Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 5119209, 14 pages
https://doi.org/10.1155/2022/5119209

Research Article

An Extendable Gaussian Mixture Model for Lane-Based Queue Length Estimation Based on License Plate Recognition Data

Chaofeng Tan,[1] **Hao Wu**,[2] **Keshuang Tang**,[2] **and Chaopeng Tan**[3]

[1] *College of Ports and Shipping Management, Guangzhou Maritime University, Guangzhou 510725, China*
[2] *The Key Laboratory of Road and Traffic Engineering of the Ministry of Education, College of Transportation Engineering, Tongji University, Shanghai 201804, China*
[3] *Department of Civil and Environmental Engineering, National University of Singapore 117576, Singapore*

Correspondence should be addressed to Chaopeng Tan; tanchaopeng@tongji.edu.cn

Received 27 September 2022; Revised 21 December 2022; Accepted 21 December 2022; Published 29 December 2022

Academic Editor: Sheng Jin

Most existing studies on queue length estimation based on license plate recognition (LPR) data require multisection LPR data. Studies based on single-section LPR data cannot ensure the accuracy and stability of the estimates when missed detections occur, which greatly limits the practicability of existing studies. Therefore, using single-section LPR data, this study proposes a lane-based queue length estimation method based on a two-dimensional Gaussian mixture model. First, the LPR data were processed to obtain the departure times and time headway of vehicles. Then, the two-dimensional Gaussian distributions of queued vehicles and nonqueued vehicles were fitted, and the expectation-maximization algorithm was adopted to solve the distribution parameters. Finally, the queuing status of each vehicle was determined, and the lane-based queue length was estimated based on the last identified queued vehicle in the cycle. The empirical results showed that the mean absolute errors (MAEs) of the proposed method were just 1.3 veh/cycle under no missed detections and 2 veh/cycle under a 20% missed detection rate, outperforming existing methods. The simulation results indicate that the proposed method can achieve accurate estimates under various traffic demands. In addition, the proposed method can be extended to real-time applications and multisection LPR systems.

1. Introduction

The queue length is widely recognized as a key performance metric for signalized intersections [1]. The lane-based queue length can reflect the capacity of traffic signals and provide queue distribution information between lanes, which greatly contributes to signal control and active queue management at signalized intersections [2, 3].

Traditional queue length estimation methods mainly use traffic flow information, such as volume, occupancy, and speed information, collected by fixed-location sensors, including loop detectors, microwave radar, and magnetometer detectors [4–7]. However, the real-world application of these methods is often constrained by the low resolution, malfunction, and high maintenance cost of the sensors [8]. Recent advancements in intelligent mobility and connected vehicle technologies have boosted many studies on traffic state parameter estimation based on mobile sensors [9–12], the majority of which focus on queue length estimation [13–19]. Nevertheless, mobile sensor data suffer from low penetration rates [20]. In addition, the current positioning accuracy of probe vehicles cannot determine the lane to which it belongs; thus, the lane-based queue length cannot be determined.

In recent years, license plate recognition (LPR) systems have been widely implemented in many cities in China to meet the needs of traffic monitoring, law enforcement, and emergency operations. For instance, more than 4,800 sets of LPR cameras were implemented on urban roads in Shanghai as of 2019, and this number is expected to increase over the years. LPR cameras are typically deployed near the stop lines of signalized intersections for real-time detection of traffic violations, such as running a red light. Additionally, LPR cameras can precisely record information, including license plate number, stop-line crossing time, and vehicle type, for numerous vehicles passing through. Furthermore, because both LPR systems and traffic signal control systems are operated by traffic management sectors in the real world, signal timing data are also available for. Thus, LPR data have great potential to facilitate the dynamic evaluation of traffic signals.

However, in the interests of data privacy, LPR data are usually managed by local traffic management sectors and are typically not available for research communities. Therefore, limited studies have been devoted to the application of LPR data in queue length estimation. Given that the LPR systems record the vehicle license plates and stop-line crossing time sequences at the intersection level, by tracking the same vehicle at multiple intersections, the path travel time of vehicles can be easily extracted. Most existing studies are based on double-section LPR data for queue length estimation. Using the Gaussian process interpolation model, Zhan et al. [21] reconstructed the equivalent arrival-departure curves based on matched vehicles recorded by two successive LPR cameras. Then, the queue length can be estimated by a car-following based simulation scheme. Combined with signal schemes, Luo et al. [22] proposed a queue length estimation method based on travel time information provided by double-section LPR data, whose key point is the intrinsic connections between the travel time of individual vehicles and the queue composition in each cycle. Although these methods can achieve reasonable accuracy in lane-based queue length estimation, their real-world applications are constrained because the upstream intersections are not equipped with LPR systems in some cases, especially for minor roads.

Meanwhile, a few studies have developed a lane-based queue length estimation method based on single-section LPR data. Fusing connected vehicle (CV) trajectories and single-section LPR data, Tan et al. [20] proposed a Bayesian method for the queue length estimation. First, the CV trajectories are used to calibrate the two-dimensional probability density distribution of queued and nonqueued vehicles. Then, the conditional probability of the queuing status of each vehicle recorded by LPR systems is derived based on the Bayesian theory. Finally, the queue length is estimated with the largest possibility. However, the requirement of CV trajectories constrains the practical application of this method. Considering that the discharging time headways between queued and nonqueued vehicles are different, Wu et al. [23] proposed a critical point analysis (CPA) method to identify the change point of the discharging headway sequence solely using single-section LPR data; thus, the queue length can be determined by the identified change point. However, this method is quite sensitive to the outliers of the discharging headway sequence, which often occurs because of the missed detections of LPR data and the existence of heavy vehicles. Later, Tang et al. [24] improved this deficiency by using multilane LPR data interchecking. Recently, Zhan et al. [25] developed a lightweight lane-based Gaussian process model for cycle maximum queue length approximation, in which three critical parameters, the saturation discharging flow with departure rate, queue clearance time, and normal departure flow with departure rate, are used to model the cumulative departure process using a Gaussian process model, and the queue length can be obtained after inferring parameters via the Markov chain Monte Carlo (MCMC) technique. However, several hyperparameters in the Gaussian process model of this method need to be carefully calibrated before estimation and because of the repeated sampling process of MCMC, this method is time-consuming, limiting its empirical application.

In summary, existing studies have not solved the problem of reliable lane-based queue length estimation based on single-section LPR data that may suffer from missed detections. This study aims to address this research gap by proposing a data-driven statistical method for lane-based queue length estimation based on single-section LPR data. First, the LPR data are preprocessed, and the discharging headway and departure time during the green phase of each recorded vehicle are extracted. Then, a two-dimensional Gaussian mixture model (GMM) is proposed to fit the distribution of queued and nonqueued vehicles, and the expectation-maximization (EM) algorithm is used to solve the parameters of the GMM. Finally, the queuing status of each recorded vehicle can be identified by the GMM, and the maximum queue length can be estimated given the last queued vehicle during the cycle. Note that, although the main goal of this study is to achieve the queue length estimation with single-section LPR systems, the proposed method is extendable to the scenario with multisection LPR systems by incorporating the travel time information into the model.

The contributions of this study can be summarized as follows: (1) A lane-based queue length estimation method is developed using single-section LPR data only, which can be further extended to scenarios with multisection LPR systems. (2) The statistical property of the proposed method makes it reliable even when the LPR data suffer from missed detections, bridging the current research gap. (3) The proposed method outperforms an existing method under the different missed detection rate.

2. Methods

2.1. Preprocessing of LPR Data. The detection process of the LPR system is shown in Figure 1. When a vehicle passed through the detection zone, the vehicle information, including the location, entrance, lane, global time, vehicle type, and license plate number, was recorded. Normally, the detection zone ranges from 5 to 15 m upstream of the stop line; thus, the first one or two vehicles may be detected during the red phase, whereas the others are detected when they pass through during the green phase. Therefore, the LPR data record the departure process of vehicles during the cycle, that is, the departure time sequence of vehicles in the cycle. Notably, as illumination changes affect the detection of the LPR cameras and the license plates of a few passenger cars may be blocked by heavy vehicles, the LPR system is susceptible to missed detections. In addition, in the real world, both the LPR and signal control systems are operated

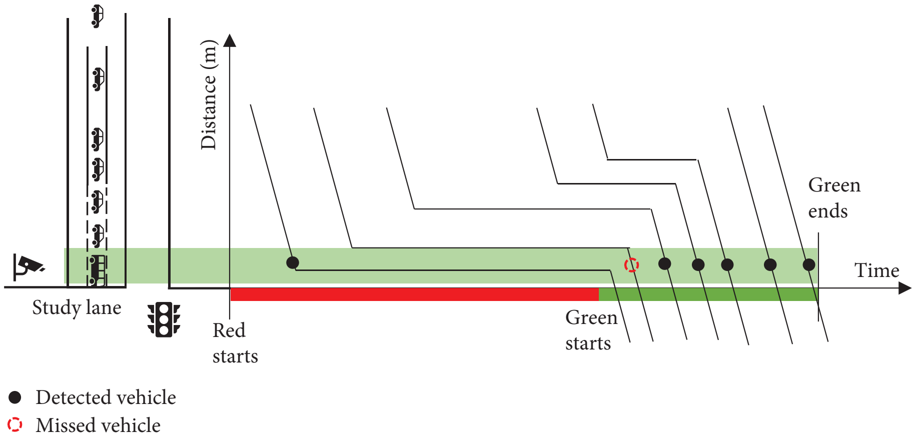


Figure 1: Detection process of LPR systems.

by traffic management sectors; thus, the corresponding signal timing data can easily be obtained. Therefore, in this study, the signal timing data were assumed to be known.

