https://doi.org/10.1016/j.trc.2019.10.006

A Bayesian Approach for Estimating Vehicle Queue Lengths at Signalized Intersections using Probe Vehicle Data

Yu Mei [a], Weihua Gu [a*], Edward C.S. Chung [a], Fuliang Li [a,b], Keshuang Tang [c]

[a] Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

[b] Didi Chuxing Technology Co. Ltd., Beijing, China

[c] College of Transportation Engineering, Tongji University, Shanghai, China

Abstract

A novel Bayesian approach is proposed for estimating the maximum queue lengths of vehicles at signalized intersections using high-frequency trajectory data of probe vehicles. The queue length estimates are obtained from a distribution estimated over several neighboring cycles via a maximum a posteriori method. An expectation maximum algorithm is proposed for efficiently solving the estimation problem. Through a battery of simulation experiments and a real-world case study, the proposed approach is shown to produce more accurate and robust estimates than two benchmark estimation methods. Fairly good accuracy is achieved even when the probe vehicle penetration rate is 2%.

Keywords: Queue length estimation; Probe vehicles; Bayesian approach; Expectation maximum algorithm

1. Introduction

Traffic signals at intersections alternatively give right of way to vehicles travelling in different directions, leading to repetitive formation and dissipation of vehicle queues in each approach. The maximum distance that a vehicle queue propagates from the stop line, often termed the "maximum queue length" in the literature (e.g. Hao et al., 2014; Li et al., 2017), has been commonly used as a measure of signal's performance and a key input for optimizing signal timing and offsets (Webster and Cobbe, 1966; Michalopoulos and Stephanopoulos, 1977a, b; Cronje, 1983; Gartner et al., 1991; Rao and Rao, 2012). Estimating this maximum queue length is a critical and challenging task for the management and control of urban traffic. In the rest of this paper, we use "queue length" to represent the maximum queue length for brevity.

Early works on queue length estimation mainly used loop detector data. Two classes of estimation methods were proposed. The first class of methods exploited the cumulative input-output diagram of vehicles in an intersection approach (Newell, 1965; Cronje, 1983; Sharma et al., 2007; Vigos et al., 2008). However, these methods overlooked the part of vehicle queues residing upstream of the detector(s). This problem was resolved by the second class of methods (Muck, 2002; Skabardonis and Geroliminis, 2008; Liu et al., 2009), which estimated queue lengths through the prediction of vehicle shockwaves. Unfortunately, loop detector measurements are known to have bias and errors (especially after many years of operation), which can greatly degrade the estimation accuracy. Worse still, failed detectors were usually not repaired or replaced in time due to the high costs (Coifman et al., 2003).

* Corresponding author.

E-mail address: weihua.gu@polyu.edu.hk

In recent decades, the growing use of probe vehicle trajectory data has created new opportunities for traffic state estimation. Probe vehicles, including connected buses, taxis and private cars, are equipped with mobile sensors that can provide high-frequency, real-time location information along their trajectories (Mei et al., 2015; Zheng and Liu, 2017; Li et al., 2017). These data are thus not constrained by location. Considering that probe vehicles usually account for only a small proportion of the traffic (termed the *penetration rate*, which is in the range of 1-10% at present), a number of queue length estimation methods have been proposed for queue length estimation. They are classified as deterministic methods and stochastic ones.

Deterministic methods were often built upon the shockwave theory of traffic (Lighthill and Whitham, 1955; Richards, 1956; Newell, 1993). Many of those studies estimated queue lengths by finding the intersection point between queue formation and discharging shockwaves in a signal cycle (Cheng et al., 2011; Cheng et al., 2012; Ramezani and Geroliminis, 2015; Li et al., 2017). However, the accuracy of this method is limited due to two reasons: (i) the intersection point between two shockwaves is sensitive to small estimation errors for the shockwaves; and (ii) the queue formation shockwave is difficult to predict accurately especially under low penetration rates, since it depends highly on the time-varying vehicle arrival headways. Other deterministic methods include Ban et al. (2011), Hao et al. (2014), and Hao and Ban (2015), in which queue lengths were derived from the travel times gauged by the probe vehicles. In addition, some works also developed data fusion methods for combining the use of loop detector data and probe vehicle data (Badillo et al., 2012; Li et al., 2013; Cai et al., 2014; Wang et al. 2017). Naturally, these deterministic methods cannot address the random factors embedded in the vehicle arrivals, penetration and distribution of probe vehicles among the traffic, driver behavior, etc.

Stochastic methods were proposed to account for those random factors. Studies in this realm include the pioneer work by Comert and Cetin (2007) and its extensions (Comert and Cetin, 2009, 2011; Comert, 2013, 2016). Although closed-form formulas were derived in the above studies for the mean and variance of queue lengths, some of their assumptions (e.g., the Poisson vehicle arrivals) may not fit the real traffic very well (Manar and Baass, 1996). The Poisson arrival assumption was relaxed by Tiaprasert et al. (2015). Yet that work still assumed a given, constant penetration rate. Hence, its estimates may be inaccurate if the penetration rate varies significantly over time, or is difficult to estimate accurately (Zheng and Liu, 2017). Hao et al. (2014) further relaxed the assumption of a given penetration rate. They estimated the queue length distribution through a complicated Bayesian network model, which requires the training and calibration of a large number of parameters using historical data. Hence, the real-world application of this method is also limited. Recently, Yin et al. (2018) integrated the shockwave theory approach in a stochastic environment for queue length estimation. However, the queue length was still derived by finding the intersection point between queue formation and discharging shockwaves. The estimates are thus still sensitive to small errors that can easily occur in the estimation of formation shockwaves.

In light of the above limitations in the literature, we propose a novel Bayesian approach for estimating vehicle queue lengths at signalized intersections. The proposed approach accounts for the stochastic and time-varying vehicle arrivals and driver behavior, instead of assuming a specific vehicle arrival process and a known penetration rate. It does not rely on the estimation of queue formation shockwaves due to the potentially large estimation errors. In addition, our approach is parsimonious and does not rely on frequent calibrations of many parameters.

In the proposed approach, the distribution of queue lengths is estimated using the lower and upper bounds of queue lengths developed for each signal cycle. A lower bound is obtained from the trajectory data of probe vehicles joining the queue, which are termed the *stopped vehicles* (Ramezani and Geroliminis, 2015). An upper bound is developed by using the data of probe vehicles that do not meet the queue (termed the *non-stopped vehicles*). Note that the upper bound of queue length was often overlooked by previous studies in this realm. Only Hao et al. (2014) and Tiaprasert et al. (2015) have exploited the upper bound information in queue length estimation, to our best knowledge. However, in the former work, the upper bound was developed with the help of a large amount of parameter training effort, while in the latter work only a loose upper bound was used. Our proposed upper bound, on the other hand, does not rely on any pre-training and is tighter than the one used in Tiaprasert et al. (2015), as we shall see momentarily.

Through a large battery of simulation experiments and a real-world case study, we demonstrate the accuracy and robustness of our estimates under both stationary and time-varying traffic condition. We also show that our approach outperforms the other two similar methods developed in recent years (Ramezani and Geroliminis, 2015; Li et al., 2017).

The rest of the paper is organized as follows: section 2 presents the estimation approach; section 3 describes the simulation tests; section 4 presents a real-world case study; section 5 discusses the choice of key model parameters in real practice; and section 6 summarizes the key findings of this paper.

2. Methodology

The assumptions, data, and an overview of our approach are described in section 2.1. The details of the approach are presented in sections 2.2-2.6, respectively.

2.1 Assumptions, data, and overview of our approach

The following assumptions are made for the proposed approach. They also define the approach's application scope:

(i) For simplicity, in this paper we focus on the through-moving vehicle queues only. The left-turning vehicles are assumed to queue up in separate lane(s) and use a separate green phase to discharge.[1] If there are more than one through-moving lanes in the approach, queued vehicles are assumed to be distributed evenly among those lanes.

(ii) Vehicle inflows do not exceed the intersection's discharging capacity[2], and the intersection is not blocked by any vehicle queue that spills over from downstream.

(iii) The signal timing is known.

(iv) The vehicle arrival rate to the intersection approach and the signal timing plan do not change significantly within an *episode*, defined as a short period of time consisting of one

[1] This assumption is valid if the left-turning queue does not spill over the left-turn pocket. If spillover occurs and the queued left-turning vehicles mix with the through-moving ones, the method introduced in this paper needs to be modified. Particularly, lane-specific vehicle position data are needed for estimating the mixed queue lengths since vehicle queues in different lanes may have different lengths.

[2] This assumption does not mean residual vehicle queues would never occur at the end of a signal cycle. Our approach can be applied as long as the queue is not ever-growing over cycles. In fact, it still performs well under near-saturated traffic, where residual queues appear frequently, as we shall see in section 3.

or several cycles.[3] Further, the queue lengths in the cycles of a specific episode are assumed to be independent random variables following the same parametric distribution. Similar assumptions were often used in the literature for simplifying the modeling work (Comert and Cetin, 2009; Taylor and Heydecker, 2014; Yin et al., 2018).

(v) Probe vehicle positions are updated at a relatively high frequency, as may occur thanks to the emerging connected vehicle technology (Zheng and Liu, 2017; Yin et al., 2018).

The probe vehicle trajectory data are illustrated in Figure 1. The left part of the figure sketches an intersection approach (note that the turning traffic is omitted here). The right part is a time-space diagram showing the trajectories of probe vehicles (only) for a cycle i in episode e. The cycles are indexed from the beginning of the study period, and we denote $\mathbb{I}^e$ as the index set of consecutive cycles that belong to episode e. The red bars located at the stop line represent effective red periods, and the gaps between the red bars represent the effective green periods. The effective red start and green start of cycle i are denoted by r^i and g^i, respectively ($r^i < g^i$). We define the space coordinate of the approach's upstream end as 0, and that of the stop line as L.

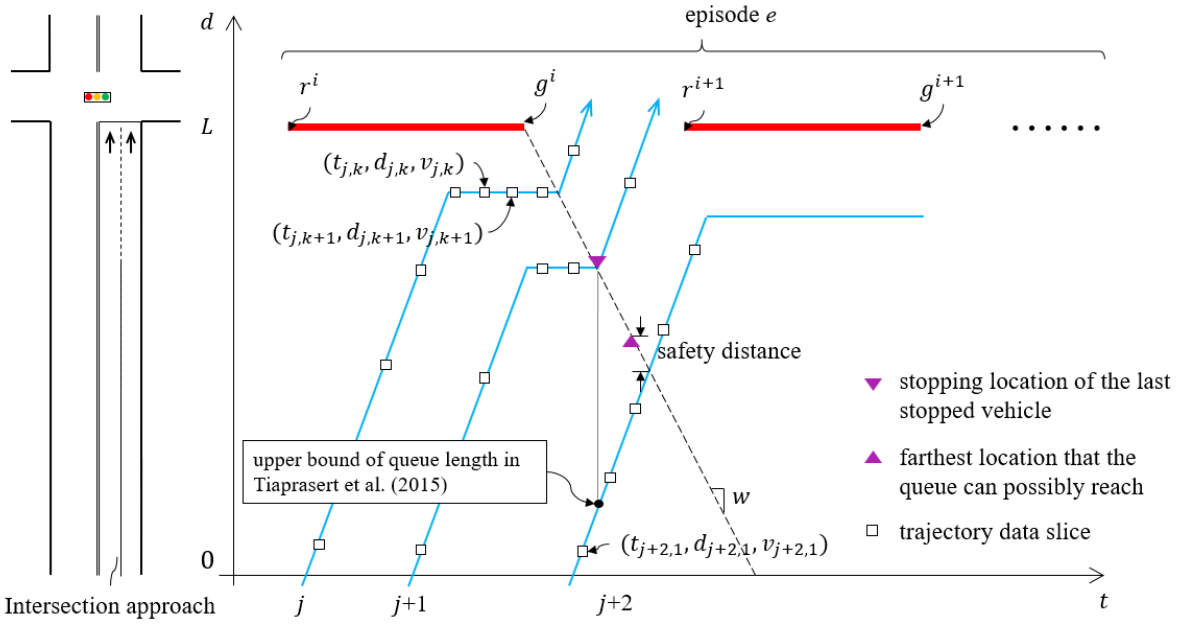


Figure 1. The intersection approach and a time-space diagram of probe vehicle trajectories

The through-moving probe vehicle trajectories are plotted as thin, blue lines in the figure. These probe vehicles are numbered in chronological order by index variable j. For each probe vehicle, the data slices are numbered in chronological order by index variable k. The k-th data slice of vehicle j contains a set of three variables: timestamp $t_{j,k}$; vehicle location $d_{j,k}$ (which is calculated by mapping the GPS position to the spatial coordinate system defined above); and instantaneous vehicle speed $v_{j,k}$. They are marked by the small squares along the trajectories.

