A Meso-to-Macro Cross-Resolution Approach for Connecting Polynomial Arrival Queue Model to Volume-Delay Function with Inflow demand-to-Capacity Ratio

Xuesong (Simon) Zhou [a]*, Qixiu Cheng [a], Xin Wu [a]*, Peiheng Li [b], Baloka Belezamo [c], Jiawei Lu [a], Mohammad Abbasi [a]

[a] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, USA.

[b] Norfolk Southern Corporation, Atlanta, GA, USA.

[c] Arizona Department of Transportation, Phoenix, AZ, USA.

* *Corresponding authors.*

Email: xzhou74@asu.com (X. Zhou), qcheng15@asu.edu (Q. Cheng), xinwu3@asu.edu (X. Wu), peiheng.li@nscorp.com (P. Li), bbelezamo@azdot.gov (B. Belezamo), jiaweil9@asu.edu (J. Lu). Mohammad_abbasi@asu.edu (M., Abbasi)

Abstract: Although the macroscopic volume-delay function (VDF) has been widely used in static traffic assignment for transportation planning, the planning community has long recognized its deficiencies as a static function in capturing traffic flow dynamics and queue evolution process. In the existing literature, many queueing-based and simulation-based dynamic traffic assignment (DTA) models involving various traffic flow parameters have been proposed to capture traffic system dynamics on different spatial scales; however, how to calibrate these DTA models could still be a challenging task in its own right, especially for real-world congested networks with complex traffic dynamics. By extending the fluid-based polynomial arrival queue (PAQ) model with quadratic inflow rates proposed by Newell (1982) and cubic inflow rates by Cheng et al. (2022), this paper attempts to propose a cross-resolution Queueing-based Volume-Delay Function (QVDF) to explicitly establish a coherent connection between (a) the macroscopic average travel delay performance function in a long-term planning horizon and (b) the mesoscopic dynamic queuing model during a single oversaturated period. By introducing two types of elasticity functional forms, this paper develops a relationship from the macroscopic inflow demand-to-capacity (D/C) ratio to the congestion duration of a bottleneck, from the congestion duration to the magnitude of speed reduction. The QVDF can be directly utilized to provide closed-form expressions for both average travel delay performance and the time-dependent speed profiles. The proposed cross-resolution QVDF provides a numerically reliable and theoretically rigorous performance function to characterize oversaturated bottlenecks at both macroscopic and mesoscopic scales.

Keywords: Mesoscopic to macroscopic modeling; Multi-resolution approach; Time-dependent delay; Polynomial arrival queue model; Volume-delay function; Travel time function

1 Introduction

Multi-Resolution Modeling (MRM) is a modeling technology that creates a family of models representing the same phenomenon or a set of questions through seamlessly integrated models with different resolutions. In transportation, the fine-grained spatial scales could cover networks, important corridors, specific roads, and detailed lane representations, and the temporal resolution refers to the time intervals (or time stamps) typically ranging from days to seconds, at which traffic states are evolved dynamically.

Many important studies have been devoted to multi-resolution modeling efforts. The pioneering work by Gazis et al. (1959) first highlighted that the fundamental diagram (e.g., Greenshields' speed-density relationship) can be linked to the microscopic car-following models (Greenshields et al., 1935; Gazis et al., 1961). The significant study by Daganzo (2006) proved that, by assuming a triangular flow-density diagram, vehicle trajectories constructed from a simplified kinematic wave model are equivalent to those generated by Newell's simple linear car-following model (Newell, 2002) and two types of cellular automata models within a certain approximation range (Nagel, 1996). A recent effort along this line includes an s-shaped three-parameter (S3) speed-density function by Cheng et al. (2021) with a macro-to-micro consistent car-following model. The latest reports by Zhou et al. (2021) and Hadi et al. (2022) provided a systematic review of MRM terminology, tools, literature review and representative case studies in traffic analysis applications. This paper attempts to propose a queueing-theoretic cross-resolution model to bridge the gap between macroscopic and mesoscopic representations.

1.1 Necessity for cross-resolution modeling to integrate static and dynamic performance functions

Large-scale network flow models are critical to assess the performance of transportation systems at different spatial scales. The models are also important tools to measure the proficiency of traffic systems at various temporal resolutions. Static Traffic Assignment (STA) is the most widely used network flow model to load OD demands and evaluate the macroscopic performance of a traffic system. On the other hand, Dynamic Traffic Assignment (DTA) has been recognized as the key building block in representing traffic dynamics under congested conditions. Many agent-based simulation tools with mesoscopic traffic flow models have been developed to capture network dynamics and queue evolution process at a finer resolution, e.g., DYNASMART (Mahmassani et al., 1992), DynaMIT (Ben-Akiva et al., 1998), etc.

The most essential functions used in STA are the volume-delay functions (VDF). Among a variety of VDFs, the Bureau of Public Roads (BPR) function, created by the US Bureau of Public Roads in 1964, plays an important role in system-wide performance evaluation. The BPR function is a simple polynomial equation to relate the relationship between demand and delay. Nonetheless, the planning community has long recognized that the static BPR function can only provide average travel time measures and is unable to capture traffic dynamics at an oversaturated bottleneck.

In DTA models, many extended queueing models are widely used to describe system dynamics with time-dependent arrival and discharge rates[1]. The early effort by Vickrey (1969) used a road pricing-oriented congestion-eliminating approach by representing traffic bottlenecks as a fluid-based point queue model. More sophisticated queueing-based models include Newell's simplified kinematic wave model (Newell, 1993a, 1993b, 1993c) that keeps track of shock wave and queue propagation using cumulative flow counts on links and cell transmission model (Daganzo, 1994, 1995a, 1995b) that adopts a "supply-demand" or

[1] In this study, the terminology of "arrival rate" is interchangeable with "inflow rate".

"sending-receiving" framework to model flow dynamics between discretized cells. Readers are referred to the book by Newell (1982) as well for more interesting applications of queueing systems. Besides, there are also many partial-differential-equation-based numerical analyses and customized simulation packages to model sophisticated interactions between the demand and supply (Behrisch et al., 2011; Marshall, 2018) at a fine resolution. In this paper, we select the term "mesoscopic" to indicate the DTA model. The microscopic simulation models capturing individual driver maneuvers, e.g., car-following, overtaking, lane changing, and gap acceptance behaviors are viewed as somewhat being distinct from the mesoscopic models, which provide a hybrid approach to modeling traffic propagation (e.g., VISSIM, Paramics, Aimsun, TRANSIMS).

There are multiple impediments associated with a loose linkage or potential inconsistency between macroscopic and mesoscopic network flow models. For example, the results of the macroscopic traffic assignment models along with the resulting link travel times could be significantly different from the link performance statistics at the mesoscopic level in terms of the low-resolution traffic flow models (e.g., spatial queue or simplified kinematic wave models used in DTA). Moreover, this inconsistency leads to an internal discrepancy of different modeling approaches and hinders tighter interconnections between different simulation/assignment components. As a result, it is necessary to provide a reasonable approximation to the complex real-world traffic flow dynamics by mixing models of the two resolutions.

Another reason to develop a cross-resolution modeling framework is that the simulation-based DTA models could be computationally intensive, especially for real-world networks with complex dynamics. In the existing literature, there are two lines of research developed to address such computational challenges.

1. A *coarse simulation approach*. For example, "Cellular Automata" (CA) has been applied to keep up with the fast computational speed necessary to simulate a whole region. Newell's simplified kinematic wave and linear car following models were also introduced by Zhou et al. (2015) for estimating emissions and fuel consumption efficiently.
2. *Parallel computing* is also adopted to utilize multiple computer processors in support of a large number of travelers on a large-scale transportation network (e.g., TRANSIMS micro-simulator). A recent effort in DTALite (Zhou and Taylor, 2014; Qu and Zhou, 2017) refer to a hybrid time-based and event-based data structure of parallel computation for mesoscopic DTA simulation.

The cross-resolution modeling approach in this paper is a further effort on the third one. It should be remarked that, if a simulation-based DTA tool is used at the mesoscopic link level, we need to clearly understand the complexity in modeling the time-dependent queue discharge rates and inflow patterns so as to reliably reproduce the queue evolution process. The ultimate hourly capacity from the macroscopic model, if adopted as the default queue discharge rate, could be a significant overestimate of its true value. If the affected queue discharge rate is modeled internally through the spatial queue propagation, then major efforts are still needed to (1) calibrate the queue discharge rate at the downstream location of each bottleneck, and (2) obtain precise time-dependent inflow patterns at different incoming links upstream of a bottleneck. The latter is in turn mainly determined by the complex time-varying OD matrix and route choice behaviors.

1.2 Objectives and contributions

Deploying a cross-resolution modeling approach, ensuring consistency across different resolutions, and addressing concerns and barriers in a full-scale implementation, are important and challenging tasks for researchers and practitioners. We would like to explicate the contributions of this research by answering the following two theoretically important questions:

1. How do we develop a simplified (cross-resolution) link performance function, which not only can be used in evaluating macro-level traffic assignment tasks but also has the capability of reflecting meso-level traffic characteristics under typical arrival flow patterns and congestion levels?
2. How do we efficiently derive meso-level (time-dependent) speed and queue length profiles which are consistent with the macroscopic link volume and average link performance from STA?

In this paper, we propose a cross-resolution link performance function, namely queueing-based volume delay function (QVDF) to connect mesoscopic Polynomial Arrival Queue (PAQ) models to macroscopic VDFs in terms of Inflow demand-to-Capacity (D/C) Ratio. The QVDF is explicitly designed to provide straightforward computational formulas for measuring both system-wide average performance over multiple days/years and time-dependent performance measures within a single oversaturated period. Its simplicity and tractability help overcome the computational challenges, which opens a new window to model and analyze dynamic traffic systems more efficiently and effectively.

The remainder of this paper is organized as follows. Section 2 systematically summerizes the building blocks of QVDF: classic VDF and PAQ Models. Section 3 introduces the fundamental concepts of the queueing-based volume-delay function (QVDF) to bridge the meso-to-macro gaps and further explores the connection between the D/C ratio in VDFs and curvature parameters of the time-dependent profile in PAQ models. With the QVDF, Section 4 connects the PAQ models and VDF models via a set of elasticity terms approximating the overall queue evolution process and establishing the meso-to-macro relationship. Two real-world case studies and their calibration results are presented in Section 5, which is followed by a range of discussions on its applications in Section 6.

2 Existing Polynomial Arrival Queue models and Volume Delay Functions

2.1 Literature review on volume-delay and link performance functions

To evaluate the link travel time performance, various VDFs were proposed at the beginning of the transportation planning discipline. The first seminal work was conducted by CATS (acronym of Chicago Area Transportation Study, 1960) in the early 1960s for the traffic assignment problem and the CATS function was explicitly stated in Muranyi (1963). This CATS function is expressed in terms of volume-to-capacity (V/C) ratio. Smock (1962, 1963) derived his VDF with an exponential function based on 'mathematical logic and trial-and-error experimentation (Boyce and Williams, 2015). Different from the CATS function, where the travel time equals the free flow time at zero volume and doubles when the volume reaches the capacity, the travel time in Smock's function is the same as the free-flow travel time when the volume is at capacity and 0.37 times of free-flow travel time at zero volume.

In 1964, the US Department of Commerce published *Traffic Assignment Manual*, where the classic BPR (1964) function was developed for the capacity restraint traffic assignment problem. The travel time obtained from the BPR function equals the free-flow travel time at zero volume and increases 15% when the volume reaches capacity. How to calibrate the parameters in BPR could be challenging since their values could significantly vary from region to region (Wu et al., 2020). Spiess (1990) proposed a set of conditions for "well behaved" VDFs to ensure that the original form and its first-order gradient are strictly increasing, and the change of congestion effects is reasonable when the capacity is reached. He recommended a collection of conical congestion functions and identified further research needs to directly estimate the parameters using observed speeds and volumes.

Davidson (1966) derived a VDF using stochastic queueing theory without a clear demonstration. This gap was filled and explained by Davidson (1978) in detail. Davidson's function uses the saturation flow rather than the capacity to calculate the travel

time and restrains the link volume below the capacity since the travel time approaches infinity when the volume approaches the capacity in his formulation. To model the travel time when the volume is near or over the capacity, a variety of modified Davidson's forms have been proposed. Akçelik (1978) modified Davidson's function by extending the function's slope linearly to yield a finite travel time under oversaturated conditions for user equilibrium and system optimality of traffic assignment problems. Tisato (1991) developed an alternative form based on Akçelik's (1978). The link performance function is time-dependent and influenced by congestion duration under oversaturated conditions. As Tisato's modification is shown to overpredict the travel time for flow near and above the capacity, Akçelik (1991) proposed an alternative time-dependent form using the coordinate transformation technique, which can be also used for intersection delay functions. Ran and Boyce (1997) used the average flow-to-capacity ratio, instead of the simplistic BPR function, to estimate link travel times and intersection delays for most types of links and intersections.

