta;\theta^{(t)}\right)}{\partial \pi_j}\right|_{\pi_j=\pi_j^{(t+1)}}=0\Leftrightarrow E_{L_{max}}^{(t)}\pi_j^{(t+1)}=E_j^{(t)}\pi_{L_{max}}^{(t+1)},\forall j=0,1,2,\ldots,L_{max}-1, \tag{B.6}$$

where $E_j^{(t)}=\sum_{i:|q_i|\leq j}^{C}\frac{\pi_j^{(t)}}{\sum_{k=|q_i|}^{L_{max}}\left(1-p^{(t)}\right)^{k-j}\pi_k^{(t)}},\forall j\in\{0,1,\ldots,L_{max}\}$. Combining with the fact $\sum_{j=0}^{L_{max}}\pi_j^{(t+1)}=1$ gives

$$\pi_j^{(t+1)}=\frac{E_j^{(t)}}{\sum_{k=0}^{L_{max}}E_k^{(t)}},\forall j=0,1,2,\ldots,L_{max}. \tag{B.7}$$

It can be easily verified that as long as $\theta^{(t)}$ satisfies the constraints, the newly generated estimate $\theta^{(t+1)}$ will satisfy the constraints as well.

Appendix C

Analytical solution of the EM algorithm in Chapter 5

The Lagrangian can be constructed as follows

$$J\left(\theta;\lambda,\nu\right)=Q(\theta;\theta^{(t)})+\lambda\left(\sum_{l=0}^{L_{max}}\pi_l-1\right)+\sum_{j=0}^{L_{max}}\nu_j\left(T_{jk}-1\right), \tag{C.1}$$

where λ and $\nu_j,\forall j=0,1,\ldots,L_{max}$ are multipliers. The first derivatives of the Lagrangian are

$$\frac{\partial J}{\partial\pi_j}=\sum_{l:l_1=j}P\left(l\mid q;\theta^{(t)}\right)\frac{1}{\pi_j}+\lambda,\forall j=0,1,\ldots,L_{max}, \tag{C.2}$$

$$\frac{\partial J}{\partial T_{jk}}=\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i:2\leq i\leq C,l_{i-1}=j,l_i=k}\frac{1}{T_{jk}}+v_j,\forall j,k=0,1,\ldots,L_{max}, \tag{C.3}$$

$$\frac{\partial J}{\partial p}=\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i=1}^{C}\left(\frac{n_i}{p}-\frac{l_i-n_i}{1-p}\right). \tag{C.4}$$

Setting the first derivatives to zero gives the update rules

$$\pi_j^{(t+1)}=\frac{\sum_{l:l_1=j}P\left(l\mid q;\theta^{(t)}\right)}{\sum_{k=0}^{L_{max}}\sum_{k:l_1=k}P\left(l\mid q;\theta^{(t)}\right)}=\frac{\sum_{l:l_1=j}P\left(l\mid q;\theta^{(t)}\right)}{\sum_{l}P\left(l\mid q;\theta^{(t)}\right)},\forall j=0,1,\ldots,L_{max}, \tag{C.5}$$

$$T_{jk}^{(t+1)}=\frac{\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i:2\leq i\leq C,l_{i-1}=j,l_i=k}1}{\sum_{k=0}^{L_{max}}\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i:2\leq i\leq C,l_{i-1}=j,l_i=k}1},\forall j,k=0,1,\ldots,L_{max}, \tag{C.6}$$

$$p^{(t+1)}=\frac{\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i=1}^{C}n_i}{\sum_{l}P\left(l\mid q;\theta^{(t)}\right)\sum_{i=1}^{C}l_i}. \tag{C.7}$$

Appendix D

Analytical solution of the EM algorithm in Chapter 6

The solutions can be obtained by setting the derivatives of $Q\left(\theta;\theta^{(k)}\right)$ to zero.

