Queue Length Estimation from Connected Vehicles with Low and Unknown Penetration Level

Hamidreza Tavafoghi*, Jared Porter*, Christopher Flores‡, Kameshwar Poolla*†, Pravin Varaiya†

***Abstract*—Queue length estimation has been a long-standing problem in transportation systems as it provides an important component for the design, operation, and performance monitoring of signalized intersections. In this paper, we present a novel estimation algorithm based on trace data from connected vehicles. In contrast to existing algorithms, our algorithm only requires a very low level of penetration rate ($\sim$1%) for connected vehicles. As such, it is already applicable given the current level of penetration in practice. Moreover, it is agnostic to the actual value of the penetration rate or any other information about an intersection. We provide verification of our algorithm via numerical simulations. We demonstrate the application of our algorithm using real-world data for four intersections.**

***Index Terms*—Queue length, signalized intersection, connected vehicles, traffic signal timing**

I. Introduction

The distribution of queue length at signalized intersections provides crucial information for performance evaluation [1], [2], design and optimization [3], [4], and real-time operation [5] of traffic signals in road networks. Traditionally, estimation algorithms are developed based on information collected via manual surveys, in-ground fixed location sensors (*e.g.* loop detectors), or cameras. These algorithms are based on two types of models: (i) input-output models [6]–[9] that attempt to estimate the queue by considering the cumulative arrivals and departures at an intersection; and (ii) shock-wave models [10], [11] that consider the dynamic process of formation and dissipation of queues at an intersection.

The main disadvantage of the traditional methods is their requirement for installation, operation, and maintenance of physical hardware at every intersection. As such, they have high capital costs, between $30,000-$60,000 per intersection. Therefore, they cannot be implemented at a large scale to cover all intersections in a network. In the U.S., it is estimated that only about 3% of intersections are instrumented and monitored in real-time, and the majority of intersections receive signal re-timing updates only once every three to five years.

Recent advancement and deployment of connected vehicle technology has created new possibilities to address the high capital cost and operational costs of traditional measurements by physical sensors such as loop detectors. In contrast with traditional measurements that provide traffic information at fixed locations, connected vehicles (CVs) provide spatio-temporal information about the trajectory of vehicles passing through an intersection, including GPS location, speed, heading. This information is typically collected by each CV in real-time and transmitted to a data center operated by the car manufacturer or by the fleet owner. For instance, currently in the U.S. and Europe, Wejo Ltd. and Otonomo Inc. provide data platforms developed from information acquired from various car manufactures (OEMs). Similarly, Uber and Didi estimate traffic conditions based on information obtained from their vehicle fleets.

As a result, there now are several studies of algorithms for queue length estimation based on CV data. The authors in [12]–[14] develop algorithms based on shock-wave models in which they try to identify critical points for each CV, marking the time it joins and leaves the queue. The estimation methods in [15], [16] are based on the travel time for CVs through an intersection. In [17]–[19], the authors develop a stochastic framework to estimate the queue length at the end of each cycle based on the location of CVs. Alternative stochastic estimation methods are proposed in [20]–[22] based on the estimation of arrival/departure processes.

A practical limitation of existing queue length estimation algorithms based on CVs is that they typically require a penetration rate $\sim$ 10% or greater to perform well; *i.e.* 10% of vehicles on road must be connected. While reaching such targets is plausible in the future, our analysis of the current penetration rate of CVs in California, using data from various car manufactures (OEMs), suggests $0.5\% - 1.5\%$ as the average penetration rate of CVs. Additionally, many estimation algorithms [17]–[21], [23], require the knowledge of penetration rate at an intersection which is difficult to obtain without knowing the ground truth for the total numbers of vehicles in advance. This is especially challenging as our analysis shows that the average penetration rate can differ considerably even between two neighboring corridors, *e.g.* $\sim$ 10% relative difference between CA-107 and CA-1 corridors.

In this paper, we propose a novel approach to estimate queue length at signalized intersections and address the above limitations. Our estimation algorithm does not require knowledge of the penetration rate and is applicable for penetration rates smaller than 1%. In contrast with existing methods that estimate the queue length on a cycle-by-cycle basis, we estimate the probability distribution and average value of queue lengths during selected time intervals, *e.g.*

*The Department of Mechanical Engineering, University of California, Berkeley, USA. Email: {tavaf, jmporter, poolla}@berkeley.edu

‡ Sensys Networks, Inc. Berkeley, USA. Email: chrisf@sensysnetworks.com

†The Department of Electrical Engineering and Computer Science, University of California, Berkeley, USA. Email: {poolla, varaiya}@berkeley.edu

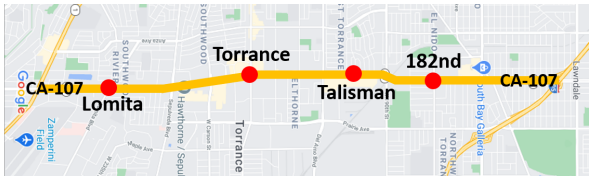

Fig. 1: Four signalized intersections along CA-107 corridor

Fig. 2: The orange rectangles depicts the location of four in-ground sensors and the set of circles in each color (red, yellow,purple) depicts the GPS samples for three individual vehicles on NB of CA-107 and Lomita Blvd Intersection.

morning peaks during weekdays. We exploit the seasonal patterns of traffic at intersections to compensate for the low value of penetration rate of CVs. As we show our algorithm is agnostic to the penetration rate and makes no additional assumption about the queue dynamics.

