Transportation Research Part C 137 (2022) 103596

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Estimating key traffic state parameters through parsimonious spatial queue models

Qixiu Cheng [a,b], Zhiyuan Liu [a,*], Jifu Guo [c], Xin Wu [d], Ram Pendyala [d], Baloka Belezamo [d,e], Xuesong (Simon) Zhou [d,*]

[a] School of Transportation, Southeast University, Nanjing, China
[b] Department of Logistics & Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[c] Beijing Transport Institute, Beijing, China
[d] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, United States
[e] Arizona Department of Transportation, United States

ARTICLE INFO

Keywords:
Traffic state estimation
Performance evaluation
Deterministic queueing model
Polynomial arrival queue
Fluid approximation
Traffic bottlenecks

ABSTRACT

As an active performance evaluation method, the fluid-based queueing model plays an important role in traffic flow modeling and traffic state estimation problems. A critical challenge in the application of traffic state estimation is how to utilize heterogeneous data sources in identifying key interpretable model parameters of freeway bottlenecks, such as queue discharge rates, system-level bottleneck-oriented arrival rates, and congestion duration. Inspired by Newell's deterministic fluid approximation model, this paper proposes a spatial queue model for over-saturated traffic systems with time-dependent arrival rates. The oversaturated system dynamics can be described by parsimonious analytical formulations based on polynomial functional approximation for virtual arrival flow rates. With available flow, density and end-to-end travel time data along traffic bottlenecks, the proposed modeling framework for estimating the key traffic queueing state parameters is able to systematically map various measurements to the bottleneck-level dynamics and queue evolution process. The effectiveness of the developed method is demonstrated based on three case studies with empirical data in different metropolitan areas, including New York, Los Angeles, and Beijing.

1. Introduction

Many regional planning organizations and transportation management authorities throughout the world face enormous challenges to mitigate heavy traffic congestion to enable a high level of services for citizens. For example, according to the 2019 Global Traffic Scorecard Report released by INRIX (2020), the traffic congestion costs Americans approximately $1,377 per driver or $88 billion in total in 2019. Estimating accurately traffic states is of great importance for traffic control and management to reduce traffic jams (Seo et al., 2017). A critical challenge in the application of traffic state estimation is how to perform congestion and bottleneck identification (CBI) (Hale et al., 2016, 2021). A number of analytical tools, including CBI tools (FHWA, 2018), are developed to analyze, visualize, and further compare traffic bottlenecks in great details. Given widely available traffic measurements, for example, from the

* Corresponding authors.
E-mail addresses: qixiu.cheng@polyu.edu.hk (Q. Cheng), zhiyuanl@seu.edu.cn (Z. Liu), guojf@bjtrc.org.cn (J. Guo), xinwu3@asu.edu (X. Wu), ram.pendyala@asu.edu (R. Pendyala), bbelezamo@adot.gov (B. Belezamo), xzhou74@asu.edu (X.(S. Zhou).

https://doi.org/10.1016/j.trc.2022.103596
Received 4 February 2021; Received in revised form 21 December 2021; Accepted 1 February 2022
Available online 14 February 2022

Regional Integrated Transportation Information System (RITIS) hosted by the University of Maryland (CATT Lab, 2021), many new approaches are developed for real-time large-scale system modeling and diverse data synthesis. By monitoring the state of the traffic system at all times, it would be possible to apply proactive traffic control actions in real-time to best utilize available unused road capacity.

Generally, traffic state estimation approaches can be divided into categories, i.e., the model-driven approach and the data-driven approach. Although the data-driven approach enables an automated process of estimating a large number of model parameters, the estimation process is data driven and might be built based on complex or black-box models without interpretability. Without embedding a traffic-flow-oriented model structure, a purely data-driven model might fail to systematically account for spatial and temporal interaction in traffic systems. In addition, parsimonious analytical models are critically needed to strive for least complex explanation for (imperfect) observations, as well as offer real-time computational efficiency for large-scale real-time network applications. Therefore, this paper aims to propose a theoretically rigorous spatial queue model to analytically estimate a number of critical traffic state parameters (e.g., queue discharge rates, system-level bottleneck-oriented arrival rates, congestion duration, and time-dependent delays and travel times) and capture system dynamics at bottlenecks with queue evolution processes.

1.1. Literature review on queueing models with time-dependent arrival rates

Before presenting the proposed model, we will first review the queueing-theoretic model for oversaturated traffic systems. In generally, congestion exists in traffic and transportation systems when the demand temporarily and spatially exceeds the supply. Vickrey (1969) developed the bottleneck model to describe traffic dynamics during rush hours. The focus of many following bottleneck-related studies (e.g., Arnott et al., 1990) is on the network equilibrium and optimal toll problem, and the arrival rate (or inflow rate) in their bottleneck model is assumed to be a step function, which may be inconsistent with empirical observations with complex nonlinear patterns. To analyze the queue evolution process, Newell (1968a, 1968b, 1968c, 1982) analytically investigated the fluid-based queues in a traffic system with time-dependent arrival rates by linear or quadratic functions. It should be remarked that, linear time-dependent arrival rates cannot fully capture nonlinear system dynamics. As for the formula using quadratic arrival rate assumptions, Newell (1968c) also raised an important modeling issue of possible negative flow rates (which is clearly contrary to realism), especially for heavily congested conditions as the arrival rates could decrease sharply on the right-hand-side of the curve. As a result, the analytical functional forms provided by Newell should be applied carefully. Other attempts along the direction include pointwise stationary approximation functions (Green et al., 2007) for service networks with $M_t/G/\infty$ queueing systems. In this study, we will extend Newell's classic fluid-based queue model and propose a family of polynomial-function-approximated arrival rates (which can be used for over-congested traffic systems) to estimate the queue profile for the system performance evaluation, including the time-dependent queue lengths and delays, and the average delays and travel times.

1.2. Numerical methods for describing queueing characteristics

With discretized time and space dimensions (such as the cell transmission model and link transmission model), dynamic traffic assignment models also need to address many computational challenges due to the introduced finer resolution. The bottleneck model has been used to investigate the dynamic traffic assignment problem (e.g., Drissi-Kaïtouni and Hameda-Benchekroun, 1992; Kuwahara and Akamatsu, 1997; Li et al., 2000). Nie and Zhang (2005) compared the bottleneck model with three other discrete link models (including the Merchant and Nemhauser (1978a, b) model, the delay function model, and the cell transmission model) in the dynamic network loading process. Ban et al. (2012) formulated the bottleneck model as a continuous-time model with differential complementarity systems. Han et al. (2013a, b) extended Vickrey's bottleneck model to a generalized Vickrey model, which allows the inflow rate to be a distribution. Jin (2015) proposed a unified approach to study the bottleneck model. However, all these bottleneck models are based on virtual queue length. In order to obtain the spatial quantities, Lawson et al. (1997) proposed an input-output diagram approach to calculate the spatial queue length and distance, and we will also use their input-output diagram approach to transfer the virtual queue length to spatial queue distance to calibrate the proposed model in this paper.

Other standard traffic modeling tools include partial-differential-equation (PDE)-based numerical analysis approaches and customized simulation packages to capture microscopic interactions between the demand and supply (Behrisch et al., 2011; Marshall, 2018) are at a very fine resolution. In particular, mesoscopic models (Mahmassani and Herman, 1984; Zhou and Taylor, 2014) are used to represent spatial extents of congestion building up and dissipating with individual agents following a macroscopic flow density relationship while ignoring detailed lane-changing and car-following behavior. From a macroscopic aggregated traffic flow modeling perspective, some studies (e.g., Ramezani et al., 2015; Han et al., 2020; Johari et al., 2021) used the macroscopic fundamental diagram to model and control traffic flows through ramp metering, signal control, and perimeter control, etc.

1.3. Objectives and potential contributions

In this paper, we propose a spatial queueing-theoretic PDE model based on a polynomial functional approximation for virtual arrival rates at bottlenecks. This simplified PDE model aims to answer the question that how to analytically estimate the queue profile and capture system dynamics (such as the time-dependent queue length, delay, and travel time, etc.) at bottlenecks with queue evolution processes, and provide a building block for traffic state estimation with heterogeneous real-world data sources. We would like to illustrate the contributions of this work: (1) The vehicular space-time trajectories during congestion are mapped to a set of dynamical queueing system equations with a family of polynomial-approximated time-dependent arrival rates. (2) A number of system

performance evaluation measures, including the time-dependent queue length, delay, and travel time, as well as the average delay and link travel time, are analytically derived. (3) For heavy congestion cases, we explicitly define the oversaturation ratio (or the queue building-up ratio) to analytically derive the queue profile with cubic arrival rate functions. (4) With some key interpretable model parameters of freeway bottlenecks, such as queue discharge rates, system-level bottleneck-oriented arrival rates, and congestion duration, the traffic state can be easily estimated based on our proposed modeling framework with heterogeneous real-world data sources.

The remainder of this paper is organized as follows: Section 2 describes the core problem with a general fluid queueing model and introduces the deterministic queueing theory by a set of dynamical system equations. The queueing-theoretic model with approximated time-dependent arrival rates and constant approximated discharge rates is formulated in Section 3, followed by the calibration methods and results in Section 4 and 5, respectively. Finally, discussions and conclusions are conducted in Sections 6 and 7, respectively.

2. Background on fluid queueing model

Consider a dynamic system with a bottleneck restricting the passing of moving vehicles; a queue forms upstream of the bottleneck when the arrival rate temporarily and spatially exceeds the discharge rate (or capacity) μ. Table 1 summarizes the notations and explanations used in this paper.

Fig. 1 illustrates the deterministic queueing model with the virtual queue evolution process: Fig. 1(a) illustrates the vehicle trajectories in the time-space plane across the queue extent along a single bottleneck. The blue curve depicts the physical queue extent, while the green line is the trajectory of a typical vehicle with a free-flow speed v_f and a speed-at-capacity v_μ. Fig. 1(b) illustrates the multisource data that can be used to calibrate the bottleneck model, including the loop detector data, the probe vehicle data, etc. Fig. 1 (c) describes the time-dependent arrival rates for the oversaturated traffic systems. The red horizontal line is a constant discharge rate μ, and the blue curve is a time-dependent virtual arrival rate function $\lambda(t)$ at the bottleneck. It is obvious that $\lambda(t_0) = \lambda(t_2) = \mu$. Because the queue dissipates at t_3, the yellow area before t_2 should be equal to the green area after t_2. Fig. 1(d) draws the time-dependent queue evolution process. The blue curve is the queue length evolution process, with a maximal queue length at time t_2 and zero queue length at times t_0 and t_3. Fig. 1(e) illustrates the cumulative counts in the system. The red line $D(t)$ with a slope of μ is the cumulative departure curve during the peak period, and the blue curve $A(t)$ is the cumulative arrival curve. Before time t_0, there is no congestion; thus, the cumulative arrival count equals the cumulative departure count, and during the peak period from t_0 to t_3, the cumulative arrival count is larger than the cumulative departure count due to the congestion effect. The vertical difference between the cumulative arrival and departure curves at time t is the queue length $Q(t)$, and due to the constant slope of the cumulative departure curve, we can easily obtain its corresponding delay $w(t)$. The model introduced here can be summarized as the *TULIP model*, where *T* represents the time-space network, *U* represents the arrival rate $\lambda(t)$ and discharge rate μ, *L* represents the queue length, and *IP* stands for the input-output diagram or cumulative arrival and departure curves.

The fluid approximated dynamic system illustrated in Fig. 1 can be formulated by a set of dynamical system equations:

Table 1
Notations and explanations used in this paper.