$$\frac{\partial Q\left(\theta;\theta^{(k)}\right)}{\partial \mu_x} = \sum_{n=1}^{N} \frac{1}{\left(\sigma^2\right)^{(k)}} \left(\mathbb{E}_{q_n^{(k)}}\left[x_n\right] - \Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right] - \mu_x\right) = 0, \tag{D.1}$$

$$\frac{\partial Q\left(\theta;\theta^{(k)}\right)}{\partial \Lambda} = \sum_{n=1}^{N} \frac{1}{\left(\sigma^2\right)^{(k)}} \left(\mathbb{E}_{q_n^{(k)}}\left[\left(x_n - \mu_x^{(k)}\right) t_n^T\right] - \Lambda\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right]\right) = 0, \tag{D.2}$$

$$\frac{\partial Q\left(\theta;\theta^{(k)}\right)}{\partial \sigma^2} = \sum_{n=1}^{N} \left(\frac{d}{\sigma^2} - \frac{1}{\sigma^4}\mathbb{E}_{q_n^{(k)}}\left[\left(x_n - \Lambda^{(k)}t_n - \mu_x^{(k)}\right)^T \left(x_n - \Lambda^{(k)}t_n - \mu_x^{(k)}\right)\right]\right) = 0, \tag{D.3}$$

$$\frac{\partial Q\left(\theta;\theta^{(k)}\right)}{\partial p} = \sum_{n=1}^{N} \mathbb{E}_{q_n^{(k)}}\left[\left(\operatorname{diag}\left(\bar{x}\left(\eta^2\right)^{(k)}\right)^{-1}\left(y_n - p x_n\right)\right)^T x_n\right] = 0, \tag{D.4}$$

$$\frac{\partial Q\left(\theta;\theta^{(k)}\right)}{\partial \eta^2} = \sum_{n=1}^{N} \left(\frac{d}{\eta^2} - \frac{1}{\eta^4}\mathbb{E}_{q_n^{(k)}}\left[\left(y_n - p^{(k)}x_n\right)^T \operatorname{diag}\left(\bar{x}\right)^{-1}\left(y_n - p^{(k)}x_n\right)\right]\right) = 0. \tag{D.5}$$

Solving the equations above yields the update rules of the parameters.

$$\mu_x^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n\right] - \Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]\right), \tag{D.6}$$

$$\Lambda^{(k+1)} = \left(\sum_{n=1}^{N}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n t_n^T\right] - \mu_x^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]^T\right)\right)\left(\sum_{n=1}^{N}\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right]\right)^{-1}, \tag{D.7}$$

$$\begin{aligned}\left(\sigma^2\right)^{(k+1)} =& \frac{1}{Nd}\sum_{n=1}^{N}\left(\mathrm{tr}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]\right) + \left(\mu_x^{(k)}\right)^T\mu_x^{(k)} + \mathrm{tr}\left(\left(\Lambda^{(k)}\right)^T\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right]\right)\right. \\ & \left. -2\left(\mu_x^{(k)}\right)^T\mathbb{E}_{q_n^{(k)}}\left[x_n\right] - 2\mathrm{tr}\left(\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[x_n t_n^T\right]^T\right) + 2\left(\mu_x^{(k)}\right)^T\Lambda^{(k)}\mathbb{E}_{q_n^{(k)}}\left[t_n\right]\right), \end{aligned} \tag{D.8}$$

$$p^{(k+1)} = \left(\sum_{n=1}^{N} y_n^T \mathrm{diag}\left(\bar{x}\left(\eta^2\right)^{(k)}\right)^{-1}\mathbb{E}_{q_n^{(k)}}\left[x_n\right]\right)\left(\sum_{n=1}^{N}\mathrm{tr}\left(\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]\mathrm{diag}\left(\bar{x}\left(\eta^2\right)^{(k)}\right)^{-1}\right)\right)^{-1}, \tag{D.9}$$