Using LPR and signal timing data, the departure time sequence (with the green phase start time as the origin) of each cycle can be obtained as $T^k = \{t_1^k, t_2^k, \ldots, t_{N^k}^k\}$, where N^k is the total number of detected vehicles during the cycle k. Then, the time headway of each vehicle can be further calculated as follows:

$$h_i^k = \begin{cases} h_s, & i = 1, \\ t_i^k - t_{i-1}^k, & i = 2, 3, \ldots, N^k, \end{cases} \quad (1)$$

where h_i^k is the time headway between the i^{th} vehicle and its leading vehicle, h_s is the saturated headway. Because the first vehicle in the cycle does not have a leading vehicle, its time headway is assumed to be h_s.

Therefore, for all vehicles detected during cycle k, the following two-dimensional dataset can be obtained:

$$\mathbf{D}^{\mathrm{k}} = \left\{\left(t_1^k, h_1^k\right), \left(t_2^k, h_2^k\right), \ldots, \left(t_{N^k}^k, h_{N^k}^k\right)\right\}. \quad (2)$$

As mentioned before, the first one or two vehicles may be detected during the red phase with a departure time less than 0. Thus, their departure times and time headways need to be corrected. Considering that vehicles detected during the red are definitely queued vehicles, if $t_i^k < 0$, we let $t_i^k = 0$ and $h_i^k = h_s$.

2.2. The Gaussian Mixture Model for the Queue Length Estimation. In this study, the queue length indicates the number of vehicles counted form the stop-line to the maximum back of queue during the entire cycle, i.e., the number of queued vehicles during the cycle.

Theoretically, queued vehicles at signalized intersections will depart with a saturated headway, while nonqueued vehicles depart with a greater time headway. However, LPR systems may have missed detections. In addition, because the detection resolution of LPR systems is 1 s, the time headway of queued vehicles is not always the saturated headway when departing the intersection. Therefore, it may introduce significant errors in classifying the queuing status of vehicles with a threshold of time headway.

Because vehicles in the same lane follow the first-in-first-out principle, queued vehicles naturally depart from the intersection before the nonqueued vehicles during the cycle. Consequently, the departure time of queued vehicles is mostly concentrated at the beginning of the green time, whereas that of nonqueued vehicles is always greater. In addition, as mentioned before, the time headway of queued and nonqueued vehicles is different, belonging to two different distributions. The former is concentrated near the saturated headway, whereas the latter is more widely distributed.

Given the above characteristics of the LPR data, this study assumes that queued and nonqueued vehicles belong to two different two-dimensional distributions of the time headway and departure time. Thus, a probabilistic clustering model, GMM, is adopted to identify the queuing status (queued or not) of vehicles based on LPR data. Therefore, the lane-based queue length can be estimated based on the last queued vehicle of each cycle. Note that, the reason we predefine 2 clusters in our method is not a purely data-driven decision, but rather a consideration of the significant differences between the LPR data characteristics of queued and nonqueued vehicles. Of course, from the data point of view, the distribution of nonqueued vehicles may be very scattered, and grouping them into one cluster may not be the optimal way, but our ultimate goal is to distinguish the queued vehicles and using 2 clusters can also serve our task very well. Besides, considering that the GMM method is essentially a classification method, the proposed method prefers to apply to cycles in the same time-of-day (TOD), in which the distribution of queued and nonqueued vehicles, especially the dimension of departure time, will not be too dispersed.

2.2.1. GMM Formulation. GMM, a probabilistic model, is a commonly used clustering algorithm. GMM assumes that the data are comprised of several Gaussian distributions, with independent means, variances, and weights. In this study, detected vehicles are assumed to follow two Gaussian distributions, indicating queued and nonqueued vehicles, respectively. For K cycles during the analysis period, the following dataset can be obtained:

$$\mathrm{X} = \left\{\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^{\mathrm{K}}\right\}. \tag{3}$$

The dataset can be rewritten into the vehicle-based form

$$\begin{aligned}\mathbf{X} &= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathbf{i}}, \ldots, \mathbf{x}_{\mathbf{N}}\} \\ &= \{(t_1, h_1), (t_2, h_2), \ldots, (t_i, h_i), \ldots, (t_N, h_N)\},\end{aligned} \tag{4}$$

where x_m is the data point of the i^{th} vehicle detected during the analysis period, $\mathbf{x}_{\mathbf{i}} = (t_i, h_i)$, and N is the total number of vehicles.

Assuming that data points in $\mathbf{X}$ follow two Gaussian distributions, the following mixture model can be used with two components to fit $\mathbf{X}$:

$$p(\mathbf{x}) = a_q \Phi\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{q}}, \sum\nolimits_{\mathbf{q}}\right) + a_{nq} \Phi\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{nq}}, \sum\nolimits_{\mathbf{nq}}\right), \tag{5}$$

where $\Phi(\cdot)$ is the two-dimensional Gaussian density function, and $\Phi(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{z}}, \sum_{\mathbf{z}}) = (2\pi|\sum_{\mathbf{z}}|^{0.5})^{-1} \exp(-0.5(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}})^T \sum_{\mathbf{z}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}}))$; $\mu_{\mathbf{z}}$ is the corresponding mean vector (1×2); $\sum_{\mathbf{z}}$ is the corresponding covariance matrix (2×2), $z \in \{q, nq\}$; the subscripts q and nq indicate the queued and nonqueued vehicles, respectively; a_z is the corresponding weighting coefficient, representing the probability that a randomly selected data point were generated by the component z, and $a_q + a_{nq} = 1$.

2.2.2. EM Algorithm. The GMM indicated by equation (5) is determined by six parameters, $\boldsymbol{\theta} = \left\{a_q, a_{nq}, \mu_{\mathbf{q}}, \mu_{\mathbf{nq}}, \sum_{\mathbf{q}}, \sum_{\mathbf{nq}}\right\}$. Generally, in frequentist probability theory, the maximum likelihood estimation, which seeks to maximize the probability of the observations given the model parameters, is used for model parameter estimation. However, a summation over two components in the log-likelihood function of the GMM is carried out, as follows:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log\left(a_q \Phi\left(\mathbf{x}_{\mathbf{i}} \mid \boldsymbol{\mu}_{\mathbf{q}}, \sum_{\mathbf{q}}\right) + a_{nq} \Phi\left(\mathbf{x}_{\mathbf{i}} \mid \boldsymbol{\mu}_{\mathbf{nq}}, \sum_{\mathbf{nq}}\right)\right). \tag{6}$$

The parameters in $\boldsymbol{\theta}$ cannot be solved analytically by setting the differential expression in equation (6) to zero. Therefore, the expectation-maximization (EM) algorithm, a numerical technique for maximum likelihood estimation, is used in this study to solve $\boldsymbol{\theta}$ in GMM.

The process of the EM algorithm for the GMM based on LPR data is described as follows:

Step 1. Initializing $\boldsymbol{\theta}$.

Two data points are randomly selected as initial means. The covariance matrix of the entire dataset is used as the initial covariance matrix for each component; the weighting coefficients remain unchanged, that is, 0.5, in every case.

Step 2. *E*-step.

Based on the current value of θ, for $\forall\, i$, the posterior probabilities of $\mathbf{x}_{\mathbf{i}}$ that belong to the queued and nonqueued vehicle distributions, respectively, are calculated.

$$\beta_z(\mathbf{x}_{\mathbf{i}}) = \frac{a_z \Phi(\mathbf{x}_{\mathbf{i}} | \boldsymbol{\mu}_{\mathbf{z}}, \sum_{\mathbf{z}})}{a_q \Phi\left(\mathbf{x}_{\mathbf{i}} | \boldsymbol{\mu}_{\mathbf{q}}, \sum_{\mathbf{q}}\right) + a_{nq} \Phi\left(\mathbf{x}_{\mathbf{i}} | \boldsymbol{\mu}_{\mathbf{nq}}, \sum_{\mathbf{nq}}\right)}, \quad z \in \{q, nq\}. \tag{7}$$

Step 3. *M*-step.