Notations	Explanations
t_0	start time of congestion period
t_1	time with maximum arrival rate
t_2	time with maximum queue length
t_3	end time of congestion period
$\bar{t}$	another root besides for t_0 and t_2 of the cubic net flow rate function
m	scale parameter, $m = (t_2 - t_0)/(t_3 - t_0)$
t_f	free flow travel time
v_f	free flow speed
v_μ	speed at capacity
γ	shape parameter for the cubic arrival rate function
μ	capacity (or discharge rate), assumed to be a constant value
D	total demand during the whole peak period
P	congestion period, $P = t_3 - t_0$
$\lambda(t)$	arrival rate function at time t
$Q(t)$	virtual queue length at time t
$Q^p(t)$	physical queue length at time t
$w(t)$	traffic delay departing at time t
w	average delay during the whole peak period
tt	average travel time during the whole peak period
$A(t)$	cumulative arrival curve at time t
$D(t)$	cumulative departure curve at time t

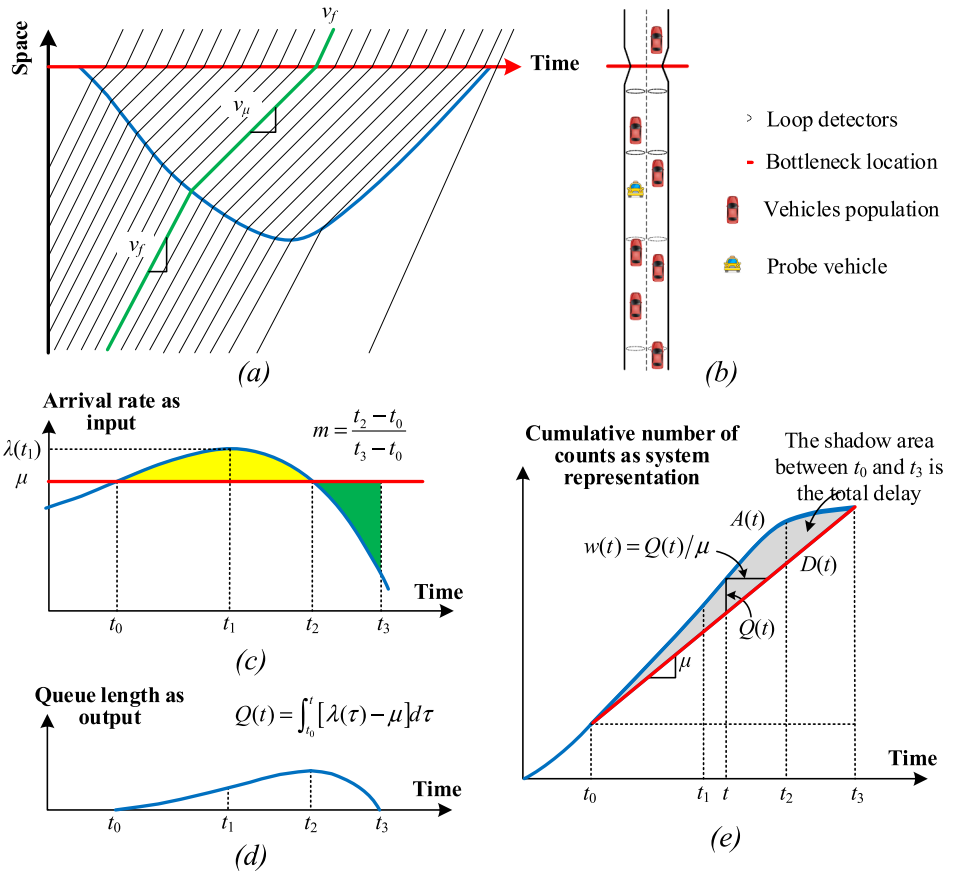


Fig. 1. Model illustration based on the fluid approximation model and spatial queue model.

$$\frac{dA(t)}{dt}=\lambda(t) \tag{1}$$

$$\frac{dD(t)}{dt}=\mu(t) \tag{2}$$

$$\frac{dQ(t)}{dt}=\lambda(t)-\mu(t) \tag{3}$$

$$\frac{dW(t)}{dt}=Q(t) \tag{4}$$

with the boundary conditions of

$$\lambda(t_0)=\mu(t_0) \tag{5}$$

$$\lambda(t_2)=\mu(t_2) \tag{6}$$

$$\lambda(t)-\mu(t)>0, \qquad t_0<t<t_2 \tag{7}$$

$$\lambda(t)-\mu(t)<0, \qquad t_2<t<t_3 \tag{8}$$

$$\frac{d\lambda(t_1)}{dt}=0 \tag{9}$$

$$Q(t_0)=0 \tag{10}$$

$$Q(t_3)=0 \tag{11}$$

where $\lambda(t)$, $\mu(t)$, and $Q(t)$ are the time-dependent arrival rate, discharge rate, and queue length at time t, respectively. In addition, $\lambda(t)-\mu(t)$ is the time-dependent net flow rate at time t. $A(t)$, $D(t)$, and $W(t)$ are the cumulative arrival count, cumulative departure count, and the total delay from time t_0 to t, respectively, and t_0, t_1, t_2, and t_3 are the first time that the arrival rate exceeds the discharge rate, the time with the maximum arrival rate, the time with the maximum queue length, and the time that the congestion dissipates, respectively. With the dynamical system equations for the fluid-approximated queueing system formulated here, we can map vehicle trajectories to the fluid approximated dynamic system as shown in Fig. 1(a).

The integrals of the first-order variable of the time-dependent arrival rate and the discharge rate are the cumulative arrival and departure counts, respectively. The integral of the first-order variable of the time-dependent net flow rate is the second-order variable of queue length, and the integral of the second-order variable of time-dependent queue length is the third-order variable of total delay. Their relationships are similar to that of the common case with the acceleration, speed, and displacement in physics.

3. Derivation of system state dynamics based on spatial queue with polynomial arrival rates

The polynomial functional form is viable to approximate the smoothly changing phenomena at different orders in the real world. Generally, the discharge rate of a bottleneck is assumed to be a constant to capture the essential cumulative input and output flow balance in the queueing system (Newell, 1982). To analyze the dynamic queueing system, we make the following two assumptions throughout this paper: (1) the virtual arrival rate $\lambda(t)$ at the bottleneck can be approximated by a polynomial function, and (2) the discharge rate $\mu(t)$ is constant. With these two assumptions, we can analytically derive a family of formulations under different orders of arrival rate functions and analyze the queueing system with time-dependent and averaged system measures. We present the core results with cubic-function-approximated arrival rates, while the results for other polynomial function approximated models are summarized in Appendix A.

3.1. Derivation of system state dynamics based on the cubic arrival rate function

First, let us compare Newell's classic model using the quadratic arrival rates with the proposed system dynamics equation with the assumption of cubic arrival rates. Increasing the order of the arrival rate function from quadratic to cubic results in another root (denoted as $\bar{t}$ in this paper), which is unobservable in traffic systems. To eliminate the unobservable $\bar{t}$, we defined a new parameter (i.e., oversaturation factor) m by the ratio between the time duration from the start of congestion to the time with maximum queue length and the whole congestion duration. With such a treatment, then we can obtain the time-dependent queue length and system dynamics with the cubic arrival rate function.

Assume that $\lambda(t)$ during a congestion period can be approximated by a cubic polynomial function, i.e., $\lambda(t) = \sum_{i=0}^{3}\gamma_i t^i$, where γ_i are the coefficients of the i-th order variables. Considering the boundary conditions of $\lambda(t_0)=\lambda(t_2)=\mu$ in the dynamical system equations, we can rewrite the time-dependent arrival rate function by the *factored form* of the net flow rate function as:

$$\lambda(t)-\mu=\gamma(t-t_0)(t-t_2)(t-\bar{t}) \tag{12}$$

where t_0 is the start time of congestion, t_2 is the time with maximal queue length, $\bar{t}$ is a root in addition to t_0 and t_2 of the cubic net flow rate function, and γ is the shape parameter.

Substitute Eq. (12) into Eq. (3), and then integrate the result to obtain the general form of time-dependent queue length as:

$$Q(t)=\int_{t_0}^{t}[\lambda(\tau)-\mu]d\tau=\int_{t_0}^{t}[\gamma(\tau-t_0)(\tau-t_2)(\tau-\bar{t})]d\tau \tag{13}$$

To derive the time-dependent queue length, we need to define the oversaturation factor m by the ratio between the time duration from the start of congestion to the time with maximum queue length and the whole congestion duration, i.e.,

$$m=\frac{t_2-t_0}{t_3-t_0},\qquad 0<m<1 \tag{14}$$

Then, with some simple algebraic operations (see Appendix B), we can obtain the time-dependent queue length as:

$$Q(t)=\gamma\cdot(t-t_0)^2\cdot\left[\frac{1}{4}(t-t_0)^2-\frac{1}{3}\cdot\left(\frac{3-4m}{4-6m}+m\right)(t_3-t_0)(t-t_0)+\frac{1}{2}\cdot\frac{(3-4m)m}{4-6m}(t_3-t_0)^2\right] \tag{15}$$

The maximum queue length $Q(t_2)$ can be derived as follows:

$$Q(t_2)=\gamma\cdot\frac{m^3(m-1)^2}{8-12m}\cdot(t_3-t_0)^4 \tag{16}$$

With the time-dependent queue length function, we can expediently calculate the time-dependent delay as follows:

$$w(t)=\frac{\gamma\cdot(t-t_0)^2}{\mu}\cdot\left[\frac{1}{4}(t-t_0)^2-\frac{1}{3}\cdot\left(\frac{3-4m}{4-6m}+m\right)(t_3-t_0)(t-t_0)+\frac{1}{2}\cdot\frac{(3-4m)m}{4-6m}(t_3-t_0)^2\right] \tag{17}$$

The total delay between time t_0 and t_3 can be calculated by integration of Eq. (15) as follows:

$$
\begin{aligned}
W(t_3) &= \int_{t_0}^{t_3} Q(t)dt \\
&= \gamma \cdot \int_{t_0}^{t_3} \left[\frac{1}{4}(t-t_0)^4 - \frac{1}{3} \cdot \left(\frac{3-4m}{4-6m} + m \right)(t_3-t_0)(t-t_0)^3 + \frac{1}{2} \cdot \frac{(3-4m)m}{4-6m}(t_3-t_0)^2(t-t_0)^2 \right] dt \\
&= \gamma \cdot \int_{0}^{t_3-t_0} \left[\frac{1}{4}u^4 - \frac{1}{3} \left(\frac{3-4m}{4-6m} + m \right)(t_3-t_0)u^3 + \frac{1}{2} \cdot \frac{(3-4m)m}{4-6m}(t_3-t_0)^2u^2 \right] du \\
&= \gamma \cdot (t_3-t_0)^5 \cdot \left[\frac{1}{20} - \frac{1}{12} \left(\frac{3-4m}{4-6m} + m \right) + \frac{1}{6} \cdot \frac{(3-4m)m}{4-6m} \right] \\
&= \gamma \cdot g(m) \cdot (t_3-t_0)^5
\end{aligned}
\tag{18}
$$

where the conversion factor $g(m)$ is:

$$
g(m) = \frac{1}{20} - \frac{1}{12} \left(\frac{3-4m}{4-6m} + m \right) + \frac{1}{6} \cdot \frac{(3-4m)m}{4-6m} \tag{19}
$$

The average delay is $w = W/D$, and the congestion duration $t_3 - t_0 = D/\mu$. Then, we can obtain the average delay function as follows:

$$
w = \frac{W}{D} = \frac{\gamma \cdot g(m)}{\mu} \cdot \left(\frac{D}{\mu} \right)^4 \tag{20}
$$

The unit of the shape parameter γ is the *vehicle per fourth-power of unit time* in the cubic form, and $g(m)$ is dimensionless since m is dimensionless.

3.2. Physical queue length and link travel time function

In the point queue model to analyze the traffic system performance, all vehicles travel at a free-flow speed for the whole road segment (i.e., link), and they may queue at the end of the link if the available discharge rate of the downstream link is restricted by its physical capacity. We first assume that the vehicles' physical lengths are zero; thus, all vehicles are accumulated at the end of the link when a queue forms, and the virtual queue length $Q(t)$ can be calculated with Eq. (15). According to Lawson et al. (1997), the physical queue length can be calculated in terms of the virtual queue length $Q(t)$, the free-flow speed v_f, and the speed at capacity v_μ based on the vehicle trajectories in the time-space plane. The mapping between the spatial queue representation and point queue is explained in Fig. 2, in which the red curve shows the physical queue extent, and the green curve is a vehicle trajectory entering the link of interest at time t and encountering the congestion at time t'. The spatial queue distance at time t' is $d(t') = v_f \cdot (t_f + t - t') = v_\mu \cdot (t_f + t - t' + w(t))$, and we can obtain that $d(t') = w(t) \cdot \left(\frac{1}{v_\mu} - \frac{1}{v_f} \right) = \frac{w(t) \cdot v_\mu}{1 - \frac{v_\mu}{v_f}}$ after eliminating $t_f + t - t'$. Since the physical queue length at time t' is the number of vehicles existing in the link from the back of the queue to the bottleneck location, we can calculate the physical queue length as:

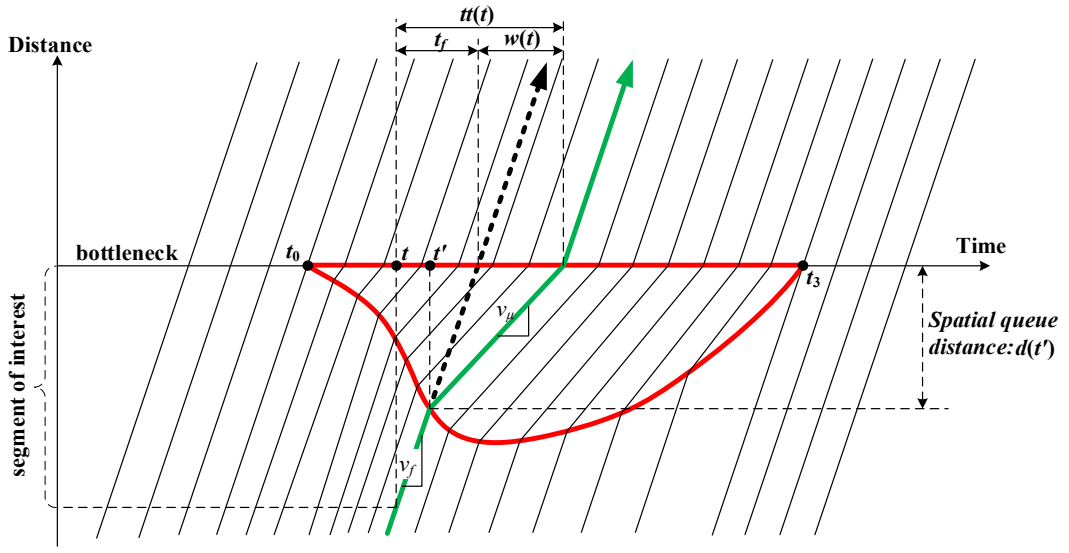


Fig. 2. Relationship between the physical queue length and virtual queue length.

$$Q^p(t^{'}) = \frac{d(t^{'})}{v_\mu} \cdot \mu = \frac{w(t) \cdot \mu}{1 - \frac{v_\mu}{v_f}} = \frac{Q(t)}{1 - \frac{v_\mu}{v_f}} \tag{21}$$

In a congested traffic system, the delay is calculated by Eq. (17), and the time-dependent travel time when entering the link of interest at time t can be obtained as:

$$tt(t) = t_f + w(t) = t_f + \frac{\gamma \cdot (t - t_0)^2}{\mu} \cdot \left[\frac{1}{4}(t - t_0)^2 - \frac{1}{3} \cdot \left(\frac{3 - 4m}{4 - 6m} + m\right)(t_3 - t_0)(t - t_0) + \frac{1}{2} \cdot \frac{(3 - 4m)m}{4 - 6m}(t_3 - t_0)^2\right] \tag{22}$$

where $tt(t)$ is the travel time for vehicles entering the link at time t.