More specifically, we propose a non-parametric estimation algorithm that is not based on estimating the parameters for either input-output models or shock-wave models. Hence, unlike many existing estimation methods, our algorithm does not depend on the approximation accuracy of a theoretical model in real-world settings. Moreover, we do not make any assumption about the penetration rate such as stationarity in order to estimate it from data. Additionally, we do not require knowledge about the signal control plan and cycle timing of traffic signal, or other local measurements at the intersection. As a result, our algorithm is applicable to almost all intersections in a network at no additional costs as it only requires mapping information (accessible via Open Street Map and Google Maps) in addition to the CV trace information.

We present a verification of our algorithm through both numerical simulations and empirical case studies. We show that our algorithm can successfully recover the queue length distribution in real-world scenarios and offer information that is valuable for performance evaluation (*e.g.* average queue length) and design and optimization of traffic signals (*e.g.* 95% percentile value of the queue length distribution). We briefly discuss one limitation of our algorithm. Since our approach does not explicitly estimate the queue length for each individual cycle, it may have limited applicability for adaptive signal timing control plans that require such real-time information. Based on our case studies, we discuss such a limitation and how it can be potentially addressed by making modifications to our algorithm.

The rest of the paper is organized as follows. We describe the data used in our real-world case studies in Section II. We present our analytical framework and estimation algorithms in Section III-A. We verify our algorithm via numerical simulations in Section IV. In Section V, we consider four signalized intersections on CA-107, and demonstrate the applicability of our algorithms in the real-world. We summarize the limitation of our algorithm, provide additional comments, and conclude in Section VI.

II. Data Description

We use two proprietary datasets for our application. The first is acquired from Sensys Networks Inc. and includes measurements from in-ground detectors located down-stream of the intersections. The second dataset is acquired from Wejo Ltd. and contains traces of connected vehicles. Both data sets cover highway CA-107 (Hawthorne Blvd) on the southwest side of Los Angeles, CA. The geographical coverage includes a stretch of 13.5 km between CA-1 (Pacific Coast Highway) and 166 St. intersections and provides data on four signalized intersections considered in this paper (see Figure 1). Both datasets comprise measurements during September, 2019.

A. Detector Dataset

The detector dataset gives measurements of vehicle counts at a lane level for every outgoing leg of the four intersections on CA-107. Figure 2 shows the location of these detectors at the intersection of CA-107 and Lomita Blvd on the northbound (NB) direction. For each sensor, the dataset records the timestamp (in seconds) of every vehicle detection during September, 2019.

B. Trace Dataset

The trace dataset contains GPS measurements for a set of connected vehicles with a sampling time of 3 (s). The vehicles are recent vehicles, manufactured by General Motors (GM). Each measurement contains (latitude, longitude) of a vehicle as well as its speed (from vehicle's speedometer) and heading direction; see Figure 2. Each CV is equipped with an enhanced GPS device that increases the accuracy of measurements to 1.5 (m) enabling us to develop a lane-level clustering as we discuss in Section V-A. Utilizing both (latitude, longitude) coordinates and the heading we project each data point to a road segment. In this paper, we mainly focus on trajectories along the northbound (NB) direction on CA-107. By projecting each GPS coordinates onto the road, we determine coordinates (y, w) with respect to the road, where y denotes the arc length position of the vehicle along the road and w denotes its lateral position across the road. The reference point for y is the intersection of CA-107 and CA-1, and thus, $y \in [0, 13.5\ (km)]$. For w the reference point is the coordinates of the left edge of the left-most lane, not including the dedicated left-turn lanes, based on satellite images. As such, w typically ranges between $[-10\ (m), 18\ (m)]$ for a road with two left-turn lane, four through lanes, and one right-turn lane.

The penetration rate of connected vehicles, denoted by α, is on average $\simeq 1.5\%$. We estimate the penetration rate by counting the number of unique connected vehicles passing

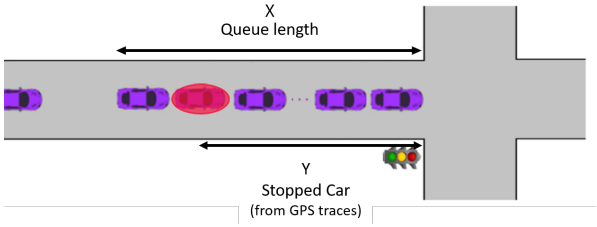

Fig. 3: Queue formation behind an intersection

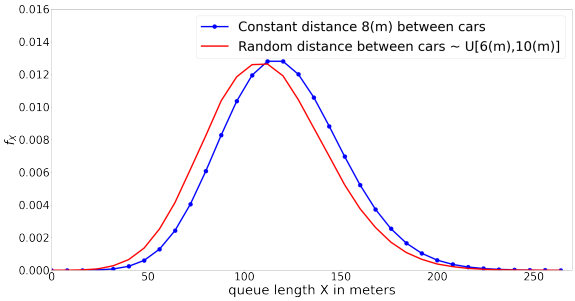

Fig. 4: Numerical simulation for the distribution of queue - length X (in meters) vs $8\ (m) \times N$ (number of cars).

through each leg of the intersection divided by the total number of vehicles counted via the vehicle detection sensors. Note that the algorithm makes no use of the penetration rate.