Given $\beta_z(\mathbf{x}_{\mathbf{i}})$ calculated in equation (7), the value of θ is updated based on the maximum likelihood estimation. The estimate of $\boldsymbol{\mu}_{\mathbf{z}}$ can be obtained by solving the following equation:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_{\mathbf{z}}} &= \sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}}) \left(2 \sum_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} - 2 \sum_{\mathbf{z}}^{-1} \mathbf{x}_{\mathbf{i}}\right), \\ &= 0.\end{aligned} \tag{8}$$

Thus, the following is obtained:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{z}} = \frac{\sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}}) \mathbf{x}_{\mathbf{i}}}{\sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}})}. \tag{9}$$

Similarly, $\sum_{\mathbf{z}}$ and a_z can be estimated

$$\begin{aligned}\widehat{\sum}_{\mathbf{z}} &= \frac{\sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}}) (\mathbf{x}_{\mathbf{i}} - \boldsymbol{\mu}_{\mathbf{z}})^T (\mathbf{x}_{\mathbf{i}} - \boldsymbol{\mu}_{\mathbf{z}})}{\sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}})}, \\ \widehat{a}_z &= \frac{\sum_{i=1}^{N} \beta_z(\mathbf{x}_{\mathbf{i}})}{N},\end{aligned} \tag{10}$$

where the subscript $z \in \{q, nq\}$.

Step 4. Steps 2 and 3 are repeated until $\ell(\boldsymbol{\theta})$ converges.

2.2.3. Queue Length Estimation. After $\boldsymbol{\theta}$ for the GMM model is obtained, given an LPR data point $\mathbf{x}_{\mathbf{i}} = (t_i, h_i)$ for arbitrary vehicle i, the probabilities that vehicle i belongs to the queued and nonqueued vehicle distribution are calculated using the posterior probabilities indicated by equation (7). If $\beta_q(\mathbf{x}_{\mathbf{i}}) > \beta_{nq}(\mathbf{x}_{\mathbf{i}})$, then vehicle i is identified as a queued vehicle; otherwise, it is a nonqueued vehicle.

Note that, in some cases, the identified queued and nonqueued vehicles may be noncontiguous, which violates the principle that queued vehicles must precede nonqueued vehicles under the first-in-first-out rule. Therefore, we need to correct the identified sequence with certain rules. Suppose that the label of queued vehicles is 1, and the label of nonqueued vehicles is 0. Then, for a sequence such as {1, . . ., 1, 0, 1, 0, . . ., 0}, we replace all the 1 s after the first 0 with 0. That is, we identify all vehicles after the first nonqueued vehicle as nonqueued vehicles.

Besides, as mentioned before, LPR systems may be susceptible to missed detections in empirical applications. If the number of queued vehicles identified by the proposed GMM are directly used as the queue length during the cycle, the queue length may be underestimated because the missed vehicles are not counted. Therefore, in this study, the departure time of the last identified queued vehicle during the cycle is used to estimate the queue length.

Given the results of the GMM, the two-dimensional Gaussian distribution of queued vehicles $\Phi(\mathbf{x} \mid \widehat{\boldsymbol{\mu}}_{\mathbf{q}}, \widehat{\sum}_{\mathbf{q}})$ is obtained. In the mean vector $\widehat{\boldsymbol{\mu}}_{\mathbf{q}} = (\widehat{t}_q, \widehat{h}_q)$, $\widehat{h}_q$ is the saturation headway for queued vehicles and $\widehat{t}_q$ is the expected departure time of queued vehicles. Assuming that the last queued vehicle identified by GMM during cycle k is $\mathbf{x}_{\mathbf{l}}^{\mathbf{k}} = (t_l^k, h_l^k)$, the queue length can be estimated as follows:

$$q^k = \lfloor \frac{t_l^k}{\widehat{h}_q} \rfloor, \tag{11}$$

where $\lfloor t_l^k/\widehat{h}_q \rfloor$ is the largest integer no more than $(t_l^k/\widehat{h}_q)$.

2.3. Real-Time Application. So far, we have described how to estimate the queue length of each cycle based on only one day of LPR data in an offline manner. Nevertheless, if historical LPR data are available, the proposed method can estimate the queue length in real time.

Traffic signal controls at intersections are commonly operated in a TOD mode in which the signal timing plans and traffic pattern, including the traffic demand level, arrival types, and queuing characteristics within each TOD are relatively consistent, although there are fluctuations between cycles. In addition, the daily traffic flow profiles of the same weekdays or weekends were comparable. Thus, in practical applications, historical LPR data during the same TOD on the same weekday can be used to calibrate the GMM and estimate the cycle-based queue length based on real-time LPR data.

A flow chart of the real-time application of the proposed method is shown in Figure 2. First, the GMM is calibrated with historical LPR data during the same TOD on the same weekday. When cycle k starts, the LPR system can detect vehicles in real time. Based on the calibrated GMM, the queuing status of each vehicle can be identified simultaneously, and when the yellow phase ends, the last queued vehicles during this cycle can be identified; eventually, the queue length of cycle k can be estimated.

2.4. The Extension to Multisection LPR Systems. Although the main goal of this study is to achieve the queue length estimation with single-section LPR systems, the proposed method can be easily extended to a scenario with multisection LPR systems.

A significant advantage of LPR data is that the license plate number of the vehicles is recorded. Thus, in a scenario with multisection LPR systems, the travel time between two sections of LPR systems of an individual vehicle can be easily extracted. Because the LPR systems are installed near the stop line, the extracted travel time consists of free-flow travel time and approach delay, as shown in Figure 3.

$$TT_i^k = FT_i^k + AD_i^k, \tag{12}$$

where TT_i^k is the travel time of vehicle i extracted from LPR systems, FT_i^k is the corresponding free-flow travel time, and AD_i^k is the corresponding approach delay.

Considering that the expected speed of different vehicles may vary, the free-flow travel time of vehicles should follow a specific distribution, for example, a Gaussian distribution. The approach delay of a queued vehicle depends on the spatial-temporal point when it joins the queue, while that of a nonqueued vehicle is 0. Thus, queued and nonqueued vehicles can be assumed to follow different travel time distributions. Therefore, for scenarios with multisection LPR systems, the travel time information of vehicles can be incorporated into the model to estimate queue length more accurately.

In particular, for vehicles with different origins (upstream straight through, left turn, or right turn), their free-flow travel time distributions are different owing to the variations in the travel distance. Nevertheless, the travel time distribution can be normalized by subtracting the minimum travel time of vehicles with different origins. Note that, the reason we use minimum travel time rather than the free-flow travel time (obtained by dividing the road length by maximum speed limit of the road) is that getting the minimum travel time is simple, convenient, and LPR data-driven, without a priori information such as road length and speed limit that cannot be obtained from LPR data.

$$NTT_i^{k,u} = TT_i^{k,u} - TT_{\min}^{k,u}, \tag{13}$$

where $NTT_i^{k,u}$ is the normalized travel time of vehicle i during cycle k, $TT_i^{k,u}$ is the corresponding travel time whose origin is u, u indicates the upstream direction, and $u \in \{s, l, r\}$; s, l, and r indicate upstream straight through, left turn, and right turn, respectively; $TT_{\min}^{k,u}$ is the minimum travel time of vehicles from u.

Thus, the two-dimensional GMM can be extended to a three-dimensional GMM using the input dataset as follows:

$$\begin{aligned} \mathbf{X} &= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathbf{i}, \ldots, \mathbf{x}_\mathbf{N}\}, \\ &= \{(t_1, h_1, NTT_1), \ldots, (t_i, h_i, NTT_i), \ldots, (t_N, h_N, NTT_N)\}. \end{aligned} \tag{14}$$

The rest of the estimation process was similar to that of the single-section case. In summary, the queue length estimation with multisection LPR systems can be easily achieved by extending the present GMM method.

