Various Methods for Queue Length and Traffic Volume Estimation Using Probe Vehicle Trajectories

Yan Zhao\textsuperscript{a}, Jianfeng Zheng\textsuperscript{b,*}, Wai Wong\textsuperscript{c}, Xingmin Wang\textsuperscript{c}, Yuan Meng\textsuperscript{b}, Henry X. Liu\textsuperscript{b,c,d}

\textsuperscript{a}Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA
\textsuperscript{b}Didi Chuxing Inc., Beijing, China
\textsuperscript{c}Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA
\textsuperscript{d}University of Michigan Transportation Research Institute, University of Michigan, Ann Arbor, MI, USA

Abstract

With the development of connected vehicle technology and the emergence of e-hailing services, a huge amount of vehicle trajectory data is being collected every day. Connected vehicles and vehicles offering e-hailing services, like mobile sensors, can provide rich information of traffic conditions. The huge amount of trajectory data could provide a new perspective on the performance measure, diagnosis, and optimization of transportation networks. Attribute to its huge amount and accessibility, trajectory data has become a potential substitute for the widely used fixed-location sensors. There has been some literature estimating traffic volume and queue length at intersections using data collected from probe vehicles. However, some of the existing models can only work when the penetration rate of probe vehicles is high enough. Some others require the prior information of the distribution of queue lengths and the penetration rate of probe vehicles, which might vary a lot both spatially and temporally and usually are not known in real life. To fill the gap, this paper proposes a series of efficient methods for the estimation of queue length, penetration rate, and traffic volume. The proposed methods have been validated by both simulation data and real-field data. The testing results show that the methods can be applied to both under-saturation and over-saturation cases, even when the penetration rate is very low. Therefore, the methods are ready for large-scale real-field applications.

Keywords: Probe vehicle, Queue length estimation, Penetration rate, Traffic volume estimation
1 Introduction and Motivation

1.1 Introduction

The past few years have witnessed the rapid development of connected vehicle technology and the emergence of e-hailing services. Connected vehicles and vehicles offering e-hailing services, like mobile sensors, can provide rich information of traffic conditions. These probe vehicles in the traffic flow, mounted with Global Positioning System (GPS) devices, record their positions with some frequency which usually varies from every 0.1 second to every a few seconds. The huge amount of data collected from probe vehicles could provide a completely new perspective on the performance measure, diagnosis, and optimization of transportation networks. Therefore, how to estimate the states of transportation networks using the trajectories of probe vehicles has become an interesting and challenging research question.

Queue length and traffic volume at signalized intersections are both important for transportation network performance measure and traffic signal control. In under-saturation cases, the goal of traffic signal control is to minimize the total delay or maximize the number of vehicles passing through the intersection per unit time. Traffic volume from different directions can be used to determine the allocation of green time. Traditionally, traffic volume is measured by fixed-location sensors such as loop detectors. Loop detectors, although can measure traffic volume quite accurately, are notorious, because of the deployment and maintenance cost. In recent years, camera-based vehicle identification systems become popular in some big cities. However, many of them do not work very well at night, or in bad weather. They might also raise privacy issues since the cameras can record the plate numbers of all the vehicles passing through the intersections. In some areas, traffic volume is measured by microwave vehicle detectors which can count vehicles on different lanes simultaneously. In general, these fixed-location sensors are expensive to install and maintain, thus they usually only cover some areas of the cities. In over-saturation cases, the main consideration for traffic signal control is to avoid spillovers which could disrupt the transportation networks. Researchers have developed different methods to estimate queue length. Most of the traditional methods apply shockwave theory to estimate queue length using event-based data collected from loop detectors (Liu et al., 2009; An et al., 2018). There are also a few studies estimating queue length by combining data from upstream and downstream loop detectors (Lee et al., 2015), or by applying Bayesian Network to mobile sensor data (Hao et al., 2014).

As for trajectory data based methods, a stream of literature uses the distribution of queue lengths, stopping positions, or time of joining queues to estimate queue length or traffic volume from a perspective of probability and statistics. In 2009, Comert and Cetin published their seminal paper on queue length estimation using the stopping positions of probe vehicles. They showed that the positions of the last probe vehicles in the queues alone were sufficient for queue length estimation. They also analyzed the quality of estimation results in cases with different penetration rates of probe vehicles (Comert and Cetin, 2009). Comert and Cetin (2011) extended their work to both spatial and temporal dimensions. Besides stopping positions, they combined the time when the probe vehicles joined the queues to estimate the queue lengths. Later in 2013, Comert studied the effect of data from stop line detection on estimation results (Comert, 2013a) and proposed a simple analytical model (Comert, 2013b). Li et al. (2013) formulated the dynamics of queue length as a state transition process and estimated queue length cycle by cycle using Kalman Filter. Tiaprasert et al. (2015) introduced discrete wavelet transform and applied it to estimate queue length cycle by cycle, under the assumption that queue length cannot be zero. Later, Comert (2016) summarized a series of estimators for queue length estimation and evaluated them systematically. Similarly to queue length estimation, Zheng and Liu (2017) applied maximum likelihood estimation to traffic volume estimation, assuming vehicle arrival at isolated intersections follows a time-varying Poisson process. They validated their model using connected vehicle data and taxi navigation data. Zhan et al. (2017) studied citywide traffic volume estimation using trajectory data, by applying machine learning techniques and traditional traffic flow theory. Rompis et al. (2018) proposed an algorithm to identify which lanes the probe vehicles were on and applied the additional lane-level information to queue length estimation.

There is also a stream of literature that applies shockwave theory to probe vehicle data (Ban et al., 2011; Cetin, 2012; Hao et al., 2015; Hao and Ban, 2015; Li et al., 2017), or combines probe vehicle data and loop detector data (Badillo et al., 2012; Cai et al., 2014; Wang et al., 2017), to estimate or predict queue length. Since these studies are not closely related to this paper methodologically, they will not be introduced in
1.2 Motivation and contribution

Most existing literature introduced above on queue length estimation focuses on cycle-by-cycle estimation, and thus requires the prior information of both penetration rate of probe vehicles and distribution of queue lengths. Nevertheless, these two critical inputs are usually not available. Although a recent study by Wong et al. (2018a) has proposed a novel method for estimating probe vehicle penetration rate solely based on probe vehicle trajectory data collected at intersections, their method cannot handle cases with zero queue. Therefore, it is very hard for the existing literature to be implemented in real field. As for traffic volume estimation, the model in (Zheng and Liu, 2017) assumes vehicle arrival in each cycle follows a time-varying Poisson process. This assumption might not be reasonable in over-saturation cases, considering that arrival process, queueing process, and departure process are all different when the traffic is over-saturated. The method also requires exact signal timing and exact positions of the stop bars, which limits its application in different scenarios. The method in (Zhan et al., 2017) requires ground truth volume data on some road segments to build a connection between their high-level features and the actual volume categories. This implies that their method depends on not only trajectory data but also other sources of data.

Estimating the states of the whole population from a small portion of it (Wong and Wong, 2015, 2016), in nature, has to build a connection between the small portion and the whole population by what they have in common. When the traffic is flowing, it is difficult to know how many regular vehicles are around probe vehicles. Consequently, it is almost impossible to estimate the penetration rate of probe vehicles in the traffic. When the vehicles are stopping at the intersections, because the empirical value of space headway in this case is usually around 7.0 m/veh or 7.5 m/veh, it can be roughly known how many vehicles are in front of the last probe vehicle in the queue. Although the number of vehicles behind the last probe vehicle is still unknown, the incomplete information provides an opportunity to estimate the penetration rate of probe vehicles. Therefore, the proposed methods in this paper take the stopping positions at intersections as the common characteristics between probe vehicles and regular vehicles.