III. Setup

A. Analytical Model

We now present the key idea underlying our estimation approach. Consider a signalized intersection as in Fig. 3. For each leg of the intersection and a time interval T (*e.g.* morning peaks), we assume that the maximum queue length X (in meters) for each traffic cycle in each lane has a probability distribution f_X where

$$\mathbb{P}\{\underline{x} \leq X \leq \overline{x}\} = \int_{\underline{x}}^{\overline{x}} f_X(x)dx.$$

Our goal is to estimate f_X from observation of GPS traces. Let α denotes the penetration level of connected vehicles. That is, the GPS traces of each vehicle is observed with probability α; in our dataset $\alpha \simeq 1.5\%$. Let Y denote the distance/position from the intersection of each stopped vehicle we observe from its GPS trace. Conditioned on the queue length X, Y is a random variable with approximate distribution

$$Y \sim \text{Uniform}[0, X] \text{ conditioned on } X.$$

So the probability distribution of Y (when we do not know X) is given by

$$f_Y(y) = \int_y^{\infty} \frac{1}{x} f_X(x)dx. \tag{1}$$

Consequently,

$$f_X(x) = -x\frac{df_Y(x)}{dx}. \tag{2}$$

Remark 1. *The simple collection of CV/probe vehicle stop positions has a sampling bias. This is because cycles with longer queues, on average, contain more probe vehicles. This sampling bias is more significant when the queue length distribution is wider. However, our approach as described above explicitly takes into account such a sampling bias.*

Remark 2. *We note that the estimation approach proposed above does not require the knowledge of the penetration rate α. Moreover, it also does not need any information about the timing plan or cycle lengths at the intersection.*

Our model requires $\frac{df_Y(x)}{dx} \leq 0$ to ensure $f_X(x) \geq 0$, *i.e.*, f_Y must be monotonically decreasing (with distance from the intersection). So the main task is to estimate a smooth and monotone curve from the empirical distribution of f_Y so that $\frac{df_Y(x)}{dx} \leq 0$ and the derivative is (relatively) smooth. We formulate this task as a quadratic program and determines it via optimization below.

Remark 3. *In constructing the empirical distribution for f_Y we assume that the position of observed stopped vehicles are independent of one another. This assumption is reasonable when the penetration level is so low that we practically observe only a few samples for each queue length X. As we show through numerical simulations, even for high penetration rate the estimation approach proposed above gives satisfactory performance.*

B. Non-Parametric Estimation

Consider the empirical histogram for f_Y. Let $B = \{b_0, b_1, ..., b_K\}$, $b_0 = 0$, denote the bin edges and $Y = \{y_1, ..., y_K\}$ denote the associated values for the histogram. Assume that $b_{i+1} - b_i = \Delta b > 0$ for all i. We fit a curve $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_K\}$ to the empirical distribution $Y = \{y_1, ..., y_K\}$ by solving the following optimization problem.

Let $z_i := \frac{y_{i+1} - y_i}{\Delta b}$, $1 \leq i < K$, denote the slope of the histogram moving from bin i to $i+1$. For the K^{th} bin, define $z_K := \frac{0 - y_K}{\Delta b}$, *i.e.* set $y_{K+1} = 0$. Then we can write $y_i = -\Delta b \sum_{j=i}^{K} z_j$. Our model requires that $z_i \leq 0$. Therefore, we estimate a smooth curve $\hat{Y} = \{\hat{y}_1, \cdots, \hat{y}_K\}$, parameterized by its slope $\hat{Z} = \{\hat{z}_1, \cdots, \hat{z}_K\}$, by solving the following optimization problem:

$$\min \sum \|y_i - \hat{y}_i\|_2^2 + \beta \sum \|\hat{z}_{i+1} - \hat{z}_i\|_2 \tag{3}$$

subject to

$$\hat{z}_i \leq 0, \tag{4}$$

$$\hat{y}_i = -\Delta b \sum_{j=i}^{K} \hat{z}_j, \tag{5}$$

where we explicitly require the fitted curve to satisfy the monotonicity condition of our model by (4). The first term in the objective function denotes the estimation error, while the second term penalizes high variations in the slope of the fitted curve to ensure a smooth curve. As such, parameter β controls the trade-off between estimation error and smoothness of the curve. We note that $\{\hat{z}_1 \Delta b, 2\hat{z}_2 \Delta b, ..., K\hat{z}_K \Delta b\}$

gives us the estimate of f_X at bin edges $x = \{b_1, b_2, \cdots, b_K\}$ via (2). The optimization problem above can be written as a quadratic program (QP),

$$\begin{aligned} &\min \ \hat{z}^T P \hat{z} + q^T \hat{z} \\ &\text{subject to} \\ &\quad G\hat{z} \leq h, \end{aligned}$$

where P, q, G, h are defined as follows:

$$P = A^T A + \beta D^T D, \tag{6}$$

$$G = I_{N\times N}, \tag{7}$$

$$h = [0, 0, \cdots, 0]^T, \tag{8}$$

$$q = -2A^T [y_1, y_2, \cdots, y_N]^T, \tag{9}$$

$$A = -\Delta b \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & 1 & \cdots & 1 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}, \tag{10}$$

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{11}$$

C. Average Queue Length

One can estimate the expected queue length $\mathbb{E}\{X\}$ using the estimated distribution for f_X. However, we show below that $\mathbb{E}\{X\}$ can be estimated directly from data, without first estimating f_X; as such, it may have the advantage of avoiding the approximation errors arising in solving the QP.

For the expected queue length $\mathbb{E}\{X\}$ we have,

$$\begin{aligned} \mathbb{E}\{X\} &= \int_0^\infty x f_X(x) dx = \int_0^\infty -x^2 \frac{f_Y(x)}{dx} dx \\ &= -x^2 f_Y(x)\big|_0^\infty + 2\int_0^\infty x f_Y(x) dx = 2\mathbb{E}\{Y\}. \end{aligned} \tag{12}$$

We note that the confidence bounds for $\{X\}$ is twice the confidence bounds for $\mathbb{E}\{Y\}$. Therefore, 95% confidence interval for the average queue length can by computed as $2\overline{Y} \pm \frac{1.96}{\#\text{observations}} s_Y$ where $\overline{Y}$ and s_Y denote the empirical mean and standard deviation for Y.