3. Evaluation

The proposed method was evaluated based on both empirical and simulation data. In the empirical evaluation, the performance of the proposed method under different miss detection rates was evaluated, and an existing method proposed by Wu et al. [23], denoted as CPA, was replicated for the comparison. The CPA method works by transforming the queue length estimation method into a change-point identification problem using the headway sequence extracted from the LPR data. Then, the queue length of each cycle is estimated after identifying the change point using the CPA method. In the simulation evaluation, the proposed

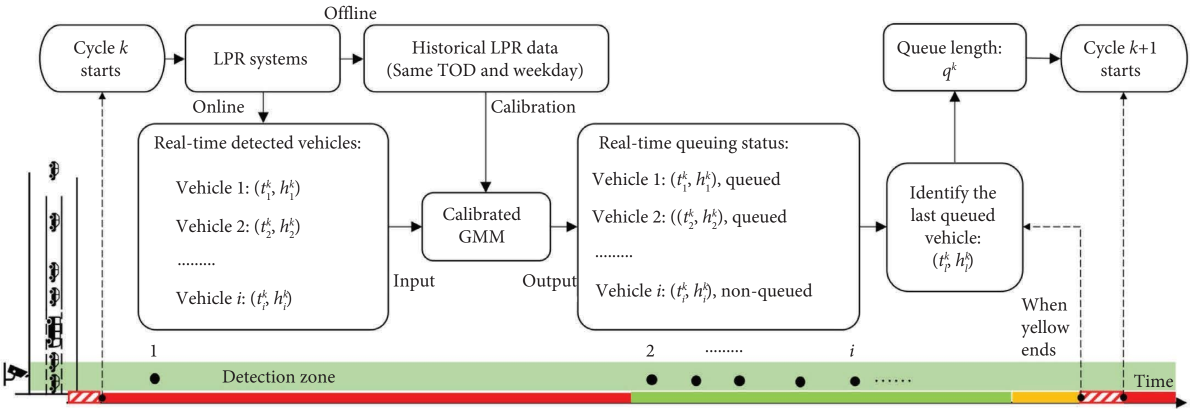

Offline
Cycle k starts
LPR systems
Historical LPR data (Same TOD and weekday)
Queue length: q^k
Cycle k+1 starts
Online
Calibration
Real-time detected vehicles:
Vehicle 1: (t_1^k, h_1^k)
Vehicle 2: (t_2^k, h_2^k)
....
Vehicle i: (t_i^k, h_i^k)
Calibrated GMM
Input
Output
Real-time queuing status:
Vehicle 1: (t_1^k, h_1^k), queued
Vehicle 2: ((t_2^k, h_2^k), queued
....
Vehicle i: (t_i^k, h_i^k), non-queued
Identify the last queued vehicle: (t_p^k, h_p^k)
When yellow ends
1
2
....
i
Detection zone
......
Time

Figure 2: The flow chart depicting the real-time application of the proposed method.

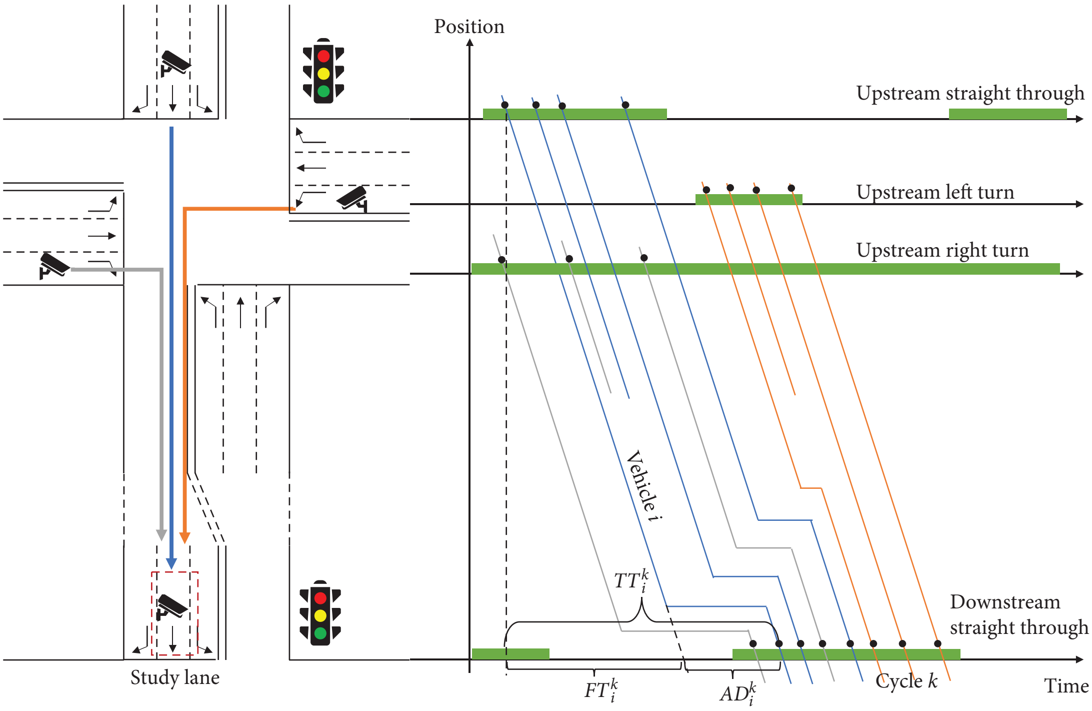


Figure 3: The illustration of vehicle detections by multisection LPR systems.

method was tested under different demand/capacities (d/c). Besides, the estimation with multisection LPR data is also evaluated.

The mean absolute error (MAE) and mean absolute percentage error (MAPE) are used to evaluate the accuracy of the queue length estimation, as given by the following:

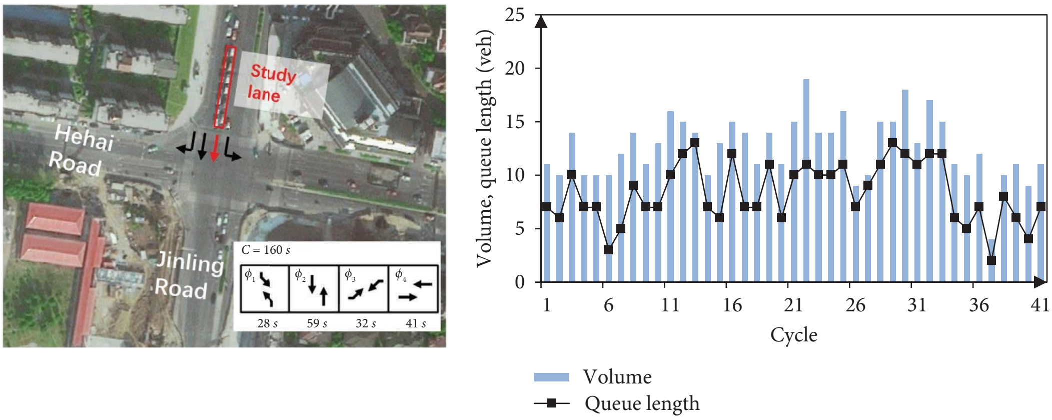
$$\mathrm{MAE} = \frac{1}{K}\sum_{k=1}^{K}\left|q^k - q_0^k\right|, \tag{15}$$

$$\mathrm{MAPE} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{q^k - q_0^k}{q_0^k}\right| \times 100\%, \tag{16}$$

where q^k is the estimated queue length of cycle k, q_0^k is the corresponding ground truth of the queue length, and K is the total number of cycles.

3.1. Empirical Evaluation

3.1.1. Study Site. The proposed method was evaluated using empirical data at the intersection of Jinling Road and Hehai Road in Changzhou City, Jiangsu Province, China. The study lane is the straight-through lane on the inner side of the north entrance, as shown in Figure 4(a). The LPR data were collected from 16:45 to 18:45 on September 7, 2020, which is a TOD of evening peak. The bus ratio was approximately 5%. The study period was operated using pretimed signal timing plans. The cycle length for the study period was 160 s, and the green phase for the study lane was approximately 60 s, which may have fluctuated in the real-world operation. In total, 41 cycles were included in the analysis period. The ground truth traffic conditions for the study lane were recorded by a field monitoring camera, from which the cycle-based traffic volume and queue length were manually



(a)

(b)

Figure 4: The study intersection for empirical cases. (a) Study lane. (b) Ground truth of the traffic volume and queue length.

extracted, as shown in Figure 4(b). Because the study period is an evening peak hour, the queue length and traffic fluctuated significantly. The average queue length was 8.4 veh, and the standard deviation was 2.8 veh.

3.1.2. Estimation Results. The preprocessed two-dimensional LPR data, indicating the departure time during the green phase and time headway for the analysis period, are shown in Figure 5(a). The distribution of the LPR data in the first and second halves of the green phase shows a completely different trend. The LPR data detected in the first half, most of which are queued vehicles, are more concentrated in terms of the time headway and mainly distributed in the range from 2 to 5 s. The LPR data detected in the second half, the majority of which are nonqueued vehicles, are widely distributed in both dimensions. Nevertheless, vehicles with time headway of 2–5 s are also observed during the period 40–60 s, which indicates that it is difficult to identify the queuing status of vehicles based on a time headway threshold only.