Since the average delay is $w = \frac{\gamma \cdot g(m)}{\mu} \cdot \left(\frac{D}{\mu}\right)^4$, we can obtain the average travel time tt during the entire peak period by:

$$tt = t_f + w = t_f \left[1 + \frac{\gamma \cdot g(m)}{\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^4\right] \tag{23}$$

which reveals that the average travel time on the link is a fourth-power polynomial function of D/μ when the arrival rates are approximated by a cubic function. Table 2 summarizes different orders of arrival rates for vehicles and their average travel time functions.

The classical BPR (acronym of the Bureau of Public Roads) link travel time function is $tt = t_f \cdot \left[1 + \alpha \cdot (V/\mu)^\beta\right]$, where V is the traffic volume per unit time during a modeling period T and the parameters are usually valued as $\alpha = 0.15$ and $\beta = 4$. From the perspective of functional form, it is interesting that the proposed average travel time functions in this paper shed light on how the BPR function can be interpreted from the queueing theory with the polynomial assumption of the arrival rates. More specifically, the parameter α in the BPR function is related to the shape parameter γ, the oversaturation factor m, and the free-flow travel time t_f. This is consistent with the experiential calibration results in the literature (Horowitz, 1991; Mannering et al., 1990) that the parameters in the BPR function vary with the capacity and speed limit (which restricts the free-flow speed and travel time). The parameter β in the BPR function is related to the order of arrival rate function plus one. The notational differences between our model and the BPR function are as follows: (1) We use the ratio of excess demand over the congestion period to the discharge rate D/μ, rather than V/μ in the BPR function. Note that D in the queuing model represents the demand over the whole congestion period, while V used in the BPR function is the traffic volume per unit time during a modeling period T. Planners should be aware of such potential inconsistent definitions for modeling periods. Besides, in the BPR function, V/μ is dimensionless, while in our model, $D/\mu = t_3 - t_0$ is the entire congestion duration in which a queue exists. To maintain the conservation of units, it can be easily derived that the unit of the shape parameter γ should be the *vehicle per fourth-power of unit time* in the cubic form. (2) The parameter μ in our model is the congestion discharge rate, which might be much smaller than the assumed maximal-flow or practical capacity μ used in the typical BPR function.

3.3. Discussion on the oversaturation factor m

In a cubic form with $\gamma < 0$, the condition $\bar{t} - t_0 \leqslant 0$ should be held; because $m = (t_2 - t_0)/(t_3 - t_0)$, we can combine this with the definition of $\bar{t}$ to obtain the range of m:

$$m \in \left(\frac{2}{3}, \frac{3}{4}\right], \quad \gamma < 0 \tag{24}$$

Similarly, the condition $\bar{t} - t_3 \geqslant 0$ should be held when $\gamma > 0$; then, we can obtain that

Table 2
The arrival rates of vehicles and corresponding travel time functions.

Arrival rate form	Arrival rate function	Average travel time function
Constant form	$\lambda(t) = \begin{cases} \pi_1 > \mu, & t_0 \leqslant t < t_2 \\ \pi_2 < \mu, & t_2 \leqslant t \leqslant t_3 \end{cases}$	$tt = t_f \cdot \left[1 + \frac{(\pi_1 - \mu)(\mu - \pi_2)}{2\mu(\pi_1 - \pi_2) \cdot t_f} \cdot \left(\frac{D}{\mu}\right)\right]$
Linear form	$\lambda(t) = -\kappa(t - t_2) + \mu, \quad \kappa > 0$	$tt = t_f \cdot \left[1 + \frac{\kappa}{12\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^2\right]$
Quadratic form	$\lambda(t) = -\xi(t - t_0)(t - t_2) + \mu, \quad \xi > 0$	$tt = t_f \cdot \left[1 + \frac{\xi}{36\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^3\right]$
Cubic form	$\lambda(t) = \gamma(t - t_0)(t - t_2)(t - \bar{t}) + \mu$	$tt = t_f \cdot \left[1 + \frac{\gamma \cdot g(m)}{\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^4\right]$

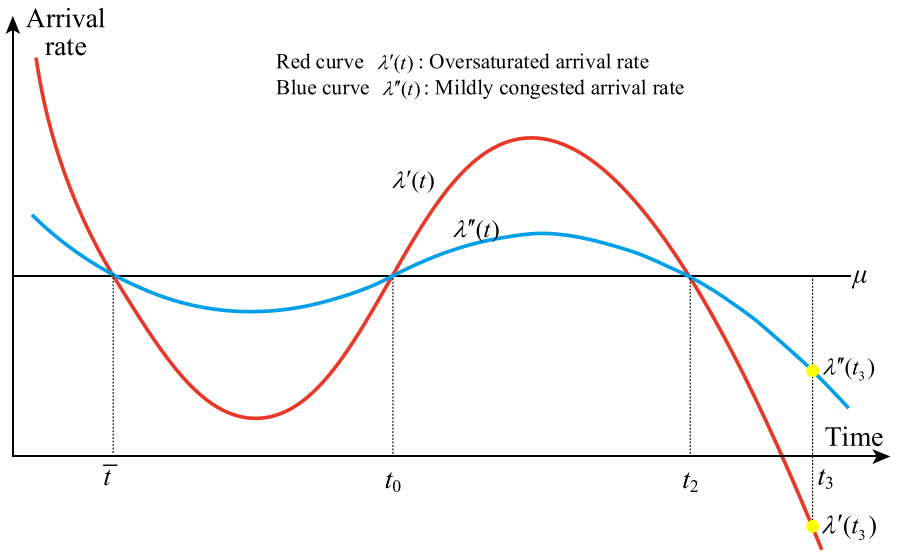


Fig. 3. Illustration of the cubic arrival rate function with a negative shape parameter (i.e., $\gamma < 0$).

$$m \in \left[\frac{1}{2}, \frac{2}{3}\right), \quad \gamma > 0 \tag{25}$$

In summary, the range of the oversaturation factor m is

$$m \in \begin{cases} \left[\frac{1}{2}, \frac{2}{3}\right), & \gamma > 0 \\ \left(\frac{2}{3}, \frac{3}{4}\right], & \gamma < 0 \end{cases} \tag{26}$$

and the range of the conversion factor $g(m)$ is

$$g(m) \in \begin{cases} \left[\frac{1}{120}, +\infty\right), & \gamma > 0 \\ \left(-\infty, -\frac{1}{80}\right], & \gamma < 0 \end{cases} \tag{27}$$

When $m = 2/3$, the shape parameter γ in the cubic arrival rate function will be zero; thus, the arrival rate function will be reduced to a quadratic form, which is consistent with the description given by Newell (1982, Chapter 2).

It is worth noting that not all of the above formulations are applicable for oversaturated dynamic queueing systems. For example, as shown in Fig. 3, $\lambda'(t)$ and $\lambda''(t)$ are the arrival rates for the overcongested and slightly congested queueing systems, respectively. For the overcongested queueing systems, the arrival rate at the end of the congestion duration $\lambda'(t_3)$ is significantly less than the discharge rate μ because $\lambda'(t)$ decreases sharply near t_3. The estimated $\lambda'(t_3)$ could very likely be a negative value, which violates the positive flow assumption. On the other hand, for a slightly congested queueing system, the arrival rate at the end of the congestion duration $\lambda''(t_3)$ is also below the discharge rate μ, but with larger values than $\lambda'(t_3)$ in the overcongested case. Therefore, it is applicable only for slight-congested queueing systems when $\gamma < 0$ with $m \in (2/3, 3/4]$, while both are applicable for slightly saturated and oversaturated queueing systems when $\gamma > 0$ with $m \in [1/2, 2/3)$. The same situations exist in the quadratic arrival rate forms.

Three special cases, including (1) $\gamma < 0$, $m = 3/4$, (2) $\gamma > 0$, $m = 1/2$, and (3) $\gamma = 0$, $m = 2/3$ are discussed as follows:

(1) Case 1: $\gamma < 0$ and $m = 3/4$

When $\gamma < 0$ and $m = 3/4$, we can see that $\bar{t} = t_0$, which means the cubic net flow function has a repeated root at $t = t_0$. Then, we can obtain the time-dependent queue length, time-dependent delay, total delay, and average delay by:

$$\begin{cases} Q(t) = \frac{1}{4}\gamma(t-t_0)^3(t-t_3) \\ w(t) = \frac{1}{4\mu}\gamma(t-t_0)^3(t-t_3) \\ W(t_3) = -\frac{1}{80}\gamma(D/\mu)^5 \\ w = -\frac{\gamma}{80\mu}\cdot(D/\mu)^4 \end{cases} \tag{28}$$

(2) Case 2: $\gamma > 0$ and $m = 1/2$

When $\gamma > 0$ and $m = 1/2$, the arrival rate function is symmetric around the point (t_2, μ), and $\bar{t} = t_3$. Then, we can obtain the time-dependent queue length, time-dependent delay, total delay, and average delay by:

$$\begin{cases} Q(t) = \frac{1}{4}\gamma(t-t_0)^2(t-t_3)^2 \\ w(t) = \frac{1}{4\mu}\gamma(t-t_0)^2(t-t_3)^2 \\ W(t_3) = \frac{1}{120}\gamma(D/\mu)^5 \\ w = \frac{\gamma}{120\mu}(D/\mu)^4 \end{cases} \tag{29}$$

(3) Case 3: $\gamma = 0$ and $m = 2/3$

When $m = 2/3$, the cubic form is not applicable because we would have $t_0 = t_3$ and $\gamma = 0$ with the condition of $Q(t_3) = 0$, which violates the definition of t_0 and t_3; then, it would be reduced to a concave quadratic form. The description for the quadratic form of the arrival rate function is given by Newell (1982, Chapter 2). This case is applicable only to slightly congested queueing systems.

4. Calibration method

In this section, a two-step calibration method is proposed to calibrate the parameters in the proposed spatial queue model. Specifically, the first-step calculates the discharge rate μ based on the observations of the cumulative departure flows from the downstream link of the bottleneck, while the second-step calibrates the shape parameter γ and the oversaturation factor m with the observations of the time-dependent queue length and/or delay. The details are presented as follows.

4.1. Calibrating the discharge rate μ

This study assumes that the discharge rate of a bottleneck is a constant, which is consistent with empirical observations in Cassidy and Bertini (1999). Combining with the boundary conditions that the queue forms at time t_0 and dissipates at time t_3, one can simply calculate the discharge rate μ by:

$$\mu = \frac{N(t_3) - N(t_0)}{t_3 - t_0} \tag{30}$$

where $N(t),\ t \in [t_0, t_3]$ is the cumulative number of vehicles from t_0 to t.

4.2. Calibrating the shape parameter γ and the oversaturation factor m

As for the calibration of γ and m in the proposed model, the objective is set to minimize the sum of squared residuals (SSR), mathematically expressed as follows:

$$\min_{(\gamma,m)} Z = \sum_{t=1}^{|P|} \left\{ (Y(t) - \widehat{Y}(t))^2 \right\} \tag{31}$$

where $|P|$ is the number of time intervals in the peak period; $Y(t)$ and $\widehat{Y}(t)$ are the estimated values and observations at time interval t, respectively. $Y(t)$ and $\widehat{Y}(t)$ can be the (virtual/physical) queue length or delay in terms of the available data. Without loss of generality, we use the virtual queue length $Q(t)$ to illustrate the calibration procedures. Thus, the objective function becomes:

$$min \underset{(\gamma,m)}{Z} = \sum_{t=1}^{|P|}\{(Q(t) - \widehat{Q}(t))^2\} \tag{32}$$

subject to

$$\lambda(t) = \gamma(t - t_0)(t - t_0 - m(t_3 - t_0))\left(t - t_0 - \frac{(3 - 4m)(t_3 - t_0)}{4 - 6m}\right) + \mu \geqslant 0 \tag{33}$$

where $Q(t)$ and $\widehat{Q}(t)$ are the estimated values and observations of the virtual queue length at time interval t, respectively. $Q(t)$ is a function of $\lambda(t)$, and constraint (33) ensures the non-negativity of $\lambda(t)$.