Remark 4. *We note that one can also determine a confidence interval for the estimated density for f_X above using Dvoretzky-Kiefer-Wolfowitz (DKW) confidence interval which provides a uniform bound, or Wilson confidence bound, which provides a point-wise bound, for the distance between the empirical distribution and the true distribution. However, given that in the QP formulation above we modify and further approximate the empirical f_Y with a monotone and smooth function, the estimated confidence bounds may have limited value.*

IV. Numerical Simulation

We numerically evaluate the ability of our proposed algorithm to recover the true distribution of queue length $f_X(x)$. The simulation setup is as follows.

We consider queue lengths during morning peak hours 7-11AM on working days for one month at an intersection. Let $\lambda(t)$ denote the average arrival rate to the intersection at time t. Assuming that the duration of red phase is T_{red}, the maximum number of cars N in the queue at the end of red phase is a Poisson random variable with parameter $\lambda_Q(t) := \int_t^{t+T_{\text{red}}} \lambda(\tau) d\tau$. We assume that the traffic pattern is similar during the morning peak. As such $\lambda_Q(t)$ is similar for all cycles during 7-11AM; we denote its average value by λ_Q. We note that the realized queue length for each cycle would be still different even if we assume that $\lambda_Q(t)$ is the same during 7-11AM; they correspond to different realizations of Poisson(λ_Q). For our numerical simulation, we set $\lambda_Q = 15$ and assume that the maximum number of cars in the queue $N \sim \text{Poisson}(15)$ for each cycle. Moreover, each traffic cycle lasts 120 (s).

Now consider a queue of vehicles at the end of a red phase. The distance between every two consecutive vehicles in the queue d_i, $1 \leq i \leq N$ is random. We assume that this distance is on average 8 (m) (including one vehicle length) and set $d_i \sim U[8 - 2\ (m), 8 + 2\ (m)]$; that is, each distance d_i may deviate from 8 (m) by up to 2 (m). Accordingly, given a realization for the number of cars in the queue N, we simulate the length of the queue in meters by assuming that the distance between successive cars is $U[8 - 2\ (m), 8 + 2\ (m)]$. Figure 4 shows the empirical distribution for queue length X (in meters). We would like to point out that the randomness in car distances d_i, makes the distribution $f_X(x)$ slightly different from $f_N([x/8])$. Most importantly, $f_X(x)$ (queue length in meters) is a continuous distribution while $f_N(n)$, where $n = [x/8]$ (number of queued cars), is a discrete distribution. Additionally, f_X is slightly left skewed compared to f_N.

For the numerical simulation, we consider three penetration levels for connected vehicles $\alpha \in \{0.5\%, 1.5\%, 5\%\}$. Figures 5-7 depict the estimated probability distribution for queue length (in meters) $\hat{f}_X$, and compares them with the true distributions f_X and f_N. Moreover, the estimated average queue length for each case $\{0.5\% : 111.4\ (m), 1.5\% : 117.4\ (m), 5\% : 122.0\ (m)\}$, along with the confidence bounds, are compared against the true sample average of queue lengths at 120.5 (m). As it can be seen, the estimation accuracy is satisfactory even for very low penetration $\alpha = 0.5\%$. Additionally, the estimation accuracy for both the average queue length and the probability distribution improves as the penetration rate α increases from 0.5% to 1.5%, and 5%.

As we discussed in the introduction, a particular value that is critical in both performance evaluation and the design of timing plan at intersections is the the queue length at specific quantiles (*e.g.* 95%). Figure 8 shows the estimation error vs. quantiles. We note that the estimation error for quantiles