The identified queuing status of vehicles by GMM and the corresponding two-dimensional Gaussian distributions of queued and nonqueued vehicles are presented in Figure 5(b). As expected, most of the vehicles that departed in the period 0–35 s are identified as queued vehicles, except for vehicles with a time headways greater than 5 s. Although the time headway of certain vehicles is as small as that of queued vehicles, almost all vehicles that departed in the period 35–60 s are identified as nonqueued vehicles.

The parameters for the GMM are presented in Table 1. The average departure time of queued vehicles is 13.4 s after the green phase starts. As mentioned earlier, the saturation headway is the average time headway of queued vehicles, which is 2.6 s. The average departure time and time headway of nonqueued vehicles are 47.6 s and 9.8 s, respectively.

Based on the above GMM results, the queue length for each cycle was estimated according to equation (11). The lane-based queue length estimation results of the GMM and CPA methods are shown in Figure 6(a). Notably, the miss detection rate of the LPR data is zero in this case. Although the ground truth queue length fluctuates, ranging from 2 veh/cycle to 13 veh/cycle during the analysis period, both the CPA method and the proposed GMM method can capture the changing trend of the queue length. The CPA method tends to overestimate the queue length in most cases, whereas the proposed method can achieve more accurate estimates. The MAE of the GMM method is 1.29 veh/cycle, which is smaller than that of the CPA method with an MAE of 1.83 veh/cycle. The MAPE of the GMM method is 19.1%, which is also less than that of the CPA method with 23.9%. The error statistics are presented in Figure 6(b). The absolute errors of both methods show that the maximum estimation error of the CPA method reaches 5 veh at cycles 17, 21, and 22, whereas that of the GMM method is 3 veh. In addition, the absolute errors of 61.0% cycles of the GMM method did not exceed 1 veh/cycle, and 90.2% cycles did not exceed 2 veh/cycle, also outperforming the CPA method.

3.1.3. Sensitive Analysis of Miss Detection Rates. The LPR systems may be susceptible to missed detection of vehicles in the empirical application. Therefore, a certain percentage of LPR data was randomly deleted to simulate the missed detection scenarios in the real world; thus, testing the robustness of the proposed method under different miss detection rates. The miss detection rates were set in the range 0–20% with a 5% interval. Notably, a 0% miss detection rate is the scenario with the original LPR data that we already tested. The experiment for each miss detection rate was repeated five times with different random seeds to ensure the reliability of the results. The queue length estimation results of both methods under different miss detection rates are shown in Figure 7.

The results show that the estimation accuracies of both the GMM and CPA methods decrease as the miss detection rates increase. However, the decrease in the GMM method is much smaller than that in the CPA method. When the miss detection rate reached 20%, the MAE of the GMM method was just 2.01 veh/cycle, increasing only 0.72 veh/cycle compared to the error of the 0% miss detection rate. On the

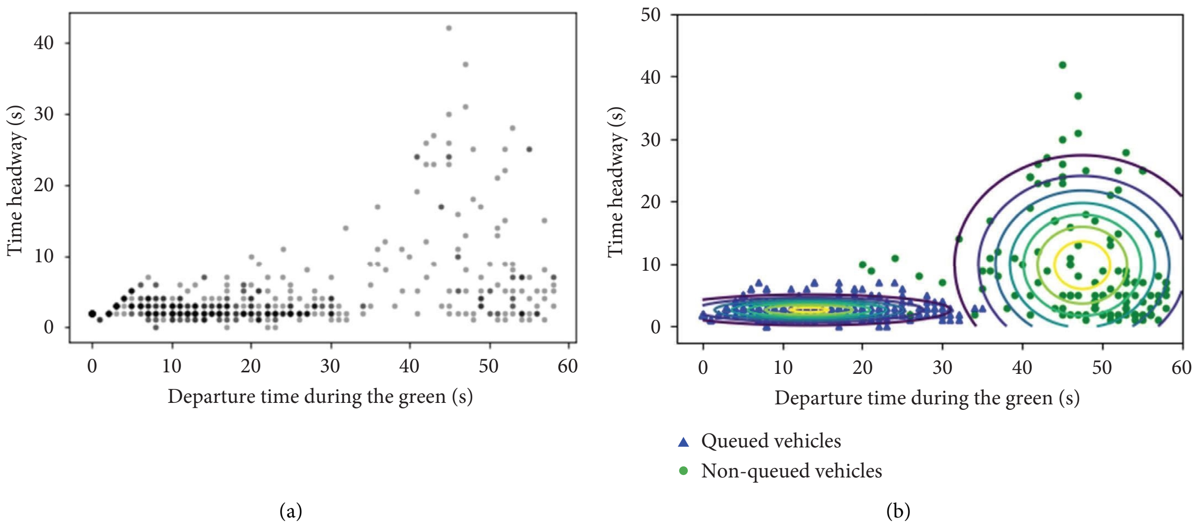


Figure 5: GMM results. (a) Vehicles detected by LPR systems. (b) Identified queuing status.

Table 1: Parameters of GMM.

Parameters	Queued vehicles ($z = q$)	Nonqueued vehicles ($z = nq$)
Total number	416	129
Mean vector ($\boldsymbol{\mu}_z$)	$\begin{bmatrix} 13.4 & 2.6 \end{bmatrix}$	$\begin{bmatrix} 47.6 & 9.8 \end{bmatrix}$
Covariance matrix ($\sum_z$)	$\begin{bmatrix} 79.0 & 0 \\ 0 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 62.5 & 0 \\ 0 & 75.8 \end{bmatrix}$
Weighting coefficient ($\boldsymbol{\alpha}_z$)	0.766	0.234

other hand, the MAE of the CPA method increased from 1.83 veh/cycle to 3.22 veh/cycle when the miss detection rate increased from 0% to 20%. The error bars indicate that the standard deviation of the experiments performed five times, and the results show that the proposed GMM method is much more stable than the CPA method with different random seeds of missed detections. In summary, the proposed GMM method is more robust than the CPA method against missed detections.

3.2. Simulation Evaluation

3.2.1. Simulation Model. The simulation model was built in VISSIM based on the intersection of the Hanjiang Road-Middle Tongjiang Road, Changzhou City, China, as shown in Figure 8(a), where the three straight-through lanes at the northbound approach were analyzed. The signal timing plan was provided by the local traffic management. The cycle length was 192 s, and the effective green time of the study lanes was 56 s. The simulation was run for 2 h (consisting of 38 cycles). The first four cycles (warm-up) and the last four cycles (for the integrity of the data) were removed from the analysis.

Four d/c (demand/capacity) values ranging from 0.4 to 1.0, with an interval of 0.2 were set in the simulation to test the performance of the proposed method under different d/c values. Figure 8(b) shows the cycle-based volume of the study lanes at four d/c values. The bus ratio in the study lanes was 2%. The ground truth cycle-based queue lengths were output by VISSIM queue detectors, which only measured the maximum queue length of the study lanes. Therefore, in this case, the final estimates of the proposed method are the maximum of the lane-based estimates of the three lanes.

3.2.2. The Estimation under Different d/c Values. The overall accuracy of the proposed method under different d/c values is shown in Figure 9. With an increase in the d/c value, the MAE increased slightly, from 1.1 veh/cycle at a d/c value of 0.4 to 1.9 veh/cycle at a d/c value of 1.0. Nevertheless, the MAPE decreased owing to the greater ground truth queue length at higher d/c values. The MAPE was 9.6% at a d/c of 0.4, and decreased to 8.0% at a d/c of 1.0.

The cycle-based estimates for different d/c values are shown in Figure 10. As shown, the overall trend of the estimates of the proposed method is consistent with that of the ground truth. In most cycles, the absolute errors of the estimates were no more than 3 veh. Notably, when the d/c value is 1.0, which is a saturated traffic condition, there are two cycles with absolute errors of seven veh and 11 veh, respectively. This is because these two cycles are oversaturated, that is, the actual queue length exceeds the capacity of the green phase, and the estimates of the proposed method are constrained by the capacity of the green phase.