With some simple calculations, one can derive that $\frac{\partial}{\partial m}\left(\frac{3-4m}{4-6m} + m\right) = \frac{3(6m^2-8m+3)}{2(2-3m)^2}$ and $\frac{\partial}{\partial m}\left(\frac{(3-4m)m}{4-6m}\right) = \frac{6m^2-8m+3}{(2-3m)^2}$. Substituting these derivations into Eq. (15), one can further calculate the first- and second-order partial derivatives of $Q(t)$ with respect to γ and m as follows:

$$\frac{\partial Q}{\partial \gamma} = (t - t_0)^2 \cdot \left[\frac{1}{4}(t - t_0)^2 - \frac{1}{3} \cdot \left(\frac{3 - 4m}{4 - 6m} + m\right)(t_3 - t_0)(t - t_0) + \frac{1}{2} \cdot \frac{(3 - 4m)m}{4 - 6m}(t_3 - t_0)^2\right] \tag{34}$$

$$\frac{\partial Q}{\partial m} = \gamma(t_3 - t_0)(t_3 - t)(t - t_0)^2 \cdot \frac{6m^2 - 8m + 3}{2(2 - 3m)^2} \tag{35}$$

$$\frac{\partial^2 Q}{\partial \gamma \partial m} = (t_3 - t)(t - t_0)^2 \cdot \frac{6m^2 - 8m + 3}{2(2 - 3m)^2} \tag{36}$$

$$\frac{\partial^2 Q}{\partial \gamma^2} = 0 \tag{37}$$

$$\frac{\partial^2 Q}{\partial m^2} = \gamma(t_3 - t_0)(t_3 - t)(t - t_0)^2 \cdot \frac{1}{(2 - 3m)^3} \tag{38}$$

To obtain the optimal parameter values in Eq. (31), one can rewrite Eq. (31) as follows:

$$min \underset{(\gamma,m)}{Z} = \sum_{t=1}^{|P|}\{(Q(t) - \widehat{Q}(t))^2\} = \sum_{t=1}^{|P|}\{Q^2(t) - 2Q(t)\widehat{Q}(t) + \widehat{Q}^2(t)\} \tag{39}$$

Substituting Eqs. (34)–(38) into Eq. (39), we can analytically derive the first- and second-order partial derivatives of Z with respect to γ and m as follows:

$$\frac{\partial Z}{\partial \gamma} = \sum_{t=1}^{|P|}\left\{2(Q(t) - \widehat{Q}(t)) \cdot \frac{\partial Q}{\partial \gamma}\right\} \tag{40}$$

$$\frac{\partial Z}{\partial m} = \sum_{t=1}^{|P|}\left\{2(Q(t) - \widehat{Q}(t)) \cdot \frac{\partial Q}{\partial m}\right\} \tag{41}$$

$$\frac{\partial^2 Z}{\partial \gamma \partial m} = \sum_{t=1}^{|P|}\left\{2(1 + Q(t) - \widehat{Q}(t)) \cdot \frac{\partial^2 Q}{\partial \gamma \partial m}\right\} \tag{42}$$

$$\frac{\partial^2 Z}{\partial \gamma^2} = \sum_{t=1}^{|P|}\left\{2(1 + Q(t) - \widehat{Q}(t)) \cdot \frac{\partial^2 Q}{\partial \gamma^2}\right\} = 0 \tag{43}$$

$$\frac{\partial^2 Z}{\partial m^2} = \sum_{t=1}^{|P|}\left\{2(1 + Q(t) - \widehat{Q}(t)) \cdot \frac{\partial^2 Q}{\partial m^2}\right\} \tag{44}$$

With the Jacobian matrix $J = \left[\frac{\partial Z}{\partial \gamma}, \frac{\partial Z}{\partial m}\right]$ and the Hessian matrix $H = \begin{bmatrix} \frac{\partial^2 Z}{\partial \gamma^2} & \frac{\partial^2 Z}{\partial \gamma \partial m} \\ \frac{\partial^2 Z}{\partial \gamma \partial m} & \frac{\partial^2 Z}{\partial m^2} \end{bmatrix}$, one can solve the problem in Eq. (39) by some nonlinear optimization algorithms, which will be introduced in the next subsection.

4.3. Nonlinear optimization algorithms

Although the optimization model in Eq. (39) seems simple, solving it to obtain a global optimum is not an easy task due to the highly nonlinear, nonconvex, and multimodal characteristics of the objective function. In this study, different algorithms, including the Newton's method, the sequential least squares programming (SLSQP) algorithm, the adaptive moment estimation (Adam) algorithm, the Bayesian optimization algorithm, are tested and compared for solving the optimization model. The details of these algorithms are summarized as follows.

- The Newton's method, also known as the Newton-Raphson method, is a recursive solution algorithm to approximate the root of a differentiable function. Thus, one can use the Newton's method to the derivative of a twice-differentiable function to obtain the root(s) of the derivative. The solution(s) may be (local) minima, maxima, or saddle points. The details on this algorithm can be found in Nocedal and Wright (2006).
- The SLSQP algorithm, also known as the sequential quadratic programming algorithm (Nocedal and Wright, 2006), is a widely used solution algorithm for constrained nonlinear optimization problems. In each iteration, it approximates the original problem by a quadratic model subject to a linearization of constraints. Based on such an approximation, one can solve for the extreme value point with the Newton's method. The quality of the solutions is very sensitive to the selection of the initial point, and the result may be a local optimum.
- The Adam algorithm (Kingma and Ba, 2014) is a stochastic optimization algorithm which only needs the first-order gradients with little memory requirements. It is designed by combining the advantages of two popular algorithms (i.e., the adaptive gradient algorithm proposed by Duchi et al. (2011) and the root mean square propagation algorithm proposed by Tieleman and Hinton (2012), and now it has been widely used for the optimization problems in deep learning. The theoretical analyses on the Adam algorithm are referred to Kingma and Ba (2014).
- The Bayesian optimization algorithm (Pelikan et al., 1999; Snoek et al., 2012) is a simulation-based optimization algorithm, and it is especially applicable to the problem where its objective function is difficult to evaluate. A surrogate function is usually built for the objective function, and uncertainty is quantified with Bayesian method and Gaussian process. Based on the surrogate function, an acquisition function can be obtained to decide where to sample. Similar to the Adam algorithm, the Bayesian optimization algorithm has been widely used in the deep learning.

Due to the complexity of the optimization problem with a highly nonlinear, nonconvex, and multimodal objective function, the corresponding multi-start versions for the Newton's method, SLSQP and Adam algorithms are also considered in this study to avoid the local optimum issue. (Note that the Bayesian optimization algorithm is a simulation-based optimization algorithm with multiple initial solutions, thus there is no need to design a multi-start version for it). Besides, a parallel grid search scheme is designed to obtain the (nearly) global optimum, which can be deemed as the benchmark for the data fitting strategy in this manuscript.

5. Estimation results on different bottlenecks

5.1. Empirical data sets and algorithm settings

The proposed performance model with a cubic arrival rate function can be calibrated with only three parsimoniously selected parameters (i.e., the discharge rate μ, the shape parameter γ, and the oversaturation factor m). We use three empirical data sets (DSs) with the descriptions in Table 3 to verify the proposed model and parameter fitting strategies. The areas in these three cases with empirical data sets are shown in Fig. 4.

The three data sets DS1–DS3 are from Los Angeles, Beijing, and New York, respectively. Set DS1 is freeway sensor data that records the traffic flow, speed, and occupancy in a 5-min interval. The original data can be accessed from http://pems.dot.ca.gov, which is a real-time freeway performance measurement system developed by the California Department of Transportation. We choose a single

Table 3
Description of empirical data sets.

Data set	Queueing system	Object	Data type and information	Collected time and location
DS1	Traffic system	Vehicle	22 freeway detectors with traffic flow, speed, and occupancy data.	Collected from the Northbound direction of I-405 freeway between absolute postmile 8.97 to 14.77 mile in Los Angeles in the month of April 2019, from 11:00 a.m. to 20:00 p.m. weekdays.
DS2	Traffic system	Vehicle	20 remote traffic microwave sensors with the traffic flow, and 47 detector locations of probe taxi (around 5% of population traffic) with averaged traffic speed data.	Collected from the west-third-ring of Beijing City on June 8, 2018, from 6:00 a.m. to 12:00 a.m. weekday.
DS3	Transportation system	Taxi	30,794 taxi trip record data with trip ID, pickup time and location, drop off time and location, etc.	Collected from the Midtown Center of Manhattan to the John F. Kennedy International Airport in New York City in the month of October 2018, provided by the New York City Taxi and Limousine Commission

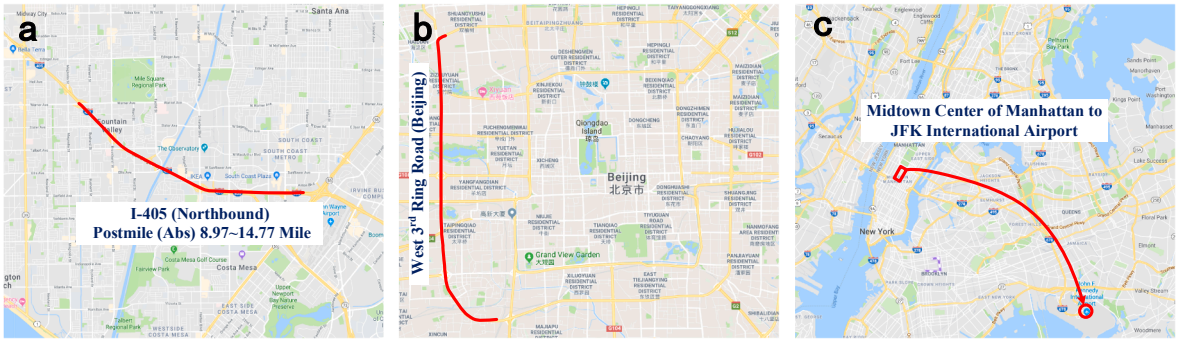


Fig. 4. Areas of the traffic queueing systems used in the calibration. (a) DS1; (b) DS2; (c) DS3.

recurrent bottleneck, which is located in the northbound direction of the I-405 freeway between absolute postmile 8.97 and mile 14.77 in Los Angeles, and the time duration covers from 11:00 a.m. to 20:00 p.m. for weekdays in the month of April 2019. The number of lanes is four at the bottleneck location in DS1. Set DS2 contains the remote traffic microwave sensor data with the traffic flow information and the probe vehicle data with the averaged traffic speed information. Although the remote traffic microwave sensor data also record the occupancy and speed information, the quality of these data are very poor; thus, we use the probe vehicle data as supplementary, high-quality data of speed information. The data are collected from the west-third-ring of Beijing City on June 8, 2018, from 6:00 a.m. to 12:00 a.m., to analyze the morning peak traffic system. There are three lanes at the bottleneck location; however, one of the lanes is a bus-only lane during the morning peak hours (7:00 a.m. $\sim$ 9:00 a.m.). Set DS3 is the New York taxi data, which contain each trip ID number, its corresponding pick-up time and location, and its drop-off time and location. The data can be accessed from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, which is free and open-source for downloading. We choose one origin–destination pair, in which passengers are picked up at the center of midtown Manhattan and dropped off at John F. Kennedy International Airport in the month of October 2018. With the pick-up and drop-off time, we can obtain the en-route travel time for each trip; then, we can calculate the delay time after subtracting the en-route travel time by a free-flow travel time, which can be assumed to be the average travel time during the nonpeak period, such as 6:00 a.m. to 6:30 a.m.

As for the solution algorithms, the configurations of parameters are set as follows. To illustrate our proposed methodology and solution algorithms, we provide the DS1 and corresponding algorithms in https://github.com/ChengTraffic/Polynomial-Arrival-Queue-PAQ.

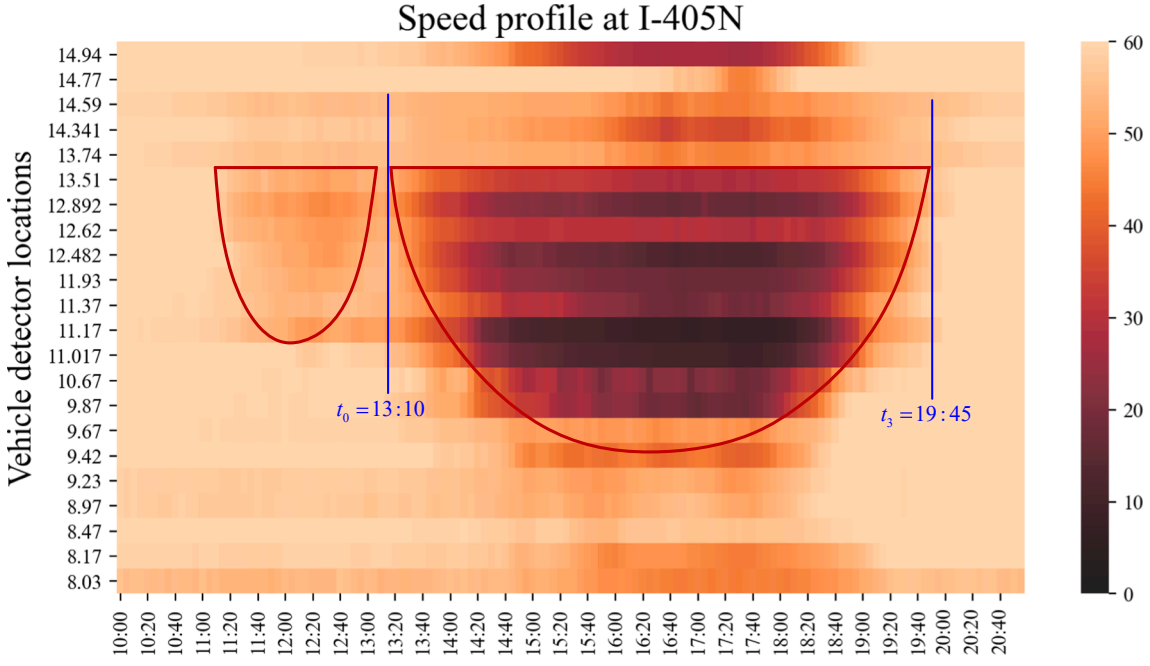


Fig. 5. Speed profile with DS1. It is clear that the bottleneck is located at Abs = 13.51 mile. In this case, we analyze only one single bottleneck with one peak period from $t_0 = 13:10$ to $t_3 = 19:45$.