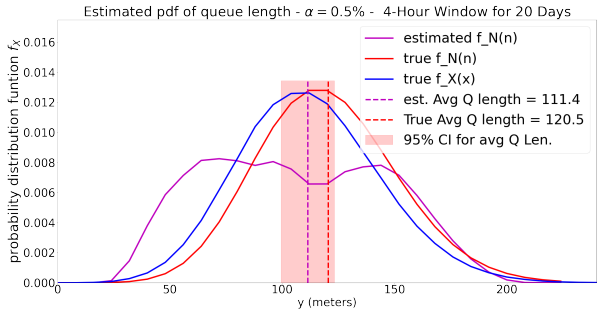

Fig. 5: Numerical simulation for penetration rate $\alpha = 0.5\%$

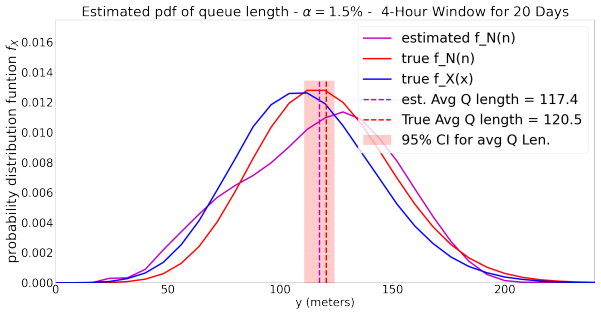

Fig. 6: Numerical simulation for penetration rate $\alpha = 1.5\%$

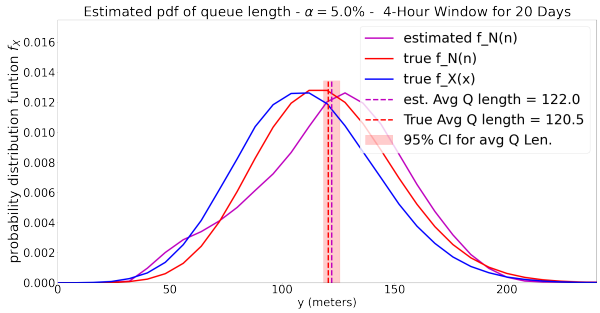

Fig. 7: Numerical simulation for penetration rate $\alpha = 5\%$

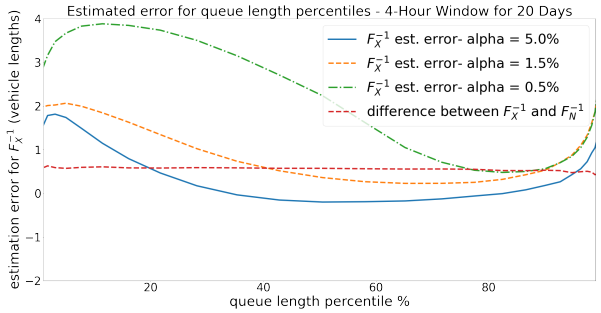

Fig. 8: Numerical simulation - the estimation error for queue length percentiles F_X^{-1}

above 60% is always less than two times the average distance between queued vehicles. Moreover, this error is comparable to the distance between the two distribution F_x (queue length in meters) and F_N (queue length in number of vehicles $\times$ average vehicle length) which describe the true queue length distribution in two different ways.

V. Empirical Verification

We demonstrate the implementation of our approach for intersections on CA-107 corridor using real data. We present the results for four signalized intersections with different characteristics: (i) Lomita Blvd & CA-107 is a major intersection with one left-turn lane, four through lanes, and no dedicated right-turn lane; (ii) Torrance Blvd & CA-107 is a major intersection with two left-turn lanes, four through lanes, and one right-turn lane; (iii) Talisman St & CA-107 is an intersection located in front of an outlet, with one left-turn lane, four through lanes, and one-right turn lane leading to the outlet; (iv) 182nd St & CA-107 is an intersection located in a less congested part of CA-107 corridor towards north, with one left-turn lane, four through lanes, and one right-turn lane. To empirically verify the result of our approach, we estimate the queue length using vehicle detection sensors located at the downstream of the intersection links as well and compare the results.

A. Lane Separation & Lane-Specific Queues

In Sections III and IV, we considered the case where the queue length is identical for all lanes. So we did not need to explicitly consider the geometry of the intersection and the number/types of lanes in each intersection leg. However, in the real world, the queue length for each lane can be different. For instance, vehicles that want to make a right/left turn queue up in the most right/left lane before they reach the right/left turn pockets, while vehicles that want to go straight through the intersection typically choose the middle lanes. Figure 9 shows the recorded position of stopped vehicles behind CA-107 & Lomita intersection. As can be seen, the relative number of observed stopped vehicles with a distance greater than 200 (m) from the intersection is noticeably lower for lane 4 compared to lane 1; this suggests that the average queue length for lane 4 is smaller than that of lane 1. Consequently, we cannot assume that the queue length distribution is identical across all lanes. Therefore, we need to first implement a lane assignment algorithm that determines the lane each stopped vehicles belong to, and then apply our estimation approach for each lane separately.

To develop our lane assignment algorithm, we consider the trace data within $30\,(m)$ before the intersection assuming that most vehicles stay within their lanes close to the intersections. Next, focusing on the lateral positions w across the road, we determine 0.5% and 99.5% percentile for the empirical distribution of projected trace data on the road. As such, we estimate the width of the road segment at the intersection. Assuming that each lane has an average width of $12\,ft \simeq 3.65\,m$, we can estimate the number of lanes. Given the number of lanes, we then utilize a Gaussian Mixture Model with tied variance to estimate the lane boundaries; see Figure 10. Moreover, we utilize additional post-processing schemes including (i) correcting for erroneous lane change estimates due to GPS noises for vehicles very close to the lane boundaries, and (ii) correcting for lane assignments inconsistent with vehicle movements, e.g. LT lane for vehicles

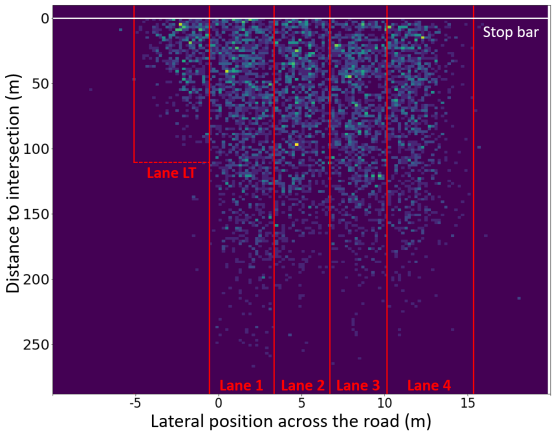

Fig. 9: 2D histogram of stopped vehicles behind CA-107 & Lomita intersection (NB direction).

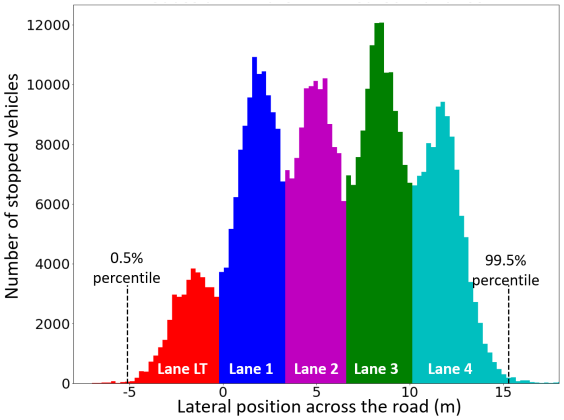

Fig. 10: Histogram of lateral position (across the road) for stopped vehicles on NB of CA-107 & Lomita intersection.