The queue length distribution in an episode is estimated via the maximum a posteriori (MAP) approach. The method takes the lower and upper bounds of queue lengths for the cycles in $\mathbb{I}^e$ as

[3] This is true even for adaptive signal control, since the transition from one signal plan to another usually takes several cycles to complete smoothly (Shelby et al., 2006).

inputs. The lower bound is determined using the stopping location of the last stopped vehicle in the cycle (vehicle $j+1$ for cycle i in Figure 1); see the reverse triangle marker on that vehicle's trajectory. The upper bound is determined by the first non-stopped vehicle in the cycle (vehicle $j+2$ for cycle i in Figure 1). Specifically, the bound is located a safety distance downstream of the intersection point between that vehicle's trajectory and the queue discharging shockwave (the dashed line in the figure), as marked by the triangle marker. Note that this upper bound is much tighter than the one used in Tiaprasert et al. (2015), which is the instantaneous position of vehicle $j+2$ when the last stopped vehicle ($j+1$) starts to discharge from the queue (see the black dot in the figure).

The calculation of upper bound requires that we know the discharging shockwave stemming from the green start. Unlike the queue formation shockwave, the discharging shockwave can be fairly accurately estimated. We denote w as the (negative) speed of discharging shockwave, which may vary over a small range due to the differences in vehicle lengths and driver behavior. The w is also assumed to follow a parametric distribution, whose parameters are updated again via a Bayesian approach.

The estimation procedure is summarized as the following five steps. They are detailed in sections 2.2-2.6, respectively.

Step 1. For episode e, identify the stopped vehicles and the non-stopped vehicles for every cycle $i \in \mathbb{I}^e$ (section 2.2).

Step 2. Using the stopped vehicle trajectories in episode e, determine the posterior distribution of discharging shockwave speed w for the episode (section 2.3).

Step 3. Using the probe vehicle trajectories and the distribution of w developed in Step 2, calculate the lower and upper bounds of queue length for each cycle $i \in \mathbb{I}^e$ (section 2.4).

Step 4. Using the bounds developed in Step 3, determine the posterior distribution of queue length for episode e via the MAP approach. Estimate the queue length, q^i, for each cycle i in the episode (section 2.5).

Step 5. Update the priors for episode $e+1$ using the posteriors obtained for episode e. Set $e \rightarrow e+1$ and return to Step 1 (section 2.6).

2.2 Identifying stopped and non-stopped vehicles

In this preparation step, the stopped vehicles are separated from the non-stopped ones for each cycle. In principle, a stopped vehicle can be identified by checking if its speed is smaller than a given threshold, denoted by ξ. However, this may create false detections since a probe vehicle may also be stopped before joining a vehicle queue formed during red phases (only the vehicles joining those queues can be classified as stopped vehicles in our approach). This may occur when the vehicle encounters a dwelling bus, a pedestrian crossing the street, or a vehicle accident. To reduce these false detections, we propose to search for the stopped vehicles only in a subset of the time-space planar for each cycle.

We first define the *target zone* of cycle i, $\mathcal{S}^i$, as the yellow shaded parallelogram in Figure 2. This zone is constructed by letting the line segment connecting (r^i, L) and (r^{i+1}, L) sweep in the time-space diagram along a line of slope w_0^e until reaching the start of intersection approach, where w_0^e denotes the prior mean of discharging shockwave speed for episode e. The target zone is thus defined as:

$$\mathcal{S}^i \equiv \{(t,d): \max(w_0^e(t-r^i)+L,0) \le d \le \min(w_0^e(t-r^{i+1})+L,L)\}. \quad (1)$$

We then define $\mathbb{V}^i$ as the set of probe vehicles with data slices contained in $\mathcal{S}^i$:

$$\mathbb{V}^i \equiv \{j: \exists k \text{ such that } (t_{j,k}, d_{j,k}) \in \mathcal{S}^i\}. \quad (2)$$

We further define a subset of $\mathcal{S}^i$ as the *discharging zone*, $\mathcal{D}^i$, as illustrated by the green shaded trapezoid in Figure 2. This trapezoidal zone is constructed by setting its top base as a $2(\varepsilon+o)$-long segment centered at (g^i, L), and the slopes of its left and right lateral sides as w^- and w^+, respectively ($w^- < w_0^e < w^+$). Here o denotes the data updating interval (which equals 1s if the updating frequency is 1 Hz, for example); ε denotes the error bound for vehicles' start-up delay[4]; and w^- and w^+ denote the minimum and maximum discharging shockwave speeds for the present episode. Specifically, in this paper we assume w in episode e follows a Gaussian distribution defined as $p(w) = \mathcal{N}(w|w_0^e, \alpha^{-1})$, where α is the prior precision (inverse of the variance). We then set $w^- = w_0^e - 3\alpha^{-1/2}$ and $w^+ = w_0^e + 3\alpha^{-1/2}$, so that $[w^-, w^+]$ is a 99.7% confidence interval for w. The definition of $\mathcal{D}^i$ thus ensures that at least one data slice of each stopped vehicle is enclosed. The formulation of $\mathcal{D}^i$ is given by:

$$\mathcal{D}^i \equiv \{(t,d): \max(w^-(t-g^i+o+\varepsilon)+L,0) \le d \le \min(w^+(t-g^i-o-\varepsilon)+L,L)\}. \quad (3)$$

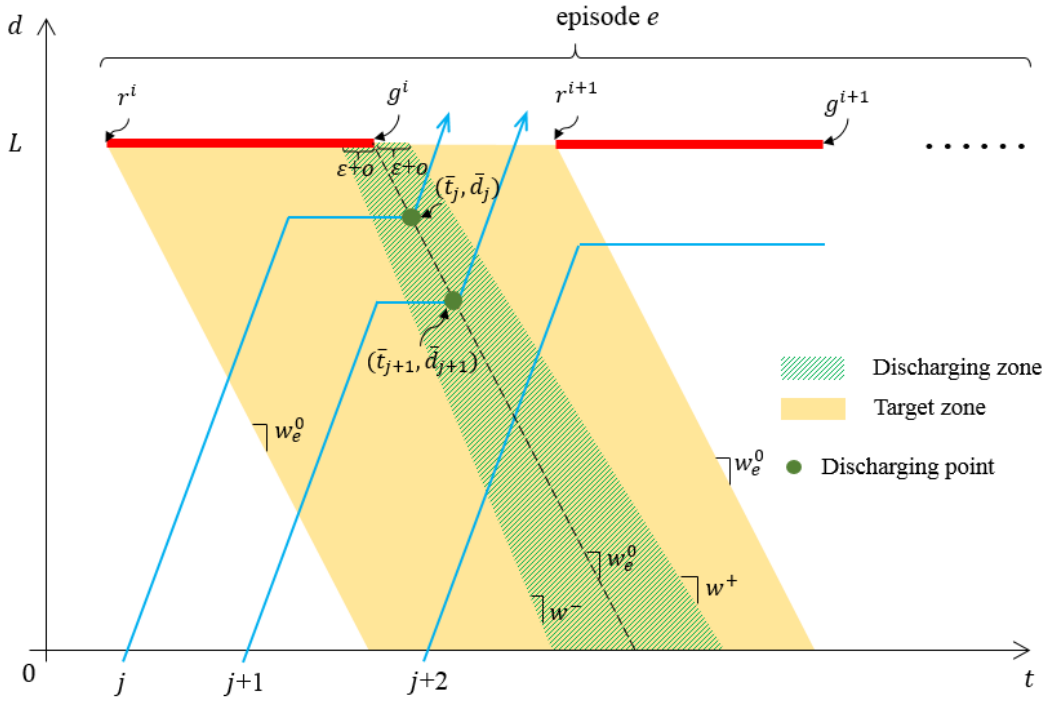


Figure 2. Vehicle identification Zones

The index sets of stopped vehicles and non-stopped vehicles in cycle i, $\mathbb{Q}^i$ and $\mathbb{M}^i$, respectively, are formulated as:

[4] The effective green start, g^i, is the sum of the actual green start time and the first queued vehicle's start-up delay (Kittelson et al. 2004). The latter is usually around 2s but varies between different vehicles. The ε is introduced to accommodate the error of this value.

$$\mathbb{Q}^i \equiv \{j: \exists(t_{j,k}, d_{j,k}) \in \mathcal{D}^i \text{ such that } v_{j,k} \leq \xi\} \quad (4)$$

$$\mathbb{M}^i \equiv \mathbb{V}^i \backslash \mathbb{Q}^i. \quad (5)$$

Note that the above method cannot eliminate all the false detections of stopped vehicles. However, it can rule out most false detections in general because $\mathcal{D}^i$ is much smaller than $\mathcal{S}^i$.

2.3 Deriving the posterior distribution of w

This step estimates the posterior distribution of w in the present episode via Bayesian regression. To this end, we first find the discharging point, $(\bar{t}_j, \bar{d}_j)$, for each stopped vehicle $j \in \mathbb{Q}^i$, which indicates the time and location when vehicle j starts to discharge from the queue in cycle i. These discharging points are illustrated by the solid green circles in Figure 2. The coordinates $(\bar{t}_j, \bar{d}_j)$ can be calculated using the vehicle's last stopped data slice $(t_{j,k_Q}, d_{j,k_Q}, v_{j,k_Q})$ recorded in $\mathcal{D}^i$ and its following moving slice $(t_{j,k_Q+1}, d_{j,k_Q+1}, v_{j,k_Q+1})$. Specifically:

$$(\bar{t}_j, \bar{d}_j) = (t_{j,k_Q+1} - \frac{d_{j,k_Q+1} - d_{j,k_Q}}{v_{j,k_Q+1}}, d_{j,k_Q}), \quad (6)$$

where $k_Q = \arg\max_k \{t_{j,k} | (t_{j,k}, d_{j,k}) \in \mathcal{D}^i, v_{j,k} \leq \xi\}$.

In practice, the discharging points in a cycle will scatter within a narrow stripe around the idealized discharging shockwave with slope w. Thus, we propose an adjusted Bayesian regression model that takes all the discharging points in the present *episode* as the data input:

$$d = w(t - g^i) + L + \epsilon, \quad (7)$$

where ϵ is a zero mean Gaussian random noise with precision β (the value of β is determined by experience).