- Newton's method. The maximum iteration is set as 10^3, and the termination tolerances on the decision variables and function value are both set as 10^{-12}.
- SLSQP. The maximum iteration is set as 10^3, and the termination tolerances on the decision variables and function value are both set as 10^{-12}.
- Adam. The maximum iteration is set as 10^3, and the hyperparameters of the step size and two exponential decay rates for the moment estimates are set as 0.001, 0.9, and 0.999, respectively.
- Bayesian optimization. The Gaussian process regression is used to describe the prior/posterior distribution of the objective function. The maximum iteration is set as 200, the number of points for each sampling is set as 10, the variance of the error term in the Gaussian process is set as 10^{-8}, and the precision tolerance of the objective value is set as 10^{-8}.

5.2. Calibration results

As for the DS1, we first draw the speed profile as shown in Fig. 5 to obtain the location of the bottleneck. It is clear that there are two bottlenecks in DS1, and we analyze only the second bottleneck, with more severe traffic congestion. Second, we draw a speed and occupancy plot (see Fig. 6) to determine t_0, t_3 and the free-flow speed v_f. The blue curves depict the speed, while the red curves depict the occupancy, which can be converted to density values through May (1990, see Chapter 7, page 193). The speed downstream of the bottleneck remained almost stable over 45 mile/hour during the peak period, while the speeds at the bottleneck and upstream of the bottleneck were reduced sharply during the peak period. Similarly, the occupancy downstream of the bottleneck reaches almost below 0.11 during the peak period, while the occupancies at the bottleneck and upstream of the bottleneck reach up to 0.25 during the peak period.

There are three measurements in DS1, including the cumulative departure count, the queue length, and the delay time. In the first-step calibration, the cumulative traffic count adjacently downstream of the bottleneck location is used as the measurement of the observed cumulative departure count. According to Eq. (30), the discharge rate in DS1 can be calculated as $\mu_{DS1} = 3936$ *veh/hour* or $\mu_{DS1} = 984$ *veh/hour/lane*.

As for the second-step calibration, the observed time-dependent delay time is calculated by the travel time of the total influence length of the bottleneck after subtracting the free-flow travel time. With regard to the time-dependent queue length, we need to first obtain the critical occupancy with the flow-occupancy plot (in which the critical occupancy for DS1 is close to 0.13), then transfer the occupancy to density, and finally calculate the observed time-dependent queue length with the density data (May, 1990).

As mentioned in Section 4, the optimization problem in Eq. (39) is highly nonlinear, nonconvex, and multimodal. Fig. 7 illustrates these characteristics with the objective value by changing the parameters of γ and m. It is clear that the objective function has multiple (local) optima, and obtaining a global optimum is very difficult. We first use the parallel grid search with multiprocessing technique by

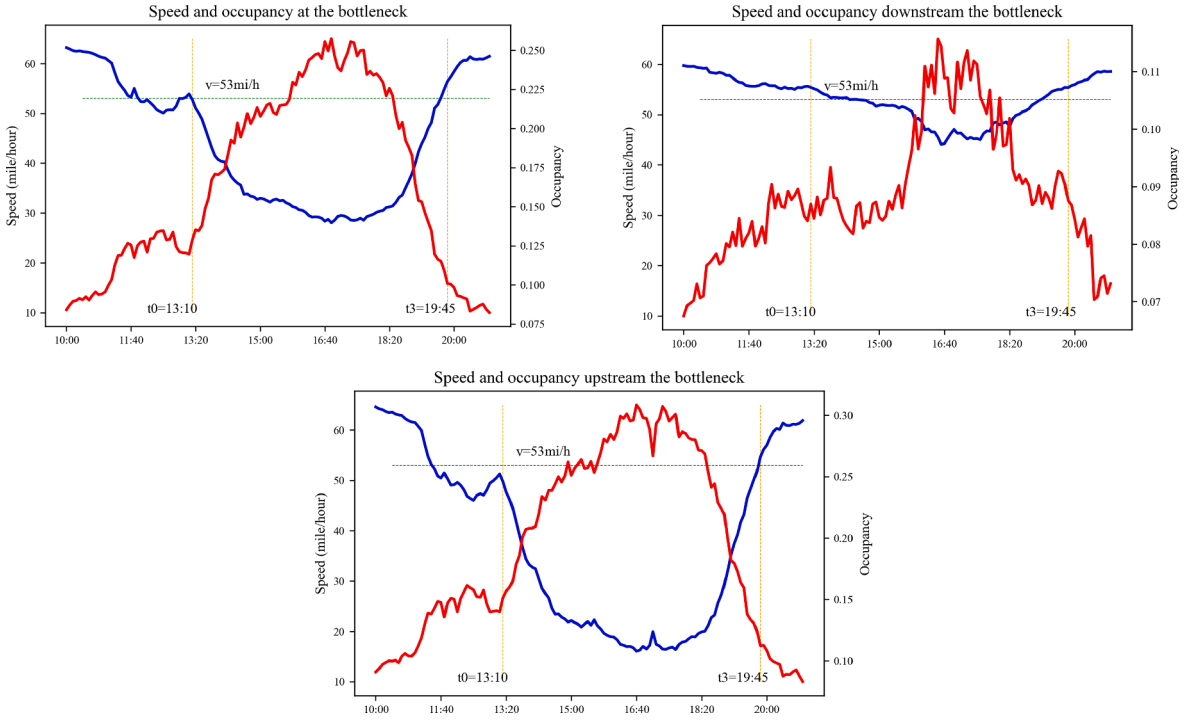


Fig. 6. Speed and occupancy with DS1.

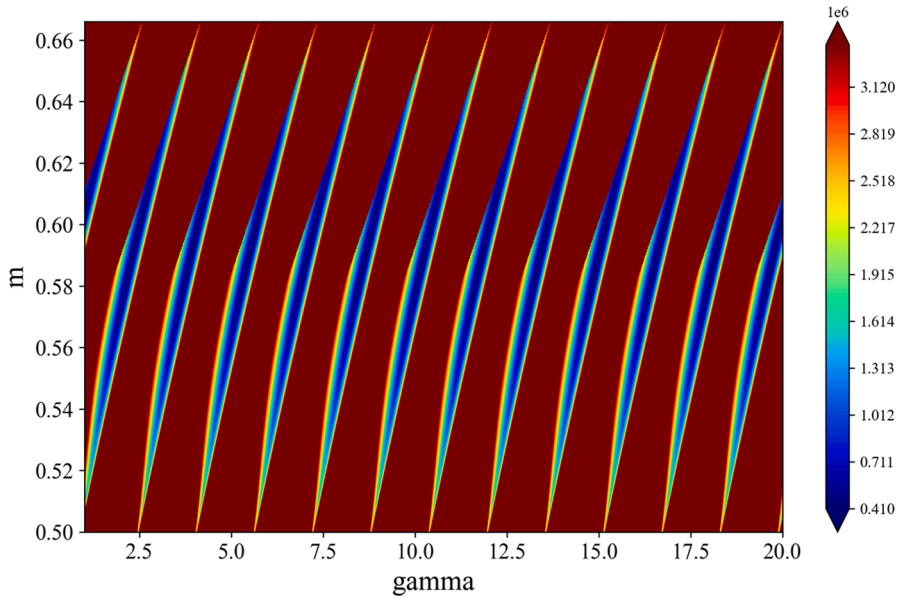


Fig. 7. Illustration of the highly nonlinear, nonconvex, and multimodal characteristics of the objective function.

changing γ from 1 to 20 with an increment of 0.0095 and m from 0.5 to 0.666 with 0.001, thus a total of 334,167 combinations are evaluated in the parallel grid search stage. The result of the minimal objective value in the parallel grid search is 409,795, which will be the benchmark for the data fitting strategies with Newton's method, SLSQP, Adam, and also the Bayesian optimization algorithms. Table 4 compares the performance of different algorithms in the second-step calibration for DS1. As we can see from it, the multi-start Adam algorithm outperforms others in terms of all the performance metrics of the relative error (RE) with respect to the minimal objective value output by the parallel grid search algorithm, the mean squared error (MSE), and the R^2, with the results of $\gamma_{DS1} = 11.536$ *veh/hour*^{4} and $m_{DS1} = 0.533$. Besides, we can also see that even without the multi-start, the Adam is the best in fitting the empirical data in DS1. The SLSQP algorithm is very sensitive to the given initial value, it may perform much worse than other algorithms if a bad initial value (which is the same to other algorithms) is chosen. The calibration result of the time-dependent queue length is shown in Fig. 8(a), and the calibrated arrival rate and discharge rate for DS1 is shown in Fig. 9(a).

For DS2, the calibration procedures are similar with that of DS1; however, there are only two measurements, namely, the cumulative departure count and the delay time, in DS2. The traffic condition is more complicated than that of DS1 because of the existence of a spatial queue spillback and temporal queue connection phenomena in DS2 (see Fig. 10). In this paper, we focus on a single bottleneck (which may have the temporal queue connection) and do not consider the queue spillback in the model. To explain the system with a temporal queue connection, we build a two-peak model, in which the queue during the first peak period does not completely dissipate at time t_3, and the new queue during the second peak appears at time t_3. The details on the two-peak model can be found in Appendix C, and it is worth noting that there is no need to guarantee the condition of $m \geqslant 0.5$ in the two-peak model. According to the comparison results of the algorithms used for fitting empirical data in DS1, we select the multi-start Adam to calibrate for the DS2. The final results for DS2 are $\mu_{DS2} = 1815$ *veh/hour* or $\mu_{DS2} = 907.5$ *veh/hour/lane*, $\gamma_{DS2} = 1126.23$ *veh/hour*^{4} and $m_{DS2} = 0.527$. The performance metrics for DS2 are MSE = 0.881 and $R^2 = 0.905$ for the time-dependent delay. The calibration result of the time-

Table 4
Comparison of the performance of different algorithms in the second-step calibration for DS1.

Algorithms	Calibration results		Performance metrics		
	γ	m	RE	MSE	R^2
Newton's method	9.999	0.547	0.088	5644.322	0.934
Multi-start Newton's method	10.510	0.542	0.039	5390.922	0.937
SLSQP	10.000	0.580	11.296	63782.352	0.257
Multi-start SLSQP	10.918	0.539	0.015	5266.788	0.939
Adam	11.240	0.536	0.003	5203.907	0.939
Multi-start Adam	11.536	0.533	0.000	5186.960	0.940
Bayesian optimization	11.906	0.529	0.020	5289.024	0.938

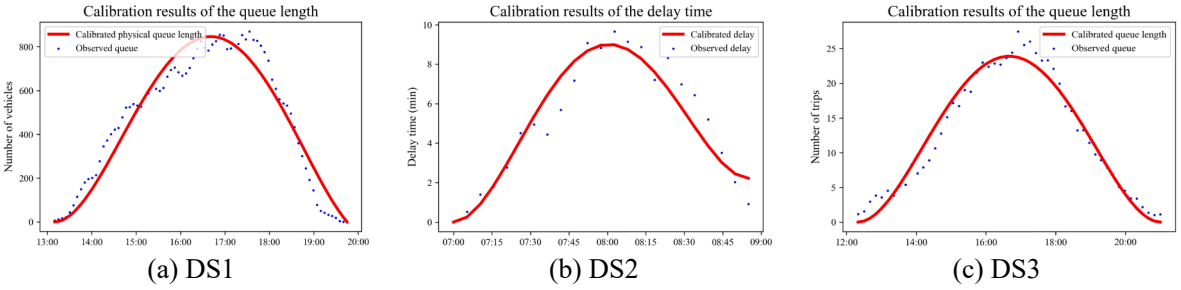


Fig. 8. Calibration results with different data sets.

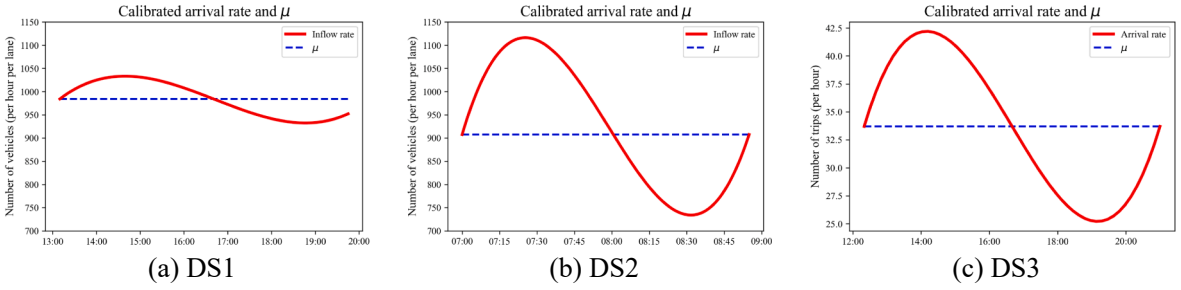


Fig. 9. Comparisons of the calibrated arrival rate and discharge rate between different data sets.