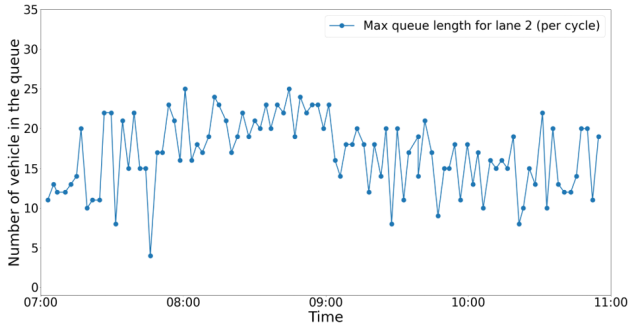

Fig. 11: Estimated queue length from detection sensors for morning peak on September 12, 2019

moving straight.

B. Estimation using Vehicle Detection Sensors

We adopt the estimation algorithm proposed in [10] based on shock-wave theory. The key idea used in the algorithm is that the discharge rate for vehicles in the queue, which is observed via sensors at the downstream of the intersection, is higher than the normal free flow of vehicles when there is no queue. As such, one can determine the time gap between consecutive vehicle detections at the downstream, and identify the time instance where the queue is cleared, and thereby estimate the number of cars in the queue.

To determine the end of the queue, we identify the first time (skipping the first vehicle in the green phase) the detection gap exceeds a given a threshold *max gap*.We choose the value of max gap based the nominal time gap between consecutive vehicles leaving the queue.

We estimate the nominal gap between detections of two vehicles leaving the queue to be approximately $3\,(s)$. This is based on our estimate of $14.5(km/h)$ for how fast the shock-wave of vehicles clearing the queue propagates backward into the intersection upstream; see our note in [24] for the detailed calculation using drone footage. Assuming that the average distance (including a vehicle length) between two queued vehicles is $8\ (m)$, vehicles leave the queue with $2\ (s)$ time gaps. Additionally, for each two consecutive vehicles, the rear vehicle has to travel the additional distance of $8\,(m)$ to reach the location of the detectors, which takes $1\ (s)$ assuming an average speed of $30\ (km/h)$.[1]

Given that the nominal gap is $\sim 3\ (s)$, we set max gap to $4.5\ (s)$, which adds a 50% margin to the nominal time gap. We note that experimenting with alternative values of $[4, 5, 6]$ for max gap, results in very similar estimates for queue lengths. Figure 11 depicts an example of the estimated queue lengths for lane 2 during the morning peak 7-11AM on September 12, 2019, using the detection data.

Remark 5. *The traffic signals on CA-107 are actuated, and thus the phase lengths vary slightly from cycle to cycle. Unfortunately, the realized phase lengths and cycle timing are not recorded and cannot incorporated in our estimation algorithm based on detections. As a result, we develop an algorithm to estimate the timing of the green phase for NB based on the vehicle detection data on all four links of the intersections. Due to space limitation we do not discuss the details of the algorithm here.*

Remark 6. *One drawback of using a max gap threshold to identify the end of the queue is its sensitivity to drivers' delays and distractions. That is, if a driver starts moving with some delay after the front vehicle leaves the queue, the recorded time gap between the two vehicles can exceed the max gap threshold. As a result, we may underestimate the queue length for that cycle. For instance, in Figures 12 for Lomita and Torrance intersections, there are unexpected small peaks for queue lengths equal at a single car length that potentially occurs because of such erroneous detections.*

C. Results & Comparison

We implement the proposed estimation algorithm based on connected vehicles for four intersections on CA-107 (see Figure 1) and compare the results with those based on vehicle detections as described in Section V-B.

[1]Alternatively, we can assume that the nominal time gap has three components as follows: (i) reaction time for the second vehicle $\sim 1\ (s)$, (ii) delay for the second vehicle to start moving $\sim 1\ (s)$, and (iii) the time needed to travel the additional distance of $8\ (m)$ which is $\sim 1\ (s)$.