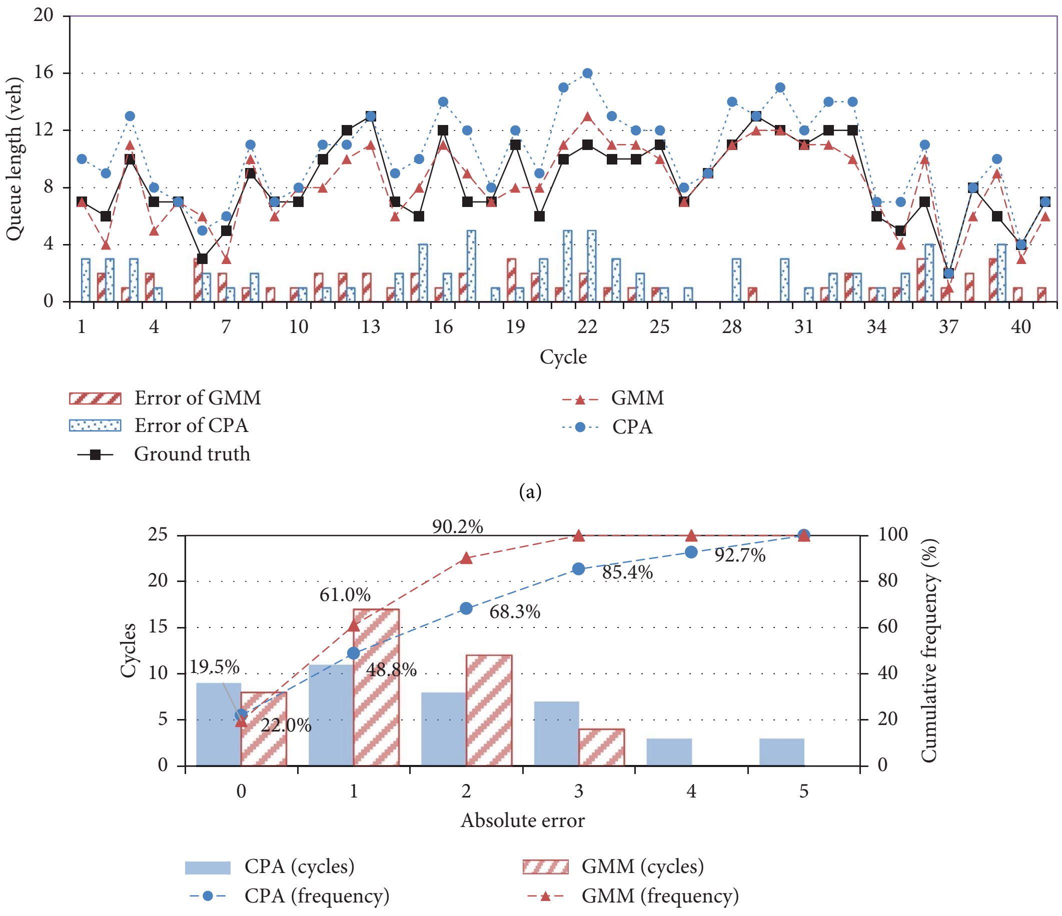


Figure 6: Lane-based queue length estimation results for cycles. (a) Cycle-based estimates. (b) Error statistics.

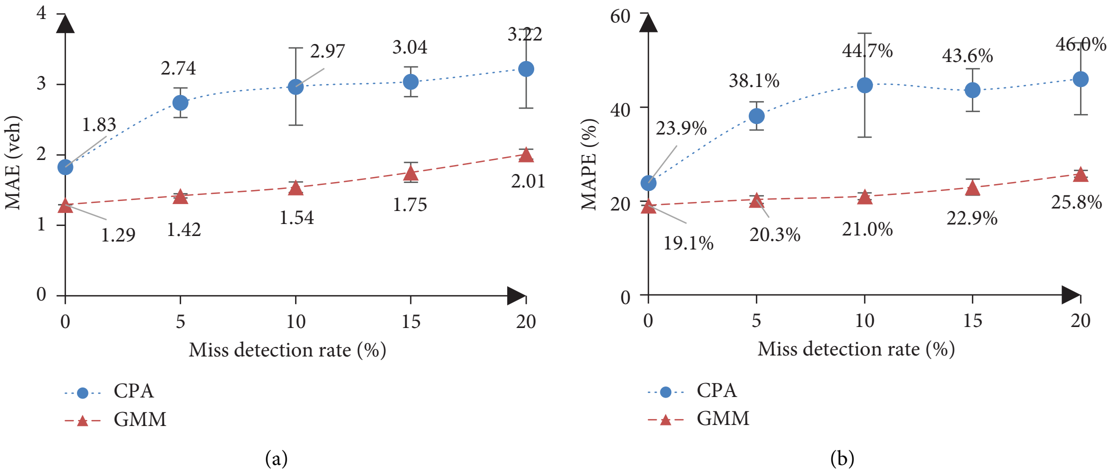
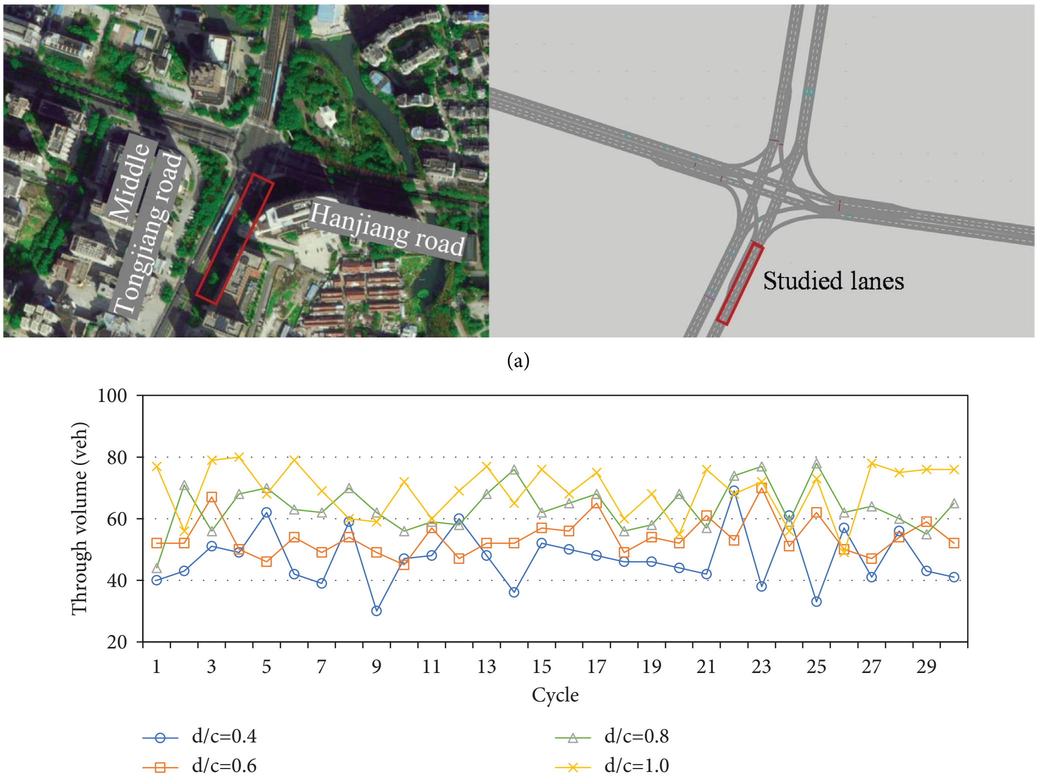


Figure 7: Estimation results under different miss detection rates. (a) MAE. (b) MAPE.

For example, in cycle 29, the maximum number of vehicles detected by LPR systems of three lanes is 27 veh, which is even smaller than the ground truth maximum queue length output by VISSIM, that is, 33 veh. In such cases, the proposed method naturally underestimated it.

3.2.3. The Estimation with Multisection LPR Data. To test the effect of the proposed method in the scenario with multisection LPR data, we added an upstream intersection at the test link. The input setting is the same as the case with $d/c = 0.6$. The GMM results with multisection LPR data are



(a)

(b)

Figure 8: Study lanes for simulation cases. (a) Studied intersection in simulation. (b) Through volumes of cycles.

presented in Figure 11. It can be seen that, although there are obvious differences in travel time between queued vehicles and nonqueued vehicles, their boundaries are blurred, which is similar to the time headway and departure time. This means that it is difficult to identify the queuing status of vehicles based on single information provided LPR data, e.g., travel time or time headway. However, after considering the distribution of three different dimensional information, that is, using as much effective information extracted from the LPR data as possible, the queuing state of the vehicles detected by the LPR systems can be better identified.