$$\left(\eta^2\right)^{(k+1)} = \frac{1}{Nd}\sum_{n=1}^{N}\sum_{i=1}^{d}\frac{1}{\bar{x}_i}\left(y_{ni}^2 - 2p^{(k)}y_{ni}\mathbb{E}_{q_n^{(k)}}\left[x_n\right]_i + \left(p^{(k)}\right)^2\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right]_{ii}\right). \tag{D.10}$$

where the five expectations can be expressed as

$$\mathbb{E}_{q_n^{(k)}}\left[x_n\right] = \begin{bmatrix} \mu_{x_n^m|x_n^o,y_n}^{(k)} \\ \\ x_n^o \end{bmatrix}, \tag{D.11}$$

$$\mathbb{E}_{q_n^{(k)}}\left[t_n\right] = \mu_{t_n|x_n^o,y_n}^{(k)}, \tag{D.12}$$

$$\mathbb{E}_{q_n^{(k)}}\left[t_n t_n^T\right] = \Sigma_{t_n|x_n^o,y_n}^{(k)} + \mu_{t_n|x_n^o,y_n}^{(k)}\left(\mu_{t_n|x_n^o,y_n}^{(k)}\right)^T, \tag{D.13}$$

$$\mathbb{E}_{q_n^{(k)}}\left[x_n x_n^T\right] = \begin{bmatrix} \Sigma_{x_n^m|x_n^o,y_n}^{(k)} + \mu_{x_n^m|x_n^o,y_n}^{(k)}\left(\mu_{x_n^m|x_n^o,y_n}^{(k)}\right)^T & \mu_{x_n^m|x_n^o,y_n}^{(k)}\left(x_n^o\right)^T \\ \\ x_n^o\left(\mu_{x_n^m|x_n^o,y_n}^{(k)}\right)^T & x_n^o\left(x_n^o\right)^T \end{bmatrix}, \tag{D.14}$$

$$\mathbb{E}_{q_n^{(k)}}\left[x_n t_n^T\right] = \begin{bmatrix} \Sigma_{x_n^m t_n|x_n^o,y_n}^{(k)} + \mu_{x_n^m|x_n^o,y_n}^{(k)}\left(\mu_{t_n|x_n^o,y_n}^{(k)}\right)^T \\ \\ x_n^o\left(\mu_{t_n|x_n^o,y_n}^{(k)}\right)^T \end{bmatrix}. \tag{D.15}$$

Bibliography

Ahmed, M., Karagiorgou, S., Pfoser, D., Wenk, C., 2015. A comparison and evaluation of map construction algorithms using vehicle tracking data. GeoInformatica 19, 601–632.

Aljamal, M.A., Abdelghaffar, H.M., Rakha, H.A., 2019. Kalman filter-based vehicle count estimation approach using probe data: A multi-lane road case study, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE. pp. 4374–4379.

An, C., Wu, Y.J., Xia, J., Huang, W., 2018. Real-time queue length estimation using event-based advance detector data. Journal of Intelligent Transportation Systems 22, 277–290.

Asif, M.T., Mitrovic, N., Dauwels, J., Jaillet, P., 2016. Matrix and tensor based methods for missing data estimation in large traffic networks. IEEE Transactions on Intelligent Transportation Systems 17, 1816–1825.

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L.D., Freeze, P.B., 2018. Missing data imputation for traffic flow speed using spatio-temporal cokriging. Transportation Research Part C: Emerging Technologies 88, 124–139.

Balakrishnan, S., Wainwright, M.J., Yu, B., et al., 2017. Statistical guarantees for the EM algorithm: From population to sample-based analysis. The Annals of Statistics 45, 77–120.

Ban, X.J., Hao, P., Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C: Emerging Technologies 19, 1133–1156.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 41, 164–171.

Cetin, M., 2012. Estimating queue dynamics at signalized intersections from probe vehicle data: Methodology based on kinematic wave model. Transportation Research Record: Journal of the Transportation Research Board 2315, 164–172.