However, unlike the existing methods that have a lot of external dependencies, the methods in this paper can estimate queue length, penetration rate of probe vehicles, and traffic volume at the same time, in both under-saturation and over-saturation cases, without knowing the exact signal timing and the exact positions of stop bars. The only data required are trajectory data of probe vehicles and basic transportation network features (such as number of lanes). Therefore, they can be applied to a much wider range of scenarios. The methods have been validated by both simulation and real-field data and thus they are practice ready.

1.3 Outline of the following sections

In Section 2, a detailed description and definition of the queue length estimation problem will be given. Depending upon the existence of probe vehicles, queues over different cycles will be categorized into two classes: observable queues (with probe vehicle) and hidden queues (without probe vehicle). Section 3 will present four different methods to estimate the total length of the observable queues; Section 4 will present two different methods to estimate the total length of hidden queues. From the results in Section 3 and 4, it can be seen that penetration rate of probe vehicles is the key in the entire estimation process. In Section 5, two intuitive methods for penetration rate estimation will be given. Then, based on the estimated penetration rate, in Section 6 how to estimate traffic volume and the distribution of queue lengths will be explained. The validation and evaluation of the proposed methods are in Section 7. Finally, there will be some discussions and conclusions in Section 8.

2 Problem Statement

2.1 Description

At a signalized intersection, when a traffic light is red, there might be queues forming behind the corresponding stop bars. In each cycle, the length of the queue, $Q$, is a random variable, following some distribution.
In a specific queue, some vehicles might be probe vehicles mounted with GPS devices (such as connected vehicles, taxis, or vehicles offering e-hailing services), as shown in Figure 1. Their trajectories can be collected and stored in a database. The probe vehicles are assumed to be homogeneously mixed in the traffic flow. Let $p$ denote the penetration rate of probe vehicles in the traffic, i.e., when arbitrarily selecting a vehicle from the traffic, its probability of being a probe vehicle is $p$.

For simplicity, like most of the relevant literature does, the average space headway between adjacent stopping vehicles are assumed to be known empirically. Therefore, not only can the probe vehicles be observed, but also the number of regular vehicles in front of the last probe vehicle can be inferred. However, the number of vehicles behind the last probe vehicle is still unknown. Denote the observed partial queue in the $i$th cycle by $q_i$, which is a sequence consisting of “0”s and “1”s. “0”s represent regular vehicles; “1”s represent probe vehicles. If $q_i$ is not empty, the last element of $q_i$ will always be “1”, because the last vehicle in any observed partial queue is always a probe vehicle. Denote the length of the observed partial queue, i.e., the number of elements in $q_i$, by $|q_i|$.

For the cycles with observed probe vehicles, denote the total queue length (number of queueing vehicles) of these cycles by $Q^{obs}$, which includes not only the probe vehicles and the regular vehicles in front of the last probe vehicles, but also the regular vehicles behind the last probe vehicles. These queues are called (partially) observable queues because the probe vehicles in them can be observed. For the cycles without any observed probe vehicles, denote the total length of the queues of these cycles by $Q^{hid}$. These queues are called hidden queues because no vehicle in them can be observed directly. The observation process is demonstrated in Figure 2. Please note that, if no probe vehicle is observed in a cycle, it does not necessarily mean that there is no queue. It is also possible that no probe vehicle is observed while a queue exists, due to low penetration rate. The goal of queue length estimation in this paper, is to find $Q^{obs} + Q^{hid}$, and the distribution of queue lengths over different cycles.

Denote the number of probe vehicles in a queue by $N$, which is a random variable. Denote the number
of probe vehicles in the $i$th cycle by $n_i$. If there exists at least one probe vehicle in a queue, denote the position of the first probe vehicle and the position of the last probe vehicle by $S$ and $T$, which are two random variables. Denote the positions of the first and the last probe vehicles in the $i$th cycle by $s_i$ and $t_i$, respectively. Define random variable $X_l$ as an indicator of queue length, i.e.,

$$X_l=\begin{cases} 
1, & Q = l \\
0, & Q \neq l.
\end{cases}$$

(1)

Denote the counts of queues with length $l$ in all the cycles by $C_l$. It can be considered as the sum of $X_l$ over different cycles. The total number of cycles can be expressed as $C = \sum_i C_l$.

### 2.2 A summary of notations

The notations mentioned above are summarized in Table 1. All the random variables are represented by capitalized letters.

| Notation | Description |
|----------|-------------|
| $Q$      | Queue length |
| $q_i$    | The observed partial queue in the $i$th cycle |
| $X_l$    | Binary variable to indicate if queue length is $l$ |
| $C_l$    | The total counts of queues with length $l$ |
| $C$      | The total number of cycles |
| $N$      | The number of probe vehicles in a queue |
| $n_i$    | The number of probe vehicles in the $i$th cycle |
| $S$      | The position of the first probe vehicle in a queue |
| $s_i$    | The position of the first probe vehicle in the $i$th cycle |
| $T$      | The position of the last probe vehicle in a queue |
| $t_i$    | The position of the last probe vehicle in the $i$th cycle |
| $Q^{obs}$| The total length of all the (partially) observable queues |
| $Q^{hid}$| The total length of the (hidden) queues without any probe vehicles |

### 3 Estimation of $Q^{obs}$

$Q^{obs}$ can be estimated through two general approaches. Method 1, 2, and 3 are based on the fact that the probe vehicles are expected to segregate regular vehicles equally between them. These methods only require the number of stopping probe vehicles in each cycle and the stopping positions of the first and the last probe vehicles in the queues, all of which are known. Therefore, the corresponding estimates are constants. Method 4, by contrast, estimates $Q^{obs}$ by Bayes’ Theorem, which relies on penetration rate $p$. Thus the estimator of $Q^{obs}$ in method 4 is a function of $p$.

#### 3.1 Method 1: Stopping positions of the first probe vehicles

Theorem 1:

For any integer $n \geq 1$,

$$E(Q \mid N = n) = E(S \mid N = n)(n + 1) - 1.$$  

(2)

The proof is in Appendix.

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

$$\sum_{i: n_i \neq 0} \mathbb{E}(Q \mid N = n_i) = \sum_{i: n_i \neq 0} \left( \mathbb{E}(S \mid N = n_i) (n_i + 1) - 1 \right).$$ (3)

$$= \sum_{i: n_i \neq 0} \mathbb{E}(S \mid N = n_i) (n_i + 1) - \sum_{i: n_i \neq 0} 1$$ (4)

$$= \sum_{j=1}^{\infty} \sum_{i: n_i = j} \mathbb{E}(S \mid N = j) (j + 1) - \sum_{i: n_i \neq 0} 1$$ (5)

$$= \sum_{j=1}^{\infty} (j + 1) \sum_{i: n_i = j} \mathbb{E}(S \mid N = j) - \sum_{i: n_i \neq 0} 1$$ (6)

Therefore, by substituting the observed average value $$\frac{\sum_{i: n_i = j} s_i}{\sum_{i: n_i = j}}$$ for the expected value $$\mathbb{E}(S \mid N = j)$$, $$\forall j \geq 1$$, $$Q^{obs}$$ can be estimated by

$$\hat{Q}^{obs}_1 = \sum_{j=1}^{\infty} (j + 1) \sum_{i: n_i = j} s_i - \sum_{i: n_i \neq 0} 1$$ (7)

$$= \sum_{j=1}^{\infty} \sum_{i: n_i = j} s_i (j + 1) - \sum_{i: n_i \neq 0} 1$$ (8)

$$= \sum_{i: n_i \neq 0} s_i (n_i + 1) - \sum_{i: n_i \neq 0} 1$$ (9)

$$= \sum_{i: n_i \neq 0} (s_i (n_i + 1) - 1)$$ (10)

### 3.2 Method 2: Stopping positions of the last probe vehicles

**Theorem 2:**

For any integer $$n \geq 1,$$

$$\mathbb{E}(Q \mid N = n) = \mathbb{E}(T \mid N = n) \frac{n + 1}{n} - 1.$$ (11)

The proof is in Appendix.