Table 2 presented the overall accuracy of the proposed method using different information combination extracted from multisections LPR data. Compared to the Model #2 using the single-section LPR data only (departure time and time headway information), Model #1 additionally used the travel time information provided by the multisection LPR data. Model #3 and #4 are other possible information combinations for multisection LPR data. As is shown, after incorporating travel time information into the GMM model, the accuracy of Model #1 performed best among all models. Comparing Models #2, #3, and #4, we found that the proposed Model #2, i.e., the two-dimensional model based on single-section LPR data, outperforms the other two possible combinations using travel time information. This justifies the necessity of using three types of information (i.e., departure time, time headway, and travel time) in the scenarios with multisection LPR data.

3.2.4. The Estimation with LPR Data in Different TODs.

As we aforementioned, the proposed method prefers to apply to cycles in the same TOD. To test the performance of our method with LPR data in different TODs with different d/c, we mixed the data between 0.4 and 0.8 d/c in the simulation for GMM calibration and estimated the lane-based queue length of two d/c, respectively. As is shown in Table 3, compared to the estimation results with GMM calibrated based on LPR data in a single TOD (denoted by single TOD), the MAE and MAPE of two TODs both increased with GMM calibrated based on LPR data in two TODs (denoted by mixed TOD). That is to say, mixing multiple TOD data for GMM calibration will weaken the performance of the method, albeit slightly, due to the differences in data distribution within different TODs. In practice, it is still recommended to use data in the same TOD for GMM calibration and queue length estimation.

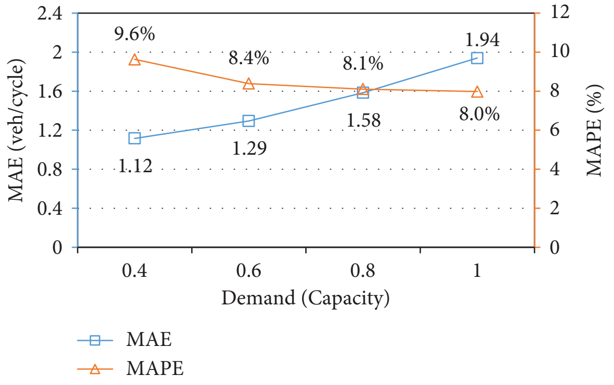
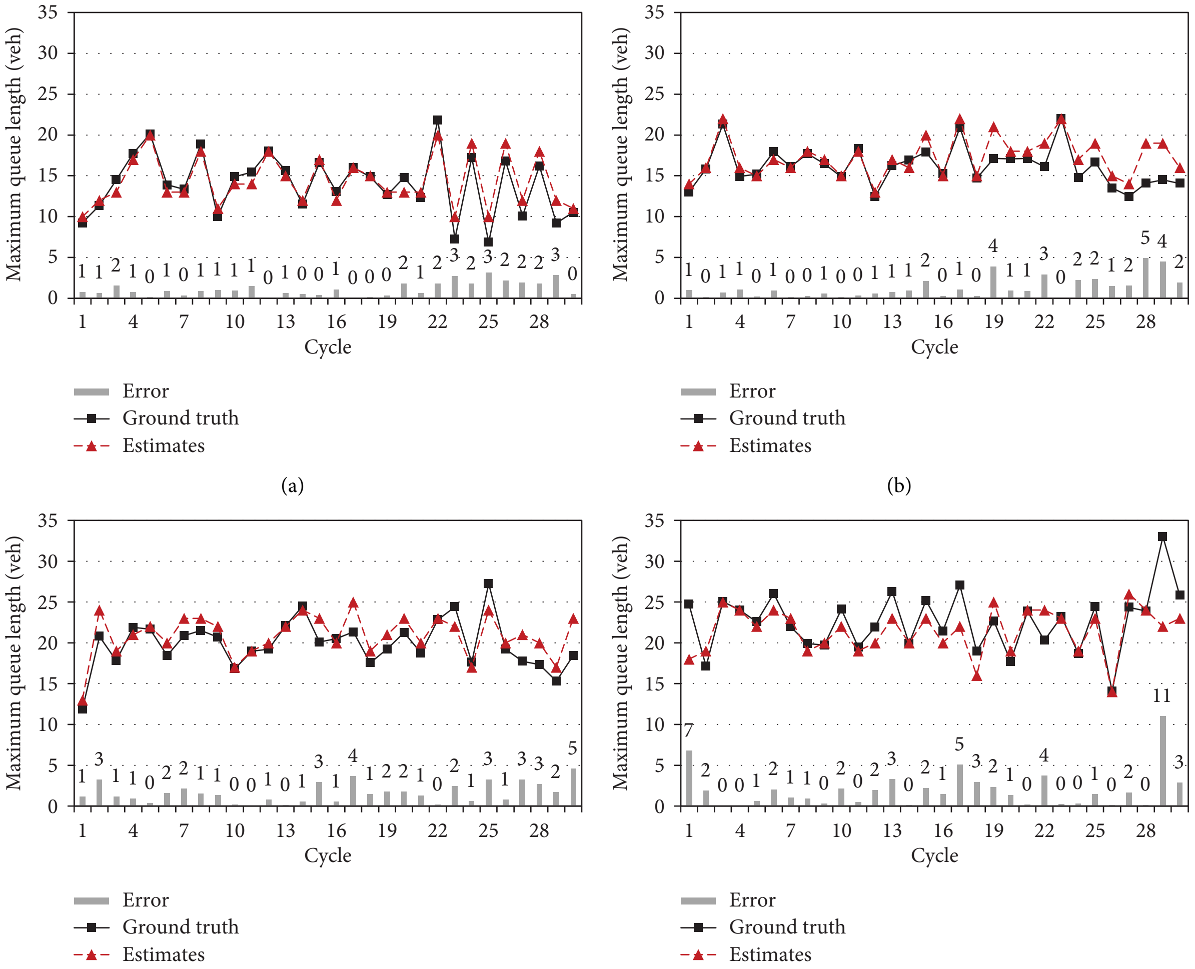


Figure 9: The overall estimation accuracy under different d/c values.



(a)

(b)

(c)

(d)

Figure 10: Cycle-based estimates under different d/c values. (a) $d/c = 0.4$. (b) $d/c = 0.6$. (c) $d/c = 0.8$. (d) $d/c = 1.0$.

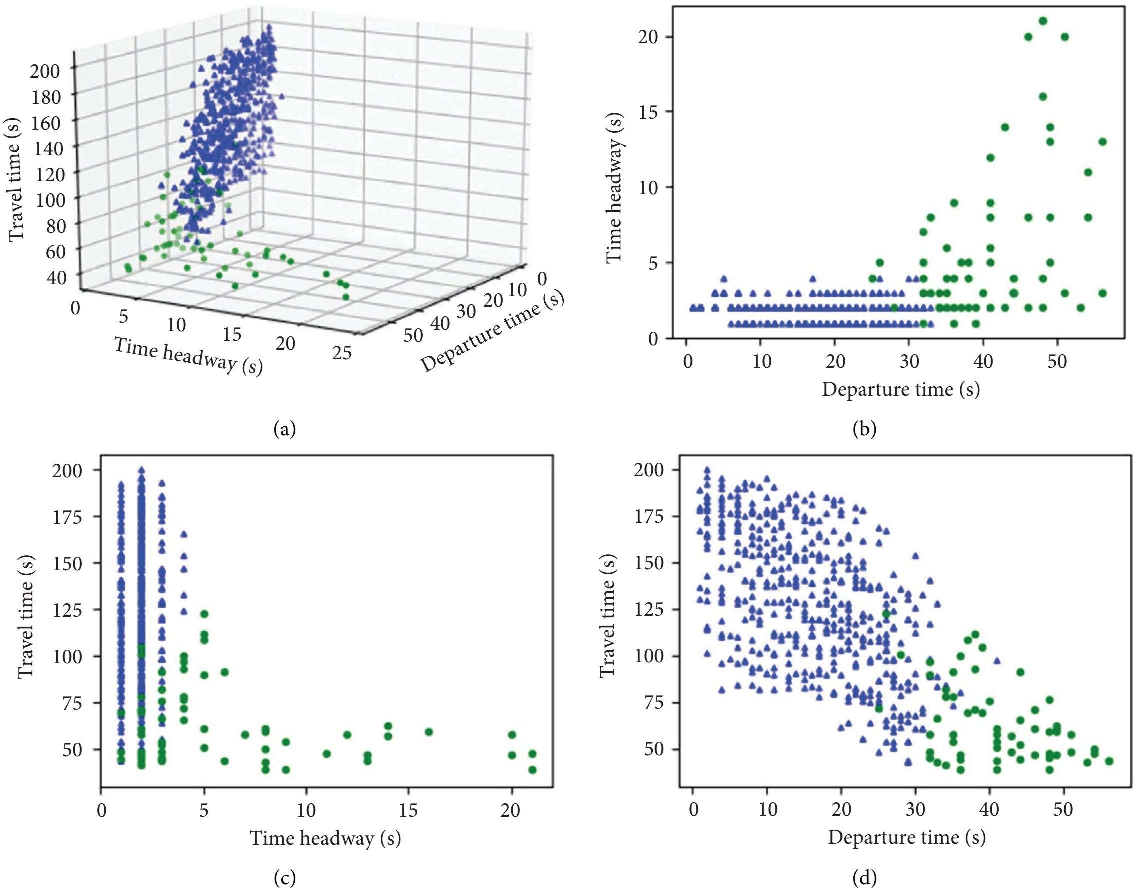


Figure 11: GMM results with multisection LPR data.