The link travel time performance functions, with respect to the number of vehicles on the link, the inflow and/or the outflow rate at a certain time, can be used in DTA problems as well. Ran et al. (1993) and Friesz et al. (1993) investigated the instantaneous DTA problem through the instantaneous link travel time functions. Carey and McCartney (2002) derived analytical solutions for travel times and outflows with a linear travel time model. Daganzo (1995) suggested that the travel time function should only depend on the number of vehicles on the link without the inflow or outflow rates. However, if the inflow and outflow rates are omitted, as claimed in Carey et al. (2003), the obtained travel times would be unrealistic since the link travel time is independent of the traffic distribution along with the link. Some other properties of the link performance functions, including uniqueness, continuity, causality, consistency, and first-in-first-out (FIFO), are discussed on the link travel time functions (Carey, 2004). Nie and Zhang investigated various FIFO-consistent link travel time functions and revealed the following observations (Nie and Zhang, 2005a, 2005b; Nie et al., 2008). Carey et al. (2014) extended the link-travel-time-based DTA models to ensure FIFO and capacity constraints, and strengthen the realism and behavioral dynamics. It should be noted that the piece-wise linear travel time functions would violate the FIFO property. The smooth and convex travel time function bounded by the linear and piece-wise linear functions are FIFO-consistent only for certain kinds of inflow profiles.

Many practitioners have the following questions related to the widely used VDFs:

1. Can we interpret the parameters in static VDFs (used in a long-term planning horizon) from a dynamic queueing perspective typically focusing on a single oversaturated period?
2. How can we fully utilize the time-dependent traffic observations over multiple days/years to calibrate the parameters of the VDF and reveal the underlying dynamic information?

Aiming to shed light on the above two questions, this next subsection presents key assumptions for a new cross-resolution travel time performance model, which is intended to establish a connection between the continuous-time fluid-based polynomial arrival queue (PAQ) model within period congestion, and the system-wide VDFs over a wide range of traffic intensity conditions.

2.2 Early efforts of Newell's fluid polynomial queueing model with quadratic arrival rates

This section introduces Newell's continuous-time fluid-based polynomial arrival queue (PAQ) model. Table 1 summarizes the notations used in the PAQ model and throughout this study.

Table 1: Symbols and definitions used in this study.

Symbols	Definitions
L	link length
t_0	start time of congestion period
t_1	time index with maximum inflow rate
t_2	time index with maximum queue length
t_3	end time of congestion period
μ	capacity (or discharge rate), assumed to be a constant value
D	total in-flow demand during the whole peak period
C	lane-based ultimate hourly capacity
V	total lane-based volume loaded on a road link during an analysis period (i.e., AM, MD, PM, or NT)
t_f	free-flow travel time
v_μ	congestion speed
v_{co}	cut-off speed[2]
v_{t2}	lowest speed on a link
γ	inflow curvature parameter used in polynomial form
$\lambda(t)$	inflow rate function at time t
$A(t)$	cumulative inflow rate at time t
$D(t)$	cumulative discharge rate at time t
$Q(t)$	queue length at time t
$w(t)$	traffic delay departing at time t
$\bar{w}$	average delay during the whole peak period
tt	average travel time during the whole peak period
α, β	parameters in the BPR-form link performance function
f_d	constant in elasticity function for mapping D/C ratio to congestion duration
n	elasticity coefficient of congestion duration in response to D/C changes, i.e., oversaturation-to-duration elasticity
f_p	constant in elasticity function for mapping congestion duration to the magnitude of speed reduction
s	elasticity coefficient of speed reduction magnitude in response to congestion duration changes, i.e., duration-to-speed reduction elasticity

[2] As the cut-off speed can be used more systematically to distinguish "congested" vs. "uncongested" states of traffic bottleneck (Hale, et al. 2016).

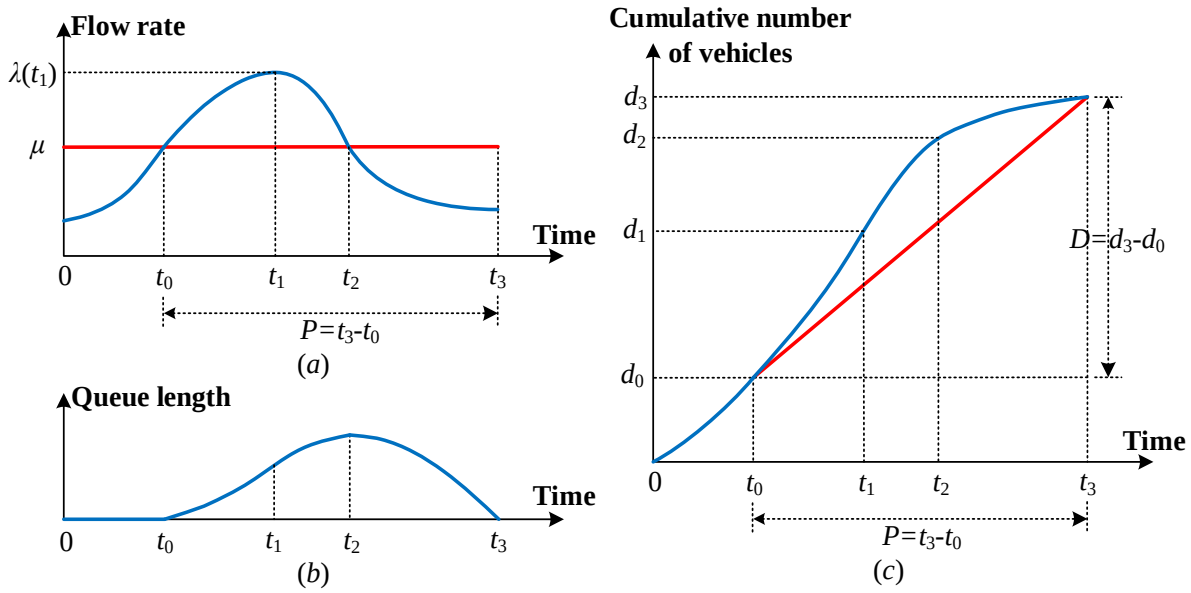


Fig. 1: Graphical illustration of Newell's PAQ model for a single congested period (Newell, 1982).

Fig. 1 illustrates the notations in Newell's model. Let t_1 be the time point with the maximum inflow rate (see Fig. 1(*a*)). Newell assumed that the inflow rate at time t_1 could be approximated by the quadratic Taylor expansion (Newell, 1982):

$$\lambda(t) = \lambda(t_1) + \lambda'(t_1) \cdot (t - t_1) + \frac{1}{2}\lambda''(t_1) \cdot (t - t_1)^2. \tag{1}$$

Given $\lambda'(t_1) = 0$, let $\gamma = -\frac{1}{2}\lambda''(t_1)$, then Eq. (1) can be transformed to

$$\lambda(t) = \lambda(t_1) - \gamma \cdot (t - t_1)^2. \tag{2}$$

Since γ describes the curvature or shape of the time-dependent inflow arrival rates, we term it as the inflow curvature parameter.

By assuming a constant discharge rate within a single queue duration, the queue discharge rate or service rate μ(where $\mu = \lambda(t_0) = \lambda(t_2)$), as shown in Fig. 1(*a*), can be estimated in terms of Eq. (2):

$$\mu = \lambda(t_1) - \gamma \cdot (t_0 - t_1)^2 = \lambda(t_1) - \gamma \cdot (t_2 - t_1)^2. \tag{3}$$

Then the two real roots, t_0 and t_2, can be obtained as follows:

$$t_0 = t_1 - \left[\frac{\lambda(t_1) - \mu}{\gamma}\right]^{\frac{1}{2}}, \qquad t_2 = t_1 + \left[\frac{\lambda(t_1) - \mu}{\gamma}\right]^{\frac{1}{2}}. \tag{4}$$

Now we can write $\lambda(t) - \mu$ in a factored form:

$$\lambda(t) - \mu = \gamma \cdot (t - t_0) \cdot (t_2 - t). \tag{5}$$

As $\lambda(t) = \mu + \gamma \cdot (t - t_0) \cdot (t_2 - t)$, it is obvious that $\lambda(t)$ passes (t_0, μ) and (t_2, μ). Then its second derivative $\lambda(t)'' = -2\gamma$, is consistent with Eq. (2). The virtual queue length at time t which equals $A(t) - D(t)$ and can be obtained:

$$Q(t) = A(t) - D(t) = \int_{t_0}^{t} [\lambda(\tau) - \mu]\, d\tau. \tag{6}$$

By substituting Eq. (5) into Eq. (6), the queue length at time t is expressed in terms of t_0, t_2, and γ:

$$Q(t) = \gamma \cdot (t - t_0)^2 \cdot \left[\frac{t_2 - t_0}{2} - \frac{t - t_0}{3}\right]. \tag{7}$$

The maximum queue length achieved at time t_2 is:

$$Q(t_2) = \frac{\gamma}{6} \cdot (t_2 - t_0)^3 = \frac{4[\lambda(t_1) - \mu]^{3/2}}{3\gamma^{1/2}}. \tag{8}$$

The queue will dissipate at time t_3, i.e., $Q(t_3) = 0$, then we can obtain t_3:

$$t_3 = t_0 + \frac{3}{2}(t_2 - t_0) = t_0 + 3(t_1 - t_0). \tag{9}$$

Therefore, Eq. (7) can also be written as follows:

$$Q(t) = \frac{\gamma}{3}(t - t_0)^2(t_3 - t). \tag{10}$$

The inflow curvature parameter γ determines the shape of queueing function with respect to *t*, the shape of derived speed, and delay profiles. The total delay between the time t_0 and t_3 can also be calculated by the area between $A(t)$ and $D(t)$ in Fig. 1(c) through integrating Eq. (10):

$$W = \frac{\gamma}{36}(t_3 - t_0)^4 = \frac{9[\lambda(t_1) - \mu]^2}{4\gamma}. \tag{11}$$

Above is the introduction of Newell's method based on the assumption of the quadratic inflow rate. With the total delay in Eq. (11), we can further derive the average delay during the congestion period from t_0 to t_3:

$$\bar{w} = \frac{W}{D} = \frac{\gamma}{36D}. \tag{12}$$

where D is the *total inflow demand volume* from t_0 to t_3. Denote the peak period as P (i.e., $P = t_3 - t_0$), then the discharge rate (or effective capacity) can be represented by D and P:

$$\mu = \frac{D}{P}. \tag{13}$$

Substituting Eq. (13) into Eq. (12) leads to the following average delay between t_0 and t_3:

$$\bar{w} = \frac{W}{D} = \frac{\gamma}{36\mu} \cdot \left(\frac{D}{\mu}\right)^3. \tag{14}$$

The quadratic model is only applicable to analyze mild traffic conditions. When the system is oversaturated, the arrival rate from the quadratic inflow model might be a (counterintuitive) negative value. Recently, Cheng et al. (2022) revisited Newell's model and described queueing systems with cubic time-dependent arrival rates. This cubic model is ideal for efficient dynamic modeling and management as it can analytically calculate the time-dependent queue length, delay, and travel time, which is introduced in the next subsection.

2.3 Latest research on fluid polynomial queueing model with cubic arrival rates

The general framework for the cubic PAQ model can be found in Cheng et al. (2022). Here, we only summarize its key points. Considering a single bottleneck where the time-dependent arrival rate can be approximated by a cubic polynomial function and the

discharge rate is assumed to be constant during a period of interest. Similar to Newell's model, the difference between the arrival rate $\lambda(t)$ and discharge rate μ, is expressed by a factored form as $\lambda(t) - \mu = \gamma(t - t_0)(t - t_2)(t - \bar{t})$, where $\bar{t}$ is an auxiliary time point representing another root in addition to t_0 and t_2, and γ is the inflow curvature parameter to be calibrated. All notations follow Newell's model. **Fig. 2** illustrates the flow rates, queue length, and cumulative flow rate evolution process over time in the cubic model (Cheng et al., 2022).