Model (7) implies that the location of a discharging point, d, follows a Gaussian distribution with mean $w(t - g^i) + L$ and precision β, i.e., $d \sim \mathcal{N}(w(t - g^i) + L, \beta^{-1})$, where t is the timestamp of the discharging point. Hence, the likelihood function of observing all the discharging points $\mathcal{C}^e \equiv \{(\bar{t}_j, \bar{d}_j) | j \in \cup_{i \in \mathbb{I}^e} \mathbb{Q}^i\}$ is given by:

$$p(\mathcal{C}^e | w) \propto \exp\left\{-\frac{\beta}{2} \sum_{i \in \mathbb{I}^e} \sum_{j \in \mathbb{Q}^i} (\bar{d}_j - w(\bar{t}_j - g^i) - L)^2\right\}. \quad (8)$$

Based on the Bayes' theorem, the posterior probability density function (PDF) of w, $p(w|\mathcal{C}^e)$, is proportional to the product of the likelihood function and the prior PDF, which is $p(w) = \mathcal{N}(w | w_0^e, \alpha^{-1}) = \sqrt{\frac{\alpha}{2\pi}} e^{-\frac{\alpha}{2}(w - w_0^e)^2}$. Hence, we have:

$$p(w|\mathcal{C}^e) \propto p(\mathcal{C}^e|w) p(w) \propto \exp\left\{-\frac{\beta}{2} \sum_{i \in \mathbb{I}^e} \sum_{j \in \mathbb{Q}^i} (\bar{d}_j - w(\bar{t}_j - g^i) - L)^2 - \frac{\alpha}{2}(w - w_0^e)^2\right\}. \quad (9)$$

Due to the conjugate property of Gaussian prior, the posterior distribution is also Gaussian. We can therefore derive the posterior distribution of w as follows:

$$p(w|\mathcal{C}^e) = \mathcal{N}\left(w|w_p^e, \left(\alpha_p^e\right)^{-1}\right) \tag{10}$$

$$w_p^e = \left(\alpha_p^e\right)^{-1}\left[\beta \sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\left(\bar{d}_j - L\right)\left(\bar{t}_j - g^i\right) + \alpha \cdot w_0^e\right] \tag{11}$$

$$\left(\alpha_p^e\right)^{-1} = \left(\beta \sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\left(\bar{t}_j - g^i\right)^2 + \alpha\right)^{-1}, \tag{12}$$

where w_p^e is the posterior mean and α_p^e is the posterior precision. The detailed derivation is relegated to Appendix B.

2.4 Calculating the lower and upper bounds of queue length

The lower bound of queue length for cycle i, denoted by l^i (in the unit of vehicles), is calculated from the location of the farthest discharging point:

$$l^i = \max_{j\in\mathbb{Q}^i}\left\{L - \bar{d}_j\right\}/s, \tag{13}$$

where s is the average jam spacing.

On the other hand, the upper bound is calculated using the intersection point between the discharging shockwave and the first non-stopped vehicle's trajectory. The first non-stopped vehicle can be identified by finding the closest data slice upstream of the shockwave, whose index is given by:

$$\left(j^i, k^i\right) = \arg\max_{j\in\mathbb{M}^i,k}\left\{d_{j,k}|d_{j,k} \le w_p^e\left(t_{j,k} - g^i\right) + L\right\}. \tag{14}$$

Note here that the posterior mean w_p^e is used for the shockwave speed. We approximate vehicle j^i's trajectory stemming from data slice $\left(j^i, k^i\right)$ by $d = v_{j^i,k^i}\left(t - t_{j^i,k^i}\right) + d_{j^i,k^i}$. By geometry, the intersection point's spatial coordinate is calculated as $\frac{w\cdot v_{j^i,k^i}}{v_{j^i,k^i}-w}\left(t_{j^i,k^i} - \frac{d_{j^i,k^i}}{v_{j^i,k^i}} - g^i + \frac{L}{w}\right)$.

Now considering that w follows a posterior distribution, the upper bound of queue length in the unit of vehicles is derived as follows:

$$u^i = \left[L - \int_{-\infty}^{0}\frac{w\cdot v_{j^i,k^i}}{v_{j^i,k^i}-w}\left(t_{j^i,k^i} - \frac{d_{j^i,k^i}}{v_{j^i,k^i}} - g^i + \frac{L}{w}\right)p(w|\mathcal{C}^e)dw - \left(s + \frac{v_{j^i,k^i}^2}{2a_{j^i}}\right)\right]/s, \tag{15}$$

where $\left(s + \frac{v_{j^i,k^i}^2}{2a_{j^i}}\right)$ is the safety distance that non-stopped vehicle j^i must keep from its preceding vehicle; and a_{j^i} denotes the maximum deceleration rate of vehicle j^i. A probe vehicle's maximum deceleration rate is sometimes furnished in the dataset. If this value is not given, it can be derived by using the speed data of that vehicle collected during the vehicle's deceleration processes.

When there are multiple lanes in the intersection approach, sometimes one may find $u^i < l^i$. This is because when adjacent lanes have different queue lengths, a non-stopped vehicle may be observed to appear downstream of a stopped vehicle. To resolve this issue, we adjust the upper bound as follows so that it is always greater than the lower bound[5]:

[5] If the queue lengths in different lanes are substantially different (i.e., assumption (i) in section 2.1 is violated), the present method may fail to find a valid upper bound. In this case, lane-specific vehicle position data are needed to obtain valid upper bounds.

$$u^i = max\left\{l^i + \delta, \int_{-\infty}^{0} \frac{w \cdot v_{j^i,k^i}}{v_{j^i,k^i} - w}\left(t_{j^i,k^i} - \frac{d_{j^i,k^i}}{v_{j^i,k^i}} - g^i + \frac{L}{w}\right) p(w|\mathcal{C}^e)dw + s + \frac{v_{j^i,k^i}^2}{2a_{j^i}}\right\}/s, \quad (16)$$

where δ is a small positive number, e.g. 0.01. The δ is added to ensure that the lower and upper bounds do not equal, so that the likelihood for observing the two bounds is not zero (note that the queue length is assumed to follow a continuous distribution). Readers may refer to the next section for details. We find from numerical experiments that the value of δ has only a trivial effect on the estimation results as long as it is sufficiently small.

2.5 Estimating queue length

In this step, we first derive the posterior PDF of the parameters of queue length distribution in the present episode from the bounds estimated in Step 3 (section 2.5.1). Those distribution parameters are then estimated using the MAP approach (section 2.5.2). Finally, the queue length estimate for each cycle in the episode is developed in section 2.5.3.

2.5.1 Deriving the posterior PDF of distribution parameters

We denote $f(x|\boldsymbol{\theta})$ as the PDF of queue length in episode e (for brevity, we omit the subscript e in distribution functions in this section), where x is the random variable representing queue length in the episode, and $\boldsymbol{\theta}$ is the parameter vector. From a Bayesian point of view, $\boldsymbol{\theta}$ is treated as a multivariate random variable. We assume its prior distribution is a multivariate Gaussian distribution given by $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_0^e, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_0^e$ is the prior mean, and $\boldsymbol{\Sigma}$ is the covariance matrix.

For cycle $i \in \mathbb{I}^e$, the likelihood function of observing the lower bound l^i and upper bound u^i is $P\{x^i \in [l^i, u^i]\} = F(u^i|\boldsymbol{\theta}) - F(l^i|\boldsymbol{\theta})$, where x^i is the queue length of cycle i, and $F(\cdot\,|\boldsymbol{\theta})$ is the CDF of queue length in episode e. Thus, the likelihood function of observing all the bounds in episode e is given by:

$$p(\boldsymbol{Y}^e|\boldsymbol{\theta}) = \prod_{i\in\mathbb{I}^e}\left(F(u^i|\boldsymbol{\theta}) - F(l^i|\boldsymbol{\theta})\right), \quad (17)$$

where $\boldsymbol{Y}^e = \bigcap_{i\in\mathbb{I}^e}\{x^i \in [l^i, u^i]\}$.

Based on the Bayes' theorem, the posterior PDF of $\boldsymbol{\theta}$ is proportional to the product of likelihood function and prior PDF, i.e.,

$$p(\boldsymbol{\theta}|\boldsymbol{Y}^e) \propto \prod_{i\in\mathbb{I}^e}\left(F(u^i|\boldsymbol{\theta}) - F(l^i|\boldsymbol{\theta})\right) \cdot p(\boldsymbol{\theta}). \quad (18)$$

2.5.2 Finding the MAP estimate of $\boldsymbol{\theta}$

One common method used for obtaining the expression of $p(\boldsymbol{\theta}|\boldsymbol{Y}^e)$ is the Markov Chain Monte Carlo (MCMC) sampling method. However, the computational cost of this method can be very high. To ensure computational efficiency, we develop a point estimate of $\boldsymbol{\theta}$ using the MAP approach. Specifically, we find the value of $\boldsymbol{\theta}$ that maximizes the posterior probability:

$$\boldsymbol{\theta}_{MAP}^e = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{Y}^e) = \arg\max_{\boldsymbol{\theta}} \prod_{i\in\mathbb{I}^e}\left(F(u^i|\boldsymbol{\theta}) - F(l^i|\boldsymbol{\theta})\right) \cdot p(\boldsymbol{\theta}). \quad (19)$$

The second equality of (19) holds because the ratio between $p(\boldsymbol{\theta}|\boldsymbol{Y}^e)$ and $\prod_{i\in\mathbb{I}^e}\left(F(u^i|\boldsymbol{\theta}) - F(l^i|\boldsymbol{\theta})\right) \cdot p(\boldsymbol{\theta})$ is a constant unrelated to $\boldsymbol{\theta}$.

To solve for $\boldsymbol{\theta}^e_{MAP}$, we first take the logarithm of the right-hand-side (RHS) of (19):

$$\boldsymbol{\theta}^e_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i\in\mathbb{I}^e} \left(\ln\left(F\left(u^i|\boldsymbol{\theta}\right) - F\left(l^i|\boldsymbol{\theta}\right)\right) + \ln p(\boldsymbol{\theta})\right). \quad (20)$$

The RHS of (20) can be solved via a gradient descent algorithm, in which the gradient of $\boldsymbol{\theta}$ can be calculated numerically. If $\boldsymbol{\theta}$ has only one or two elements**,** a brute-force search can also be used to rapidly find the optimal solution.

In practice, queue length is often assumed to follow a distribution in the exponential family, e.g. negative binomial distribution, Gaussian distribution, and gamma distribution (Comert and Cetin, 2009; Taylor and Heydecker, 2014). For these distributions, (20) can be solved efficiently using the expectation maximization (EM) algorithm. The details of this solution algorithm are relegated to Appendix C.

2.5.3 Estimating queue lengths for each cycle

The queue length estimate of cycle i, denoted as q^i, is then obtained by the following equation:

$$q^i = \min\left(\max\left(E[x|\boldsymbol{\theta}^e_{MAP}], l^i\right), u^i\right), \quad i \in \mathbb{I}^e, \quad (21)$$

Equation (21) basically uses the two queue length bounds of cycle i to correct the mean queue length, $E[x|\boldsymbol{\theta}^e_{MAP}]$, which is developed from the estimated distribution $f(x|\boldsymbol{\theta}^e_{MAP})$.