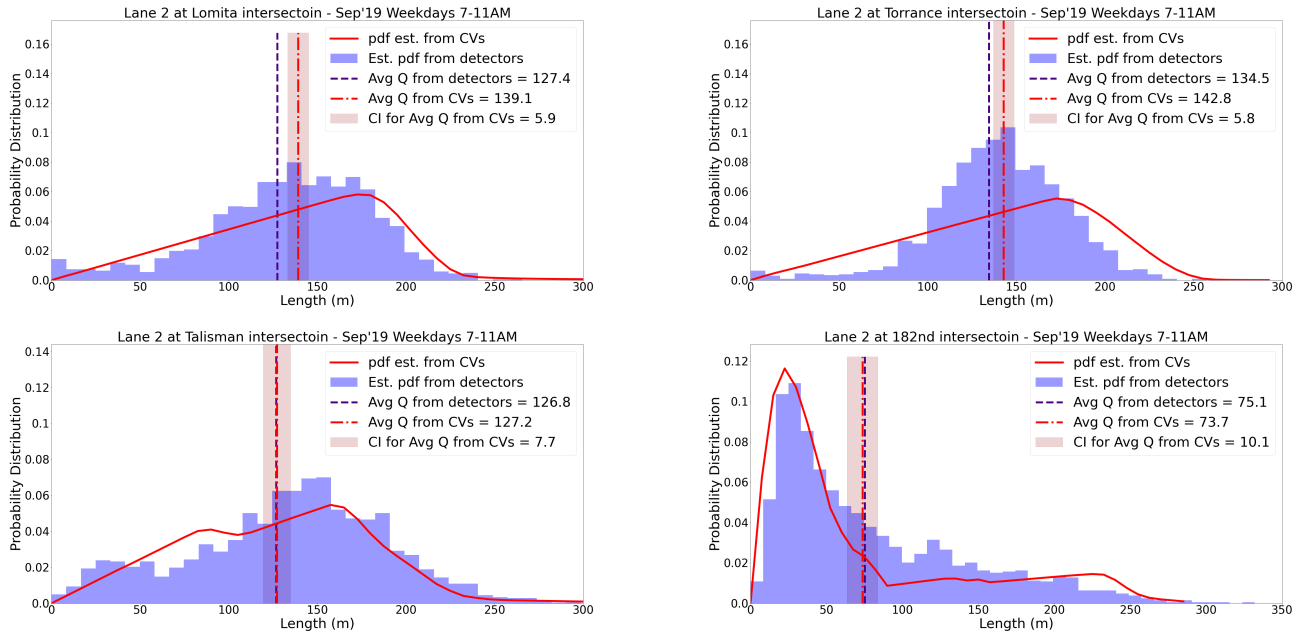

Fig. 12: Estimated queue length distribution on lane 2 (NB) for intersections on CA-107 from connected vehicles with $\alpha \simeq 1.5\%$ vs. queue length estimation from in-ground vehicle detection sensors.

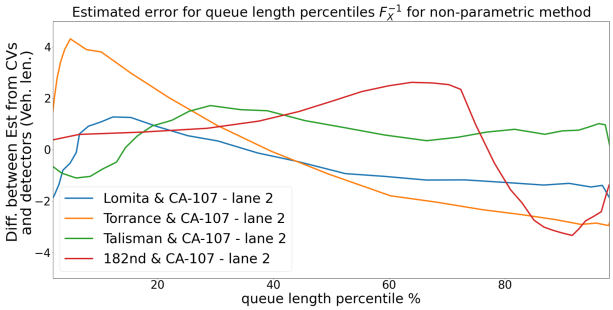

Fig. 13: The difference between estimated queue length percentiles F_X^{-1} using connected vehicles vs. detectors.

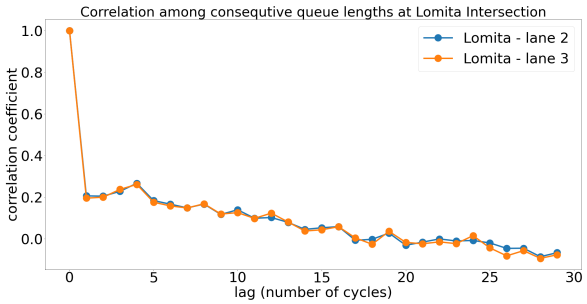

Fig. 14: Auto-correlation of queue length at CA-107 & Lomita intersection.

We set $\beta = 150$ and utilize the dataset during morning-peak (7-11AM) of business days in September, 2019, consisting of approximately 2000 traffic cycles. Using satellite images of the intersections from Google Maps, we set the average distance between queued vehicles (including a vehicle length) to $8.2\ m$. Due to space limitations, we make the comparison only for lane 2 on NB direction at each intersection; we note that vehicles queuing in lane 1 (left-most lane) and lane 4 (right-most lane) may end up turning at the intersection, and thus do not end up being detected by the in-ground detection sensors.

Figure 12 depicts the estimated probability distribution as well as the average queue length using our non-parametric estimation algorithm. We note that both estimates are very close to the results using the in-ground detection sensors. Notably, the difference in the average queue length is less than or equal to an average distance between two queued vehicles; we note that the difference for Lomita intersection is slightly greater than $8.2\,m$, which as we argued in Remark 6, is potentially due to an underestimation of queue length using the in-ground detection sensors.

Figure 13 shows the difference between estimated percentiles F_X^{-1} using connected vehicles vs. detection sensors. As can be seen, the difference in queue length percentiles for 90%-98% is less than or equal to twice the average distance between two queued vehicles. Hence the empirical results suggest that the estimation method we propose in this paper can generate information about queue lengths that is accurate enough for both performance monitoring and the design of time-of-day traffic cycle plans.

VI. Additional Discussion & Conclusion

The simulation and empirical results demonstrate the effectiveness of the estimation algorithm we propose based on trace data from connected vehicles with very low penetration levels. A major advantage of estimation from trace data is the full coverage it provides at all intersections without any hardware equipment and traffic interruption for their installation and maintenance. Additionally, it has negligible marginal costs since the trace data is being collected by various OEMs.

As such, the algorithm proposed here can already be utilized to monitor the performance of intersections across the road network, and to estimate the queue lengths necessary for the design and optimization of time-of-day traffic cycle plans for each intersection.

The main disadvantage of our approach is its limited ability to provide accurate real-time queue length measurements for each cycle individually. However, we note that the existing algorithms for cycle-by-cycle estimation of queue length rely heavily on the assumption that they observe at least one CV in each cycle. At a penetration rate of 1.5% with an average queue length of 15 cars ($\sim 120m$), we do not observe any CV in 80% of cycles. A recent work by [22] suggests to address such an issue by utilizing information from historical trends. As such, our estimation algorithm can provide information to estimate such recent historical trends. More specifically, one can use our algorithm to estimate the distribution and/or average queue length during a narrower recent window of time (*e.g.* past few hours) in order to be used in such a hybrid scheme as suggested in [22].