Chen, M., Chien, S.I., 2001. Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based. Transportation Research Record: Journal of the Transportation Research Board 1768, 157–161.

Chen, R., Levin, M.W., 2019. Traffic state estimation based on kalman filter technique using connected vehicle v2v basic safety messages, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE. pp. 4380–4385.

Chen, X., He, Z., Chen, Y., Lu, Y., Wang, J., 2019a. Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. Transportation Research Part C: Emerging Technologies 104, 66–77.

Chen, X., He, Z., Sun, L., 2019b. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. Transportation Research Part C: Emerging Technologies 98, 73–84.

Cheng, Y., Qin, X., Jin, J., Ran, B., 2012. An exploratory shockwave approach to estimating queue length using probe trajectories. Journal of intelligent transportation systems 16, 12–23.

Comert, G., 2013a. Effect of stop line detection in queue length estimation at traffic signals from probe vehicles data. European Journal of Operational Research 226, 67–76.

Comert, G., 2013b. Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals. Transportation Research Part B: Methodological 55, 59–74.

Comert, G., 2016. Queue length estimation from probe vehicles at isolated intersections: Estimators for primary parameters. European Journal of Operational Research 252, 502–521.

Comert, G., Cetin, M., 2009. Queue length estimation from probe vehicle location and the impacts of sample size. European Journal of Operational Research 197, 196–202.

Comert, G., Cetin, M., 2011. Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data. IEEE Transactions on Intelligent Transportation Systems 12, 563–573.

Coogan, S., Flores, C., Varaiya, P., 2017. Traffic predictive control from low-rank structure. Transportation Research Part B: Methodological 97, 1–22.

Cookson, G., 2018. INRIX Global Traffic Scorecard. Technical Report.

Cui, Y., Jin, B., Zhang, F., Han, B., Zhang, D., 2017. Mining spatial-temporal correlation of sensory data for estimating traffic volumes on highways, in: Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, ACM. pp. 343–352.

Darroch, J., 1964. On the traffic-light queue. The Annals of Mathematical Statistics , 380–388.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological) , 1–38.

Duan, Y., Lv, Y., Liu, Y.L., Wang, F.Y., 2016. An efficient realization of deep learning for traffic data imputation. Transportation Research Part C: Emerging Technologies 72, 168–181.

Dunn, M.R., Ross, H.W., Baumanis, C., Wall, J., Lammert, J., Duthie, J., Ruiz Juri, N., Machemehl, R.B., 2019. Data-driven methodology for prioritizing traffic signal retiming operations. Transportation Research Record: Journal of the Transportation Research Board 2673, 104–113.

Feng, S., Wang, X., Sun, H., Zhang, Y., Li, L., 2018. A better understanding of long-range temporal dependence of traffic flow time series. Physica A: Statistical Mechanics and its Applications 492, 639–650.

Feng, Y., Head, K.L., Khoshmagham, S., Zamanipour, M., 2015. A real-time adaptive signal control in a connected vehicle environment. Transportation Research Part C: Emerging Technologies 55, 460–473.

Forney, G.D., 1973. The Viterbi algorithm. Proceedings of the IEEE 61, 268–278.

Geroliminis, N., Haddad, J., Ramezani, M., 2012. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. IEEE Transactions on Intelligent Transportation Systems 14, 348–359.

Gordon, R.L., 2010. Traffic signal retiming practices in the United States. volume 409. Transportation Research Board.

Goulart, J.d.M., Kibangou, A., Favier, G., 2017. Traffic data imputation via tensor completion based on soft thresholding of Tucker core. Transportation Research Part C: Emerging Technologies 85, 348–362.

Guo, Q., Li, L., Ban, X.J., 2019. Urban traffic signal control with connected and automated vehicles: A survey. Transportation Research Part C: Emerging Technologies 101, 313–334.