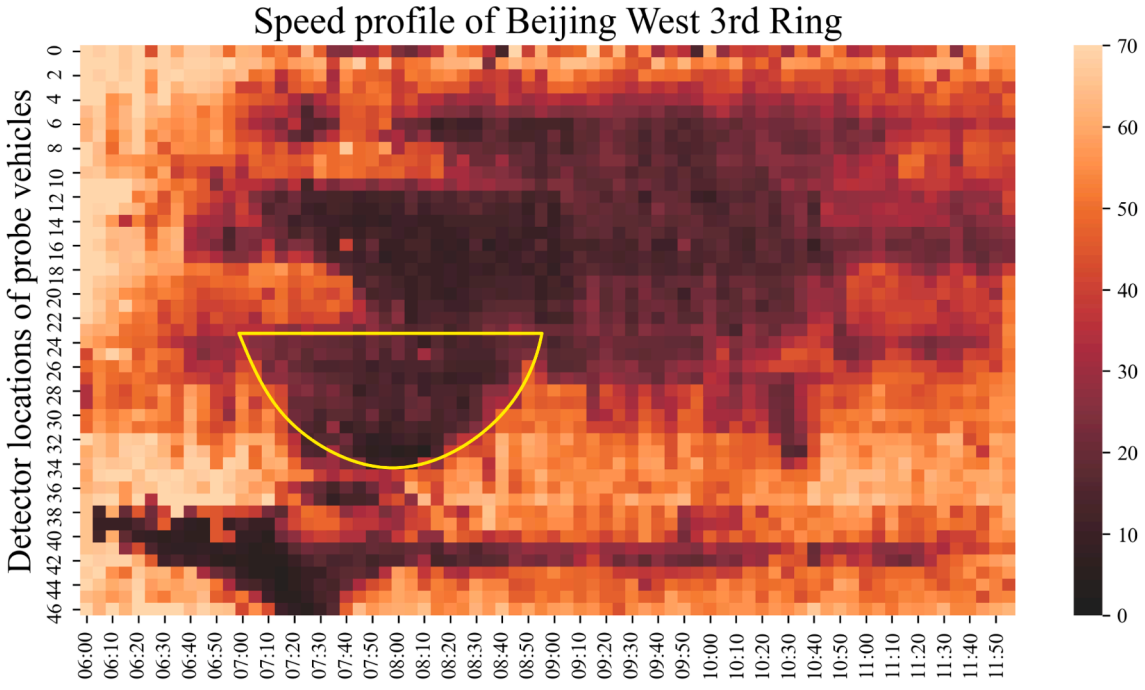
dependent delay is shown in Fig. 8(b), and the calibrated arrival rate and discharge rate for DS2 is shown in Fig. 9(b). When comparing Fig. 9(a) with 9(b), it is clear that the discharge rate in DS1 shows smaller variation than that in DS2, as DS1 covers highway roads with a higher speed limit, but the elevated urban expressways in DS2 more frequently exhibit stop-and-go phenomena. For the particular period of analysis, the effective discharge rate during heavy congestion in DS1 is higher than that in DS2, but both values are significantly lower than the theoretical maximum flow capacity, typically due to prevailing traffic density at bottlenecks, complex road geometry, and driving behavior. In DS2, the queue does not dissipate at the end of the first peak period as a bus-only lane opens to general-purpose vehicles at 9 AM, which attracts another wave of demand.

For DS3, we first choose one origin–destination pair (from the center of midtown Manhattan to John F. Kennedy International Airport) and draw the cumulative arrival and departure curves and then shift the cumulative arrival curve to the right to reach the cumulative departure curve (with the moving distance taken as the free-flow travel time). The obtained new curve is the virtual cumulative arrival curve. Based on the virtual cumulative arrival curve and the cumulative departure curve, we can calculate the observed time-dependent delay time. Similarly, we use the multi-start Adam algorithm to calibrate for the DS3. The final results for DS3 are $\mu_{DS3} = 33.705$ *trip/hour*, $\gamma_{DS3} = 0.271$ *veh/hour*^{4}, and $m_{DS3} = 0.500$. The performance metrics for DS3 are MSE $= 0.877$ and $R^2 = 0.908$ for the time-dependent delay. The calibration result of the time-dependent queue length is shown in Fig. 8(c), and the calibrated arrival rate and discharge rate for DS3 is shown in Fig. 9(c).

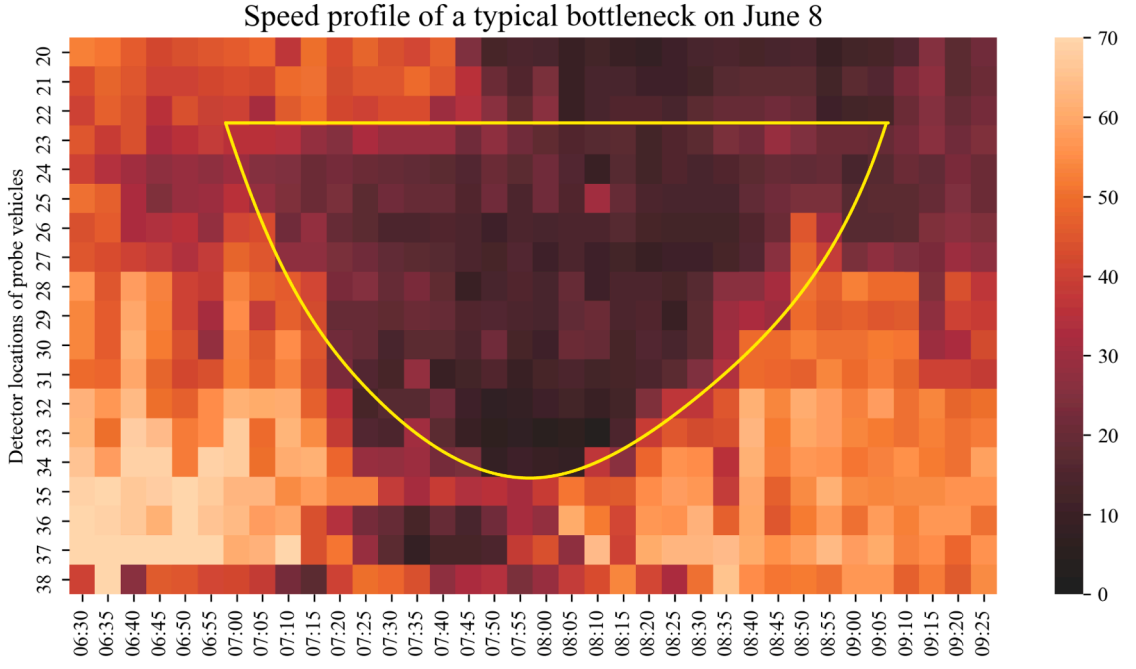
One of the very important features of our proposed model is the quantification of the peak demand $\lambda(t)$ over the supply μ at the signature timestamp t_1 for this oversaturated process, which can be denoted as the system utilization ratio $\rho = \lambda(t_1)/\mu$. The results show that $\rho_{DS1} = 1.059$, $\rho_{DS2} = 1.232$, and $\rho_{DS3} = 1.252$. Congestion in the Los Angeles data set DS1 is associated with long-period accumulation of excess demand even though it only marginally exceeds the discharge rate. On the other hand, the congestion in the Beijing case DS2 is most likely due to a sharp surge of the incoming flow, which holds the full potential for demand spreading strategies.

6. Discussion on queueing state dynamics in demand and supply curves

In real-world oversaturated dynamic queueing systems, the demand and supply curves may have many distinct ways of pattern dynamics. Fig. 11 depicts the different approximation forms of the arrival rate function, and Fig. 12 depicts different patterns of the cumulative departure curve in dynamic oversaturated queueing systems. Based on the framework used in the above study, one could select one of the most likely demand and supply patterns that match real-world observations and could utilize critical control points



(a) Speed profile of Beijing West 3rd Ring on June 8, 2018



(b) Speed profile of a typical bottleneck

Fig. 10. Speed profile with DS2. The upper figure is the speed profile of the West 3rd Ring (from the south to the north direction) of Beijing City in the morning of June 8, 2018. We do collect raw data across 2 weeks, but the data from this typical weekday have been systematically verified across different sections to ensure that all the related loop detectors are working properly. The bottom figure is a typical bottleneck. This case is much more complicated than the case in DS1.

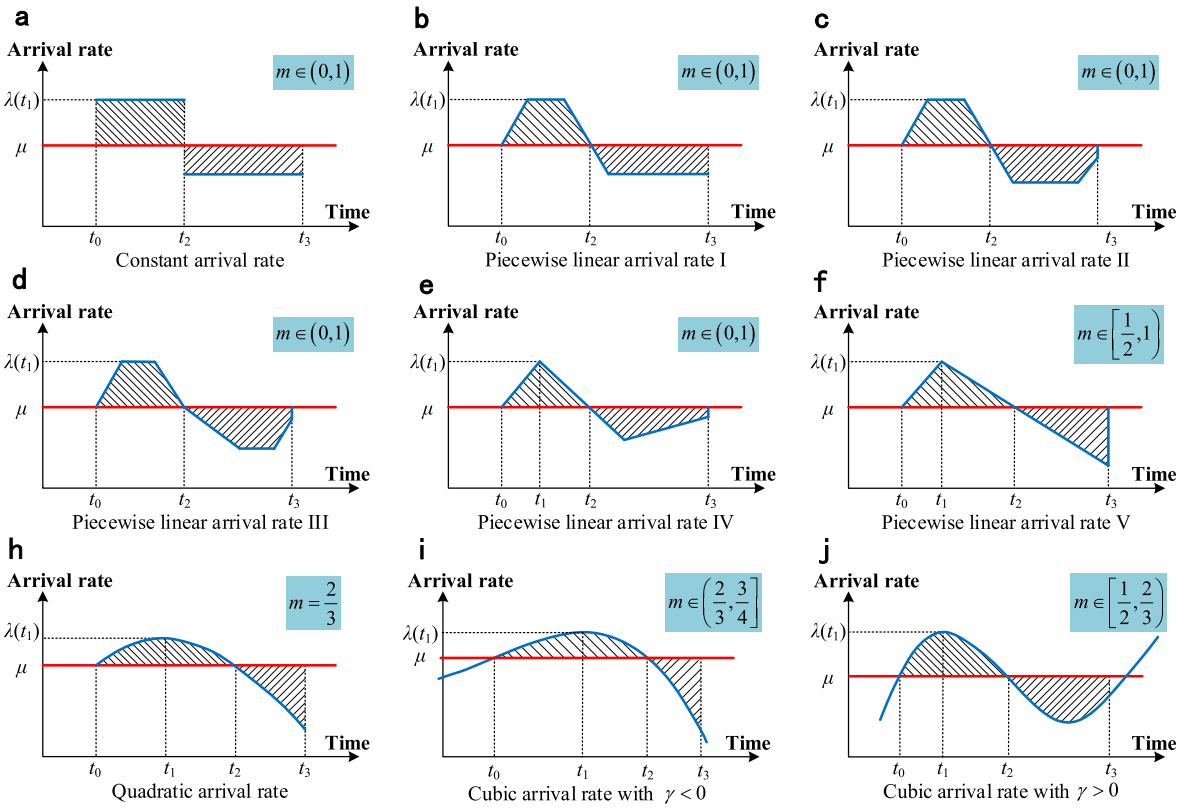


Fig. 11. Different patterns of the arrival rate function in dynamic oversaturated queueing systems. The red horizontal line is the constant discharge rate μ, and the blue line or curve is the arrival rate function $\lambda(t)$ with diverse patterns. The shadow area between t_0 and t_2 is the maximal queue length; thus, the shadow area before and after t_2 should be equal, indicating a flow conservation condition. Specifically, (a) depicts the arrival rate with step constants; (b)~(f) approximate the arrival rate by different types of piecewise linear functions. (g) draws the arrival rate by a quadratic function, and it is symmetric at $t = t_1$. The symmetric quadratic form is adopted to analyze the mildly congested traffic system in the literature (Newell, 1982). (h)-(i) approximate the arrival rate by a cubic function, which is a discussion focus in this paper. The cubic arrival rate function can be used to analyze the asymmetry of the arrival rate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

along the time horizon t_0, t_1, t_2, t_3, and the oversaturation factor m to analytically derive the resulting queueing and travel time. From the decision makers' perspective, we should not only collect more measurements to observe the queue extent and average delay as part of the system performance but also understand the root causes of traffic congestion by uncovering and identifying the underlying demand flow dynamics.

7. Conclusion

Recognizing the needs for interpretable models in traffic state estimation applications, we propose a queueing-theoretic model for oversaturated traffic systems with time-dependent demand rates, and average demand-delay function forms are established from time-dependent queueing systems based on the polynomial functional approximation for virtual arrival rates. The space-time trajectories during congestion are mapped to a set of dynamical queueing system equations with a family of polynomial-approximated time-dependent arrival rates.