2.6 Updating the priors

For episode $e+1$, the prior mean shockwave speed is set to the posterior mean in episode e, i.e. $w_0^{e+1} = w_p^e$. The prior precision of shockwave speed α is unchanged. In addition, the prior mean of $\boldsymbol{\theta}$ is set to the posterior mean in episode e, i.e. $\boldsymbol{\mu}_0^{e+1} = \boldsymbol{\theta}^e_{MAP}$, while the covariance matrix $\boldsymbol{\Sigma}$ is also unchanged. We choose not to update α and $\boldsymbol{\Sigma}$ because updating them will always increase the precision, or equivalently, reduce the (co-)variance. The smaller the variance of a prior distribution, the greater the prior's effect on the estimation result will be (Bishop, 2006). Thus, after updating the prior precision or (co-)variance for a number of episodes, the estimates would depend heavily on the prior information, and thus fail to trace any changes in traffic state over time.

3. Simulation Experiments

The performance of our approach is first examined via a battery of simulation experiments. Section 3.1 describes the simulation setup; section 3.2 presents the estimation results; and section 3.3 compares the performance of our approach against two similar methods proposed in previous studies.

3.1 Simulation set up and parameter values

All the simulation experiments were conducted in Aimsun 8.1 on a personal computer with Intel Core i7-4970 CPU @ 3.60 GHz and 16G RAM. A signalized intersection with a 300m two-lane approach was created in the simulation. Only through-moving traffic was assumed. The signal's cycle length and effective green period were fixed at 100 and 40 seconds, respectively.

Aimsun generated stationary Poisson vehicle arrivals at the start of the approach, and adjusted their arrival times to ensure the headway between consecutive vehicles in each lane is never less than 2s.

Given a 2s saturation headway, the saturation flow for the intersection approach is $\frac{40s}{100s} \times \frac{3600s}{2s/vehicle} \times 2lanes = 1440$ vehicles/hour. Three Poisson arrival rates were considered in our experiments, which are equal to 40% (light traffic), 70% (medium traffic), and 97% (heavy traffic) of the saturation flow, respectively. Note that when the arrival flow is 97% of saturation flow, residual queues will frequently occur at the end of green periods. In addition, we specified seven penetration rates for probe vehicles: 2%, 4%, 6%, 8%, 10%, 15% and 20%. We simulated all the 21 scenarios combining the three vehicle flows and the seven penetration rates. In each scenario, 20 simulation runs were executed using different random seeds. Each simulation run contains 100 signal cycles (i.e., about 2 hours and 46 minutes).

Table 1 summarizes the other parameter values used in the simulation experiments. Particularly, the initial prior mean and the prior precision of w are set as $w_0^1 = -5$ m/s and $\alpha = 1$ (m/s)$^{-2}$, respectively, to represent the realistic range of w for typical urban traffic (Izadpanah et al., 2009). The β is set to 0.01 m^{-2}, meaning that the standard deviation of the noise term ϵ in (7) is 10m. For the parametric distribution of queue lengths, we choose the gamma distribution, which has a shape parameter and a scale parameter. The initial prior mean vector of the two distribution parameters, $\boldsymbol{\mu}_0^1$, is set speculatively to $\begin{bmatrix}5\\1\end{bmatrix}$, $\begin{bmatrix}10\\1\end{bmatrix}$, and $\begin{bmatrix}15\\1\end{bmatrix}$ for scenarios with light, medium, and heavy traffic, respectively (meaning that the average queue length would be 5, 10, and 15 vehicles, respectively). Speculative values are used since we assume there is not much prior knowledge about the traffic state. The prior variances of the two parameters are set to some intermediate-level values (5^2 and 1^2, respectively, appearing in the diagonal of matrix $\boldsymbol{\Sigma}$). More will be discussed in section 5 on how the choices of $\boldsymbol{\mu}_0^1$ and $\boldsymbol{\Sigma}$ affect estimation results. In addition, ε is conservatively set to 5s.

Table 1. Parameter values for simulation experiments

Parameter	Value	Definition
w_0^1	-5 m/s	Initial prior mean of discharging shockwave speed.
α	1 (m/s)$^{-2}$	Prior precision of discharging shockwave speed.
β	0.01 m^{-2}	Precision of the noise term ϵ in regression model (7).
$\boldsymbol{\mu}_0^1$	$\begin{bmatrix}5\\1\end{bmatrix}$, $\begin{bmatrix}10\\1\end{bmatrix}$, $\begin{bmatrix}15\\1\end{bmatrix}$	Initial prior mean vector of queue length distribution parameters for light, medium, and heavy traffic.
$\boldsymbol{\Sigma}$	$\begin{bmatrix}5^2 & 0\\0 & 1^2\end{bmatrix}$	Prior covariance matrix of queue length distribution parameters.
ε	5 s	Error bound for vehicles' start-up delay.
s	6.5 m	Vehicle jam spacing.
ξ	1 m/s	Speed threshold for stopped vehicles.
o	1 s	Data updating interval.

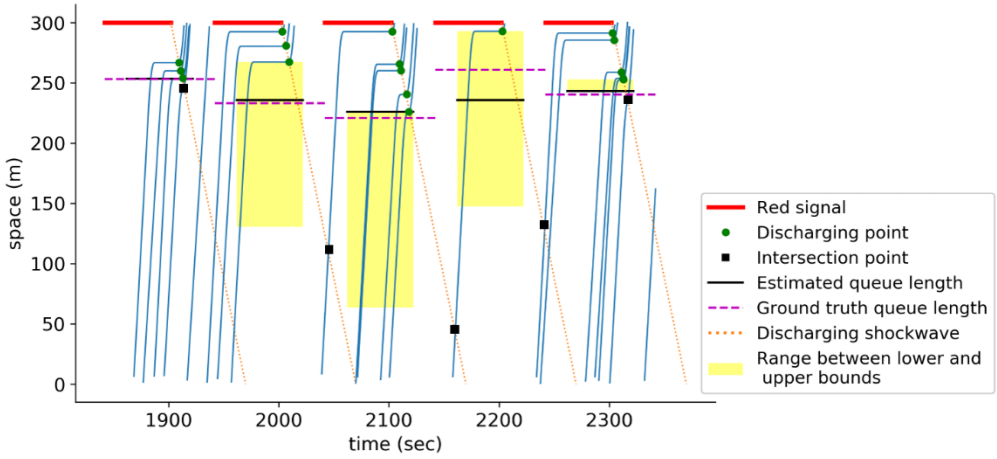
We further specified that each episode consists of 5 cycles. More analysis on the choice of episode length is also furnished in section 5.

Estimations were performed in Python 3 on the same computer. The estimation for each episode took less than 1s to complete.

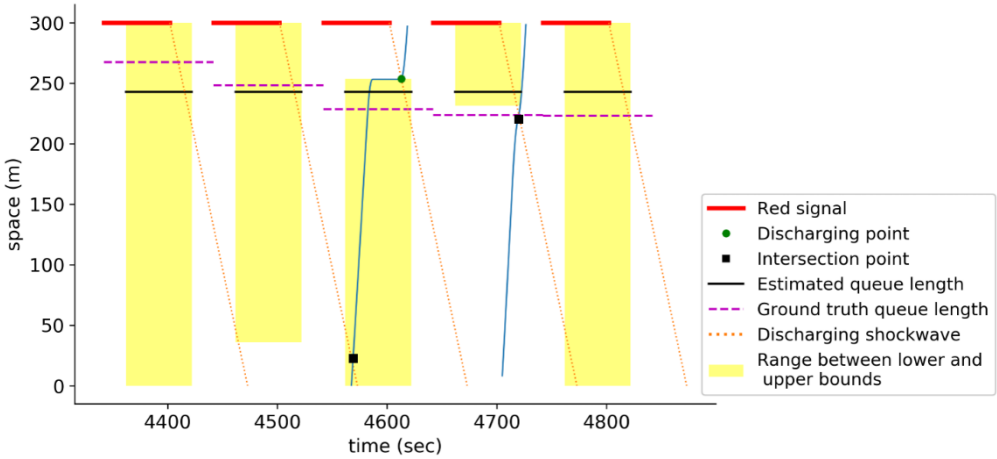
3.2 Estimation results

Figures 3a and 3b illustrate the estimation results in time-space diagrams for two typical episodes with

different penetration rates, 20% and 2%, respectively. The saturation ratio is set to 70% for both figures. The blue lines in the figures represent probe vehicle trajectories. The dotted lines represent estimated discharging shockwaves. The large, green dots on the shockwaves mark the discharging points of stopped vehicles, and the black squares mark the intersection points between the discharging shockwave and the first non-stopped vehicle trajectory in each cycle. The yellow shaded boxes indicate the range between the lower and upper bounds in each cycle. The black solid bars indicate the estimated queue lengths, and the magenta dashed bars represent the ground truth queue lengths.



(a) High penetration rate (20%)



(b) Low penetration rate (2%)

Figure 3. Queue length estimates of two typical episodes

Figure 3a shows that the estimates are very close to the ground truth for the 1st, 2nd, 3rd and 5th cycles, with errors less than a vehicle. The only large error (about three vehicles) occurs in the 4th cycle. Specifically, in the 1st cycle the lower bound almost overlaps with the upper bound since a non-stopped vehicle is observed to follow closely the last stopped vehicle. In the 2nd and the 4th cycles, the estimated queue length is equal to the episode's expected queue length since the latter is contained between the lower and upper bounds. In the 3rd and 5th cycles, however, the estimate is attained at the lower and upper bounds, respectively, because the expected queue length resides beyond the bounds. Note in the last cycle that the estimated upper bound is a little smaller than the ground truth. This false estimation occurred because the queue lengths in the two lanes were slightly unequal, and a non-stopped vehicle in one lane happened to be detected to appear downstream of a stopped vehicle in the other lane.

In Figure 3b, on the other hand, only two probe vehicles are contained over 5 cycles. The first one provides an upper bound for the 2nd cycle and a lower bound for the 3rd cycle, and the second one provides an upper bound for the 4th cycle. All the other bounds in this episode are attained at the ends of the intersection approach. Note that all the queue length estimates in this episode are the same since the expected queue length lies within all the bounds. Compared against the ground truth, these estimates are still fairly good, although not as accurate as those in Figure 3a. Also note that the expected queue length largely depends on the two tight bounds: the lower bound in the 3rd cycle and the upper bound in the 4th cycle. This means that some rough estimates can still be obtained by our approach even if only a few probe vehicle trajectories are captured.

3.3 Comparison against two previous methods

We compare the estimation errors of our method against two previous methods proposed in Ramezani and Geroliminis (2015) and Li et al. (2017), respectively. They are similar to our method in that they did not rely on assumptions regarding the vehicle arrival process and a known penetration rate. Thus, the three methods are more suitable to be applied under more dynamic traffic conditions. In Ramezani and Geroliminis, the queue length was estimated by the position of the last stopped vehicle in a cycle; i.e. their estimate is the lower bound in our method. In Li et al., the estimate was obtained by calculating the intersection point between discharging shockwave and queue formation shockwave. Non-stopped vehicles were ignored in both methods, and no estimate is derived if no stopped vehicle is detected in a cycle.

The following three performance metrics are used for comparison between the three methods under each of the 21 scenarios:

i) The success rate, defined as the proportion of cycles for which a queue length estimate is successfully obtained.