Haight, F.A., 1959. Overflow at a traffic light. Biometrika 46, 420–424.

Hao, P., Ban, X.J., Guo, D., Ji, Q., 2014. Cycle-by-cycle intersection queue length distribution estimation using sample travel times. Transportation research part B: methodological 68, 185–204.

Hao, P., Ban, X.J., Whon Yu, J., 2015. Kinematic equation-based vehicle queue location estimation method for signalized intersections using mobile sensor data. Journal of Intelligent Transportation Systems 19, 256–272.

He, T., Bao, J., Li, R., Ruan, S., Li, Y., Tian, C., Zheng, Y., 2018. Detecting vehicle illegal parking events using sharing bikes' trajectories., in: KDD, pp. 340–349.

Heidemann, D., 1994. Queue length and delay distributions at traffic signals. Transportation Research Part B: Methodological 28, 377–389.

Igbinosun, L.I., Omosigho, S.E., 2016. Traffic flow model at fixed control signals with discrete service time distribution. Croatian Operational Research Review 7, 19–32.

Ilin, A., Raiko, T., 2010. Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research 11, 1957–2000.

Jain, P., Kar, P., et al., 2017. Non-convex optimization for machine learning. Foundations and Trends® in Machine Learning 10, 142–336.

Jenelius, E., Koutsopoulos, H.N., 2017. Urban network travel time prediction based on a probabilistic principal component analysis model of probe data. IEEE Transactions on Intelligent Transportation Systems 19, 436–445.

Kawasaki, Y., Hara, Y., Kuwahara, M., 2019. Traffic state estimation on a two-dimensional network by a state-space model. Transportation Research Part C: Emerging Technologies. doi:10.1016/j.trc.2019.03.016.

Lai, Y.C., Huang, S.Y., 2017. Accurate traffic flow estimation in urban roads with considering the traffic signals, in: International Conference on Internet of Vehicles, Springer. pp. 41–52.

Lavrenz, S.M., Day, C.M., Smith, W.B., Sturdevant, J.R., Bullock, D.M., 2016. Assessing longitudinal arterial performance and traffic signal retiming outcomes. Transportation Research Record: Journal of the Transportation Research Board 2558, 66–77.

Lee, S., Wong, S.C., Li, Y.C., 2015. Real-time estimation of lane-based queue lengths at isolated signalized junctions. Transportation Research Part C: Emerging Technologies 56, 1–17.

van Leeuwaarden, J.S., 2006. Delay analysis for the fixed-cycle traffic-light queue. Transportation Science 40, 189–199.

Li, F., Tang, K., Yao, J., Li, K., 2017. Real-time queue length estimation for signalized intersections using vehicle trajectory data. Transportation Research Record: Journal of the Transportation Research Board 2623, 49–59.

Li, J., Zhou, K., Shladover, S.E., Skabardonis, A., 2013a. Estimating queue length under connected vehicle technology: Using probe vehicle, loop detector, fused data. Transportation Research Record: Journal of the Transportation Research Board 2366, 17–22.

Li, L., Li, Y., Li, Z., 2013b. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transportation Research Part C: Emerging Technologies 34, 108–120.

Li, Y., Fu, K., Wang, Z., Shahabi, C., Ye, J., Liu, Y., 2018. Multi-task representation learning for travel time estimation, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM. pp. 1695–1704.

Liu, H.X., Ban, J.X., Ma, W., Mirchandani, P.B., 2007. Model reference adaptive control framework for real-time traffic management under emergency evacuation. Journal of urban planning and development 133, 43–50.

Liu, H.X., Wu, X., Ma, W., Hu, H., 2009. Real-time queue length estimation for congested signalized intersections. Transportation research part C: emerging technologies 17, 412–427.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., 2009. Map-matching for low-sampling-rate GPS trajectories, in: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM. pp. 352–361.