For heavy congestion cases, we explicitly define the oversaturation ratio (or the queue building-up ratio) to analytically derive the system state dynamics equations with cubic arrival rate functions. With parsimonious selection of a small number of parameters, this proposed model can be easily calibrated with real-world data. Calibration results with different data sets, including the open-source PeMS data set, the open-source New York taxi data set, and the Beijing probe vehicle and sensor data set, validated the effectiveness of the proposed model. The proposed model in this paper analytically reveals the system evolution process and makes the extremely

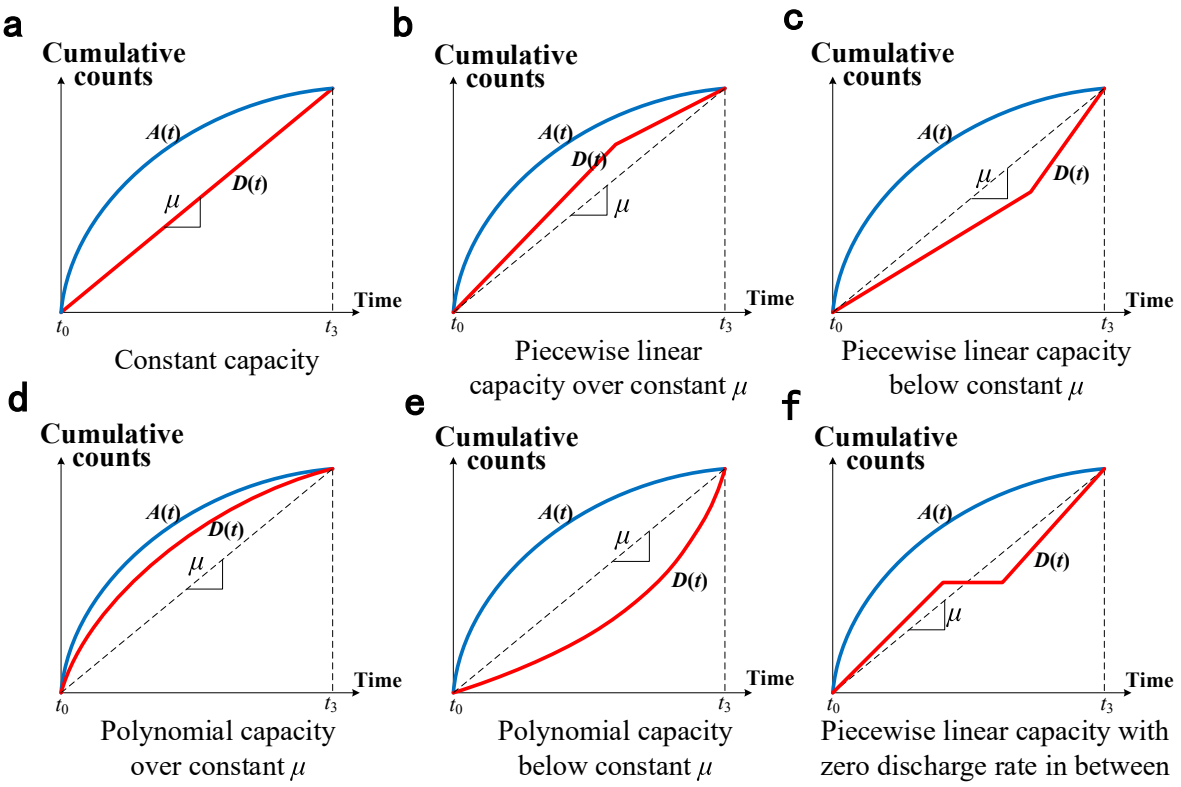


Fig. 12. Different patterns of the cumulative departure curve in dynamic oversaturated queueing systems. The red line or curve $D(t)$ is the cumulative departure curve, the blue curve $A(t)$ is the cumulative arrival curve, and the dashed line is an imaginary cumulative departure curve with a slope of μ. (a) approximates the discharge rate by a constant value of μ; thus, the cumulative departure curve is a line with a slope of μ; (b)~(c) approximate the cumulative departure curve by piecewise linear curves upon and below the dashed line with a slope of μ, respectively; (d)~(e) approximate the cumulative departure curve by polynomial curves upon and below the dashed line with a slope of μ, respectively; (f) approximates the cumulative departure curve by a piecewise linear curve with a zero-discharge rate in between. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

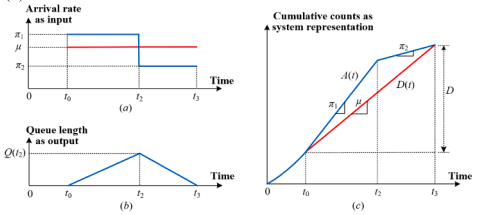
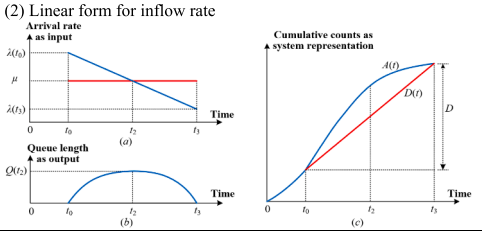
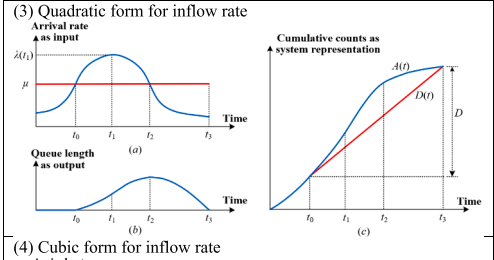
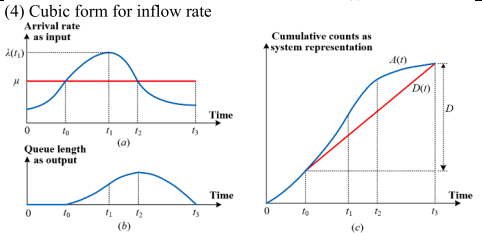
complex system more observable with approximated analytical formulations.

With this line of simplified analytical approach, decision makers can systematically interpret/explain the key queueing parameters, and further optimize an integrated set of demand- and supply-side congestion mitigation strategies for complex and oversaturated dynamic queueing systems at different scales. In the future, more advanced queueing models (e.g., Huang et al., 2016, 2017; Smith et al., 2019; Jin, 2021) should be compared with our proposed model in terms of the computational efficiency and solution accuracy. This work can be extended from a single bottleneck to a network-wide bathtub model (Jin, 2020; Vickrey, 2019, 2020) to investigate the overall system performance. In addition, automatically identifying bottlenecks is critical for the proposed methodology to be reliable and replicable in estimating parameters. As for the task of automatically identifying bottlenecks, a tentative idea is to construct the space-time speed profile (e.g., the speed profile in Fig. 5), and then identify the bottleneck location and congestion period through image recognition with deep residual learning approach (He et al., 2016).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Summary of the results for queueing system performance

Graphical illustration of queue evolution	Analytical formulation	
(1) Constant form for inflow rates 	Arrival rate	$\lambda(t)=\begin{cases}\pi_1>\mu,\ t_0\le t<t_2\\ \pi_2<\mu,\ t_2\le t\le t_3\end{cases}$
	Oversaturation factor	$m=\dfrac{\pi_2}{\pi_1}$
	Queue length	$Q(t)=\begin{cases}(\pi_1-\mu)(t-t_0),\ t_0\le t<t_2\\ (\mu-\pi_2)(t_3-t),\ t_2\le t\le t_3\end{cases}$
	Time-dependent delay	$w(t)=\begin{cases}(\pi_1-\mu)(t-t_0)/\mu,\ t_0\le t<t_2\\ (\mu-\pi_2)(t_3-t)/\mu,\ t_2\le t\le t_3\end{cases}$
	Average delay	$w=\dfrac{(\pi_1-\mu)(\mu-\pi_2)}{2\mu(\pi_1-\pi_2)}\cdot\left(\dfrac{D}{\mu}\right)$
(2) Linear form for inflow rate 	Arrival rate	$\lambda(t)=-\kappa(t-t_2)+\mu,\ \kappa>0$
	Oversaturation factor	$m=1/2$
	Queue length	$Q(t)=\dfrac{\kappa}{2}(t-t_0)(t_3-t)$
	Time-dependent delay	$w(t)=\dfrac{\kappa}{2\mu}(t-t_0)(t_3-t)$
	Average delay	$w=\dfrac{\kappa}{12\mu}\cdot\left(\dfrac{D}{\mu}\right)^2$
(3) Quadratic form for inflow rate 	Arrival rate	$\lambda(t)=-\xi(t-t_0)(t-t_2)+\mu,\ \xi>0$
	Oversaturation factor	$m=2/3$
	Queue length	$Q(t)=\dfrac{\xi}{3}(t-t_0)^2(t_3-t)$
	Time-dependent delay	$w(t)=\dfrac{\xi}{3\mu}(t-t_0)^2(t_3-t)$
	Average delay	$w=\dfrac{\xi}{36\mu}\cdot\left(\dfrac{D}{\mu}\right)^3$
(4) Cubic form for inflow rate 	Arrival rate	$\lambda(t)=\gamma(t-t_0)(t-t_2)(t-\bar{t})+\mu$
	Oversaturation factor	$m\in\begin{cases}[1/2,2/3),\ \gamma>0\\ (2/3,3/4],\ \gamma<0\end{cases}$
	Queue length	$Q(t)=\gamma\cdot(t-t_0)^2\cdot\left[\dfrac{1}{4}(t-t_0)^2-\dfrac{1}{3}\cdot\left(\dfrac{3-4m}{4-6m}+m\right)(t_3-t_0)(t-t_0)+\dfrac{1}{2}\cdot\dfrac{(3-4m)m}{4-6m}(t_3-t_0)^2\right]$
	Time-dependent delay	$w(t)=\dfrac{\gamma\cdot(t-t_0)^2}{\mu}\cdot\left[\dfrac{1}{4}(t-t_0)^2-\dfrac{1}{3}\cdot\left(\dfrac{3-4m}{4-6m}+m\right)\cdot(t_3-t_0)(t-t_0)+\dfrac{1}{2}\cdot\dfrac{(3-4m)m}{4-6m}(t_3-t_0)^2\right]$
	Average delay	$w=\dfrac{W}{D}=\dfrac{\gamma\cdot g(m)}{\mu}\cdot\left(\dfrac{D}{\mu}\right)^4$, where $m=\dfrac{t_2-t_0}{t_3-t_0}$ and $g(m)=\dfrac{1}{20}-\dfrac{1}{12}\left(\dfrac{3-4m}{4-6m}+m\right)+\dfrac{1}{6}\cdot\dfrac{(3-4m)m}{4-6m}$

Appendix B. Derivation of the time-dependent queue length

Set $u=\tau-t_0$; then, $\tau=t_0$ corresponds to $u=0$, $\tau=t$ corresponds to $u=t-t_0$, and $d\tau=du$. Therefore, we can derive the time-dependent queue length function as follows:

$$\begin{aligned}Q(t)&=\int_{t_0}^{t}[\gamma(\tau-t_0)(\tau-t_2)(\tau-\bar{t})]d\tau\\&=\gamma\cdot\int_0^{t-t_0}[u(u+t_0-t_2)(u+t_0-\bar{t})]du\\&=\gamma\cdot(t-t_0)^2\cdot\left[\frac{1}{4}(t-t_0)^2+\frac{1}{3}(2t_0-t_2-\bar{t})(t-t_0)+\frac{1}{2}(t_0-t_2)(t_0-\bar{t})\right]\end{aligned}\tag{62}$$

Because the queue will dissipate at time t_3, we can calculate t_3 by setting $Q(t_3)=0$, which gives t_3 as follows:

$$\frac{1}{4}(t_3-t_0)^2+\frac{1}{3}(2t_0-t_2-\bar{t})(t_3-t_0)+\frac{1}{2}(t_0-t_2)(t_0-\bar{t})=0 \tag{63}$$

Then, we can obtain the relationship between t_0, t_2, t_3 and $\bar{t}$ as:

$$\bar{t}-t_0=\frac{3(t_3-t_0)^2-4(t_2-t_0)(t_3-t_0)}{4(t_3-t_0)-6(t_2-t_0)} \tag{64}$$

Denote the oversaturation factor m by the ratio between the time duration from the start of congestion to the time with maximum queue length and the whole congestion duration, i.e.,

$$m=\frac{t_2-t_0}{t_3-t_0},\qquad 0<m<1 \tag{65}$$

then we can obtain the time-dependent queue length function after substituting Eqs. (64) and (65) into Eq. (62):

$$\begin{aligned}Q(t)&=\gamma\cdot(t-t_0)^2\cdot\left[\frac{1}{4}(t-t_0)^2+\frac{1}{3}(2t_0-t_2-\bar{t})(t-t_0)+\frac{1}{2}(t_0-t_2)(t_0-\bar{t})\right]\\&=\gamma\cdot(t-t_0)^2\cdot\left[\frac{1}{4}(t-t_0)^2-\frac{1}{3}\left(\frac{3(t_3-t_0)^2-4(t_2-t_0)(t_3-t_0)}{4(t_3-t_0)-6(t_2-t_0)}+t_2-t_0\right)(t-t_0)+\frac{1}{2}(t_2-t_0)\left(\frac{3(t_3-t_0)^2-4(t_2-t_0)(t_3-t_0)}{4(t_3-t_0)-6(t_2-t_0)}\right)\right]\\&=\gamma\cdot(t-t_0)^2\cdot\left[\frac{1}{4}(t-t_0)^2-\frac{1}{3}\cdot\left(\frac{3-4m}{4-6m}+m\right)(t_3-t_0)(t-t_0)+\frac{1}{2}\cdot\frac{(3-4m)m}{4-6m}(t_3-t_0)^2\right]\end{aligned} \tag{66}$$

Appendix C. Two-peak model

In some cases, the queue during the first peak period does not completely dissipate at time t_3, and the new queue during the second peak appears at time t_3. For example, in the Beijing data set analyzed in this paper, there are two lanes open to general-purpose vehicles and one lane open only to buses during 07:00 AM to 09:00 AM, while after 09:00 AM, the bus-only lane opens to general-purpose vehicles and attracts another wave of demand.