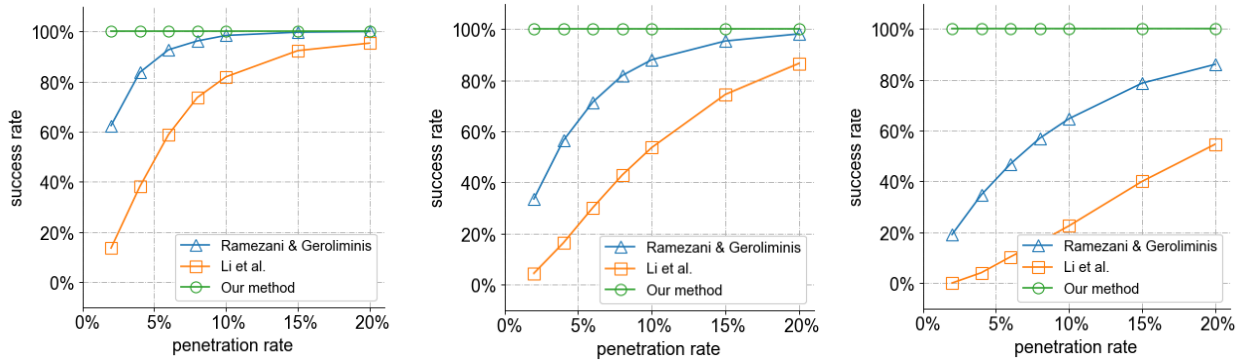
ii) The mean absolute error (MAE), defined as $\text{MAE} = \frac{\sum_{i\in\mathbb{S}}|q^i-\tilde{q}^i|}{|\mathbb{S}|}$, where q^i and $\tilde{q}^i$ denote the queue length estimate and the ground-truth queue length of cycle i, respectively; and $\mathbb{S}$ denotes the index set of cycles for which a queue length estimate is obtained. The smaller the MAE, the more accurate the estimates are.

iii) The standard deviation of absolute error (SDAE), defined as $\text{SDAE} = \sqrt{\frac{\sum_{i\in\mathbb{S}}\left(|q^i-\tilde{q}^i|-\text{MAE}\right)^2}{|\mathbb{S}|-1}}$. The smaller the SDAE, the more robust the estimates are.

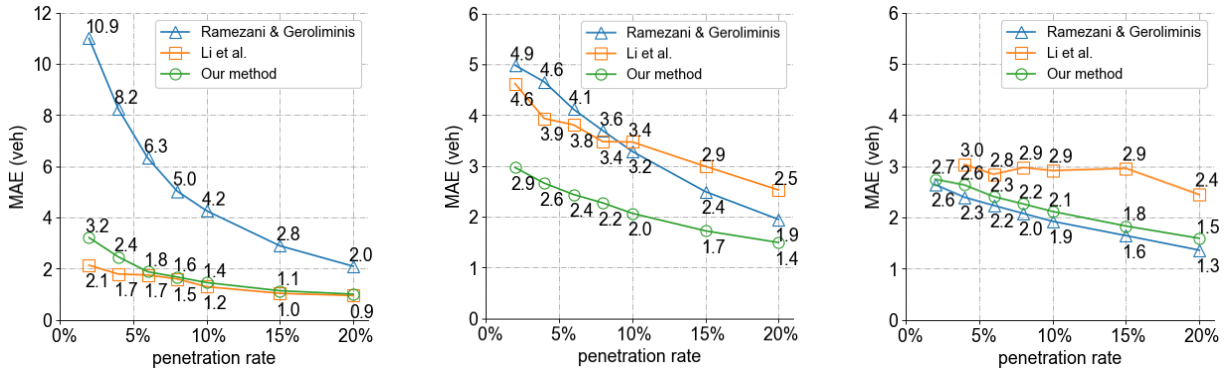
The success rate, MAE, and SDAE are plotted against the penetration rate for heavy, medium, and light traffic in Figures 4a-i. In each of the nine figures, our estimation results are shown by green curves with circular markers; results of Ramezani and Geroliminis' method are shown by blue curves with triangular markers; and results of Li et al.'s method are shown by orange curves with square markers.

First note in Figures 4a-c that our approach renders a 100% success rate for all the scenarios tested, while the two previous methods have much low success rates especially under low penetration rates. For example, under 2% penetration, Ramezani and Geroliminis' method generated estimates for only 20-60% of the cycles, and Li et al.'s method has an even lower success rate of only 0-15%. This is because the latter method requires at least two stopped vehicles to be detected in a cycle.[6]

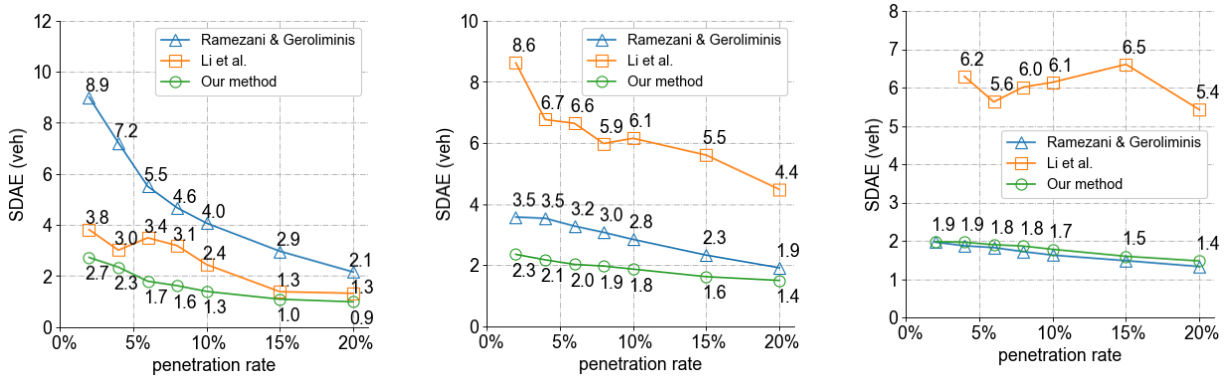
[6] One may argue that the two previous methods can also be modified such that when a cycle does not contain



(a) Success rate under heavy traffic **(b) Success rate under medium traffic** **(c) Success rate under light traffic**



(d) MAE under heavy traffic **(e) MAE under medium traffic** **(f) MAE under light traffic**



(g) SDAE under heavy traffic **(h) SDAE under medium traffic** **(i) SDAE under light traffic**

Figure 4. Comparison of three methods under various vehicle flows and penetration rates

Despite the 100% success rate, the estimation accuracy is not compromised. See in Figures 4d-f that the MAE of our estimates is consistently low, regardless of the traffic volume. In contrast, Ramezani and Geroliminis' method yielded slightly lower MAE only under light traffic, while Li et al.'s method performed well only under heavy traffic. This is as expected. For example, when the traffic is light, the ground-truth queue length is usually small and close to Ramezani and Geroliminis' lower bound estimate, if a stopped vehicle is detected (note however that the success rate under light traffic is quite low for that method; see Figure 4c). On the other hand, vehicle arrivals are by-and-large uniform under heavy traffic, and thus the queue formation shockwave estimated by Li et al.'s method would be fairly accurate.

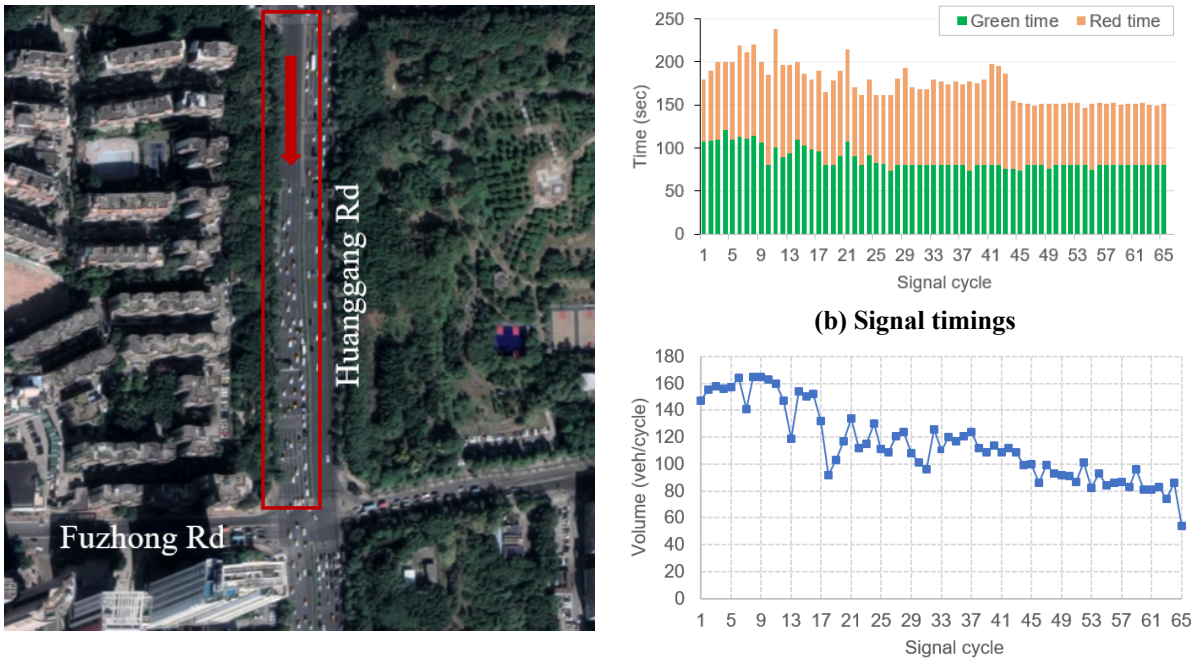
---

enough probe vehicle data, the queue length is estimated using the average of estimates in some neighboring "successful" cycles. However, this may render the estimation accuracy of their methods considerably lower than that shown in Figures 4d-f. Nevertheless, this possibility was not discussed and tested in the cited works.

Finally, Figures 4g-i reveal that the SDAE of our method is also consistently lower than that of the two previous methods, save for the light traffic case where our SDAE is slightly higher than that of Ramezani and Geroliminis'. Our low SDAE is also insensitive to both traffic volume and penetration rate. This manifests that our method can produce more reliable and robust queue length estimates across all the cycles under various operating conditions.

4. A Real-world Case Study

To examine the performance of our method under more dynamic traffic scenarios with time-varying signal timing, vehicle flow and penetration rate, a case study is conducted using traffic data collected at a real intersection approach. The site is the southbound approach of the intersection of Fuzhong Rd. and Huanggang Rd. in Shenzhen, China, as shown in Figure 5a. We focus on the four through-moving lanes in the approach. The left-turning and right-turning traffic use separate lanes. The signal timing and probe vehicle trajectory data were collected from 9:30AM to 12:30PM on Thursday, April 13, 2017 by Didi Chuxing Technology Co. Ltd. The 3-hour study period consists of 65 adaptively controlled signal cycles, and their green and red periods are plotted in Figure 5b. Figure 5c illustrates the traffic volume of each cycle, counted from a video taken on-site during the same period. The traffic volume waned over time, showing the transition from a morning peak period to an off-peak period.



(a) Bird's-eye view of the intersection approach **(b) Signal timings** **(c) Traffic volumes**

Figure 5. The case study site

The penetration rate of probe vehicles varied in the range of 5-10% during the study period, resulting in an average of 10 probe vehicles per cycle. The vehicle positions were updated every 3 seconds (i.e. $o = 3$s). We still specify that each episode contains 5 cycles (thus there are totally 13 episodes), and that the queue length in an episode follows a gamma distribution. The initial prior mean of queue length distribution parameters, $\boldsymbol{\mu}_0^1$, is set as $\begin{bmatrix}18\\1\end{bmatrix}$. Other parameters are set to the same values as in Table 1.

Figure 6 compares the estimates (the orange solid curve) against the ground-truth queue lengths (the blue dashed curve, which are also measured from the on-site video) for all the 65 cycles. The ground-truth queue length fluctuated roughly between 20-30 vehicles during the first 45 cycles, and then dropped to 10-20 vehicles in the last 20 cycles. The figure shows that our estimates match the true queue lengths fairly well. No significant bias is found. The MAE of our estimates is only 1.7 vehicles.