Luo, X., Liu, B., Jin, P.J., Cao, Y., Hu, W., 2019. Arterial traffic flow estimation based on vehicle-to-cloud vehicle trajectory data considering multi-intersection interaction and coordination. Transportation Research Record: Journal of the Transportation Research Board 2673, 68–83.

Marlin, B., 2008. Missing data problems in machine learning. Ph.D. thesis.

McNeil, D.R., 1968. A solution to the fixed-cycle traffic light problem for compound poisson arrivals. Journal of Applied Probability 5, 624–635.

Mei, Y., Gu, W., Chung, E.C., Li, F., Tang, K., 2019. A bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data. Transportation Research Part C: Emerging Technologies 109, 233–249.

Meng, C., Yi, X., Su, L., Gao, J., Zheng, Y., 2017a. City-wide traffic volume inference with loop detector data and taxi trajectories, in: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM. p. 1.

Meng, F., Wong, S.C., Wong, W., Li, Y.C., 2017b. Estimation of scaling factors for traffic counts based on stationary and mobile sources of data. International journal of intelligent transportation systems research 15, 180–191.

Merris, R., 2003. Combinatorics. volume 67. John Wiley & Sons.

Miller, A.J., 1963. Settings for fixed-cycle traffic signals. Journal of the Operational Research Society 14, 373–386.

Mung, G.K., Poon, A.C., Lam, W.H., 1996. Distributions of queue lengths at fixed time traffic signals. Transportation Research Part B: Methodological 30, 421–439.

Nakata, T., Takeuchi, J.i., 2004. Mining traffic data from probe-car system for travel time prediction, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 817–822.

National Highway Traffic Safety Administration, 2018. 2017 fatal motor vehicle crashes: overview.

Newell, G., 1971. Applications of queueing theory. Chapman and Hall, London.

Newell, G.F., 1965. Approximation methods for queues with application to the fixed-cycle traffic light. SIAM Review 7, 223–240.

Newson, P., Krumm, J., 2009. Hidden Markov map matching through noise and sparseness, in: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, ACM. pp. 336–343.

Ohno, K., 1978. Computational algorithm for a fixed cycle traffic signal and new approximate expressions for average delay. Transportation Science 12, 29–47.

Olszewski, P., 1990. Modelling of queue probability distribution at traffic signals, in: International Symposium on Transportation and Traffic Theory, 11th, 1990, Yokohama, Japan.

Olszewski, P.S., 1994. Modeling probability distribution of delay at signalized intersections. Journal of Advanced Transportation 28, 253–274.

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

Parkinson, B.W., Enge, P.K., 1996. Differential GPS. Global Positioning System: Theory and applications 2, 3–50.

Patire, A.D., Wright, M., Prodhomme, B., Bayen, A.M., 2015. How much GPS data do we need? Transportation Research Part C: Emerging Technologies 58, 325–342.

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. PPCA-based missing data imputation for traffic flow volume: A systematical approach. IEEE Transactions on Intelligent Transportation Systems 10, 512–522.

Qu, L., Zhang, Y., Hu, J., Jia, L., Li, L., 2008. A BPCA based missing value imputing method for traffic flow volume data, in: 2008 IEEE Intelligent Vehicles Symposium, IEEE. pp. 985–990.

Ramezani, M., Geroliminis, N., 2012. On the estimation of arterial route travel time distribution with Markov chains. Transportation Research Part B: Methodological 46, 1576–1590.

Ramezani, M., Geroliminis, N., 2015. Queue profile estimation in congested urban networks with probe data. Computer-Aided Civil and Infrastructure Engineering 30, 414–432.

Ran, B., 2013. Use of cellphone data in travel survey and transportation planning. Urban transport of China 1, 72–81.

Ran, B., Tan, H., Wu, Y., Jin, P.J., 2016. Tensor based missing traffic data completion with spatial-temporal correlation. Physica A: Statistical Mechanics and its Applications 446, 54–63.