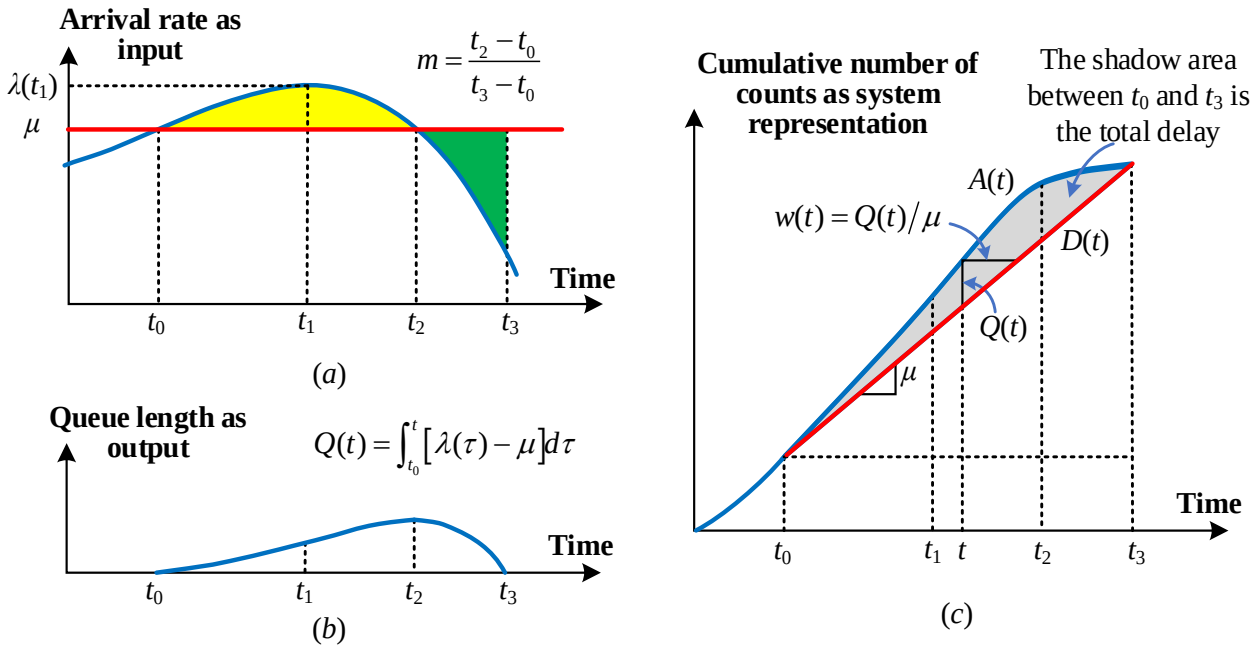


Fig. 2: Illustration of (*a*) the arrival and discharge rates, (*b*) the mesoscopic queue evolution process, and (*c*) the cumulative arrival and departure flow counts (Cheng et al., 2022).

Given the assumptions on the cubic arrival rate and constant discharge rate, the time-dependent queue length $Q(t)$, we will have the time-dependent delay $w(t)$, the average delay $\bar{w}$, and the average travel time tt as follows (Cheng et al., 2022):

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

$$w(t) = \frac{\gamma(t - t_0)^2}{\mu} \left[\begin{array}{c} \frac{1}{4}(t - t_0)^2 - \frac{1}{3}\left(\frac{3 - 4m}{4 - 6m} + m\right) \cdot (t_3 - t_0) + \\ \frac{1}{2}\frac{(3 - 4m)m}{4 - 6m} \cdot (t_3 - t_0)^2 \end{array} \right], \tag{16}$$

$$\bar{w} = \frac{W}{D} = \frac{\gamma \cdot g(m)}{\mu} \cdot \left(\frac{D}{\mu}\right)^4, \tag{17}$$

$$tt = t_f + \bar{w} = t_f \left[1 + \frac{\gamma \cdot g(m)}{\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^4\right], \tag{18}$$

where

$$\gamma > 0, \tag{19}$$

$$m = \frac{t_2 - t_0}{t_3 - t_0}, \frac{1}{2} \le m < \frac{2}{3}, \tag{20}$$

$$g(m) = \frac{1}{20} - \frac{1}{12}\left(\frac{3-4m}{4-6m} + m\right) + \frac{1}{6}\frac{(3-4m)\cdot m}{4-6m}, g(m) \ge \frac{1}{120}. \tag{21}$$

Similar to Newell's quadratic model, inflow curvature parameter γ determines the shapes of arrival rates and queueing functions Table 2 summarizes the functional forms and key parameters of travel delay functions under different arrival rate patterns (Cheng et al., 2022). A common feature is that they are all functions of inflow demand volume *D*, discharge rates μ, and curvature parameters γ. The inflow curvature parameters vary over polynomial arrival rate functions of different orders, namely, π_1 and π_2 for the constant form, κ for the linear form, and γ for quadratic and cubic forms.

In this paper, we will focus on the cubic form arrival rate function and its inflow curvature parameter γ. Note that $\gamma < 0$ is only applicable to mild traffic conditions and the model reduces to Newell's quadratic model when $\gamma = 0$. We recommend $\gamma > 0$, and it applies to both mild and oversaturated conditions.

Table 2: Average travel delay functions for bottlenecks based on constant discharge rate μ (Cheng et al., 2022).

Forms	Arrival rate functions	Average travel delay functions for bottlenecks
Constant form	$\lambda(t) = \begin{cases} \pi_1 > \mu, t_0 \le t < t_2 \\ \pi_2 < \mu, t_2 \le t \le t_3 \end{cases}$	$\bar{w} = \frac{(\pi_1 - \mu)\cdot(\mu - \pi_2)}{2\mu(\pi_1 - \pi_2)} \cdot \left(\frac{D}{\mu}\right)$
Linear form	$\lambda(t) = -\kappa(t - t_2) + \mu, \kappa > 0$	$\bar{w} = \frac{\kappa}{12\mu} \cdot \left(\frac{D}{\mu}\right)^2$
Quadratic form	$\lambda(t) = -\gamma(t - t_0)(t - t_2) + \mu, \gamma > 0$	$\bar{w} = \frac{\gamma}{36\mu} \cdot \left(\frac{D}{\mu}\right)^3$
Cubic form	$\lambda(t) = \gamma(t - t_0)(t - t_2)(t - \bar{t}) + \mu$	$\bar{w} = \frac{\gamma \cdot g(m)}{\mu} \cdot \left(\frac{D}{\mu}\right)^4$

3 Framework to derive Queueing-based Volume-Delay Function (QVDF)

This section introduces fundamental concepts of the QVDF. The critical point to develop such a function is to connect two important variables: inflow demand-to-capacity (D/C) ratio and inflow curvature parameter γ in PAQ models, via a set of intermediate variables (e.g., discharge rate and congestion duration).

3.1 Basic concepts of Queueing-based Volume-Delay Function (QVDF)

3.1.1 Volume-to-capacity (V/C) ratio vs. inflow demand-to-capacity (D/C) ratio

From the perspective of macroscopic stationary-state analysis, traffic demand is usually defined as vehicles planning to pass a roadway section (i.e., inflow demand). If the demand does not exceed the capacity, the measured flow rate in the field is the traffic demand rate (e.g., within a one-hour window). However, if the traffic demand rate exceeds the capacity, inflow demand will be heldand its true value is usually unobservable. The measured flow rate only represents the discharge rate at or under the capacity.

Understanding the inflow demand requires key insights from some existing studies to distinguish "volume" and "demand". Huntsinger and Rouphail (2011) stated that the V/C ratio should be replaced with demand-over-capacity by assuming: *demand* =

capacity + queue length. Mtoi and Moses (2014) computed the "demand above capacity" as *demand = capacity + (capacity – measured flow).* Dowling et al. (2016) also pointed out that the extra demand is the presence of queueing at a specific location. Focusing on the inability of the standard BPR function to accommodate traffic exceeding capacity, Small (1983) developed a duration-dependent function to quantify the average travel delay over a "congestion duration" and found it well approximates the travel delay pattern during the afternoon peak on an 11-mile freeway segment in San Francisco Bay area. By utilizing both flow and speed data from MAG, Wu et al. (2020) defined the D/C ratio based on the speed at capacity and congestion duration. Belezamo (2020) further examined the trade-offs among the aforementioned demand definitions.

In this paper, we define period volume and inflow demand as follows:

1. Period volume (V) is the total lane-based volume loaded on a road link during an analysis period.
2. Inflow demand (D) is the queued volume or queued demand which represnts the total volume with travel speed under a specific cut-off speed (v_{co}).

The time period when speed is slower than v_{co} is then defined as "congestion duration". For instance, if $v_{co} = 45\text{miles/hour}$, the inflow demand D is the total volume within the "congestion duration" with a speed lower than 45 miles/hour. We introduce a queued demand factor (QDF) to covert V to D, which represents the percentage of congested flows within the entire analysis period:

$$D = V \cdot QDF, \tag{22}$$

where $D \leq V$ and $0 \leq QDF \leq 1$.

The hourly maxium flow rate per lane when the level of service is under E (Branston, 1976; HCM, 2010) is used in this study as the "ultimate hourly capacity" in D/C.. The reciprocal of QDF is similar to the hour-to-period factor in Wu et al. (2020) to convert the ultimate hourly capacity to the period-based capacity C_p for the entire assignment period (e.g. AM or PM):

$$\frac{D}{C} = \frac{V \cdot QDF}{C} = \frac{V}{C_p}. \tag{23}$$

3.1.2 Congestion duration, cut-off-speed, and average discharge rate

One critical concept shared by both the D/C ratio and PAQ models is "congestion duration". In PAQ models, congestion duration P indicates the peak period from t_0 to t_3 (i.e., $P = t_3 - t_0$), where arrival rate $\lambda(t)$ is higher than the average discharge rate μ. According to the definition, μ should be lower than the ultimate capacity C. Noted that congestion duration P is dependent on two correlated factors: the time-dependent arrival rate and the average discharge rate.

Compared to the V/C ratio in the BPR function, the D/C ratio provides a queue-theoretic measure consistent with the congestion dynamics. As shown in Fig.3, μ and D/C can be connected P, i.e., $D = P\mu$ where P is assumed to be an exogenous parameter in our model. In the future study, the inflow demand model could be extended to generate P using departure time choices and μ can be extended to incorporate dynamic patterns, i.e., $\mu(t)$ depending on the loading and unloading process of traffic. More details regarding the dynamic aspects of the macroscopic fundamental diagram (MFD) can be found in Mahmassani et al. (2013).

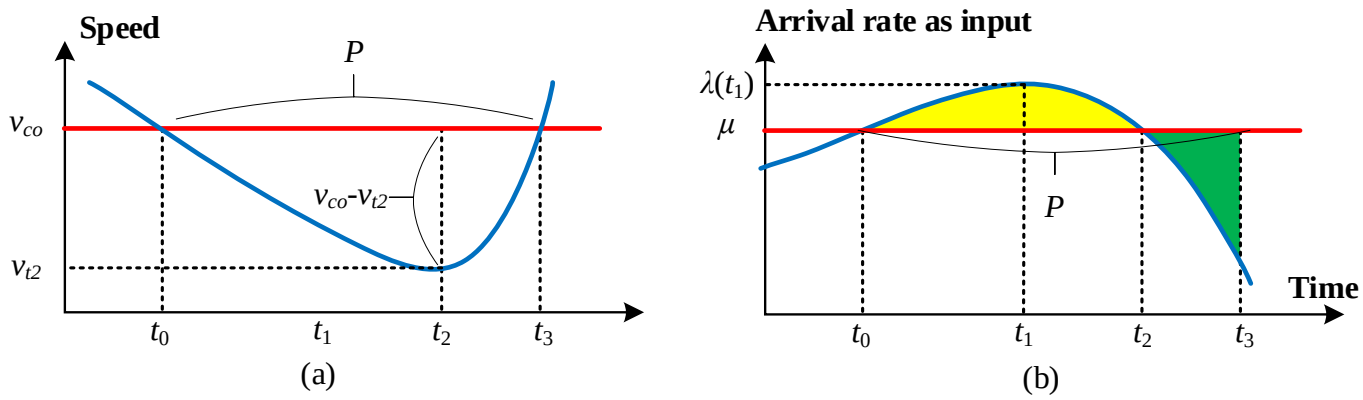


Fig. 3: The relationship among time-dependent speed, arrival rates, cut-off speed, and average discharge rate

3.2 Outline of deriving QVDF

3.2.1 Major steps towards queue-oriented link performance function QVDF

We propose the following steps to obtain the queue-based VDF for each link or each facility type using the modeling elements mentioned in Sections 2 and 3.1.

Step 1. [**D/C ratio –> congestion duration]** Introduce a power function with oversaturation-to-duration elasticity parameter n to describe the relation between D/C ratio and P. This power function can be used to estimate P with D/C after calibration.

Step 2. [**Congestion duration –> Speed reduction ratio**] Introduce a power function with duration-to-speed reduction elasticity parameter s to map congestion duration and speed reduction ratio. The ratio is defined by the cut-off-speed v_{co} and the lowest speed v_{t2}, and v_{t2} is directly related to the longest delay in the queue, as shown in Fig. 6. v_{t2} and the longest waiting time w_{t2} can be estimated with P via the calibrated power function.

Step 3. [**Speed reduction ratio –> Inflow curvature parameter**] With the cubic PAQ model, the inflow curvature γ can be represented by $P$and w_{t2}. γ then becomes a function of D/C.

Step 4. [**Inflow curvature parameter –> Time-dependent queue length/ average link performance]** Derive both time-dependent queue length and average link performance during the congestion duration using the D/C ratio

Fig. 4 shows the calculation procedure to develop the QVDF, as well as other components.