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

$$\sum_{i: n_i \neq 0} \mathbb{E}(Q \mid N = n_i) = \sum_{i: n_i \neq 0} \left( \mathbb{E}(T \mid N = n_i) \frac{n_i + 1}{n_i} - 1 \right).$$ (12)

Following the similar derivations with method 1, by substituting the observed average value $$\frac{\sum_{i: n_i = j} t_i}{\sum_{i: n_i = j}}$$ for the expected value $$\mathbb{E}(T \mid N = j)$$, $$\forall j \geq 1$$, $$Q^{obs}$$ can be estimated by

$$\hat{Q}^{obs}_2 = \sum_{i: n_i \neq 0} \left( t_i \frac{n_i + 1}{n_i} - 1 \right).$$ (13)

### 3.3 Method 3: Stopping positions of the first and the last probe vehicles

**Theorem 3:**

For $$n \geq 1,$$

$$Q^{obs}$$ can be estimated by
\[
\mathbb{E}(Q \mid N = n) = \mathbb{E}(S \mid N = n) + \mathbb{E}(T \mid N = n) - 1, \quad (14)
\]

\[
\mathbb{E}(Q \mid N \geq 1) = \mathbb{E}(S \mid N \geq 1) + \mathbb{E}(T \mid N \geq 1) - 1. \quad (15)
\]

The proof is in Appendix. (It can be easily proved by applying Theorem 1 and Theorem 2.)

Theorem 3 states that the expected length of an observable queue can be obtained from the expected stopping positions of the first and the last probe vehicles. Based on Theorem 3, given the numbers of probe vehicles over different cycles, the expected total length of observable queues can be expressed as

\[
\sum_{n_i \neq 0} \mathbb{E}(Q \mid N = n_i) = \sum_{n_i \neq 0} (\mathbb{E}(S \mid N = n_i) + \mathbb{E}(T \mid N = n_i) - 1).
\]

(16)

Therefore, by substituting the observed average values \(\sum_{i: n_i \neq 0} \mathbb{E}(S \mid N = n_i)\) and \(\sum_{i: n_i \neq 0} \mathbb{E}(T \mid N = n_i)\) for the expected values \(\mathbb{E}(S \mid N = n_i)\) and \(\mathbb{E}(T \mid N = n_i), \forall j \geq 1\), respectively, \(Q_{obs}^{\text{obs}}\) can be estimated by

\[
Q_{obs} = \sum_{n_i \neq 0} (s_i + t_i - 1).
\]

(17)

### 3.4 Method 4: Bayes’ Theorem

Given the observations of probe vehicles’ stopping positions in each cycle, a natural way to estimate \(Q_{obs}^{\text{obs}}\) is to apply Bayes’ Theorem.

For the \(i\)th cycle, given the observed partial queue \(q_i\), according to Bayes’ Theorem, the probability of the queue length \(Q\) being \(l\) is

\[
P(Q = l \mid q_i) = \frac{P(Q = l)p(q_i \mid Q = l)}{\sum_{j=0}^{\infty} P(Q = j)p(q_i \mid Q = j)}. \quad (18)
\]

The conditional expectation of queue length is thus

\[
\mathbb{E}(Q \mid q_i) = \sum_{l=1}^{\infty} P(Q = l \mid q_i)l = \sum_{l=1}^{\infty} \frac{P(Q = l)p(q_i \mid Q = l)}{\sum_{j=0}^{\infty} P(Q = j)p(q_i \mid Q = j)}l. \quad (19)
\]

Given all the observed partial queues, the conditional expectation of the total length of the observable queues is

\[
\sum_{i:|q_i| \neq 0} \mathbb{E}(Q \mid q_i) = \sum_{i:|q_i| \neq 0} \sum_{l} \frac{P(Q = l)p(q_i \mid Q = l)}{\sum_{j=0}^{\infty} P(Q = j)p(q_i \mid Q = j)}l
\]

\[
= \sum_{i:|q_i| \neq 0} \sum_{l} \frac{\mathbb{E}(C_i)p^{n_i} (1-p)^{l-n_i}}{\sum_{j=0}^{\infty} \mathbb{E}(C_j)p^{n_i} (1-p)^{l-n_i}}l \quad \quad (20)
\]

\[
= \sum_{i:|q_i| \neq 0} \sum_{l} \frac{\mathbb{E}(C_i)p^n(1-p)^{l-n_i}}{\sum_{j=0}^{\infty} \mathbb{E}(C_j)p^n(1-p)^{l-n_i}}l \quad \quad (21)
\]

Equation (20) is equivalent to equation (21) due to the fact that \(\frac{\mathbb{E}(C_i)}{\mathbb{E}(C_j)} = \frac{P(Q = l)}{P(Q = j)}, \forall j, l\), as long as the denominators are not zero. Equation (21) is reformulated to equation (22), because

\[
P(q_i \mid Q = l) = \begin{cases} 0, & l < |q_i| \\ p^n(1-p)^{l-n_i}, & l \geq |q_i| \end{cases}. \quad (24)
\]

In equation (23), \(\mathbb{E}(C_i), \forall i \geq 1\) can be calculated approximately from the distribution of stopping positions, in the following way. Please note that \(\mathbb{E}(C_0)\), the expected counts of queues of zero lengths, never appears in equation (23).
3.4.1 Estimate $\mathbb{E}(C_l)$

Suppose the distribution of stopping positions of all the vehicles (probe vehicles and regular vehicles) is given, it implicitly reflects the distribution of queue lengths over different cycles. As demonstrated in Figure 3, if the maximal queue length is 6, then the counts of stopping vehicles at position 6 (the sixth stopping position behind the stop bar) is equal to the counts of queues of length 6; similarly, the counts of stopping vehicles at position 5 is equal to the total counts of queues of length 5 or 6; the counts of stopping vehicles at position 4 is equal to the total counts of queues of length 4, 5, or 6... As a result, the distribution of stopping positions has a decreasing trend with respect to the distance to stop bar.

![Figure 3: The relationship between the distributions of queue lengths and stopping positions](image)

Based on the reasoning above, once the distribution of stopping positions of all the vehicles is given, $\mathbb{E}(C_l)$ can be approximated by the difference between the counts of stopping vehicles at position $l$ and position $l+1$. Nevertheless, only the stopping positions of probe vehicles are observable. This is not a big issue though. Since probe vehicles are homogeneously mixed with regular vehicles, their stopping positions in the queues follow the same distribution. Figure 4 is an example that shows the distribution of stopping positions of all vehicles and the distribution of stopping positions of probe vehicles respectively, drawn from a Poisson process with average arrival rate $\lambda = 3$ and penetration rate $p = 0.1$. One can see that the shapes of two distributions are very close to each other.