Schrank, D., Eisele, B., Lomax, T., 2019. 2019 urban mobility report.

Seo, T., Bayen, A.M., 2017. Traffic state estimation method with efficient data fusion based on the Aw-Rascle-Zhang model, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1–6.

Seo, T., Bayen, A.M., Kusakabe, T., Asakura, Y., 2017. Traffic state estimation on highway: A comprehensive survey. Annual reviews in control 43, 128–151.

Seo, T., Kusakabe, T., 2018. Traffic state estimation using small imaging satellites and connected vehicles. Transportation research procedia 34, 4–11.

Seo, T., Kusakabe, T., Asakura, Y., 2015. Estimation of flow and density using probe vehicles with spacing measurement equipment. Transportation Research Part C: Emerging Technologies 53, 134–150.

Shahrbabaki, M.R., Safavi, A.A., Papageorgiou, M., Papamichail, I., 2018. A data fusion approach for real-time traffic state estimation in urban signalized links. Transportation Research Part C: Emerging Technologies 92, 525–548.

Shang, J., Zheng, Y., Tong, W., Chang, E., Yu, Y., 2014. Inferring gas consumption and pollution emission of vehicles throughout a city, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 1027–1036.

Shiryayev, A., 1984. Probability. Translated from the Russian by RP Boas. Graduate texts in Mathematics 95, 336.

Skabardonis, A., Geroliminis, N., 2008. Real-time monitoring and control on signalized arterials. Journal of Intelligent Transportation Systems 12, 64–74.

Srinivasan, K.K., Jovanis, P.P., 1996. 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, 15–22.

Tak, S., Woo, S., Yeo, H., 2016. Data-driven imputation method for traffic data in sectional units of road links. IEEE Transactions on Intelligent Transportation Systems 17, 1762–1771.

Takenouchi, A., Kawai, K., Kuwahara, M., 2019. Traffic state estimation and its sensitivity utilizing measurements from the opposite lane. Transportation Research Part C: Emerging Technologies 104, 95–109.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.J., Li, F., 2013. A tensor-based method for missing traffic data completion. Transportation Research Part C: Emerging Technologies 28, 15–27.

Tang, X., Gong, B., Yu, Y., Yao, H., Li, Y., Xie, H., Wang, X., 2019. Joint modeling of dense and incomplete trajectories for citywide traffic volume inference, in: The World Wide Web Conference, ACM. pp. 1806–1817.

Tettamanti, T., Demeter, H., Varga, I., 2012. Route choice estimation based on cellular signaling data. Acta Polytechnica Hungarica 9, 207–220.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 611–622.

Turner, S.M., Eisele, W.L., Benz, R.J., Holdener, D.J., 1998. Travel time data collection handbook. Technical Report. United States. Federal Highway Administration.

Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13, 260–269.

Viti, F., Van Zuylen, H.J., 2010. Probabilistic models for queues at fixed control signals. Transportation Research Part B: Methodological 44, 120–135.

Wang, S., Huang, W., Lo, H.K., 2019. Traffic parameters estimation for signalized intersections based on combined shockwave analysis and Bayesian network. Transportation Research Part C: Emerging Technologies 104, 22–37.

Wolshon, B., Urbina, E., Wilmot, C., Levitan, M., 2005a. Review of policies and practices for hurricane evacuation. i: Transportation planning, preparedness, and response. Natural hazards review 6, 129–142.

Wolshon, B., Urbina Hamilton, E., Levitan, M., Wilmot, C., 2005b. Review of policies and practices for hurricane evacuation. ii: Traffic operations, management, and control. Natural Hazards Review 6, 143–161.

Wong, W., Shen, S., Zhao, Y., Liu, H.X., 2019a. On the estimation of connected vehicle penetration rate based on single-source connected vehicle data. Transportation Research Part B: Methodological 126, 169–191.