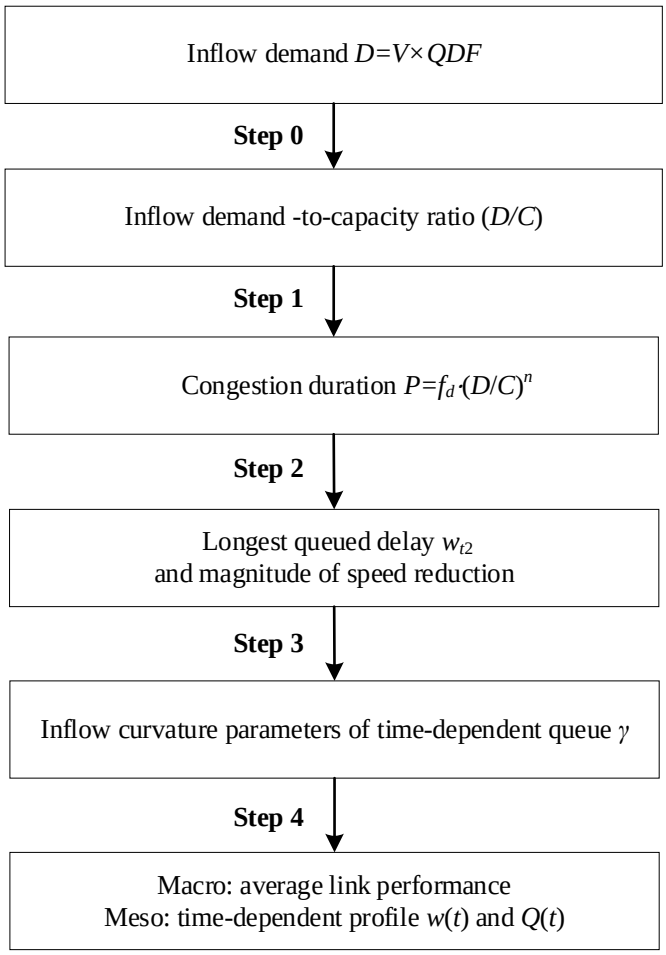


Fig. 4: Major steps towards queue-oriented link performance function QVDF

3.2.2 Requirements for a well-behaved queue-based volume-delay function

Based on our definition of the inflow D/C ratio, we extend the basic requirements proposed by Spiess (1990) to develop a well-behaved QVDF. The additional requirements are listed as follows:

1. [**Consistency of congestion demand**] The inflow demand D should be consistent with the congestion duration P and discharge rate μ in the queueing dynamics, that is, $D = P\mu$.
2. [**Consistency of time-averaged delay**] The average speed within the congestion duration should be consistent with the time-averaged delay.
3. [**Consistency between macroscopic parameters and mesoscopic parameters**] The inflow curvature parameter γ in PAQ models can be highly sensitive to the underlying queue evolution (in terms of length and speed) case by case. There could be a family of values for the mesoscopic parameters with respect to different locations or different days/years. A QVDF needs to use the aggregated data to systematically calibrate macroscopic parameters that can be shared to links of the same facility type and area type.
4. [**First-In-First-Out**] The time-dependent travel time in the underlying mesoscopic vehicular fluid model should satisfy the FIFO property and capacity constraints.

Requirement 1 is satisfied according to the proposed definition of P. Requirement 2 is ensured by the formulas of average

travel delay $\bar{w}$, i.e., (12) and (17) in the quadratic and cubic PAQ models. The guarantee of Requirement 3 will be discussed in detail in the next section.

Now, we demonstrate how the FIFO requirement can be complied by mapping virtual point queue length to mesoscopic physical queue length and linking space-time vehicle trajectories with spatial queue distances. The PAQ models are point queue models presuming that vehicles are zero-length and queueing at the bottleneck until there is enough receiving capacity downstream. The number of those zero-length vehicles waiting at a bottleneck is taken as the virtual queue length. Fig. 5 shows the relation between the virtual queue length and the physical queue length.

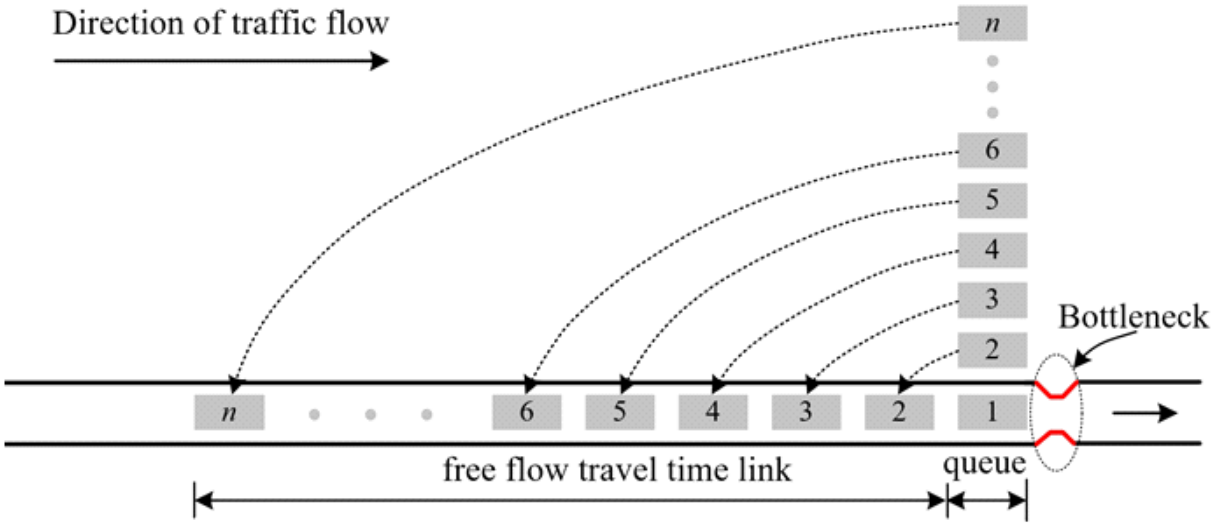


Fig. 5: Relation between the virtual queue and the spatial queue length (the virtual queue is accumulated at the bottleneck without physical vehicle length, while the physical queue is lined along the road with physical vehicle length)

Furthermore, the trajectories with spatial queue distance following the FIFO principle are shown in Fig. 6. The spatial queue is upstream to the bottleneck. In Fig. 6, the vehicle reaches the tail of the physical queue at time t', changes the speed from the cut-off speed v_{co} to congestion speed v_{μ}, arrives at the downstream node at time $t + tt(t)$. $tt(t)$ is the travel time, which leads to the virtual waiting time $w(t)$ in point queue and spatial queue distance $d(t')$ or physical queue length $Q^p(t')$. According to the input-output diagram proposed by Lawson et al. (1997), the virtual queue length from Eq. (6) can be obtained using the physical queue length through Eq. (24) too:

$$Q(t) = Q^p(t') \cdot \left(1 - \frac{v_{\mu}}{v_{co}}\right), \tag{24}$$

where t' is the time when a vehicle encounters the tail of the queue (see Fig. 6 for illustration). Consequently, one can compute the physical queue length via analytically computing the virtual queue length from Eq. (6) and its derivations, and then map it to the mesoscopic vehicle trajectories following the FIFO principle.

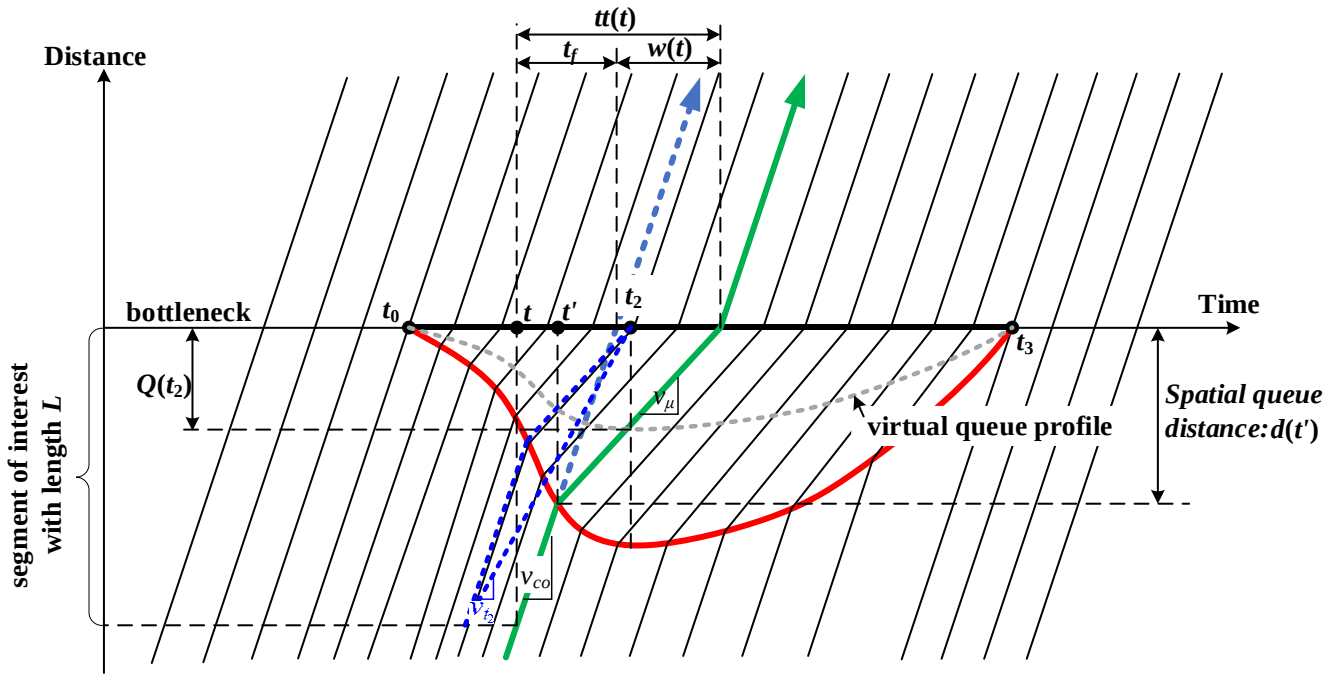


Fig 6: Illustration of the mesoscopic physical vehicle trajectory for a vehicle entering the upstream node of the link at time t, adapted from Lawson et al. (1997) and Cheng et al. (2022). It is worth noting that $tt(t_2) = \frac{L}{v_{co}} + w_{t2}$, so the link travel time is also determined by the length of the segment of interest, in addition to the cutoff speed and the delay time due to traffic congestion.

3.3 Comparison among PAQ models, BPR function, and QVDF

Table 3: Differences and connections among fluid-based PAQ, BPR, and QVDF.

	PAQ (fluid queue)	BPR	QVDF
Temporal horizon of interest	Within-day queue dynamics for a single period	Long-term congestion evolution	Consistent mapping between long-term queue evolution and within-day queue dynamics
Spatial aspect of interest	Single bottleneck	Many links from the same category of facility type and area type	Capturing link-specific characteristics and aggregated features for the same facility type
Supply parameters	Discharge rate μ within congestion duration P	Ultimate capacity C for the design year	Ultimate capacity C and derived average discharge rate
Inflow demand parameter	D: congestion demand	V: period-based volume (e.g., 3 hours from 6:00 AM to 9:00 AM)	Inflow demand $D = V \cdot QDF$
Key parameters	Nonlinear inflow curvature parameter γ for each instance of queue profile	Constant α and exponent term β for multiple days or years in long-term planning applications	n: discharge-rate-efficiency factor to link μ with D/C s: speed reduction elasticity factor to link γ with D/C
Time-dependent queue dynamics	Continuous-time queue profile for a single period	Not applicable	A family of time-dependent queue profiles for a feasible range of D/C

Table 3 compares PAQ models, the BPR function, and our proposed QVDF. Many subtle differences and definitional gaps

exist between the PAQ models intended for short-term queueing profiles and the BPR function used for long-term planning applications. Our proposed QVDF attempts to integrate them, while establishing consistency between macroscopic parameters and mesoscopic parameters.

4 Linking Polynomial Arrival Queue with Macroscopic Volume-Delay Function

This section elaborates the four derivation steps presented in the previous section. We introduce two elasticity factors and the cubic PAQ model to systematically link the temporal queue evolution with the traffic performance measures in a long-term planning process.

4.1 Link congestion duration and D/C ratio

In general, the elasticity term of a function shows the relative percentage change of a dependent variable due to a relative percentage change from an independent variable (e.g., demand as a function of ticket price). We assume a power function between congestion duration P and traffic intensity in terms of the D/C ratio.

$$P = f_d \cdot \left(\frac{D}{C}\right)^n \tag{25a}$$

where $n \geq 1$ is the oversaturation-to-duration elasticity and f_d is a congestion duration constant in response to D/C changes. The formula implicitly assumes the relationship between ultimate capacity C and average discharge rate μ. When $\frac{D}{C} = 1$, f_d represents the baseline value of congestion duration. When $n = 1$, $P = f_d \cdot (D/C)$ and $\frac{D}{P} = \frac{C}{f_d} = \mu$.