![Figure 4: The similarity between the distribution of all vehicles and probe vehicles](image)

This can be explained analytically as well. By multiplying the denominator and the numerator by $p$ at...
the same time, equation 23 can be reexpressed as
\[
\sum_{i: q_i \neq 0} \mathbb{E}(Q | q_i) = \sum_{i: q_i \neq 0} \sum_{l=|q_i|}^\infty \frac{p \mathbb{E}(C_l)}{\sum_{j=|q_i|}^\infty b \mathbb{E}(C_j) (1-p)^{j-l}}.
\]
(25)

Although the stopping positions of probe vehicles cannot be used to approximate \(\mathbb{E}(C_l)\) directly, it can be used to approximate \(p \mathbb{E}(C_l)\). For instance, the difference between the counts of stopping probe vehicles at position 1 and position 2 is an approximation of \(p \mathbb{E}(C_1)\); the difference between the counts of stopping probe vehicles at position 2 and position 3 is an approximation of \(p \mathbb{E}(C_2)\)... Denote the approximation of \(p \mathbb{E}(C_l)\) by \(\hat{C}_l\).

Then, by replacing \(p \mathbb{E}(C_l)\) by its approximation \(\hat{C}_l\) in equation 25, \(Q^{obs}\) can be estimated by
\[
\hat{Q}^{obs}(p) = \sum_{i: q_i \neq 0} \frac{\hat{C}_l}{\sum_{j=|q_i|}^\infty \hat{C}_j (1-p)^{j-l}}.
\]
(26)
The notation \(\hat{Q}^{obs}(p)\) reflects the fact that it is a function of penetration rate \(p\).

However, due to randomness, sometimes the distribution of probe vehicles’ stopping positions does not follow the decreasing trend. For instance, in Figure 4 at position 8, there are more stopping probe vehicles than position 7. When this happens, it cannot be directly used to calculate \(\hat{C}_l\), because otherwise some of the values will be negative. In this case, \(\hat{C}_l\) can be calculated by solving the following optimization problem.

3.4.2 Calculate \(\hat{C}_l\)

Denote the observed counts of stopping probe vehicles at position \(l\) by \(\bar{c}_l\). A convex optimization problem can be formulated to calculate \(\bar{C}_l\) as follows.

\[
\begin{align*}
\text{minimize} & \quad \sum_i w_i \left( \sum_{l=i}^{\bar{c}_l} \hat{C}_l - \bar{c}_l \right)^2 \\
\text{subject to} & \quad \hat{C}_l \geq 0, \forall l = 1, 2, \ldots
\end{align*}
\]
(27)

Either 27 or 25 can be selected as the objective function of the optimization problem. Objective function 27 minimizes the sum of squared deviations of cumulative estimates \(\sum_{l=i}^{\bar{c}_l} \hat{C}_l\) from the observed counts \(\bar{c}_l\) of probe vehicles at all stopping positions \(i\); objective 28 focuses on the percentage deviations instead of the absolute values. \(w_i\) and \(v_l\) represent the corresponding weights of each squared error term. Constraint 29 states that \(\hat{C}_l\), an approximation of \(p \mathbb{E}(C_l)\), should be non-negative.

The idea behind the optimization problem is to make sure the non-negativity of \(\hat{C}_l\), while preserving the information provided by observation to the largest extent. Or, from another perspective, it tries to modify the distribution of probe vehicles’ stopping positions to the least extent, such that non-negative \(\hat{C}_l\) could be obtained. Once \(\hat{C}_l\) is obtained, plugging \(\hat{C}_l\) into equation 26 yields \(\hat{Q}^{obs}(p)\).

4 Estimation of \(Q^{hid}\)

After estimating \(Q^{obs}\), the following question is how to estimate \(Q^{hid}\), as there is no observation at all in the corresponding cycles. Fortunately, the fact that no probe vehicle is observed also contains information.

In last section, four different methods of estimating \(Q^{obs}\) are presented. In method 1, 2, and 3, the estimators do not rely on \(p\), thus the corresponding results are constants. Method 4 applies Bayes’ Theorem which requires \(p\), therefore the estimator \(\hat{Q}^{obs}(p)\) is a function of \(p\).

In this section, two methods of estimating \(Q^{hid}\) will be shown. Method 1 utilizes the ratio between the probability of being observed and the probability of being hidden for each queue, to estimate the total length of hidden queues. Method 2, like method 4 in last section, applies Bayes’ Theorem directly, but it requires that the total number of cycles is given as it appears in the final expression.
4.1 Method 1: probability of being observable / being hidden

Recall that random variable $X_l = 1$ if queue length $Q = l$, otherwise $X_l = 0$, as defined by equation (1). For those cycles with probe vehicles observed, the expected counts of queues of length $l$ can be expressed as

\[
\sum_{i: |q_i| \neq 0} \mathbb{E}(X_l | q_i) = \sum_{i: |q_i| \neq 0} P(Q = l | q_i).
\] (30)

For a queue of length $l$, the probability of being hidden (without any probe vehicle) is $P(N = 0 | Q = l) = (1-p)^l$; the probability of being observed (with at least one probe vehicle) is $P(N \geq 1 | Q = l) = 1 - (1-p)^l$. Therefore, in all the hidden queues, the expected counts of queues of length $l$ can be estimated by

\[
P(N = 0 | Q = l) \sum_{i: |q_i| \neq 0} \mathbb{E}(X_l | q_i) = \frac{(1-p)^l}{1 - (1-p)^l} \sum_{i: |q_i| \neq 0} P(Q = l | q_i) \sum_{i: |q_i| \neq 0} P(Q = l | q_i).
\] (31)

Thus, the expected total length of the hidden queues can be estimated by

\[
\sum_{l=1}^{\infty} \frac{(1-p)^l}{1 - (1-p)^l} \sum_{i: |q_i| \neq 0} P(Q = l | q_i) l = \sum_{i: |q_i| \neq 0} \sum_{l=1}^{\infty} \frac{(1-p)^l}{1 - (1-p)^l} P(Q = l | q_i) l
\]
\[
= \sum_{i: |q_i| \neq 0} \sum_{l=1}^{\infty} \frac{(1-p)^l}{1 - (1-p)^l} \sum_{j=0}^{\infty} P(Q = j) P(q_i | Q = j) l
\]
\[
= \sum_{i: |q_i| \neq 0} \sum_{l=1}^{\infty} \frac{(1-p)^l}{1 - (1-p)^l} \sum_{j=|q_i|}^{\infty} p\mathbb{E}(C_j) (1-p)^j l.
\] (32)

Equation (32) is equivalent to equation (33) because of Bayes’ Theorem shown by equation (18). Equation (33) is reformulated to equation (34) following the same reasoning with the reformulation from equation (20) to equation (25) in last section.