Due to the highly dynamic feature of this case study, it is interesting to know how the priors in our method update over time to capture the changing traffic state. Figure 7a plots the prior mean shockwave speed w_0^e over the 13 episodes, showing that w_0^e fluctuates only in a small range from -5.1 to -4.1 m/s. On the other hand, the prior mean queue length $E[x|\boldsymbol{\mu}_0^e]$ exhibits a clear decreasing trend after episode 10, as shown in Figure 7b. Note how this decreasing trend closely matches that of the ground-truth queue length curve in Figure 6. For example, a peak of the ground-truth queue length curve occurs in cycles 41-45 (i.e. episode 9) in Figure 6, which is echoed in Figure 7b by a peak of the prior mean queue length curve in episode 10.

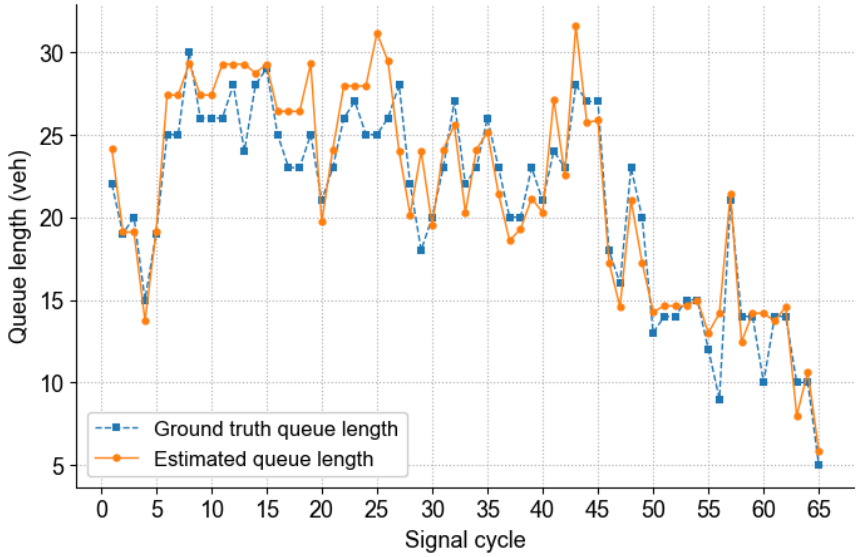
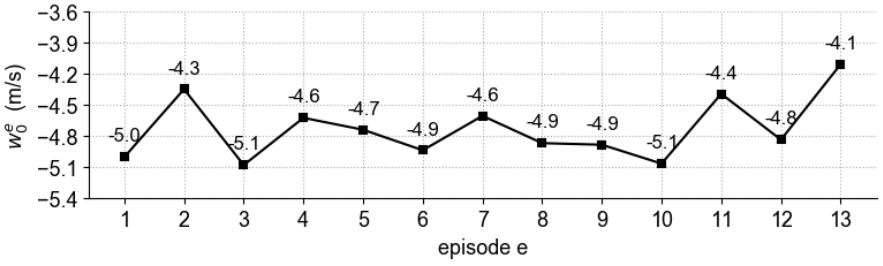
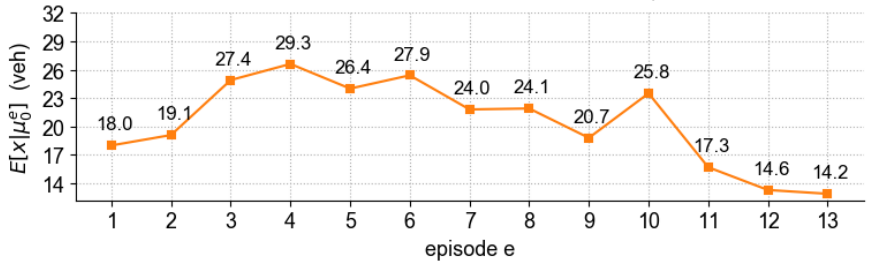


Figure 6. Comparison between the queue length estimates and the ground truth

The good estimation accuracy may be partially attributed to the high frequency of probe vehicles (about 10 probe vehicles per cycle). Thus, to examine our method's performance for real traffic with low probe vehicle frequencies, we conducted a sensitivity analysis using the same dataset of the studied case. Specifically, we kept the signal timings and traffic volumes unchanged, and assumed that only a randomly sampled proportion, ρ, of the actual probe vehicles were observed. The sampling rate ρ was set to 10%, 20%, 40%, 60%, and 80%. For each sampling rate, the sampling was repeated for 20 times with different random seeds. We then estimated the queue lengths using our method and the two previous methods for each sampling rate. The resulting success rate, MAE and SDAE are plotted against the average number of probe vehicles per cycle in Figures 8a-c, respectively. Note that the points for 10 probe vehicles per cycle are for the real case (i.e. $\rho = 100\%$). The figures reveal that our method performed fairly good even when only one probe vehicle is detected per cycle: the MAE is only 3.0 and the SDAE is 2.4. They also show that our method prevailed over the previous two consistently and significantly.

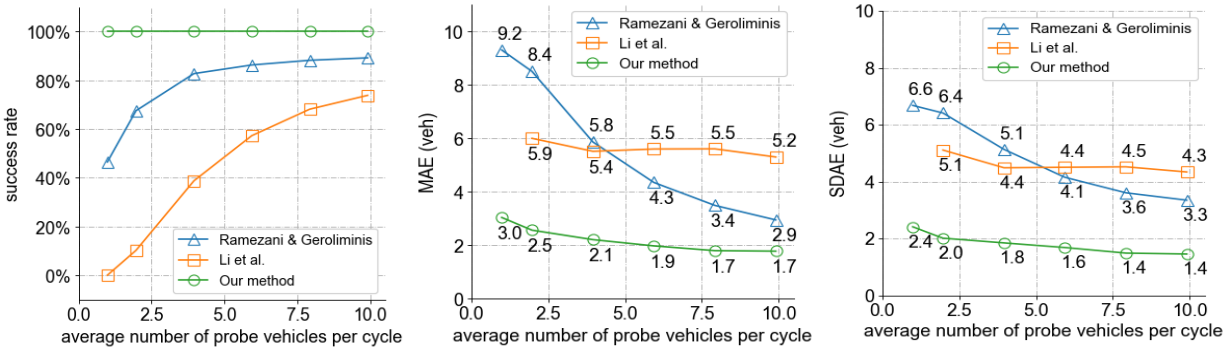


(a) The prior mean shockwave speed, w_0^e



(b) The prior mean queue length, $E[x|\mu_0^e]$

Figure 7. Updating of the prior means



(a) Success rate (b) MAE (c) SDAE

Figure 8. Comparison of three methods under various probe vehicle frequencies

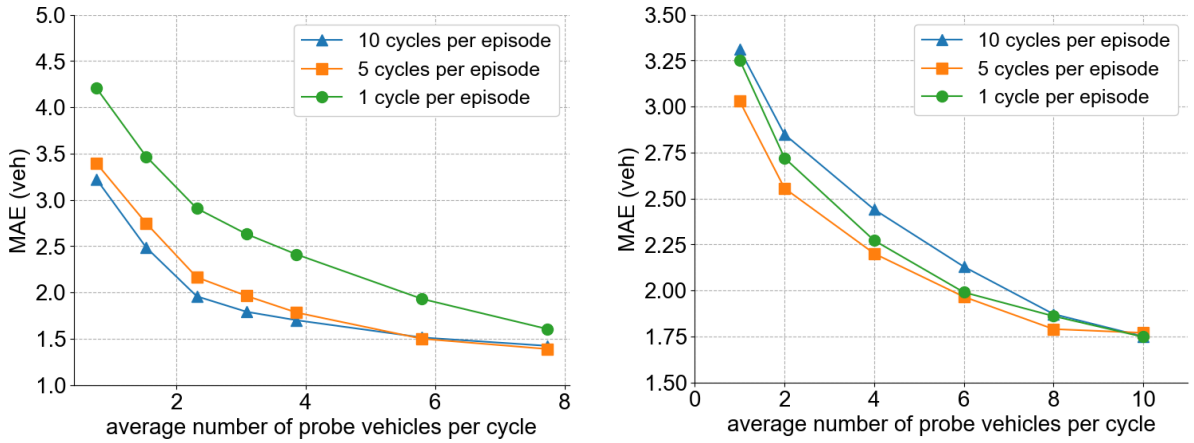
5. Choice of Episode Length and Prior Parameter Values

In this section we discuss how to set the values for some key model parameters, including the number of cycles in an episode, the queue length distribution parameters' initial prior mean vector, $\boldsymbol{\mu}_0^1$, and prior covariance matrix, $\boldsymbol{\Sigma}$.

The appropriate episode length depends on the number of probe vehicles that are contained in an episode, and how fast the traffic state changes. If an episode is too short and contains only few probe vehicle trajectories, the estimation accuracy may be low. On the other hand, in a too long episode, assumption (iv) in section 2.1 may be violated, and our method would fail to capture the change of traffic state over time. For illustration, we compare the MAE of estimation results when each episode contains 1, 5, and 10 cycles for simulation experiments under the heavy-traffic scenario in Figure 9a, and for the real-world case in Figure 9b.

The two figures reveal different results. In Figure 9a, estimates with 10-cycle episodes (the blue curve) exhibit the lowest MAE, followed by estimates with 5-cycle episodes (the orange curve), and then by those with 1-cycle episodes (the green curve). This is because the simulated scenario has a fixed signal plan and a stationary traffic flow, and hence the distribution of queue length does not vary

largely over time. As a result, a longer episode containing more probe vehicles will yield more accurate estimates. In Figure 9b, however, estimates with 10-cycle episodes have the greatest MAE, because in the real case the traffic condition varied over time, and a 10-cycle episode cannot capture this time-varying traffic condition well. In practice, we recommend using episodes of 3-8 cycles long. Practitioners can also calibrate this value using historical data to attain a better estimation performance.