Furthermore, the average discharge rate should be less than the capacity, i.e., $\mu \leq C$. To ensure the capacity constraint, the following condition should be satisfied: $D/C \leq P$ It implies that the D/C ratio is the lower bound of congestion duration P, which leads to Eq. (25a):

$$P = \max\left[f_d \cdot \left(\frac{D}{C}\right)^n, \frac{D}{C}\right]. \tag{25b}$$

To make discharge rate μ as a function of D/C ratio, we have

$$\mu = \min\left(\frac{D}{P}, C\right) = \min\left[\frac{D}{f_d\left(\frac{D}{C}\right)^n}, C\right] = \min\left[\frac{C}{f_d\left(\frac{D}{C}\right)^{n-1}}, C\right]. \tag{26a}$$

Eq. (26a) implies that higher traffic intensity (i.e., the D/C ratio) results in lower throughputs (i.e., discharge rate). Eq. (26b) indicates that a longer congestion duration leads to a loss in capacity utilization.

$$\mu = \min\left[\frac{C}{f_d^{\frac{1}{n}} P^{\frac{n-1}{n}}}, C\right] \tag{26b}$$

When both $n = 1$ and $f_d = 1$, the estimated discharge rate μ has a constant value of C.

4.2 Linking magnitude of speed reduction and congestion duration

As shown in Fig. 3 (a), we attempt to connect P and the speed reduction in the PAQ model (i.e., $v_{co} - v_{t_2}$). Therefore we first define the magnitude of speed reduction (MSR):

$$MSR = \frac{v_{co} - v_{t_2}}{v_{t_2}} \tag{27a}$$

We further establish a relationship between congestion duration P and MSR. Note that v_{t_2} and P in Eqs. (27a) and (27b) are directly observable.

$$MSR = f_p \cdot (P)^s \tag{27b}$$

where $s \geq 1$ is the duration-to-speed reduction elasticity factor and f_p is an MSR reduction constant in response to changes of congestion duration. The boundary condition is satisfied for $MSR = 0$ as $P = 0$ and $v_{t_2} = v_{co}$. When $P = 1$, f_p represents the baseline value for speed reduction magnitude, which corresponds to the lowest speed ratio of $\frac{v_{t_2}}{v_{co}} = \frac{1}{1+f_p}$.

4.3 Linking inflow curvature parameter and magnitude of speed reduction

Next, we create a connection between w_{t_2}(i.e., the longest delay at t_2 compared with v_{co}) and P

$$w_{t_2} = \frac{L}{v_{t_2}} - \frac{L}{v_{co}} = \frac{L}{v_{co}} \cdot \left(\frac{v_{co} - v_{t_2}}{v_{t_2}}\right) = \frac{L}{v_{co}} \cdot MSR = \frac{L}{v_{co}} \cdot f_p \cdot (P)^s \tag{28a}$$

If we use the cubic PAQ model with $m = 1/2$, then we have

$$w(t) = \frac{\gamma}{4\mu} \cdot (t - t_0)^2 \cdot (t - t_3)^2.$$

By assuming $\gamma > 0$ and $m = 1/2$ in Eqs. (16) and (17), we have $t_0 = t_2 - P$ and $t_3 = t_2 + P$ according to Cheng et al. (2022). Then the longest time-dependent delay at t_2 becomes available as follows:

$$w_{t_2} = \frac{\gamma}{4\mu} \cdot (t_2 - t_0)^2 \cdot (t_2 - t_3)^2 = \frac{\gamma}{4\mu} \cdot \left(\frac{P}{2}\right)^4 = \frac{\gamma}{64\mu} \cdot P^4. \tag{28b}$$

From Eq. (28b), γ can be represented as a function w_{t_2}.

$$\gamma = \frac{64\mu \cdot w_{t_2}}{(P)^4} \tag{29a}$$

Derive the parameter γ as a composite function of D/C, P, and μ from Eq. (28a) and Eq. (28b),

$$\gamma = 64\mu \cdot \frac{L}{v_{co}} \cdot f_p (P)^{s-4} \tag{29b}$$

If $\mu < C$, then we have

$$\mu = \frac{C}{f_d \left(\frac{D}{C}\right)^{n-1}},$$

$$P = f_d \cdot \left(\frac{D}{C}\right)^n.$$

They lead to

$$\gamma = 64C \cdot \frac{L f_p {f_d}^{s-5}}{v_{co}} \cdot \left(\frac{D}{C}\right)^{ns-5n+1}. \tag{30a}$$

If $\mu = C$, then we can use the settings of $n = 1$, $f_d = 1$, and

$$\mu = C,$$

$$P = \frac{D}{C}.$$

Then, we have

$$\gamma = 64C \cdot \frac{Lf_p}{v_{co}} \cdot \left(\frac{D}{C}\right)^{s-4}. \tag{30b}$$

When $n = 1$ and $s = 4$, γ will always be a constant.

4.4 Time-dependent delay and average speed during congestion duration.

4.4.1 Generate time-dependent delay

We have the ordinary differential equation within congested space-time regimes for each link. By assuming dynamic arrival rates, departure rates, queue length process, and introducing elasticity parameters n and s, we can capture the relationship between the congestion duration P and the average discharge rate μ. Besides, we correlate congestion duration and maximum virtual queues (at the lowest speed) through regression analysis.

If the cubic model is adopted with $m = 0.5$, we will have the following time-dependent queue and delay.

$$Q(t) = \frac{\gamma}{4}(t - t_0)^2 \cdot (t - t_3)^2$$

$$w(t) = \frac{\gamma}{4\mu}(t - t_0)^2 \cdot (t - t_3)^2$$

Time-dependent speed can be derived as follows:

$$v(t) = \frac{L}{\frac{L}{v_{co}} + w(t)}. \tag{31}$$

4.4.2 Generate average speed during congestion duration

With the cubic PAQ model and $m = 0.5$, the average delay during a congestion duration can be calculated by the following equation (see Cheng et al., 2022):

$$\bar{w} = \frac{\gamma}{120\mu} \cdot P^4 = \frac{\gamma}{120\mu} \cdot \left(\frac{D}{\mu}\right)^4. \tag{32}$$

Similar to Eq. (31), we have the average speed during the congestion duration as

$$\bar{v} = \frac{L}{\frac{L}{v_{co}} + \bar{w}}. \tag{33}$$

This leads to an important property that the ratio of average waiting time and longest waiting time (i.e., $\theta = \bar{w}/w_{t_2}$) is a constant value. Given $m = 0.5$, dividing Eq. (31) by Eq. (28b) gives us

$$\theta = \frac{\bar{w}}{w_{t_2}} = \frac{64}{120} = \frac{8}{15}.$$

Eq. (33) can be transformed in terms of speed reduction factor, v_{co}/v_{t_2}, which is in turn a function of D/C ratio below. This leads to a BPR-like link performance function.

$$\begin{aligned}\bar{v} &= \frac{L}{\frac{L}{v_{co}}+\bar{w}} = \frac{L}{\frac{L}{v_{co}}+\theta w_{t_2}} = \frac{L}{\frac{L}{v_{co}}+\theta\cdot\left(\frac{L}{v_{t_2}}-\frac{L}{v_{co}}\right)} = \\ &= \frac{v_{co}}{1+\theta\left(\frac{v_{co}-v_{t_2}}{v_{t_2}}\right)} = \frac{v_{co}}{1+\theta\cdot\left[f_p\cdot(P)^s\right]} \\ &= \frac{v_{co}}{1+\theta\left[f_p\cdot {f_d}^s\cdot\left(\frac{D}{C}\right)^{n\cdot s}\right]} = \frac{v_{co}}{1+\alpha\left(\frac{D}{C}\right)^{\beta}}\end{aligned} \tag{34}$$

where $\alpha = \theta f_p {f_d}^s$ and $\beta = ns$. An alternative form can be written as $\bar{w} = \frac{L}{v_{co}}\cdot\alpha\left(\frac{D}{C}\right)^{\beta}$.

Denote D/C ratio as x, we can consider the proposed QVDF as a link delay function $f(x) = \alpha x^{\beta}$, and revisit the basic well-behaved VDF requirements proposed by Spiess (1990) as follows. .

1. $f(x)$ should be strictly increasing for a feasible range of x. This is the necessary condition for the traffic assignment to converge to a unique solution.
2. $f(0) = 0$ and $f(1) = \text{constant}\ \alpha$. The conditions ensure compatibility with the BPR form.
3. $f'^{(x)} = \alpha\beta x^{\beta-1}$ exists and is strictly increasing. This ensures the convexity of the congestion function, which is important when calculating marginal costs in system optimum assignment.
4. $f'(1) = \alpha\beta$, and the exponent β indicates how sudden the congestion effects change under the oversaturated condition.
5. $f'(x) =< M\beta$, where M is a positive constant. The steepness of the congestion curve is limited so that the derived average delay is reasonable.
6. $f'(0) \geq 0$, which guarantees the uniqueness of the link volumes.
7. The evaluation of $f(x)$ should be computationally efficient.

To interpret coefficient β from a queue evolution perspective, it can be viewed as a joint effect and complex demand-supply interactions of the relationship between congestion duration and capacity effectiveness (on the supply side) and the magnitude of speed reduction and congestion duration. The latter factor can be better linked to maximum delay, preferred arrival time, and schedule delay from the departure time choice modeling on the demand side.

5 Calibration and Results

In this section, we conduct two case studies to evaluate our proposed QVDF. The first study utilizes the daily time-dependent time-mean speed and link counts in a freeway corridor over an entire year. The locations are viewed as different modeling "links". The second one considers a single oversaturated bottleneck on a 6-mile corridor and treats it as a whole modeling element or "link". Four months of sensor data are applied in the second case study. By using these data sets, which are also available at https://github.com/asu-trans-ai-lab/QVDF, we hope to demonstrate how the time-dependent queue dynamics and time-dependent dependent speed profile estimation can be modeled consistently from the macroscopic link performance with the two studies through the following two statges.

.

Preprocessing stage: Collect the time-dependent traffic counts and speed over multiple days . Let the dataset be ***K***. The records of

each day are expressed as $k \in \boldsymbol{K}$. Considering an analysis period $\boldsymbol{T}$, for each $k \in \boldsymbol{K}$ and timestamp $t \in \boldsymbol{T}$, we have the time-dependent traffic counts $q_{t,k}^{\text{obs}}$ and speed $v_{t,k}^{\text{obs}}$ with a specified time interval (e.g., 15 minutes). Furthermore, let v_{co} be the cut-off speed to define the congestion duration. Before the calibration of the link performance function, we should prepare the following data and parameters:

1) **[Ultimate hourly capacity, free-flow speed, and cut-off speed]** Ultimate capacity C, free-flow speed t_f, and cut-off speed v_{co} of each link (or link type) can be obtained from a traffic flow model after calibration using the observed speed and volume data.
2) **[Congestion duration]** Find the lowest speed $v_{t2}(k)$ at time $t_2(k) \in \boldsymbol{T}$ for each $k \in \boldsymbol{K}$. Take $t_0(k) = t_3(k) = t_2(k)$ as initial values if there are speeds strictly below v_{co}. Then, extend the congestion duration by decreasing $t_0(k)$ and increasing $t_3(k)$ until $v_{t3,k}^{\text{obs}} \geq v_{co}$ and $v_{t1,k}^{\text{obs}} \geq v_{co}$. The congestion duration on day k is obtained as $P_k^{\text{obs}} = t_3(k) - t_0(k)$.
3) **[Period volume, inflow demand, queued demand factor]** Determine period volume $V_k = \sum_{t \in \boldsymbol{T}} q_{t,k}^{\text{obs}}$, inflow demand $D_k^{\text{obs}} = \sum_{t_0(k) \leq t \leq t_3(k)} q_{t,k}^{\text{obs}}$, and queued demand factor $QDF_k = V_k / D_k^{\text{obs}}$ for each day $k \in \boldsymbol{K}$ using the corresponding defitions in Section.3.1.1.

Calibration stage: Calibrate the following four parameters in the proposed *QVDF*.

1) [$\boldsymbol{n}$ **and** $\boldsymbol{f_d}$ **for oversaturation-to-duration elasticity]** f_d and n in the power function of Eq. (25a). can be calibrated using P_k^{obs} and D_k^{obs}/C for each $k \in K$.
2) [**s and** $\boldsymbol{f_p}$ **for duration-to-speed reduction elasticity**] Given v_{co} and P_k^{obs}, $v_{t2}(k)$, for each $k \in K$, we can compute $\boldsymbol{MSR_k}$, and then calibrate f_p and s in the power function of Eq. (27b).
3) [α and β **in QVDF link performance function**] With 1) and 2), α and β can be obtained via $\alpha = \theta f_p {f_d}^s$ and $\beta = ns$, where $\theta = \overline{w}/w_{t_2}$.

5.1 Case study 1: *Individual locations along a freeway corridor in Phoenix, Arizona*

This section presents a typical freeway bottleneck in Phoenix equipped with loop, that continuously collect traffic speed and counts. More details on the collected data can be found in Belezamo (2020).

5.1.1 Dataset description

We use four loop detectors on the I-10 westbound corridor. The analysis period is from 7:00-21:00 fordays from Jan 1st to Dec 31st, 2016. Traffic counts and speed data are collected every 15 minutes, where the HOV and ramp lanes are not considered. **Fig. 8** shows the locations of the four detectors and their identification numbers (IDs). **Table 4** provides additional information over the four links includinglane configurations, and so on.

We take **7:00-21:00 as one analysis period** to show the entire within-day flow dynamics in this case study, which is different from a common analysis period (e.g., AM, MD, or PM). The reason is that the I-10 corridor sometimes has extensive traffic congestion throughout from MD to PM.