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

\[
\hat{Q}^{\text{hid}}(p) = \sum_{i: |q_i| \neq 0} \sum_{l=0}^{\infty} \frac{(1-p)^l}{1 - (1-p)^l} \sum_{j=|q_i|}^{\infty} \hat{C}_i (1-p)^j l.
\] (35)

where $\hat{C}_i, \forall l \geq 1$ could be found using the method shown in Section 3

4.2 Method 2: Bayes’ Theorem

According to equation (19), similarly to equation (25), given the fact that no probe vehicle is observed in those cycles, the expected total length of the hidden queues can be expressed as

\[
\sum_{i: |q_i| = 0} \mathbb{E}(Q | q_i) = \sum_{i: |q_i| = 0} \sum_{l=0}^{\infty} \frac{p\mathbb{E}(C_i) P(q_i | Q = l)}{p\mathbb{E}(C_j) P(q_i | Q = j)} l.
\] (36)

Please note that the inner summation over $l$ starts from 0, because $P(q_i | Q = 0) = 1 \neq 0$ if $|q_i| = 0$. Therefore, $\hat{C}_0$, an estimate of $p\mathbb{E}(C_0)$, should be calculated.

An estimator of $Q^{\text{hid}}$ can be given by

\[
\hat{Q}^{\text{hid}}_2(p) = \sum_{i: |q_i| = 0} \sum_{l=0}^{\infty} \frac{\hat{C}_i P(q_i | Q = l)}{C_j P(q_i | Q = j)} l,
\] (37)

\[
= \sum_{i: |q_i| = 0} \sum_{l=0}^{\infty} \frac{\hat{C}_i}{C_j (1-p)^j} l.
\] (38)

The calculation of $\hat{C}_i, \forall l \geq 1$, has been shown in Section 3 Here shows how to find $\hat{C}_0$. 
4.2.1 Calculate $\hat{C}_0$

In all the queues, the expected counts of empty queues (with zero length) is

$$E(C_0) = C - \sum_{l=1}^{\infty} E(C_l).$$

(39)

Therefore,

$$pE(C_0) = pC - \sum_{l=1}^{\infty} pE(C_l).$$

(40)

An estimator of $pE(C_0)$ can be easily defined as

$$\hat{C}_0 = pC - \sum_{l=1}^{\infty} \hat{C}_l,$$

(41)

Therefore, all the parameters except $p$ on the right-hand side of equation (38) are known. If stop line detection (such as loop detectors) data are given, $\hat{C}_0$ can be more easily obtained.

5 Estimation of Penetration Rate

In last section, two methods of estimating $Q^{hid}$ are presented. Both of the estimators rely on $p$. Apparently, $p$ is the key because only if $p$ is known, can the total queue length be estimated. In this section, two different methods of estimating penetration rate $p$ will be shown. Method 1 is based upon the equivalence between different estimators of $Q^{obs}$ and $Q^{hid}$. Method 2 exploits the fact that the portion of probe vehicles in the queues should be close to penetration rate $p$.

5.1 Method 1

When estimating $Q^{obs}$, method 1, 2, and 3 can generate constant results, whereas the estimator given by method 4 is a function of $p$. Since they all estimate the same variable $Q^{obs}$, it implies an easy way to estimate penetration rate $p$ by letting

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

(42)

In this equation, there is only one unknown variable $p$. Therefore, solving the equation will yield an estimate of penetration rate $p$.

Similarly, when estimating $Q^{hid}$, the estimators given by method 1 and method 2 rely upon only one unknown variable $p$. Therefore, the following equation can be used to estimate $p$ as well.

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

(43)

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

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

(44)

As long as it is an equation with a single unknown variable $p$, solving it gives an estimate of $p$. Both the left-hand side and the right-hand side can be regarded as estimators of the total length of all the queues.

5.2 Method 2

Here is another different way to estimate $p$.

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

(45)

where $Q^{probe}$ denotes the total number of probe vehicles in all the queues, thus $Q^{probe} = \sum_i n_i$. The left-hand side of equation (45) could be interpreted as an estimate of the portion of probe vehicles in all the queueing vehicles. The right-hand side is penetration rate, or the probability of being a probe vehicle when
arbitrarily selecting a vehicle from the traffic. Similarly, solving this equation with only one single unknown variable yields an estimate of $p$.

In practice, when solving $p$ by equation (42), (43), (44), or (45), one can search $p$ from an upper bound to 0 with a small step size, until the difference between the left-hand side and the right-hand side reaches certain criteria. For instance, an upper bound can be $Q_{\text{probe}} \sum_i |q_i|$.

6 Estimation of Queue Length and Traffic Volume

6.1 Queue length estimation

Once penetration rate $p$ is obtained, all of the estimators of $Q_{\text{obs}}$ and $Q_{\text{hid}}$ can be calculated. The total queue length can then be estimated by summing up $Q_{\text{obs}}^i(p)$ and $Q_{\text{hid}}^j(p)$, $\forall i = 1, 2, 3$, $\forall j = 1, 2$. The distribution of queue lengths can also be obtained. Since $\hat{C}_l$ is an approximation of $p E(C_l)$, the expected counts of queues of length $l$ can be estimated by $\hat{C}_l p$.

A potential extension is to estimate queue length cycle by cycle (or even in real time), as illustrated in some literature such as (Comert and Cetin, 2009). The prior information of penetration rate and the distribution of queue lengths, as required by those models, can be obtained by applying the methods in this paper to data collected during a relatively long period of time (for instance a few weeks). In other words, the methods in this paper could enable the practical applications of some existing studies.

6.2 Traffic volume estimation

Traffic volume can also be estimated once $p$ is known. Since the current traffic signal control algorithms do not distinguish probe vehicles and regular vehicles, for each movement at an intersection, the penetration rate of probe vehicles among the stopping vehicles and the penetration rate of probe vehicles among the non-stopping vehicles are the same. Therefore, the total traffic volume can be easily scaled up (Wong et al., 2018b; Wong and Wong, 2018) by

$$\hat{V}_{\text{all}} = \frac{V_{\text{probe}}}{p},$$

where $V_{\text{probe}}$ denotes the volume of probe vehicles for a specific movement.

7 Validation and Evaluation

7.1 Simulation

7.1.1 Input data and ground truth

First, a distribution of queue lengths, for instance Poisson distribution, is pre-determined. Then in each cycle, a sample (queue length) is drawn from the distribution. The queue length represents the length of the whole queue, which is ground truth. For each vehicle in the queue, their types (probe vehicle or regular vehicle) are determined randomly based on penetration rate which varies from 0.01 to 0.99 in different scenarios. According to the vehicle types, the observed partial queues (queues until the last probe vehicles) could be extracted. The input to the proposed methods is the observed partial queues over different cycles.

7.1.2 Results

Some of the existing literature assumes the vehicle arrival process to be Poisson process (Li et al., 2017; Zheng and Lin, 2017). The proposed method does not assume the queue length follow Poisson distribution, thus is adaptive to any distribution of queue lengths. For demonstration purposes, when testing the proposed methods by simulation, the input data are drawn from a Poisson distribution with $\lambda = 10$ for 1,000 cycles.

As can be seen from the derivations in the previous sections, the estimation of penetration rate is a key step. Therefore, the focus of the tests by simulation is on penetration rate. Figure (5) shows the results for four different submethods. The horizontal axes represent the ground truth of penetration rates, whereas the
vertical axes represent the estimated values. The diagonal in red is for reference. It can be seen that the dots in blue are very close to the diagonals, which implies that the method can estimate penetration rate very accurately. Generally, the submethods of method 2 outperforms the submethods in method 1.