Table 2: Estimation results under different information combination.

#	Information used*	MAE (veh)	MAPE (%)
1	*D*, *H*, *T* (multisection)	1.09	6.7
2	*D*, *H* (single-section)	1.29	8.4
3	*H*, *T*	2.26	15.3
4	*D*, *T*	4.10	25.4

*D, departure time; H, time headway; T, travel time.

Table 3: Estimation results with LPR data in different TOD.

TOD with different d/c	MAE		MAPE (%)	
	Single TOD	Mixed TOD	Single TOD	Mixed TOD
0.4	1.12	1.59	9.6	14.3
0.8	1.58	1.62	8.1	8.1
Average	1.35	1.61	8.9	11.2

4. Conclusions and Future Work

Using single-section LPR data as input, this study proposes a two-dimensional GMM method to estimate the lane-based queue length cycle by cycle. Single-section LPR data can only provide the departure information of detected vehicles, including the time headway and departure time during the green phase. The proposed method fully exploits the information provided by the LPR data and transforms the queue length estimation problem into a data clustering problem. Then, the vehicle queuing status is identified by the GMM method, and the lane-based queue length is estimated using the departure time of the last identified queued vehicle during the cycle. In particular, the model was also extended to real-time applications and the scenario with multisection LPR systems, which prepares the method for practical application in various scenarios.

The proposed method is evaluated based on both empirical and simulation data to demonstrate its advantages. The empirical results show that the proposed GMM method shows higher accuracy and stability than the existing method at different miss detection rates. During the two-hour analysis period, the MAE of the GMM method was only 1.29 veh/cycle when there were no missed detections. Even with a 20% miss detection rate, the MAE of the GMM method was only 2.01 veh/cycle. The simulation results demonstrate that the GMM method can achieve accurate estimates under different levels of the traffic demand, and the estimation accuracy can be further improved after considering travel time information in multisection scenarios.

A number of research avenues could be explored in the future: (1) Presently, constrained by the departure information of detected vehicles provided by single-section LPR data, only the queue length estimation for undersaturated conditions was carefully investigated. Future work can focus on the queue length estimation for oversaturated conditions by fusing upstream fixed detectors or using multisection LPR data. (2) CV data are another emerging data source for the dynamic evaluation of traffic signals. Even though the current penetration rate is low, the CVs can still provide a portion of the arrival information of traffic flows. Therefore, future work can fuse CV data with LPR data to achieve better queue length performance. (3) Although such information was not considered in this study, LPR systems can also record vehicle types. Future work can explore ways to improve the accuracy of the estimation by considering vehicle types.

Data Availability

The license plate recognition data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Authors' Contributions

The study was conceptualized and designed by Chaofeng Tan, Keshuang Tang, and Chaopeng Tan. The data were collected by Chaofeng Tan, Hao Wu, and Chaopeng Tan. Analysis and interpretation of results were carried out by Chaofeng Tan and Chaopeng Tan. Draft manuscript was prepared by Chaofeng Tan, Hao Wu, Keshuang Tang, and Chaopeng Tan. All the authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

This research was supported by the Shanghai Municipal Science and Technology Major Project and the Fundamental Research Funds for the Central Universities (grant number: 2021SHZDZX0100), the Youth Talent Project of University Scientific Research Project of Guangzhou Education Bureau (grant numbers: J20210210/202032836), the Higher Education Teaching Reform Project of Guangdong Province Education Department (grant numbers: C2206001112/C2106001257), the Young Innovative Talents Project of Guangdong General Universities (grant number: 2021KQNCX073), and the Guangzhou Education Science Planning 2021 (grant number: 202113690).

References

[1] Transportation Research Board, *Highway Capacity Manual 2010*, p. 1207, 5th edition, Transportation Research Board, National Research Council, Washington, DC, USA, 2010.

[2] Y. Lu and G. L. Chang, "Stochastic model for estimation of time-varying arterial travel time and its variability with only link detector data," *Transportation Research Record*, vol. 2283, no. 1, pp. 44–56, 2012.

[3] S. Tolami, B. Mehran, and B. Hellinga, "Delay and queue length estimation at signalized intersections using archived automatic vehicle location and passenger count data from transit vehicles," in *Proceedings of the Presented at the the 94th Annual Meeting of Transportation Research Board*, Washington, DC, USA, January 2015.

[4] 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*, vol. 17, no. 4, pp. 412–427, 2009.

[5] 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*, vol. 2035, no. 1, pp. 69–80, 2007.

[6] A. Skabardonis and N. Geroliminis, "Real-time monitoring and control on signalized arterials," *Journal of Intelligent Transportation Systems*, vol. 12, no. 2, pp. 64–74, 2008.

[7] G. Vigos, M. Papageorgiou, and Y. Wang, "Real-time estimation of vehicle-count within signalized links," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 1, pp. 18–35, 2008.

[8] X. Jeff 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*, vol. 19, no. 6, pp. 1133–1156, 2011.

[9] 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*, vol. 16, no. 1, pp. 12–23, 2012.

[10] C. Tan, J. Yao, X. Ban, and K. Tang, "Cumulative flow diagram estimation and prediction based on sampled vehicle trajectories at signalized intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11325–11337, 2022.

[11] K. Tang, C. Tan, Y. Cao, J. Yao, and J. Sun, "A tensor decomposition method for cycle-based traffic volume estimation using sampled vehicle trajectories," *Transportation Research Part C: Emerging Technologies*, vol. 118, Article ID 102739, 2020.

[12] C. Tan, J. Yao, X. Ban, and K. Tang, "Joint estimation of multiphase traffic demands at signalized intersections based on connected vehicle trajectories," 2022, https://arxiv.org/abs/2210.10516.

[13] G. Comert, "Queue length estimation from probe vehicles at isolated intersections: estimators for primary parameters," *European Journal of Operational Research*, vol. 252, no. 2, pp. 502–521, 2016.

[14] G. Comert and M. Cetin, "Queue length estimation from probe vehicle location and the impacts of sample size," *European Journal of Operational Research*, vol. 197, no. 1, pp. 196–202, 2009.

[15] F. Li, K. Tang, J. Yao, and K. Li, "Real-time queue length estimation for signalized intersections using vehicle trajectory data," *Transportation Research Record*, vol. 2623, no. 1, pp. 49–59, 2017.

[16] Y. Mei, W. Gu, E. C. Chung, F. Li, and K. Tang, "A Bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 109, pp. 233–249, 2019.

[17] M. Ramezani and N. Geroliminis, "Queue profile estimation in congested urban networks with probe data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 6, pp. 414–432, 2015.

[18] 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*, vol. 22, no. 1, pp. 91–106, 2021.

[19] 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*, vol. 2672, no. 45, pp. 253–264, 2018.

[20] C. Tan, L. Liu, H. Wu, Y. Cao, and K. Tang, "Fuzing license plate recognition data and vehicle trajectory data for lane-based queue length estimation at signalized intersections," *Journal of Intelligent Transportation Systems*, vol. 24, no. 5, pp. 449–466, 2020.

[21] 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*, vol. 57, pp. 85–102, 2015.

[22] X. Luo, D. Ma, S. Jin, Y. Gong, and D. Wang, "Queue length estimation for signalized intersections using license plate recognition data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 3, pp. 209–220, 2019.

[23] H. Wu, J. Yao, L. Liu, Y. Cao, and K. Tang, "Left-turn spillback identification based on License Plate Recognition data," in *Proceedings of the Presented at the the 98th Annual Meeting of Transportation Research Board*, Washington, DC, USA, January 2019.

[24] K. Tang, H. Wu, J. Yao, C. Tan, and Y. Ji, "Lane-based queue length estimation at signalized intersections using single-section license plate recognition data," *Transportmetrica B: Transport Dynamics*, vol. 10, no. 1, pp. 293–311, 2022.

[25] X. Zhan, R. Li, and S. V. Ukkusuri, "Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102660, 2020.