(a) Simulated scenarios under 97% saturation ratio (b) The real-world case (with different sampling rates)

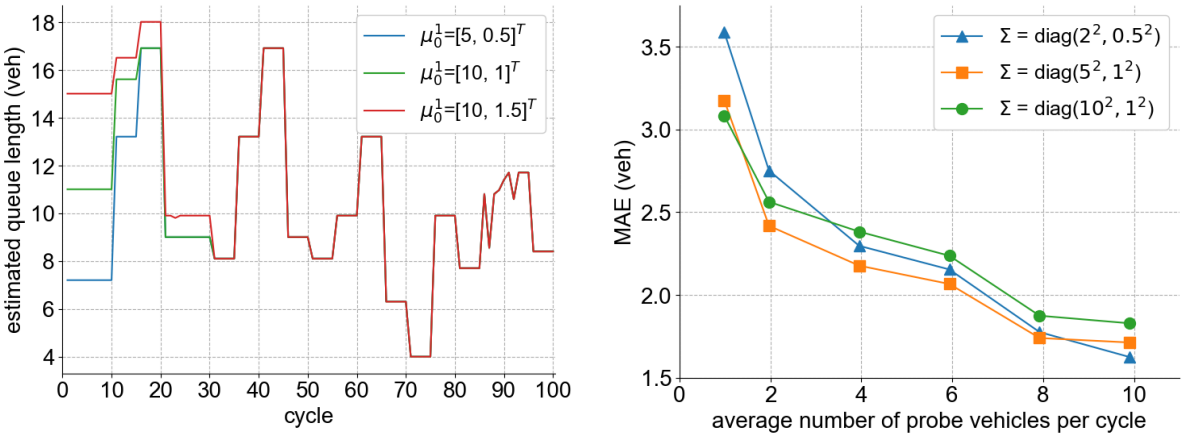
Figure 9. Effect of episode length on the estimation accuracy

Sensitivity analyses were also conducted with respect to $\boldsymbol{\mu}_0^1$ and $\boldsymbol{\Sigma}$. Figure 10a plots, for a randomly generated simulation experiment under medium traffic and 2% penetration, how the queue length estimates vary over cycles when different values of $\boldsymbol{\mu}_0^1$ are used. Here $\boldsymbol{\Sigma}$ is set to $\begin{bmatrix} 5^2 & 0 \\ 0 & 1^2 \end{bmatrix}$. The figure shows that notable differences between the queue length estimates under the three $\boldsymbol{\mu}_0^1$ values examined are mostly observed only for the first a few cycles. For the remaining cycles in the estimation period, the difference is usually less than 1 vehicle. Similar results were also observed for other simulation experiments and the real-world case study. This is as expected, since the effect of initial prior mean information is diluted quickly during the iterative estimation process. Figure 10b plots the MAE against the average number of probe vehicles per cycle for different values of $\boldsymbol{\Sigma}$ in the real-world case, where $\boldsymbol{\mu}_0^1$ is set to $\begin{bmatrix} 18 \\ 1 \end{bmatrix}$. The figure reveals that the difference in MAE produced with various values of $\boldsymbol{\Sigma}$ is also modest (no greater than 0.5 vehicle in the figure). Therefore, the accuracy of our approach seems to be quite robust to the two prior parameters. In practice, we recommend using similar values as we used in this paper for those parameters (if gamma distribution is used for queue lengths) in the absence of further knowledge about the traffic condition. The prior covariances can be set relatively higher if the traffic condition is known to vary quickly (so as to reduce the impact of prior means), and relatively lower if the traffic is stationary. One can also calibrate $\boldsymbol{\mu}_0^1$ and $\boldsymbol{\Sigma}$ together with the episode length if the historical data are available.

6. Conclusions

A Bayesian method is developed to estimate the maximum queue lengths in a signalized intersection approach. We propose a MAP formulation for estimating the distribution of queue lengths in an episode consisting of several neighboring cycles, which takes the lower and upper bounds of queue lengths as inputs. The prior means of queue length distribution parameters are updated episode by

episode to capture the time-varying traffic condition. Both the queue length and the discharging shockwave speed (which is used to derive the upper bounds of queue lengths) are modeled stochastically. An EM algorithm is proposed to find the solution to the MAP problem efficiently.



(a) Effect of μ_0^1 (for a simulation experiment under medium traffic and 2% penetration)

(b) Effect of Σ (for the real-world case)

Figure 10. Effects of prior queue length distribution parameters

Our method does not rely on any assumption regarding vehicle arrivals. Instead, we assume the queue lengths in an episode follow a parametric distribution. In practice, using a common distribution of the exponential family is recommended, for instance a Gaussian or gamma distribution.

Through a large array of simulation experiments and a real-world case study, the performance of our method is shown to be pretty good. In most cases tested in this paper, our estimates exhibit a MAE of only 1-3 vehicles, superior to two previous methods reported in the literature; see Figures 4d-f and 8b. Recall that the low MAE is achieved when estimates were obtained for all the cycles, no matter whether a cycle contains probe vehicle data or not. In contrast, the two previous methods have quite low success rates especially under low penetration rates. In addition, our method furnishes more robust estimates, as manifested by the consistently low MAE and SDAE for varying traffic condition and probe vehicle frequencies; see Figures 4d-i and 8b-c. The real-world case study also verified that our approach works well under highly dynamic real traffic condition.

For practical applications, the episode length, the initial prior means and the prior covariance matrix of queue length distribution parameters should be selected by the practitioners. Calibration is not a must, but is recommended if historical data are available.

Limitations of our approach are also noted. For example, it cannot be directly applied to the scenario where a left-turning vehicle queue spills over its storage bay to mix with the through-moving queue, or when different lanes have significantly different queue lengths. We are currently developing novel algorithms to address these problems. In addition, our approach may not perform well when the vehicle inflow varies largely over cycles with a low penetration rate. In this case, using long episodes would overlook the temporal variation in traffic, while using short episodes would produce inaccurate estimates due to the low probe vehicle frequency. Potential solutions to this issue are also under exploration (a promising one is to integrate nearby intersections' probe vehicle data into the estimation approach). Another potential extension of the present work is to integrate the signal timing estimation (Hao et al., 2012, and Du et al., 2019) into our queue length estimation approach, so that

637 the integrated approach can be applied where the signal timing information is unavailable.

638 Acknowledgements

639 This research was supported by General Research Funds (No. 15217415 and 15224317) provided by
640 the Research Grants Council of Hong Kong and a startup fund provided by the Hong Kong
641 Polytechnic University. We thank the Didi Chuxing Technology Co. Ltd. for providing the probe
642 vehicle data and signal timings for the case study.

643 Appendix A. Table of notation

Notation	Description
L	Length of the intersection approach
e	Episode number
i	Cycle number
j	Probe vehicle number
k	Data slice number of a specific vehicle
d	Vehicle location
t	Time
$d_{j,k}$	Vehicle location in data slice k for vehicle j
$t_{j,k}$	Timestamp in data slice k for vehicle j
$v_{j,k}$	Instantaneous speed in data slice k for vehicle j
j^i	Number of the first non-stopped vehicle in cycle i
k^i	Number of the data slice used to identify the first non-stopped vehicle j^i
a_{j^i}	Maximum deceleration rate of vehicle j^i
s	Jam spacing
$\bar{d}_j$	Location of discharging point for stopped vehicle j
$\bar{t}_j$	Timestamp of discharging point for stopped vehicle j
r^i	Effective red start time of cycle i
g^i	Effective green start time of cycle i
l^i	Lower bound of queue length in cycle i
u^i	Upper bound of queue length in cycle i
$\mathcal{S}^i$	Target zone of cycle i
$\mathcal{D}^i$	Discharging zone of cycle i
$\mathcal{C}^e$	Set of all the discharging points in episode e
$\mathbb{I}^e$	Set of numbers of cycles that belong to episode e
$\mathbb{V}^i$	Set of numbers of probe vehicles in cycle i
$\mathbb{Q}^i$	Set of numbers of stopped vehicles in cycle i
$\mathbb{M}^i$	Set of numbers of non-stopped vehicles in cycle i
$\mathbb{S}$	Set of numbers of cycles for which queue lengths are successfully estimated
w	Discharging shockwave speed
$[w^-, w^+]$	99.7% confidence interval of discharging shockwave speed
w_0^e	Prior mean of discharging shockwave speed of episode e
w_p^e	Posterior mean of discharging shockwave speed of episode e
α	Prior precision of discharging shockwave speed
α_p^e	Posterior precision of discharging shockwave speed in episode e
o	Data updating interval
ε	Error bound for vehicles' start-up delay
ξ	Speed threshold to identify stopped vehicles

ϵ	Zero-mean Gaussian random noise for regression of discharging shockwave speed
β	Precision of the Gaussian noise ϵ
δ	A small positive number to ensure the lower and upper queue length bounds do not equal
$\boldsymbol{\theta}$	Parameter vector of the queue length distribution
$\boldsymbol{\theta}^e_{MAP}$	MAP estimator of $\boldsymbol{\theta}$ in episode e
$\boldsymbol{\mu}^e_0$	Prior mean of $\boldsymbol{\theta}$ in episode e
$\boldsymbol{\Sigma}$	Covariance matrix of $\boldsymbol{\theta}$
$f(x\|\boldsymbol{\theta})$	PDF of queue length
$F(x\|\boldsymbol{\theta})$	CDF of queue length
x	Random variable representing a queue length in an episode
x^i	Random variable representing the queue length in cycle i
$\boldsymbol{X}^e$	The set of true queue lengths in episode e
$\boldsymbol{Y}^e$	The set of observed lower and upper queue length bounds in episode e
N^e	The number of cycles in episode e
q^i	Queue length estimate in cycle i
$\tilde{q}^i$	Ground-truth queue length in cycle i
ρ	Sampling rate

Appendix B. Derivation of the posterior distribution of w

To find the explicit form of the posterior distribution $p(w|\mathcal{C}^e)$, we take the logarithm of the PDF:

$$\ln\big(p(w|\mathcal{C}^e)\big) = -\frac{\beta}{2}\sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\big(\bar{d}_j - w\big(\bar{t}_j - g^i\big) - L\big)^2 - \frac{\alpha}{2}(w - w_0^e)^2 + C_0, \quad \text{(B1)}$$

where C_0 is a constant.

For Gaussian distribution, the variance (inverse of the precision) can be obtained by checking the second-order term's coefficient in the logarithm. The second-order term of the logarithm is $-\frac{1}{2}\Big(\beta\sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\big(\bar{t}_j - g^i\big)^2 + \alpha\Big)w^2$. Hence the posterior precision is given by:

$$\big(\alpha_p^e\big)^{-1} = \Big(\beta\sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\big(\bar{t}_j - g^i\big)^2 + \alpha\Big)^{-1}. \quad \text{(B2)}$$

Similarly, the posterior mean can be found by checking the first-order term's coefficient, which is $\big[\beta\sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\big(\bar{d}_j - L\big)\big(\bar{t}_j - g^i\big) + \alpha\cdot w_0^e\big]$. The mean is then calculated by multiplying the above coefficient with the precision:

$$w_p^e = \big(\alpha_p^e\big)^{-1}\big[\beta\sum_{i\in\mathbb{I}^e}\sum_{j\in\mathbb{Q}^i}\big(\bar{d}_j - L\big)\big(\bar{t}_j - g^i\big) + \alpha\cdot w_0^e\big]. \quad \text{(B3)}$$

Appendix C. The EM algorithm for solving for $\boldsymbol{\theta}^e_{MAP}$

We write the queue length PDF following the definition of exponential-family distributions (Bishop, 2006):

$$f(x|\boldsymbol{\theta}) = h(x)g(\boldsymbol{\theta})\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(x)\}, \quad \text{(C1)}$$

where $\boldsymbol{\eta}(\boldsymbol{\theta})$ represents the natural parameters of the distribution; $h(x)$ and $\boldsymbol{T}(x)$ are functions of x; and $g(\boldsymbol{\theta})$ is added to ensure the PDF integrates to 1.

To apply the EM algorithm, we first note that the true queue lengths in episode e, $\boldsymbol{X}^e = \{x^i | i \in \mathbb{I}^e\}$, can be regarded as latent variables. The logarithm of the complete-data likelihood function (i.e. the likelihood of observing both $\boldsymbol{X}^e$ and the bounds $\boldsymbol{Y}^e$) can be written as:

$$\ln p(\boldsymbol{X}^e, \boldsymbol{Y}^e | \boldsymbol{\theta}) = \sum_{i \in \mathbb{I}^e} \ln f\left(x^i \middle| \boldsymbol{\theta}\right) = N^e \ln g(\boldsymbol{\theta}) + \sum_{i \in \mathbb{I}^e} \left(\ln h\left(x^i\right) + \boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}\left(x^i\right)\right), \quad \text{(C2)}$$

where $N^e = |\mathbb{I}^e|$ is the number of cycles in episode *e*.