In order to demonstrate the impact of sample size on estimation accuracy, data of 100 cycles, 200 cycles, 500 cycles, 1,000 cycles are used in four rounds of tests. The method evaluated is submethod \( \frac{q_{\text{probe}}}{q_{\text{obs}} + q_{\text{hid}}(p)} = p \) of Method 2. By comparing the results of different sample sizes in Figure 6, it can be concluded that the larger the sample size is, the more accurate the estimation tends to be.

Figure 5: The results of some methods using simulation data
Figure 6: The estimation results of penetration rates with different sample sizes

7.2 Real-field data

The proposed methods are also tested using real-field data. First of all, an area in Suzhou, Jiangsu Province, China is selected, because it is covered by Didi Chuxing’s e-hailing services and most of the intersections in this area are covered by the local government’s camera-based vehicle automatic identification system. The trajectories of the vehicles offering Didi Chuxing’s e-hailing services (penetration rate usually ranges from 5% to 15%) are the input data, while the counts from the cameras are considered as the ground truth of traffic volume. Since the proposed methods cannot deal with right-turn movements (no regular queues), and the lanes with mixed movements (for instance a lane with both left-turn vehicles and through vehicles), only a part of the movements in this area are studied. The selected movements in the area are shown in Figure 7a and Figure 7b.
7.2.1 Input data

The original data are the GPS trajectories of vehicles offering Didi Chuxing’s e-hailing services. The trajectory data contain timestamp, longitude, latitude, speed, etc. The data of the weekdays from May 8, 2018 to May 21, 2018 are used (15 days in total). At the preprocessing stage, the trajectories are mapped to the transportation network by map matching, using a method like (Newson and Krumm, 2009). Then, for each movement, a “snapshot” of its corresponding trajectories is taken every five seconds. From the “snapshots”, the positions of all the stopping probe vehicles are identified and the positions of the regular vehicles before the last probe vehicles are inferred by assuming constant gaps between queueing vehicles. Average distance between two adjacent stopping vehicles in the same lane is empirically set to be 7.5 m/veh for peak hours and 8.0 m/veh for off-peak hours. One thing worth to mention is that, the observed partial queues during green phase should be excluded because they could make penetration rate underestimated. For instance, if a (partial) queue \((0, 0, 1, 0, 0, 0, 1)\) was observed, and five seconds later it became \((0, 0, 0, 0, 0, 0, 1)\), then the latter queue should be excluded because it was discharging.
In the studied area, most movements are corresponding to multiple lanes. However, the accuracy of trajectory data cannot reach lane level, thus it is almost impossible to identify on which lanes the vehicles stopped. Therefore, the stopping vehicles are randomly assigned to different lanes of the corresponding movements. The final input to the proposed method is a list of observed (partial) queues. The random assignment process is repeated for many times to get an average estimate.

7.2.2 Ground truth of traffic volume

Over the roads at the studied intersections, camera-based automatic vehicle identification systems are installed. The cameras can identify which vehicle passes which lane at what time. Nevertheless, not all the vehicles can be successfully identified, even at day time. Since all the trajectories of the e-hailing vehicles operated by DiDi Chuxing are known, the identification rate of these vehicles can be served as an estimate of the real accuracy of the cameras. Using this method, the identification rates of the cameras are estimated. The identification rates are then used to project the direct vehicle counts given by the cameras to their real “ground truth”. Figure 8 shows the changes of the accuracy of three cameras in granularity of half an hour during May 8, 2018 to May 15, 2018. The accuracy majorly ranged from 80% to 100% at day time. At night, the performance of cameras became unstable over different days. In order to ensure the quality of “ground truth”, here the focus is on the estimation of traffic volume during day time.
Figure 8: Accuracy of the cameras

### 7.2.3 Results

Figure 9 shows the traffic volume estimation results for all the 22 through movements shown in Figure 7a. The three studied time slots, 8 am to 9 am, 13 pm to 14 pm, and 18 pm to 19 pm, are of morning peak hours, off-peak hours, and evening peak hours respectively. The used measure of estimation accuracy is mean absolute percentage error (MAPE). The estimation results show that the method applied \( \frac{Q_{\text{probe}}}{Q_{2i,1} + Q_{2i,2}(p)} = p \) can estimate traffic volume very accurately, which is sufficient for most signal control and performance measure applications.

Figure 10 shows the results for all the 31 left-turn movements shown in Figure 7b. The estimation results for the three studied time slots, although not as good as the results of the through movements, are satisfactory. One possible explanation of the undermined performance is that the traffic volume of left-turn movements is much smaller than through movements, therefore, the number of samples that can be used for estimation is much smaller.
Figure 9: Results for through movements
Figure 10: Results for left-turn movements
In general, the figures above have shown the good quality of estimation results. Figure 11 shows the ground-truth penetration rates for all the selected movements during the three studied time slots. In most cases, penetration rates are in the range of 5% to 15%. This is a profound evidence that the proposed method can estimate traffic volume accurately in terms of mid-term or long-term signal control and performance measure, even under low penetration rates of probe vehicles. Therefore, it is ready for large-scale real-field applications.

Figure 11: The penetration rates of probe vehicles in all the selected movements
8 Discussions and Conclusions

8.1 Advantages of the proposed methods

1. Unlike many existing studies that assume vehicle arrival follows Poisson processes, the proposed methods do not assume the type of stochastic process of vehicle arrival or the type of distribution of the queue lengths. This implies a few advantages of the proposed methods.

   (a) They can be applied in both under-saturation and over-saturation cases. In over-saturation cases, the queueing processes are different from the arrival processes or departure processes due to the overflow queues. In other words, assuming Poisson process may not be reasonable in over-saturation cases.

   (b) They do not require the exact positions of stop bars. For a specific movement, penetration rate of probe vehicles in any segment of its corresponding lanes is still $p$, therefore it can be estimated by only using the data in a road segment which may not start from a stop bar.

   (c) They do not necessarily rely on signal timing information. As long as the penetration rate of probe vehicles in the queue samples is not biased, the methods can generate accurate results. For example, when the proposed methods are tested using real-field data, the “snapshots” of trajectories taken every five seconds can be used as an input to the model after removing the “snapshots” during green phase, without knowing about signal timing.

2. At all different levels of penetration rates, the methods can generate estimation results with good quality, as long as a sufficient amount of data is given. The more data are given, the more accurate the estimation results tend to be. As a result, the proposed methods have an edge over the shockwave theory based methods which usually require relatively high penetration rate of probe vehicles.

3. The methods can be combined with other sources of data to improve estimation accuracy. For instance, when estimating $pE(C_0)$, $E(C_0)$ can be estimated by loop detector data.

8.2 Limitations of the proposed methods

1. The proposed methods rely on stopping positions to estimate queue length, penetration rate of probe vehicles, and traffic volume, thus may not be used for non signalized intersections, or right-turn movements. To deal with the problem, in engineering practice, one can assume the penetration rates in adjacent movements or at adjacent intersections are close and thus use them as substitutes.

2. When estimating traffic volume for a movement, the proposed methods require that the movement is not in a shared left-through lane or right-through lane. The reason is that if a lane has both left-turn (right-turn) and through vehicles, its queueing pattern might not be the same with other left-turn (right-turn) lanes or other through lanes. Since the accuracy of current trajectory data cannot reach lane level, it is difficult to separate the vehicles that stopped on this kind of lanes from other vehicles of the same movement.