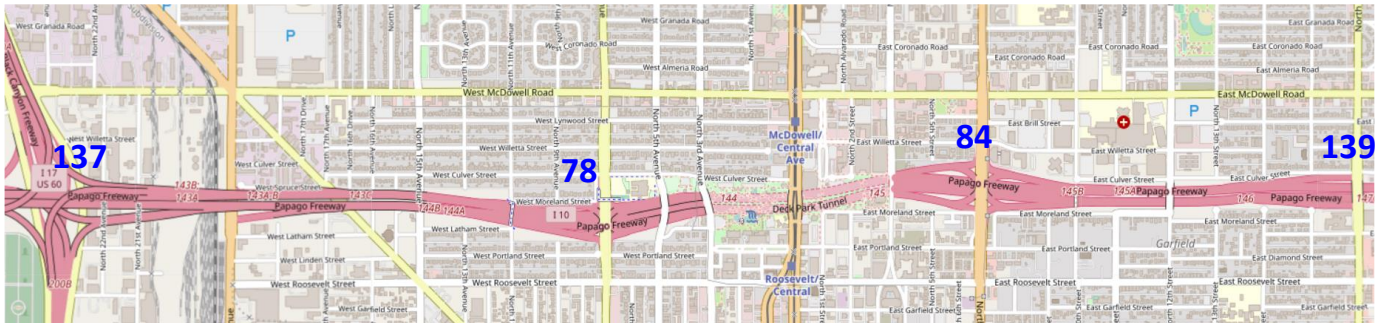


Fig. 8: Phoenix I-10 freeway corridor and the four detectors

Table 4: Detectors and lane configuration data.

Detector ID	Road order	Milepost	Number of general-purpose lanes	Length
139	1	146.823	5	1.11 miles
84	2	145.681	4	1.14 miles
78	3	144.641	4	1.04 miles
137	4	143.346	3	1.30 miles

5.1.2 Traffic flow models

We adopt the 3-parameter traffic flow model proposed by Cheng et al., (2021) and calibrate its fundamental diagram (in terms of volume-speed curve) for each link with the following five parameters: ultimate capacity, speed at capacity, critical density, free-flow speed, and maximum flow inertia coefficient. Note that Speeds higher than the speed limit provided by ADOT for each link are preempted from the calibration process. Their calibrated values are summarized in **Table 5** while **Fig. 9** shows the calibrated fundamental diagram along with observations of each link. The four fundamental diagrams have different patterns. Take link 137 for instance, its calibrated capacity of is 1380.8 vehicles/ hour/ lane, which is lower than the other three links.

Table 5. Calibrated parameters of the fundamental diagram for each link

Detector ID	Ultimate capacity (vehicles/lane/hour)	Speed at capacity (miles/hour)	Critical density (vehicles/mile)	Free-flow speed (miles/hour)	Maximum flow inertia coefficient
139	1709.2	51.6	33.1	70.0	4.5
84	1816.5	53.4	34.0	70.0	5.1
78	1586.6	55.9	28.4	69.6	6.3
137	1380.6	50.0	27.6	70.0	4.1

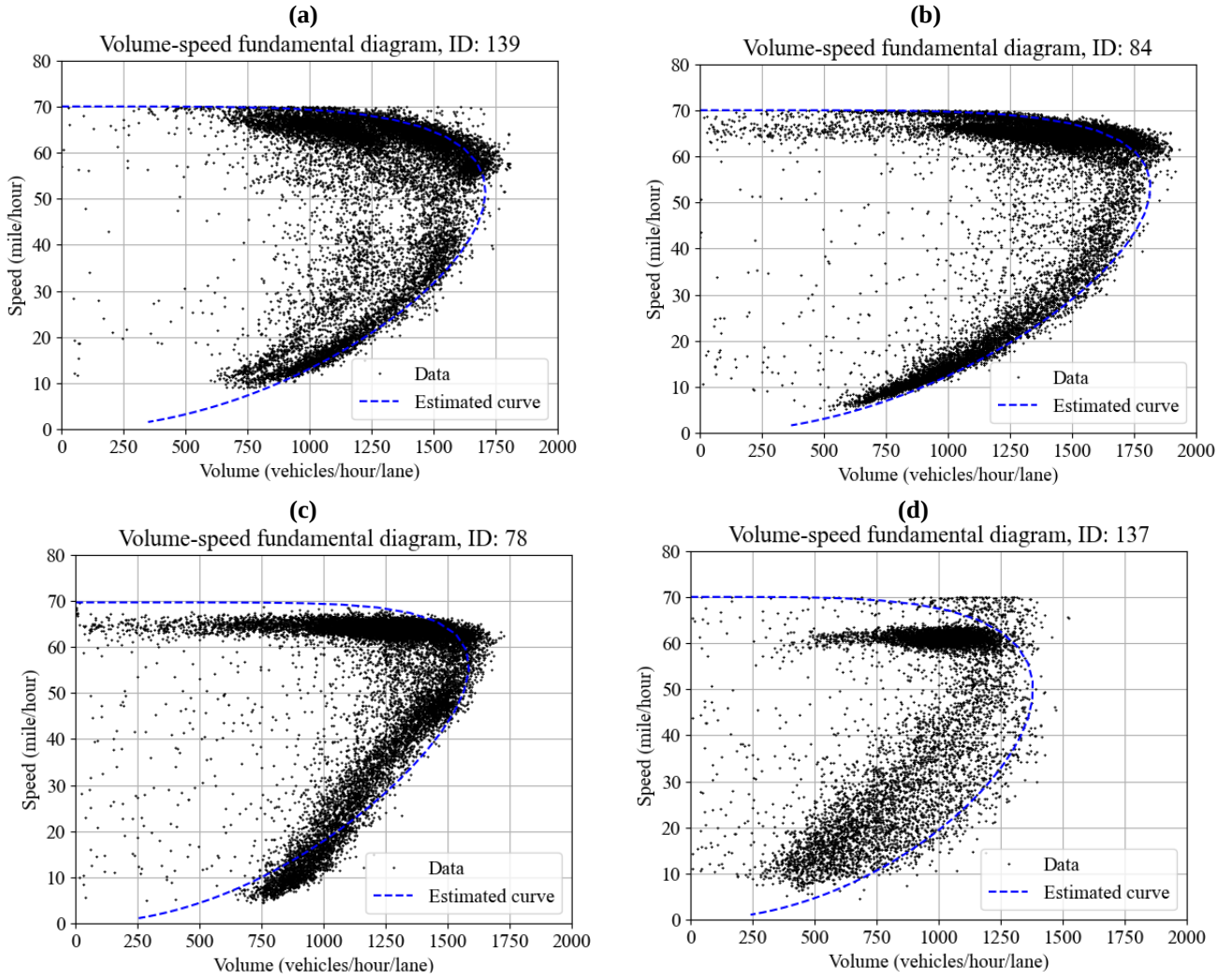


Fig.9 Volume-speed scatters of entire year's data (from 7:00 to 21:00) and calibrated traffic flow model

5.1.3 Observed congestion duration and D/C ratio

Now, we organize the dataset to obtain four important parameters up to decide: (a) congestion duration, (b) queue demand factor, (c) D/C ratio, and (d) inflow demand. In addition to ultimate hourly capacity and free-flow speed, a cutoff speed needs to be specified before determing congestion duration. In this case study, we choose 49 miles/hour across different links, which is 70% of the free-flow speed (70 mph).

Table 6. Congstion statistics for each link

Detector ID	Total number of days with valid records	Number of days with $P > 0$	The ratio of congested days (including weekdays and weekends)	Average congestion duration when $P > 0$
139	300	239	80%	4.52 hours
84	325	255	78%	4.77 hours
78	325	263	81%	4.63 hours
137	325	235	72%	3.96 hours
Total	1,275	992	78%	4.48 hours

Congestion duration P, inflow demand, D/C ratio, and QDF are then calculated following the steps detailed in the preprocessing stage at the beginning of Section 5. Their distributions are ploted in **Fig.10 (a), (b), (c),** and **(d)** respectively. Note that **Fig.10(a)** shows the distribution of P for $P > 0$ only. **Table 6** summarizes congestion information for each link.

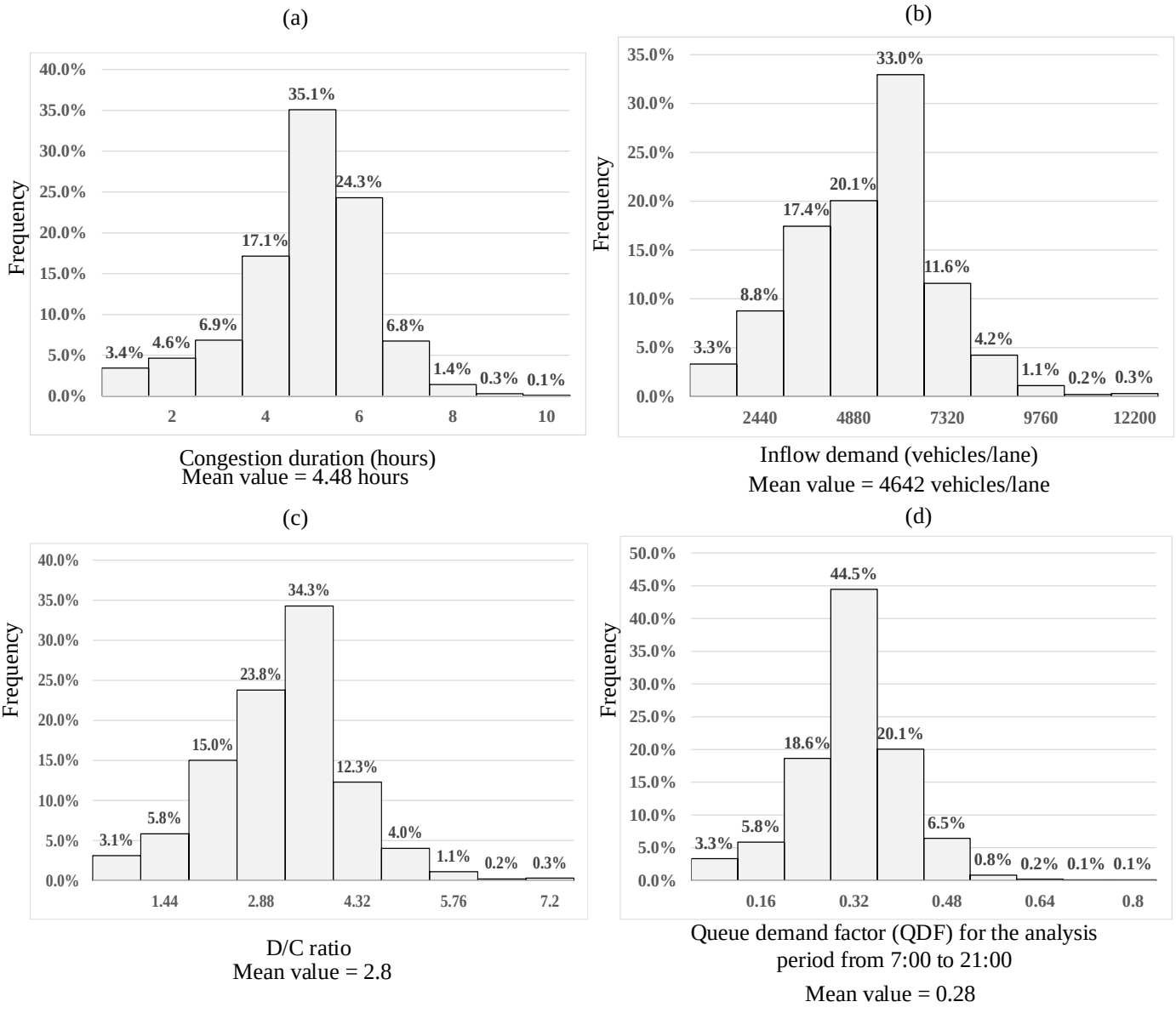


Fig.10 Distributions of congestion duration, inflow demand, D/C ratio, and QDF for analysis period from 7:00 to 21:00

5.1.4 Calibration of parameters

This section implements the first two steps of the calibration process in section 3.2.1. Calibrated parameters are summarized in **Table 7** with the following validations. The oversaturation-to-duration elasticity n has an average of 1.11and indicates that P is elastic in response to the D/C changes. The duration-to-speed reduction elasticity s is more significant as its mean value is 1.6475. The average of f_d is 1.415, which implies P of 1.415 hours with D/C =1. Furthermore, f_p is 0.225 on average indicating $v_{t_2} =$

$\frac{1}{1+f_p} v_{co} = 0.8163 \times 49 = 40$ mph when $P = 1$.