The EM algorithm recursively executes the expectation (E) step and the maximization (M) step until the solution converges. The E step first finds the posterior distribution of the latent variables when the parameter vector $\boldsymbol{\theta}$ is fixed as $\boldsymbol{\theta}^{old}$:

$$p\left(\boldsymbol{X}^e \middle| \boldsymbol{Y}^e, \boldsymbol{\theta}^{old}\right) = \frac{\prod_{i \in \mathbb{I}^e} f(x^i | \boldsymbol{\theta}^{old})}{\prod_{i \in \mathbb{I}^e} \left(F(u^i | \boldsymbol{\theta}^{old}) - F(l^i | \boldsymbol{\theta}^{old})\right)}. \quad \text{(C3)}$$

This posterior distribution is used to calculate the expectation of the complete-data log likelihood with respect to $\boldsymbol{X}^e$. This expectation, denoted as $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right)$, is given by:

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \boldsymbol{E}_{\boldsymbol{X}^e \sim p\left(\boldsymbol{X}^e | \boldsymbol{Y}^e, \boldsymbol{\theta}^{old}\right)}[\ln p(\boldsymbol{X}^e, \boldsymbol{Y}^e | \boldsymbol{\theta})] = N^e \ln g(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})^T \sum_{i \in \mathbb{I}^e} \boldsymbol{\gamma}^i + C_1, \quad \text{(C4)}$$

where $\boldsymbol{\gamma}^i$ is the conditional expectation of $\boldsymbol{T}(x^i)$, given by $\boldsymbol{\gamma}^i = \frac{\int_{l^i}^{u^i} \boldsymbol{T}(x^i) f(x^i | \boldsymbol{\theta}^{old}) dx^i}{F(u^i | \boldsymbol{\theta}^{old}) - F(l^i | \boldsymbol{\theta}^{old})}$; $C_1 = \sum_{i \in \mathbb{I}^e} \frac{\int_{l^i}^{u^i} \ln h(x^i) f(x^i | \boldsymbol{\theta}^{old}) dx^i}{F(u^i | \boldsymbol{\theta}^{old}) - F(l^i | \boldsymbol{\theta}^{old})}$ is a term unrelated to $\boldsymbol{\theta}$.

In the M step, we maximize the sum of $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right)$ and the logarithm of $p(\boldsymbol{\theta})$ (prior PDF) with respect to $\boldsymbol{\theta}$, while $\boldsymbol{\gamma}^i$'s are fixed; i.e.,

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) + \ln p(\boldsymbol{\theta}) = \boldsymbol{E}_{\boldsymbol{X}^e \sim p\left(\boldsymbol{X}^e | \boldsymbol{Y}^e, \boldsymbol{\theta}^{old}\right)}[\ln p(\boldsymbol{X}^e, \boldsymbol{Y}^e | \boldsymbol{\theta})] + \ln p(\boldsymbol{\theta})$$

$$= N^e \ln g(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})^T \sum_{i \in \mathbb{I}^e} \boldsymbol{\gamma}^i - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_0^e)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0^e) + C_2, \quad \text{(C5)}$$

where C_2 is also a term unrelated to $\boldsymbol{\theta}$.

Maximization of the RHS of (C5) is usually tractable, even though the original problem may not be. Specifically, the first order condition of the RHS of (C5) with respect to $\boldsymbol{\theta}$ yields the following equation for the optimal solution $\boldsymbol{\theta}^*$:

$$\boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\mu}_0^e) - N^e \frac{1}{g(\boldsymbol{\theta}^*)} \frac{\partial g(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \sum_{i \in \mathbb{I}^e} \boldsymbol{\gamma}^i. \quad \text{(C6)}$$

For common distributions in the exponential family, e.g. negative binomial, Gaussian, and gamma, the partial derivatives $\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $\frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ can be derived analytically. Thus $\boldsymbol{\theta}^*$ can be easily computed numerically or analytically (Bishop, 2006).

The EM algorithm then sets $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^*$ and proceeds to the next iteration. The algorithm stops when the values of $\boldsymbol{\theta}^*$ converge under a certain tolerance level.

References

Badillo, B.E., Rakha, H., Rioux, T.W., and Abrams, M., 2012. Queue length estimation using conventional vehicle detector and probe vehicle data. The 15th International IEEE Conference on Intelligent Transportation Systems, 1674-1681.

Ban, X.J., Hao, P., and Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C 19(6), 1133-1156.

Bishop, C.M., 2006. Pattern recognition and machine learning. Information Science and Statistics. Springer, Heidelberg.

Cai, Q., Wang, Z., Zheng, L., Wu, B., and Wang, Y., 2014. Shock wave approach for estimating queue length at signalized intersections by fusing data from point and mobile sensors. Transportation Research Record 2422, 79-87.

Cheng, Y., Qin, X., Jin, J., Ran, B., and Anderson, J., 2011. Cycle-by-cycle queue length estimation for signalized intersections using sampled trajectory data. Transportation Research Record 2257, 87-94.

Cheng, Y., Qin, X., Jin, J., and Ran, B., 2012. An exploratory shockwave approach to estimating queue length using probe trajectories. Journal of Intelligent Transportation Systems 16(1), 12-23.

Comert, G., and Cetin, M., 2007. Queue length estimation from probe vehicle location: Undersaturated conditions. Transportation Research Board 86th Annual Meeting, Washington DC, United States, 07-2558.

Comert, G., and Cetin, M., 2009. Queue length estimation from probe vehicle location and the impacts of sample size. European Journal of Operational Research 197(1), 196-202.

Comert, G., and 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(2), 563-573.

Comert, G., 2013. Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals. Transportation Research Part B 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(2), 502-521.

Coifman, B., Dhoorjaty, S., and Lee, Z.H., 2003. Estimating median velocity instead of mean velocity at single loop detectors. Transportation Research Part C 11(3-4), 211-222.

Cronje, W.B., 1983. Derivation of equations for queue length, stops, and delay for fixed-time traffic signals. Transportation Research Record 905, 93-95.

Du, Z., Yan, X., Zhu, J., and Sun, W., 2019. Signal timing parameters estimation for intersections using floating car data. Transportation Research Record. https://doi.org/10.1177/0361198119844756.

Gartner, N.H., Assman, S.F., Lasaga, F., and Hou, D.L., 1991. A multi-band approach to arterial traffic signal optimization. Transportation Research Part B 25(1), 55-74.

Hao, P., Ban, X.J., Bennett, K.P., Ji, Q., and Sun, Z., 2012. Signal timing estimation using sample intersection travel times. IEEE Transactions on Intelligent Transportation Systems, 13(2), 792-804.

Hao, P., Ban, X.J., Guo, D., and Ji, Q., 2014. Cycle-by-cycle intersection queue length distribution estimation using sample travel times. Transportation Research Part B 68, 185-204.

Hao, P., and Ban, X.J., 2015. Long queue estimation for signalized intersections using mobile data. Transportation Research Part B 82, 54-73.

Izadpanah, P., Hellinga, B., and Fu, L. 2009. Automatic traffic shockwave identification using

754 vehicles' trajectories. Transportation Research Board 88th Annual Meeting, Washington DC, United
755 States, 09-0807.
756 Kittelson, W.K., Courage, K.G., Kyte, M.D., List, G.F., Roess, R.P., and Sampson, W.M.,
757 2004. Highway Capacity Manual Applications Guidebook. National Cooperative Highway Research
758 Program.
759 Li, F., Tang, K., Yao, J., and Li, K., 2017. Real-Time Queue Length Estimation for Signalized
760 Intersections Using Vehicle Trajectory Data. Transportation Research Record 2623, 49-59.
761 Li, J. Q., Zhou, K., Shladover, S., and Skabardonis, A., 2013. Estimating queue length under
762 connected vehicle technology: Using probe vehicle, loop detector, and fused data. Transportation
763 Research Record 2356, 17-22.
764 Lighthill, M.J., and Whitham, G.B., 1955. On kinematic waves. I. Flood movements in long
765 rivers. II. A theory of traffic flow on long crowded roads. In: Proceedings of the Royal Society
766 (London) A 229, 281-345.
767 Liu, H.X., Wu, X., Ma, W., and Hu, H., 2009. Real-time queue length estimation for
768 congested signalized intersections. Transportation Research Part C 17(4), 412-427.
769 Manar, A., and Baass, K., 1996. Traffic platoon dispersion modeling on arterial streets.
770 Transportation Research Record 1566, 49-53.
771 Mei, Y., Tang, K., and Li, K., 2015. Real-time identification of probe vehicle trajectories in
772 the mixed traffic corridor. Transportation Research Part C 57, 55-67.
773 Michalopoulos, P. G., and Stephanopoulos, G., 1977. Oversaturated signal systems with
774 queue length constraints - I: Single intersection. Transportation Research 11(6), 413-421.
775 Michalopoulos, P. G., and Stephanopoulos, G., 1977. Oversaturated signal systems with
776 queue length constraints - II: Systems of intersections. Transportation Research, 11(6), 423-428.
777 Muck, J., 2002. Using detectors near the stop-line to estimate traffic flows. Traffic
778 Engineering and Control, 43, 429-434.
779 Newell, G.F., 1965. Approximation methods for queues with application to the fixed-cycle
780 traffic light. SIAM Review 7(2), 223-240.
781 Newell, G.F., 1993. A simplified theory of kinematic waves in highway traffic, part I~III.
782 Transportation Research Part B 27(4), 281-313.
783 Ramezani, M., and Geroliminis, N., 2015. Queue profile estimation in congested urban
784 networks with probe data. Computer-Aided Civil and Infrastructure Engineering 30(6), 414-432.
785 Rao, A.M., and Rao, K.R., 2012. Measuring Urban Traffic Congestion - A
786 Review. International Journal for Traffic and Transport Engineering 2(4), 286-305.
787 Richards, P.I., 1956. Shockwaves on the highway. Operations Research 4(1), 42-51.
788 Sharma, A., Bullock, D., and Bonneson, J., 2007. Input-output and hybrid techniques for real-
789 time prediction of delay and maximum queue length at signalized intersections. Transportation
790 Research Record 2035, 69-80.
791 Shelby, S., Bullock, D., and Gettman, D., 2006. Transition methods in traffic signal
792 control. Transportation Research Record 1978, 130-140.
793 Skabardonis, A., and Geroliminis, N., 2008. Real-time Monitoring and Control on Signalized
794 Arterials. Journal of Intelligent Transportation Systems 12(2), 64-74.
795 Taylor, N.B., and Heydecker, B.G., 2014. The effect of green time on stochastic queues at
796 traffic signals. Transportation Planning and Technology 37(1), 3-19.
797 Tiaprasert, K., Zhang, Y., Wang, X.B., & Zeng, X., 2015. Queue length estimation using
798 connected vehicle technology for adaptive signal control. IEEE Transactions on Intelligent
799 Transportation Systems 16(4), 2129-2140.
800 Vigos, G., Papageorgiou, M., and Wang, Y., 2008. Real-time estimation of vehicle-count

801 within signalized links. Transportation Research Part C 16(1), 18-35.
802 Wang, Z., Cai, Q., Wu, B., Zheng, L., and Wang, Y., 2017. Shockwave-based queue
803 estimation approach for undersaturated and oversaturated signalized intersections using multi-source
804 detection data. Journal of Intelligent Transportation Systems 21(3), 167-178.
805 Webster, F.V., and Cobbe, B.M., 1966. Traffic signals. Road Research Technical Paper 56,
806 Road Research Laboratory, London.
807 Yin, J., Sun, J., and Tang, K., 2018. A Kalman filter based queue length estimation method
808 with low-penetration mobile sensor data at signalized intersections. Transportation Research Record
809 2672, 253-264.
810 Zheng, J., and Liu, H.X., 2017. Estimating traffic volumes for signalized intersections using
811 connected vehicle data. Transportation Research Part C 79, 347-362.