3. Average space headway between adjacent stopping vehicles might vary spatially and temporally. However, because of the accuracy of trajectory data, it is difficult to accurately calculate it for each movement and for each time slot. When testing the proposed methods using data collected in Suzhou, the average distances are set empirically. In the future, if data with higher accuracy are accessible, the values should be estimated independently.

In summary, the first limitation is due to the methodology of the proposed method. However, the second and the third limitations are due to the quality of data that are available. They could be overcome if data with higher accuracy can be collected and accessed in the future.
8.3 Contributions of this paper

1. This paper proposes a series of novel methods to estimate the distribution of queue lengths, penetrate rate of probe vehicles, and traffic volume, using probe vehicle trajectories. The methods work well even when the penetration rate of probe vehicles is very low.

2. The proposed methods are adaptive to different scenarios. They work in both under-saturation cases and over-saturation cases. They do not require the information of signal timing and the exact locations of stop bars. The tests by both simulation and real-field data have shown that the estimation results are accurate enough for mid-term or long-term signal control and performance measure, thus they are ready for large-scale real-field applications.

3. The methods can be used to calculate the prior information of penetration rate and distribution of queue lengths, which could enable some methods in the existing literature to estimate queue length cycle by cycle.
References

An, C., Wu, Y.-J., Xia, J., Huang, W., 2018. Real-time queue length estimation using event-based advance detector data. Journal of Intelligent Transportation Systems 22 (4), 277–290.

Badillo, B. E., Rakha, H., Rioux, T. W., Abrams, M., 2012. Queue length estimation using conventional vehicle detector and probe vehicle data. In: Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on. IEEE, pp. 1674–1681.

Ban, X. J., Hao, P., Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C: Emerging Technologies 22 (4), 277–290.

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

Comert, G., 2013a. Effect of stop line detection in queue length estimation at traffic signals from probe vehicles data. European Journal of Operational Research 226 (1), 67–76.

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

Comert, G., 2016. Queue length estimation from probe vehicles at isolated intersections: Estimators for primary parameters. European Journal of Operational Research 252 (2), 502–521.

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

Comert, G., Cetin, M., 2011. Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data. IEEE Transactions on Intelligent Transportation Systems 12 (2), 563–573.

Hao, P., Ban, X. J., 2015. Long queue estimation for signalized intersections using mobile data. Transportation Research Part C: Emerging Technologies 56, 1–17.

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

Li, J., Zhou, K., Shladover, S. E., Skabardonis, A., 2013. Estimating queue length under connected vehicle technology: Using probe vehicle, loop detector, fused data. Transportation Research Record 2366 (1), 17–22.

Liu, H. X., Wu, X., Ma, W., Hu, H., 2009. Real-time queue length estimation for congested signalized intersections. Transportation research part C: emerging technologies 17 (4), 412–427.

Newson, P., Krumm, J., 2009. Hidden markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, pp. 336–343.
Rompis, S. Y., Cetin, M., Habtemichael, F., 2018. Probe vehicle lane identification for queue length estimation at intersections. Journal of Intelligent Transportation Systems 22 (1), 10–25.

Tiaprasert, K., Zhang, Y., Wang, X. B., Zeng, X., 2015. Queue length estimation using connected vehicle technology for adaptive signal control. IEEE Transactions on Intelligent Transportation Systems 16 (4), 2129–2140.

Wang, Z., Cai, Q., Wu, B., Zheng, L., Wang, Y., 2017. Shockwave-based queue estimation approach for undersaturated and oversaturated signalized intersections using multi-source detection data. Journal of Intelligent Transportation Systems 21 (3), 167–178.

Wong, W., Shen, S., Liu, H. X., 2018a. On the estimation of penetration rate based on single-source connected vehicle data, unpublished Manuscript.

Wong, W., Wong, S., 2015. Systematic bias in transport model calibration arising from the variability of linear data projection. Transportation Research Part B: Methodological 75, 1–18.

Wong, W., Wong, S., 2016. Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection. Transportation Research Part B: Methodological 88, 72–92.

Wong, W., Wong, S., 2018. Unbiased estimation methods of nonlinear transport models based on linearly projected data, (in press).

Wong, W., Wong, S., Liu, H. X., 2018b. Bootstrap standard error estimations of nonlinear transport models based on linearly projected data, (in press).

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

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

Lemma 1

\[ \sum_{k=0}^{m} C_{n+k}^m = C_{n+m+1}^{n+1}, \quad (A.1) \]

where \( m, n \in \mathbb{N} \).

Proof:

1. Base Case: When \( m = 0 \),

\[ \sum_{k=0}^{m} C_{n+k}^m = C_{n}^n = 1 = C_{n+1}^{n+1} = C_{n+m+1}^{n+1}, \quad (A.2) \]

2. Induction: Assume equation (A.1) holds for a given natural number \( m \in \mathbb{N} \), i.e.,

\[ \sum_{k=0}^{m} C_{n+k}^m = C_{n+m+1}^{n+1}, \quad (A.3) \]

Then for \( m + 1 \),

\[ \sum_{k=0}^{m+1} C_{n+k}^m = \sum_{k=0}^{m} C_{n+k}^m + C_{n+1}^m \]

\[ = C_{n+m+1}^{n+1} + C_{n+m+1}^{n+1} \]

\[ = \frac{(n + m + 1)!}{(n+1)!m!} + \frac{(n + m - 1)!}{n!(m+1)!} \]

\[ = \frac{(n + m + 1)!(m+1)}{(n+1)!(m+1)!} + \frac{(n + m + 1)!(n + 1)}{(n+1)!(m+1)!} \]

\[ = \frac{(n + m + 2)!}{(n+1)!(m+1)!} \]

\[ = C_{n+m+2}^{n+1}. \quad (A.9) \]

Equation (A.1) holds for \( m + 1 \) as well. Therefore, Lemma 1 is proved by induction.

Theorem 1: Stopping position of the first vehicle

\[ \mathbb{E}(S \mid N = n, Q = l) = \frac{l + 1}{n + 1}, \quad (A.10) \]

\[ \mathbb{E}(Q \mid N = n) = \mathbb{E}(S \mid N = n)(n + 1) - 1, \quad (A.11) \]

where \( n \geq 1 \).

Proof:
\[ E(S \mid N = n, Q = l) = \sum_{i=1}^{l-n+1} P(S = i \mid N = n, Q = l)i \]  
\[ = \sum_{i=1}^{l-n+1} \frac{nC_{i-n}^{l-1}A_{i-1}^{l-1}A_{i}^{l-i}}{A_{i}^{l}} \]  
\[ = \sum_{i=1}^{l-n+1} \frac{nA_{i}^{n-1}}{A_{i}^{l}} \]  
\[ = \frac{n}{A_{l}^{n}} \sum_{i=1}^{l-n+1} A_{i}^{n-1}i \]  
\[ = \frac{n}{A_{l}^{n}} \sum_{j=0}^{l-n} A_{n+j-1}^{n-1}(l - n + 1 - j) \]  
\[ = \frac{n}{A_{l}^{n}} \sum_{j=0}^{l-n} A_{n+j-1}^{n-1}(l + 1) - \frac{n}{A_{l}^{n}} \sum_{j=0}^{l-n} A_{n+j-1}^{n-1}(n + j) \]  
\[ = (l + 1) \sum_{j=0}^{l-n} \frac{(n + j - 1)!(l - n)!n!}{j!(n - 1)!} - \frac{n}{A_{l}^{n}} \sum_{j=0}^{l-n} A_{n+j}^{n} \]  
\[ = \frac{l + 1}{C_{l}^{n}} \sum_{j=0}^{l-n} C_{n+j-1}^{n-1} - \frac{n}{C_{l}^{n}} \sum_{j=0}^{l-n} C_{n+j}^{n} \]  
\[ = (l + 1) \frac{C_{l}^{n}}{C_{l}^{n}} - n \frac{C_{l+1}^{n+1}}{C_{l}^{n}} \]  
\[ = (l + 1) - n \frac{l + 1}{n + 1} \]  
\[ = \frac{l + 1}{n + 1} \]