Table 7. Calibrated coefficients and elasticity factors

Detector ID	The first step calibration, Eq. (25)		The second step calibration, Eq. (27)		Parameters for BPR-like average speed function, Eq. (34)	
	f_d	n	f_p	s	α	β
139	1.43	1.06	0.35	1.14	0.28	1.21
84	1.37	1.14	0.23	1.64	0.21	1.87
78	1.13	1.22	0.11	2.10	0.07	2.55
137	1.73	1.02	0.21	1.71	0.29	1.75

Fig.11 shows the scatters and the calibrated curves between the D/C ratio and P for each link. Since each P corresponds to multiple D/C ratios with large disturbances, we calculate mean value of D/C ratios for each congestion duration (i.e., red circles in **Fig.11**) to ensure the exogeneity of D/C ratio over P the calibration stage . **Fig.12** illustrates the scatters and the calibrated curves between the magnitude of the speed reduction and P. Similarly, each P has multiple speed reduction points with large disturbances. The mean value of speed reduction $v_{co}/v_{t2} - 1$ for each P (red circles in **Fig.12**) is used for the second step of calibration. **Fig.13** shows calibrated curves between D/C ratio and average speed within their congestion durations.

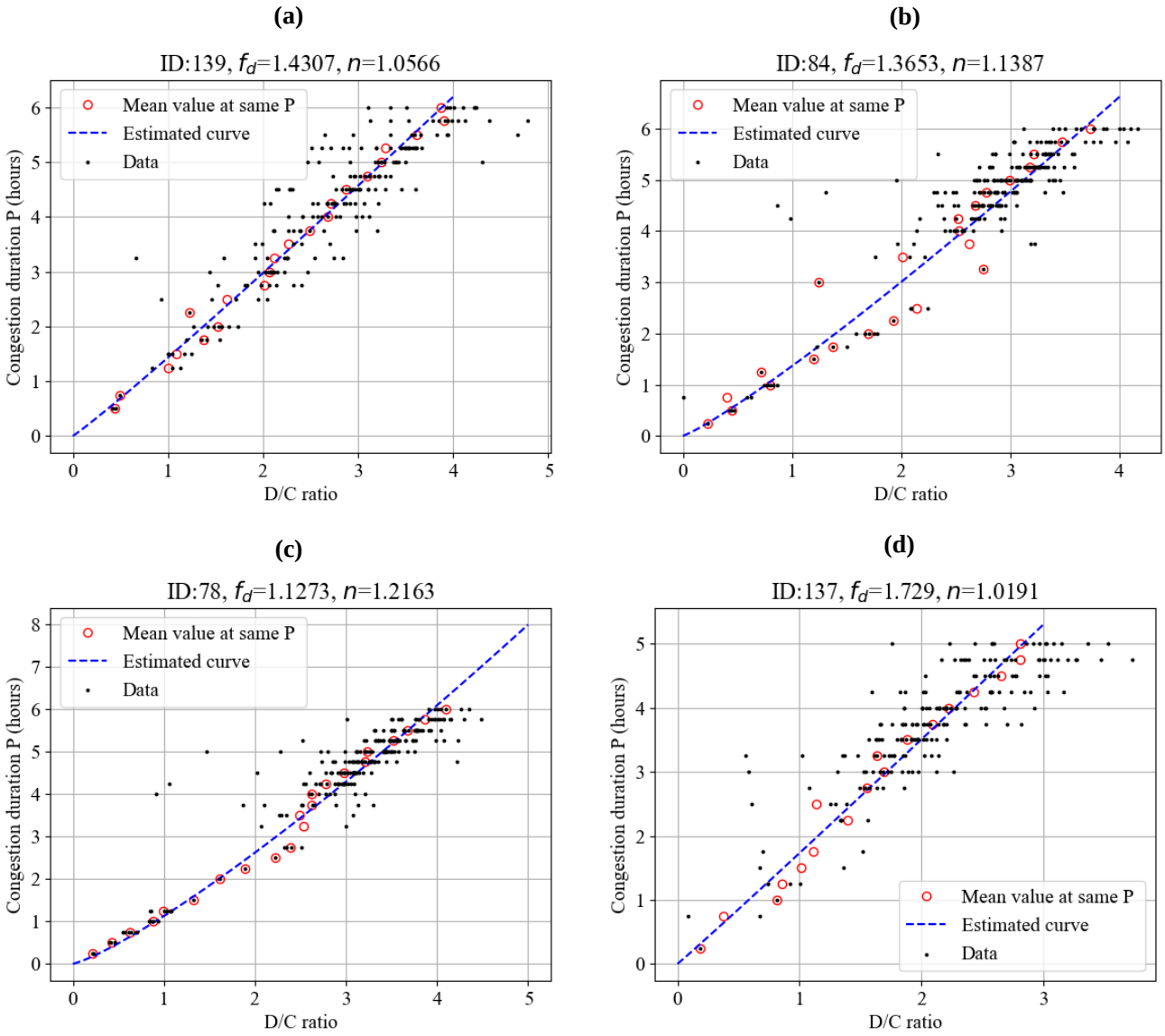


Fig.11 Calibration results of step 1 calibration that links D/C to congestion duration

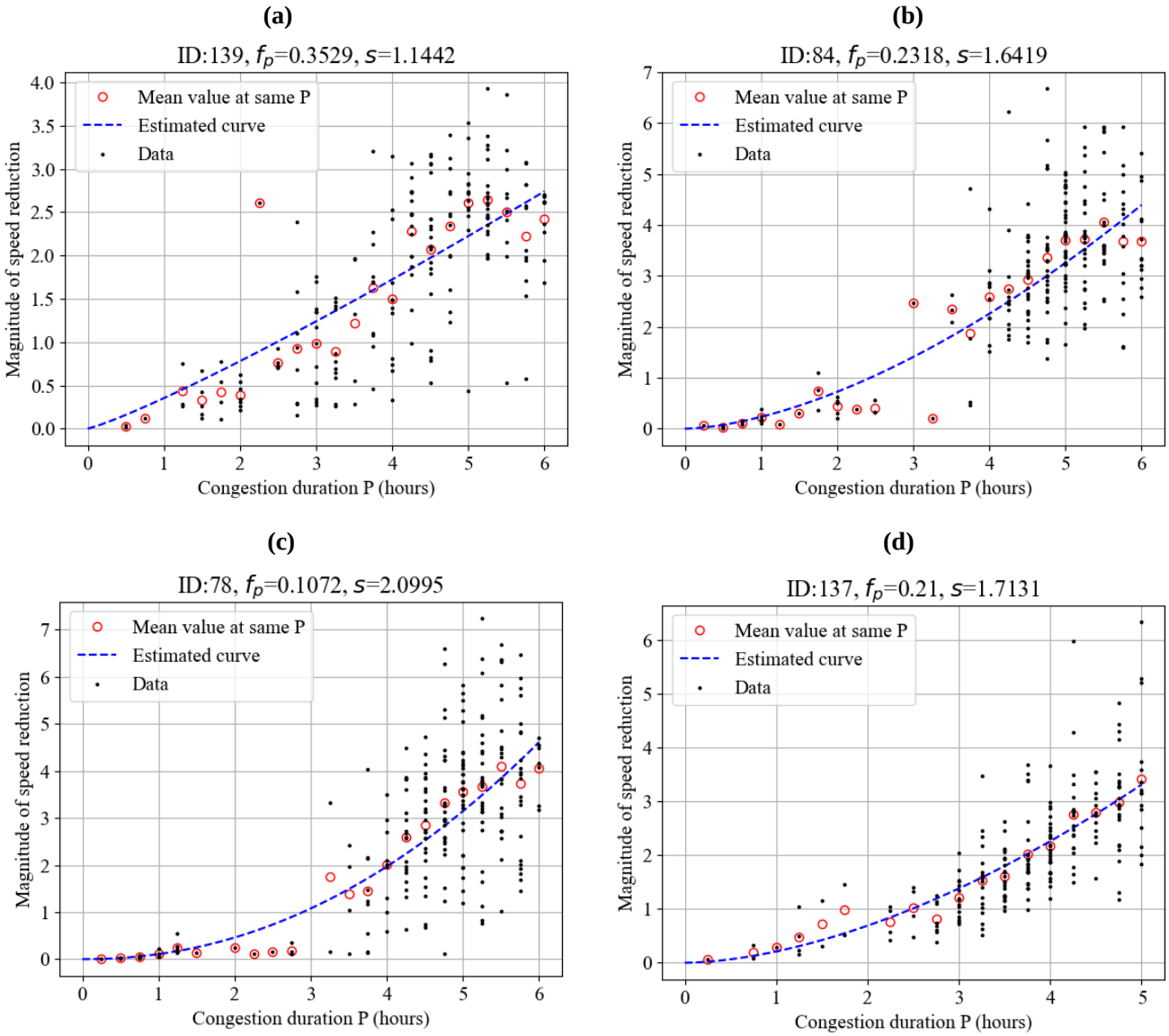


Fig.12 Calibration results of step 2 calibration that links congestion duration to the magnitude of speed reduction

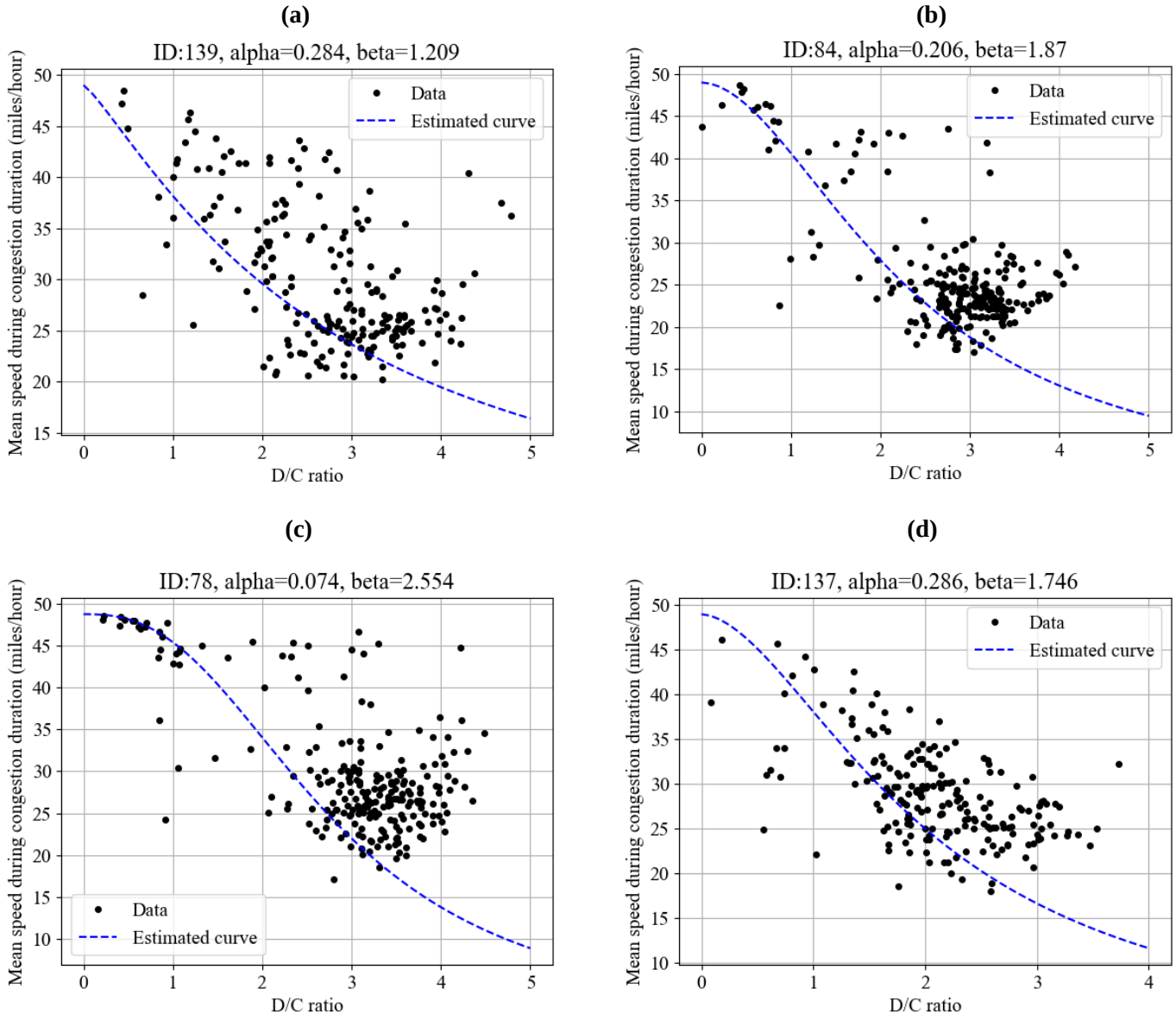


Fig.13 Curves between D/C ratio and average speed during congestion duration

5.1.5 Derive average link performance and time-dependent speed profiles for different weekdays

While parameter f_d, f_p, s, n, as well as α, and β measure the average link performance in the long-term planning horizon, we derive inflow curvature parameter γ to reflect time-dependent link performance. Different weekdays will use different γ in this case study. **Table 8** shows different inflow curvature parameters for different links and weekdays. In all these links, γ is decreased with the increase of the D/C ratio. When $n = 1$ and $s = 4$, the curvature parameter will be a constant, as shown in **Fig. 14**. **Fig. 15-18** show the derived time-dependent speed profiles of Link 84 (from Monday to Thursday). Our proposed method can derive different dynamic speed patterns according to specific given D/C ratios. In the figures, we compare our derived time-dependent speed with the observed time-dependent speed profile. The observed speed at each timestamp is the average value of valid samples over different days.