We also assume that the arrival rate at time t during each period can be approximated by a cubic polynomial function, i.e., $\lambda(t)=\sum_{i=0}^{3}\gamma_i t^i$, where γ_i are the coefficients of the i-th order variables. Denote $t_0^1(t_0^2)$, $t_2^1(t_2^2)$, $\bar{t}^1(\bar{t}^2)$, and $\mu_1(\mu_2)$ as the time that the arrival rate exceeds the discharge rate for the first (second) time, the time with a maximum queue length during the first (second) peak period, the time that the arrival rate rises to the discharge rate for the second (third) time, and the constant discharge rate during the first (second) peak period, respectively (see Fig. B1 for a detailed model illustration). It is clear that $\bar{t}^1=t_0^2$. The time-dependent arrival rate function can be approximated by the following piecewise cubic function:

$$\lambda(t)=\begin{cases}\gamma_1\left(t-t_0^1\right)\left(t-t_2^1\right)\left(t-\bar{t}^1\right)+\mu_1, & t\in\left[t_0^1,\bar{t}^1\right]\\ \gamma_2\left(t-t_0^2\right)\left(t-t_2^2\right)\left(t-\bar{t}^2\right)+\mu_2, & t\in\left[t_0^2,t_3^2\right]\end{cases} \tag{67}$$

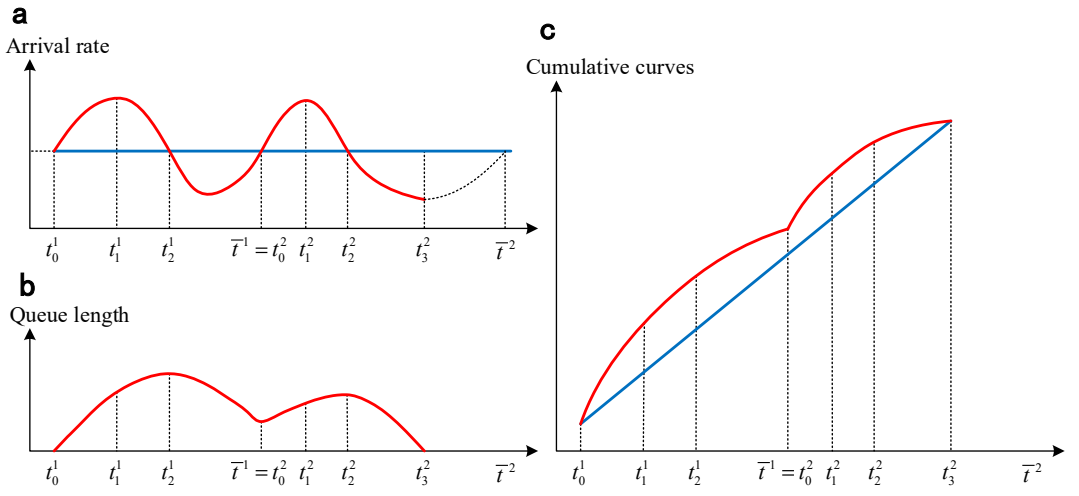


Fig. B1. Illustration of the two-peak model. (a) Illustration of the arrival rate in the two-peak model. (b) Illustration of the virtual queue length in the two-peak model. (c) Illustration of the cumulative arrival and departure curves in the two-peak model.

When $\mu_1 = \mu_2$, we can obtain the net flow function as

$$\lambda(t) - \mu = \begin{cases} \gamma_1 \left(t - t_0^1\right)\left(t - t_2^1\right)\left(t - \bar{t}^1\right), & t \in \left[t_0^1, \bar{t}^1\right] \\ \gamma_2 \left(t - t_0^2\right)\left(t - t_2^2\right)\left(t - \bar{t}^2\right), & t \in \left[t_0^2, t_3^2\right] \end{cases} \tag{68}$$

The queue length can be calculated by $Q(t) = \int_{t_0^1}^{t} [\lambda(t) - \mu] dt$. Thus, we have the queue length function as

$$Q(t) = \begin{cases} \gamma_1 \left(t - t_0^1\right)^2 \left[\frac{1}{4}\left(t - t_0^1\right)^2 + \frac{1}{3}\left(2t_0^1 - t_2^1 - \bar{t}^1\right)\left(t - t_0^1\right) + \frac{1}{2}\left(t_0^1 - t_2^1\right)\left(t_0^1 - \bar{t}^1\right) \right], & t \in \left[t_0^1, \bar{t}^1\right] \\ Q(\bar{t}^1) + \gamma_2 \left(t - t_0^2\right)^2 \left[\frac{1}{4}\left(t - t_0^2\right)^2 + \frac{1}{3}\left(2t_0^2 - t_2^2 - \bar{t}^2\right)\left(t - t_0^2\right) + \frac{1}{2}\left(t_0^2 - t_2^2\right)\left(t_0^2 - \bar{t}^2\right) \right], & t \in \left[t_0^2, t_3^2\right] \end{cases} \tag{69}$$

With the time-dependent queue length function, we can derive the time-dependent delay and the total delay functions as in Section 3.1, which are omitted here. In this two-peak model, the observed time indexes are t_0^1, $\bar{t}^1 = t_0^2$, and t_3^2, while the parameters to be calibrated are μ, γ_1, γ_2, t_2^1, t_2^2, and $\bar{t}^2$.

References

Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. J. Urban Econ. 27 (1), 111–130.

Ban, X.(., Pang, J.-S., Liu, H.X., Ma, R., 2012. Continuous-time point-queue models in dynamic network loading. Transp. Res. Part B 46 (3), 360–380.

Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D., 2011. SUMO-Simulation of Urban MObility: An Overview. IARIA SIMUL2011 Third International Conference on Advances in System Simulation.

Cassidy, M.J., Bertini, R.L., 1999. Some traffic features at freeway bottlenecks. Transp. Res. Part B 33 (1), 25–42.

CATT Lab, 2021. "RITIS CATT Lab" web page. Accessed 03/10/2021. https://www.cattlab.umd.edu/?portfolio=ritis.

Drissi-Kaïtouni, O., Hameda-Benchekroun, A., 1992. A dynamic traffic assignment model and a solution algorithm. Transp. Sci. 26 (2), 119–128.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (7).

Green, L.V., Kolesar, P.J., Whitt, W., 2007. Coping with time-varying demand when setting staffing requirements for a service system. Prod. Oper. Manag. 16, 13–39.

FHWA, 2018. Congestion and Bottleneck Identification (CBI) Software Tool User's Guide (NO. FHWA-HRT-18-071).

Hale, D., Chrysikopoulos, G., Kondyli, A., Ghiasi, A., 2021. Evaluation of data-driven performance measures for comparing and ranking traffic bottlenecks. IET Intell. Transp. Syst. 15 (4), 504–513.

Hale, D., Jagannathan, R., Xyntarakis, M., Su, P., Jiang, X., Ma, J., Hu, J., Krause, C., 2016. Traffic bottlenecks: Identification and Solutions (No. FHWA-HRT-16-064).

Han, K.e., Friesz, T.L., Yao, T., 2013a. A partial differential equation formulation of Vickrey's bottleneck model, Part I: Methodology and theoretical analysis. Transp. Res. Part B 49, 55–74.

Han, K.e., Friesz, T.L., Yao, T., 2013b. A partial differential equation formulation of Vickrey's bottleneck model, Part II: Numerical analysis and computation. Transp. Res. Part B 49, 75–93.

Han, Y.u., Ramezani, M., Hegyi, A., Yuan, Y., Hoogendoorn, S., 2020. Hierarchical ramp metering in freeways: an aggregated modeling and control approach. Transp. Res. Part C 110, 1–19.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Horowitz, A.J., 1991. Delay-volume relations for travel forecasting based upon the 1985 Highway Capacity Manual.

Huang, W., Viti, F., Tampère, C.M.J., 2016. Repeated anticipatory network traffic control using iterative optimization accounting for model bias correction. Transp. Res. Part B 67, 243–265.

Huang, W., Viti, F., Tampère, C.M.J., 2017. An iterative learning approach for anticipatory traffic signal control on urban networks. Transportmetrica B 5 (4), 402–425.

INRIX, 2020. 2019 Global Traffic Scorecard Report.

Jin, W.-L., 2015. Point queue models: a unified approach. Transp. Res. Part B 77, 1–16.

Jin, W.-L., 2020. Generalized bathtub model of network trip flows. Transp. Res. Part B 136, 138–157.

Jin, W.-L., 2021. A link queue model of network traffic flow. Transp. Sci. 55 (2), 436–455.

Johari, M., Keyvan-Ekbatani, M., Leclercq, L., Ngoduy, D., Mahmassani, H.S., 2021. Macroscopic network-level traffic models: Bridging fifty years of development toward the next era. Transp. Res. Part C 131, 103334.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kuwahara, M., Akamatsu, T., 1997. Decomposition of the reactive dynamic assignments with queues for a many-to-many origin-destination pattern. Transp. Res. Part B 31 (1), 1–10.

Lawson, T.W., Lovell, D.J., Daganzo, C.F., 1997. Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. Transp. Res. Rec. 1572 (1), 140–147.

Li, J., Fujiwara, O., Kawakami, S., 2000. A reactive dynamic user equilibrium model in network with queues. Transp. Res. Part B 34 (8), 605–624.

Mahmassani, H., Herman, R., 1984. Dynamic user equilibrium departure time and route choice on idealized traffic arterials. Transp. Sci. 18 (4), 362–384.

Mannering, F., Jones, B., Garrison, D.H., Sebranke, B., Janssen, L., 1990. Generation and assessment of incident management strategies, volume 3, Seattle-area incident management - microcomputer traffic simulation results.

Marshall, N.L., 2018. Forecasting the impossible: The status quo of estimating traffic flows with static traffic assignment and the future of dynamic traffic assignment. Res. Transp. Bus. Manag. 29, 85–92.

May, A.D., 1990. Traffic flow fundamentals. Prentice Hall, Inc., New Jersey.

Merchant, D.K., Nemhauser, G.L., 1978a. Optimality conditions for a dynamic traffic assignment model. Transp. Sci. 12, 183–199.

Merchant, D.K., Nemhauser, G.L., 1978b. A model and an algorithm for the dynamic traffic assignment problems. Transp. Sci. 12, 200–207.

Newell, G.F., 1968a. Queues with time-dependent arrival rates: I. The transition through saturation. J. Appl. Probab. 5 (02), 436–451.

Newell, G.F., 1968b. Queues with time-dependent arrival rates: II. The maximum queue and the return to equilibrium. J. Appl. Probab. 5 (03), 579–590.
Newell, G.F., 1968c. Queues with time-dependent arrival rates: III. A mild rush hour. J. Appl. Probab. 5 (03), 591–606.
Newell, G.F., 1982. Applications of queueing theory, second ed. Chapman and Hall Ltd, New York.
Nie, X., Zhang, H.M., 2005. A comparative study of some macroscopic link models used in dynamic traffic assignment. Networks Spat. Econ. 5 (1), 89–115.
Nocedal, J., Wright, S., 2006. Numerical optimization. Springer Science & Business Media.
Pelikan, M., Goldberg, D.E., Cantú-Paz, E., 1999. BOA: The Bayesian optimization algorithm. In: Proceedings of the genetic and evolutionary computation conference GECCO-99, vol. 1, pp. 525–532.
Ramezani, M., Haddad, J., Geroliminis, N., 2015. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. Transp. Res. Part B 74, 1–19.
Seo, T., Bayen, A.M., Kusakabe, T., Asakura, Y., 2017. Traffic state estimation on highway: A comprehensive survey. Annu. Rev. Control 43, 128–151.
Smith, M., Huang, W., Viti, F., Tampère, C.M.J., Lo, H.K., 2019. Quasi-dynamic traffic assignment with spatial queueing, control and blocking back. Transp. Res. Part B 122, 140–166.
Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. Adv. Neural Inf. Process. Syst. 25.
Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. 4 (2), 26–31.
Vickrey, W., 1969. Congestion Theory and Transport Investment. Am. Econ. Rev. 59, 251–260.
Vickrey, W., 2019. Types of Congestion Pricing Models. Econ. Transp. 20, 1–3.
Vickrey, W., 2020. Congestion in midtown Manhattan in relation to marginal cost pricing. Econ. Transp. 21, 1–6.
Zhou, X., Taylor, J., 2014. DTAlite: A queue-based mesoscopic traffic simulator for fast model evaluation and calibration. Cogent Eng. 1 (1), 961345.