Lemma 1 is applied when equation (A.19) is converted to (A.20). Then, based on the results above,
\[
E(S \mid N = n) = \sum_{i=n} P(S = i \mid N = n)i 
\]
(A.23)

\[
= \sum_{i=1}^{l-n+1} \sum_{i=1}^{l+n-1} P(S = i \mid N = n, Q = l) P(Q = l \mid N = n)i
\]
(A.24)

\[
= \sum_{i=n}^{l-n+1} P(S = i \mid N = n, Q = l) P(Q = l \mid N = n)i
\]
(A.25)

\[
= \sum_{i=n} P(Q = l \mid N = n) \sum_{i=1}^{l-n+1} P(S = i \mid N = n, Q = l)i
\]
(A.26)

\[
= \sum_{i=n} P(Q = l \mid N = n) E(S \mid N = n, Q = l)
\]
(A.27)

\[
= \sum_{i=n} P(Q = l \mid N = n) \frac{l+1}{n+1}
\]
(A.28)

\[
= \frac{1}{n+1} \sum_{i=n} P(Q = l \mid N = n)(l+1)
\]
(A.29)

\[
= \frac{1}{n+1} (E(Q \mid N = n) + 1).
\]
(A.30)

This is equivalent to

\[
E(Q \mid N = n) = E(S \mid N = n)(n+1) - 1.
\]
(A.31)

**Theorem 2: Stopping positions of the last vehicle**

\[
E(T \mid N = n, Q = l) = \frac{n+1}{n+1},
\]
(A.32)

\[
E(Q \mid N = n) = E(T \mid N = n) \frac{n+1}{n} - 1,
\]
(A.33)

where \( n \geq 1 \).

**Proof:**
\( \mathbb{E}(T \mid N = n, Q = l) = \sum_{i=n}^{l} P(T = i \mid N = n, Q = l)i \)  
(A.34)

\[ = \sum_{i=n}^{l} nC_{l-n}^{i-i}A_{i-1}^{i-i}A_{i-1}^{i-i} \]
(A.35)

\[ = \sum_{i=n}^{l} nA_{i-n}^{i-n} \]
(A.36)

\[ = n \sum_{i=n}^{l} A_{i}^{i} \]
(A.37)

\[ = n \sum_{i=n}^{l} \frac{C_{n}^{i}}{C_{i}^{i}} \]
(A.38)

\[ = \sum_{i=n}^{l} \frac{C_{n}^{i}}{C_{i}^{i}} \]
(A.39)

\[ = \sum_{i=n}^{l} \frac{C_{n}^{i+1}}{C_{i}^{i}} \]
(A.40)

\[ = \sum_{i=n}^{l} \frac{C_{n+1}^{i+1}}{C_{i}^{i}} \]
(A.41)

Then, based on the results above,

\[ \mathbb{E}(T \mid N = n) = \sum_{i=n}^{l} P(T = i \mid N = n)i \]
(A.42)

\[ = \sum_{i=n}^{l} \sum_{l=i}^{l} P(T = i \mid N = n, Q = l)P(Q = l \mid N = n)i \]
(A.43)

\[ = \sum_{l=n}^{n} \sum_{i=n}^{l} P(T = i \mid N = n, Q = l)P(Q = l \mid N = n)i \]
(A.44)

\[ = \sum_{l=n}^{n} P(Q = l \mid N = n) \sum_{i=n}^{l} P(T = i \mid N = n, Q = l)i \]
(A.45)

\[ = \sum_{l=n}^{n} P(Q = l \mid N = n)\mathbb{E}(T \mid N = n, Q = l) \]
(A.46)

\[ = \sum_{l=n}^{n} P(Q = l \mid N = n)n \frac{l + 1}{n + 1} \]
(A.47)

\[ = \frac{n}{n + 1} \sum_{i=n}^{l} P(Q = l \mid N = n)(l + 1) \]
(A.48)

\[ = \frac{n}{n + 1} \left( \mathbb{E}(Q \mid N = n) + 1 \right). \]  
(A.49)

This is equivalent to

\[ \mathbb{E}(Q \mid N = n) = \mathbb{E}(T \mid N = n) \frac{n + 1}{n} - 1. \]  
(A.50)

**Theorem 3: Stopping positions of the first and the last CVs**

\[ \mathbb{E}(Q \mid N \geq 1) = \mathbb{E}(S \mid N \geq 1) + \mathbb{E}(T \mid N \geq 1) - 1 \]  
(A.51)
Proof:
First of all,

\[ P(S = i \mid N \geq 1, Q = l) = p(1 - p)^{i-1} \tag{A.52} \]

\[ P(T = l - i + 1 \mid N \geq 1, Q = l) = p(1 - p)^{l-(l-i+1)} = p(1 - p)^{i-1} = P(S = i \mid N \geq 1, Q = l) \tag{A.53} \]

Then,

\[ \mathbb{E}(S \mid N \geq 1) = \sum_{i=1}^{l} P(S = i \mid N \geq 1)i \tag{A.54} \]

\[ = \sum_{i=1}^{l} \sum_{l=i}^{l} P(S = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)i \tag{A.55} \]

\[ = \sum_{i=1}^{l} \sum_{l=i}^{l} P(T = l - i + 1 \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)i \tag{A.56} \]

\[ = \sum_{i=1}^{l} \sum_{j=1}^{l} P(T = j \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)(l - j + 1) \tag{A.57} \]

\[ = \sum_{i=1}^{l} \sum_{j=1}^{l} P(T = j \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)(l - j + 1) \tag{A.58} \]

\[ = \sum_{i=1}^{l} \sum_{j=1}^{l} P(T = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)(l - i + 1) \tag{A.59} \]

\[ \mathbb{E}(T \mid N \geq 1) = \sum_{i=1}^{l} P(T = i \mid N \geq 1)i \tag{A.60} \]

\[ = \sum_{i=1}^{l} \sum_{l=i}^{l} P(T = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)i \tag{A.61} \]

\[ = \sum_{i=1}^{l} \sum_{l=i}^{l} P(T = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)i \tag{A.62} \]

\[ = \sum_{i=1}^{l} \sum_{l=i}^{l} P(T = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)i \tag{A.63} \]

Therefore,

\[ \mathbb{E}(S \mid N \geq 1) + \mathbb{E}(T \mid N \geq 1) - 1 = \sum_{i=1}^{l} \sum_{l=i}^{l} P(T = i \mid N \geq 1, Q = l)P(Q = l \mid N \geq 1)(l - 1) + 1 \tag{A.64} \]

\[ = \sum_{l} P(Q = l \mid N \geq 1)(l - 1) + 1 \tag{A.65} \]

\[ = \sum_{l} P(Q = l \mid N \geq 1)l \tag{A.66} \]

\[ = \mathbb{E}(Q \mid N \geq 1) \tag{A.67} \]

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