Wong, W., Wong, S., 2019. Unbiased estimation methods of nonlinear transport models based on linearly projected data. Transportation Science 53, 665–682.

Wong, W., Wong, S.C., Liu, H.X., 2019b. Bootstrap standard error estimations of nonlinear transport models based on linearly projected data. Transportmetrica A: Transport Science 15, 602–630.

Work, D.B., Tossavainen, O.P., Blandin, S., Bayen, A.M., Iwuchukwu, T., Tracton, K., 2008. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices, in: 2008 47th IEEE Conference on Decision and Control, IEEE. pp. 5062–5068.

Wu, C.J., et al., 1983. On the convergence properties of the EM algorithm. The Annals of Statistics 11, 95–103.

Yao, J., Li, F., Tang, K., Jian, S., 2019. Sampled trajectory data-driven method of cycle-based volume estimation for signalized intersections by hybridizing shockwave theory and probability distribution. IEEE Transactions on Intelligent Transportation Systems .

Yoon, J., Noble, B., Liu, M., 2007. Surface street traffic estimation, in: Proceedings of the 5th international conference on Mobile systems, applications and services, ACM. pp. 220–232.

Zhan, X., Zheng, Y., Yi, X., Ukkusuri, S.V., 2016. Citywide traffic volume estimation using trajectory data. IEEE Transactions on Knowledge and Data Engineering 29, 272–285.

Zhan, X., Zheng, Y., Yi, X., Ukkusuri, S.V., 2017. Citywide traffic volume estimation using trajectory data. IEEE Transactions on Knowledge & Data Engineering 2, 272–285.

Zhao, Y., Liu, H.X., 2020. Maximum likelihood estimation of probe vehicle penetration rates and queue length distributions from probe vehicle data, in: Transportation Research Board 99th Annual Meeting.

Zhao, Y., Shen, S., Liu, H.X., 2020a. A hidden Markov model for the estimation of dependent queues using probe vehicle data, in: Transportation Research Board 99th Annual Meeting.

Zhao, Y., Yan, X., Liu, H.X., 2020b. A data fusion framework for traffic volume reconstruction from low-rank structures, in: Transportation Research Board 99th Annual Meeting.

Zhao, Y., Yan, X., Liu, H.X., 2020c. A probabilistic model for traffic volume reconstruction based on data fusion, in: Transportation Research Board 99th Annual Meeting.

Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., Liu, H.X., 2019a. Estimation of queue lengths, probe vehicle penetration rates, and traffic volumes at signalized intersections using probe vehicle trajectories. Transportation Research Record: Journal of the Transportation Research Board 2673, 660–670.

Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., Liu, H.X., 2019b. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transportation Research Part C: Emerging Technologies 107, 70–91.

Zheng, F., Jabari, S.E., Liu, H.X., Lin, D., 2018. Traffic state estimation using stochastic Lagrangian dynamics. Transportation Research Part B: Methodological 115, 143–165.

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

Zheng, J., 2016. Data-Driven Applications for Connected Vehicle Based Traffic Signal Systems. Ph.D. thesis.

Zheng, J., Liu, H.X., 2017. Estimating traffic volumes for signalized intersections using connected vehicle data. Transportation Research Part C: Emerging Technologies 79, 347–362.

Zheng, J., Liu, H.X., 2020. DASCOS: Dynamic Area-wide Signal Control Optimization System Using Trajectory Data. Technical Report.

Zheng, Y., Zhang, L., Xie, X., Ma, W.Y., 2009. Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the 18th international conference on World wide web, ACM. pp. 791–800.

Zhuang, Y., Ke, R., Wang, Y., 2018. Innovative method for traffic data imputation based on convolutional neural network. IET Intelligent Transport Systems 13, 605–613.

van Zuylen, H.J., Zheng, F., Chen, Y., 2010. Using probe vehicle data for traffic state estimation in signalized urban networks, in: Traffic Data Collection and its Standardization. Springer, pp. 109–127.