Moreover, we argue that during peak hours when the traffic flow tends to be more predictable, the performance of a well-optimized time-of-day traffic cycle plan is as good as an adaptive traffic cycle control. The authors in [25] found that in fact that the time-of-day traffic cycle plan performs better than the adaptive traffic cycle control for a test site in Anaheim, CA, which is located only 35-40 km away from the intersections on CA-107. While we do not aim to provide a similar detailed comparison as in [25] here, we investigate the correlation among consecutive queue lengths as a proxy for the potential value of adaptive traffic cycle control. We note that currently, all intersection on CA-107 employs a coordinated actuated time-of-day cycle plan. Figure 14 shows the auto-correlation among consecutive queue lengths (estimated from detection sensors) at CA-107 & Lomita Blvd intersections. As it can be seen, the realized queue length at the end of each cycle includes very limited additional information about the queue lengths and traffic during the next few cycles, beyond the information already contained in the historical distribution. This suggests that the use of adaptive traffic cycle has potentially very limited positive impact on the performance of the traffic signals during peak hours with normal traffic patterns that do not deviate significantly from the historical distribution.

VII. Acknowledgements

This research was supported by National Science Foundation EAGER award 1839843. We are grateful to Sensys Networks Inc. and Wejo, Ltd. for providing us data on detection data, traffic signals timing plan, and traces of connected vehicles on CA-107 corridor. We are grateful to Amine Haoui (Sensys Networks Inc.) and Jo Birch (Wejo Ltd.) for their invaluable comments and suggestions.

References

[1] H. C. Manual, "Hcm2010," *Transportation Research Board, National Research Council, Washington, DC*, p. 1207, 2010.

[2] K. N. Balke, H. A. Charara, and R. Parker, "Development of a traffic signal performance measurement system (tspms)," Texas Transportation Institute, Texas A & M University System College, Tech. Rep., 2005.

[3] T.-H. Chang and J.-T. Lin, "Optimal signal timing for an oversaturated intersection," *Transportation Research Part B: Methodological*, 2000.

[4] P. B. Mirchandani and N. Zou, "Queuing models for analysis of traffic adaptive signal control," *IEEE Transactions on Intelligent Transportation Systems*, 2007.

[5] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

[6] A. Sharma, D. M. Bullock, and J. A. Bonneson, "Input–output and hybrid techniques for real-time prediction of delay and maximum queue length at signalized intersections," *Transportation Research Record*, 2007.

[7] N. Geroliminis and A. Skabardonis, "Prediction of arrival profiles and queue lengths along signalized arterials by using a markov decision process," *Transportation Research Record*, 2005.

[8] X. Zhan, R. Li, and S. V. Ukkusuri, "Lane-based real-time queue length estimation using license plate recognition data," *Transportation Research Part C: Emerging Technologies*, 2015.

[9] Z. Amini, R. Pedarsani, A. Skabardonis, and P. Varaiya, "Queue-length estimation using real-time traffic data," in *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016.

[10] H. X. Liu, X. Wu, W. Ma, and H. Hu, "Real-time queue length estimation for congested signalized intersections," *Transportation research part C: emerging technologies*, 2009.

[11] A. Skabardonis and N. Geroliminis, "Real-time monitoring and control on signalized arterials," *Journal of Intelligent Transportation Systems*, 2008.

[12] Y. Cheng, X. Qin, J. Jin, and B. Ran, "An exploratory shockwave approach to estimating queue length using probe trajectories," *Journal of intelligent transportation systems*, 2012.

[13] F. Li, K. Tang, J. Yao, and K. Li, "Real-time queue length estimation for signalized intersections using vehicle trajectory data," *Transportation Research Record*, 2017.

[14] J. Yin, J. Sun, and K. Tang, "A kalman filter-based queue length estimation method with low-penetration mobile sensor data at signalized intersections," *Transportation Research Record*, 2018.

[15] X. J. Ban, P. Hao, and Z. Sun, "Real time queue length estimation for signalized intersections using travel times from mobile sensors," *Transportation Research Part C: Emerging Technologies*, 2011.

[16] P. Hao and X. Ban, "Long queue estimation for signalized intersections using mobile data," *Transportation Research Part B: Methodological*, 2015.

[17] G. Comert and M. Cetin, "Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data," *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[18] G. Comert, "Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals," *Transportation Research Part B: Methodological*, 2013.

[19] ——, "Queue length estimation from probe vehicles at isolated intersections: Estimators for primary parameters," *European Journal of Operational Research*, 2016.

[20] P. Hao, Z. Sun, X. J. Ban, D. Guo, and Q. Ji, "Vehicle index estimation for signalized intersections using sample travel times," *Procedia-Social and Behavioral Sciences*, 2013.

[21] P. Hao, X. J. Ban, D. Guo, and Q. Ji, "Cycle-by-cycle intersection queue length distribution estimation using sample travel times," *Transportation research part B: methodological*, 2014.

[22] C. Tan, J. Yao, K. Tang, and J. Sun, "Cycle-based queue length estimation for signalized intersections using sparse vehicle trajectory data," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[23] K. Tiaprasert, Y. Zhang, X. B. Wang, and X. Zeng, "Queue length estimation using connected vehicle technology for adaptive signal control," *IEEE Transactions on Intelligent Transportation Systems*, 2015.

[24] H. Tavafoghi, J. Porter, K. Poolla, and P. Varaiya, "Report for UAS4T competition," *23rd IEEE International Conference on Intelligent Transportation*. [Online]. Available: www.hamidtavaf.github.io/UAS4T.pdf

[25] I. Chia, X. Wu, S. S. Dhaliwal, J. Thai, and X. Jia, "Evaluation of actuated, coordinated, and adaptive signal control systems: A case study," *Journal of Transportation Engineering, Part A: Systems*, 2017.