Table 8: Calibrated inflow curvature parameters for different links and weekdays (sorted by D/C ratio)

Detector ID	Days of week	Mean of D/C ratio	Mean of P (hours)	Mean of v_{t2} (miles/hour)	Mean of $\bar{v}$ (miles/hour)	γ
139	Friday	3.82	5.90	13.27	20.12	3.56
	Thursday	3.19	4.87	15.50	22.76	6.21
	Wednesday	3.06	4.67	16.03	23.36	7.04
	Tuesday	2.74	4.16	17.49	24.99	9.88
	Monday	2.25	3.38	20.25	27.89	18.09
	Saturday	1.69	2.49	24.49	31.95	44.00
	Sunday	0.65	0.91	37.21	41.92	819.45
84	Friday	3.75	6.16	8.78	14.23	5.26
	Wednesday	3.04	4.85	11.97	18.49	9.52
	Thursday	3.03	4.83	12.03	18.57	9.62
	Tuesday	2.92	4.63	12.67	19.38	10.69
	Monday	2.79	4.39	13.49	20.39	12.16
	Saturday	1.07	1.48	34.00	39.67	180.60
	Sunday	0.71	0.93	40.69	44.19	577.90
78	Friday	3.86	5.83	9.12	14.70	5.39
	Thursday	3.44	5.07	11.50	17.88	7.22
	Wednesday	3.42	5.03	11.65	18.07	7.34
	Tuesday	3.20	4.65	13.18	19.98	8.64
	Monday	3.07	4.42	14.23	21.25	9.60
	Saturday	1.22	1.43	39.68	43.45	99.69
	Sunday	0.70	0.73	46.19	47.35	406.85
137	Friday	2.64	4.65	12.49	19.15	8.32
	Thursday	2.33	4.09	14.64	21.76	11.17
	Wednesday	2.29	4.02	14.96	22.14	11.65
	Monday	2.11	3.70	16.44	23.83	14.05
	Tuesday	2.08	3.64	16.77	24.20	14.62
	Saturday	0.73	1.25	37.49	42.11	172.74
	Sunday	0.56	0.95	41.05	44.41	321.08

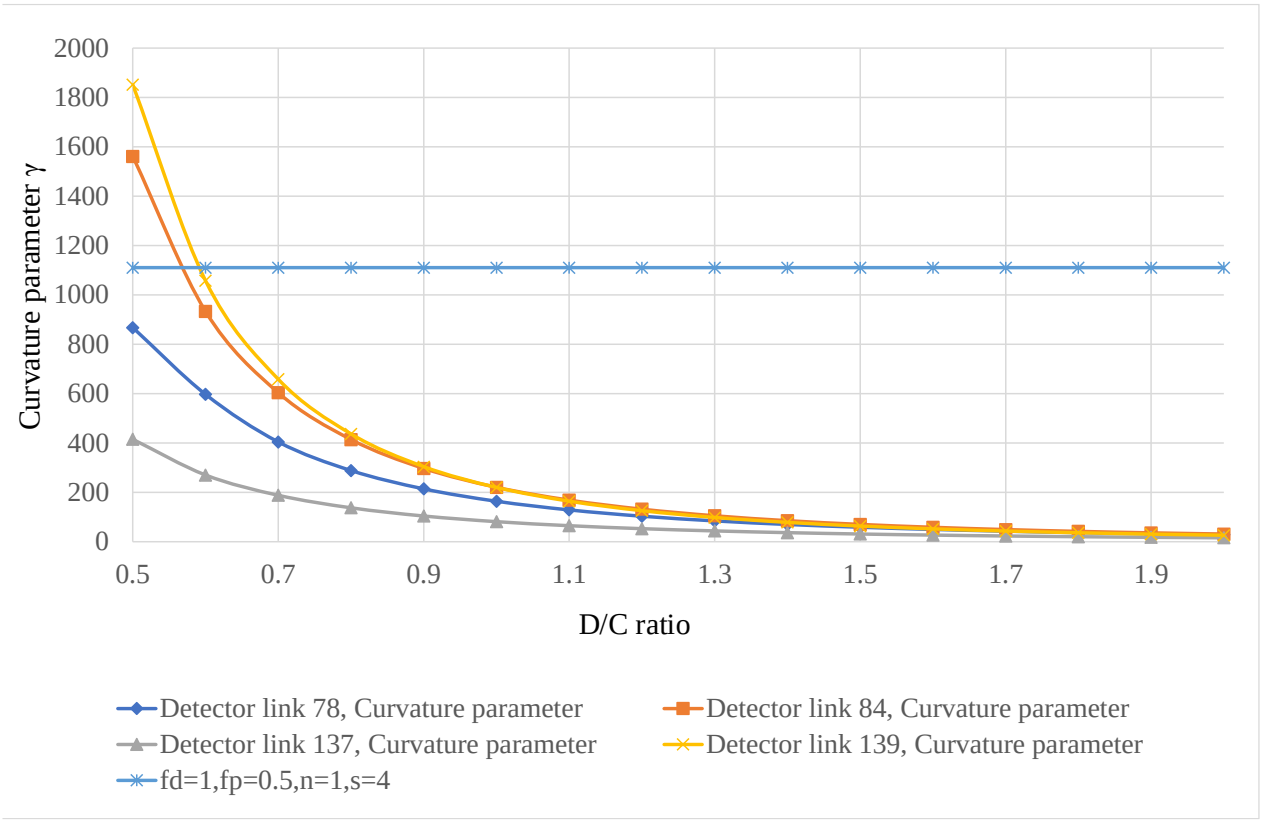


Fig. 14 Curvature parameter with the change of the D/C ratio

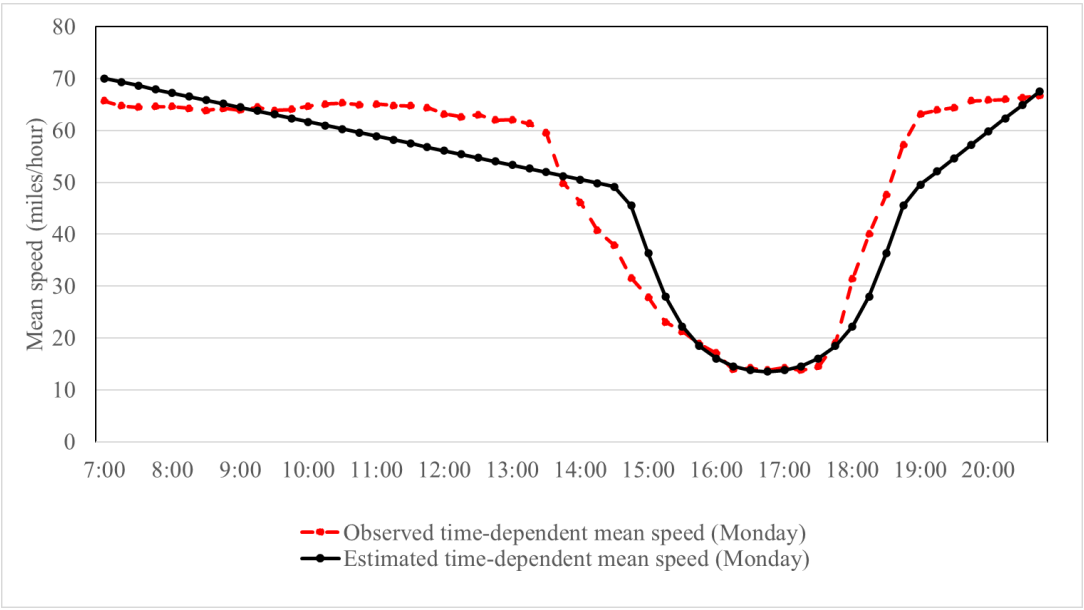


Fig. 15 Observed time-dependent mean speed and estimated speed at detector ID 87 on Monday

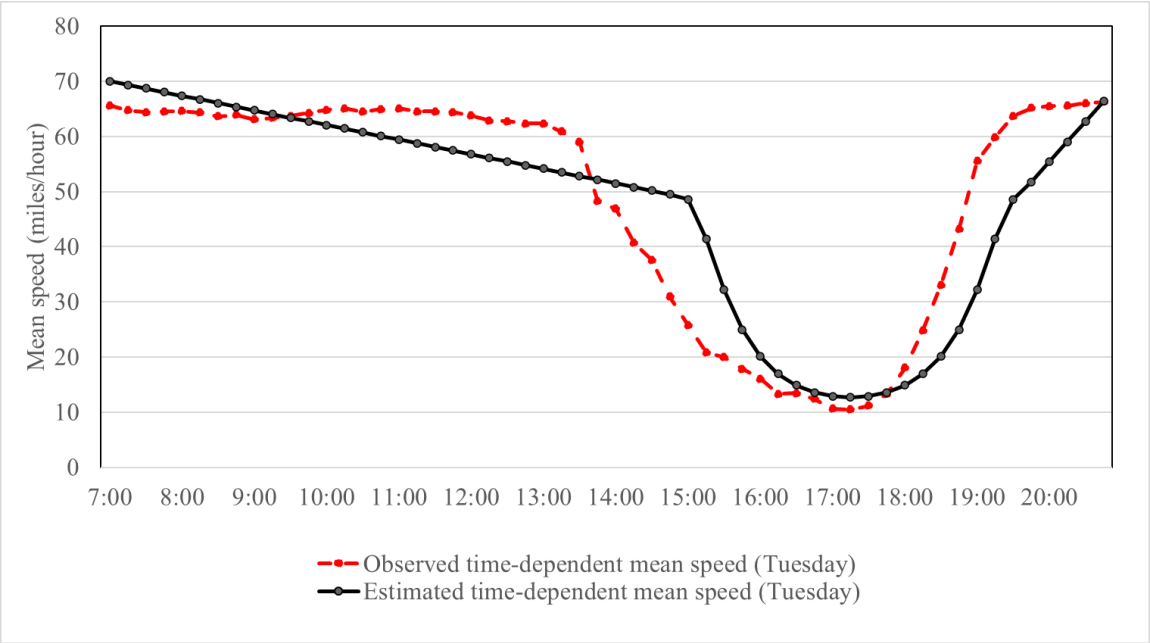


Fig. 16 Observed time-dependent mean speed and estimated speed at detector ID 87 on Tuesday

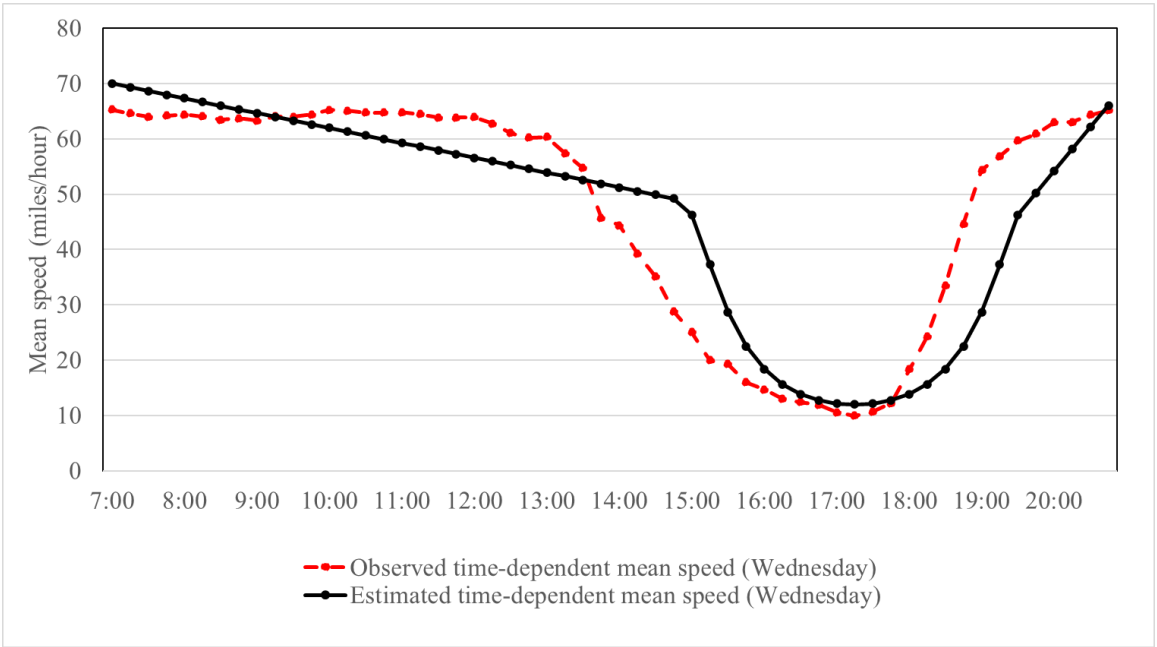


Fig. 17 Observed time-dependent mean speed and estimated speed at detector ID 87 on Wednesday

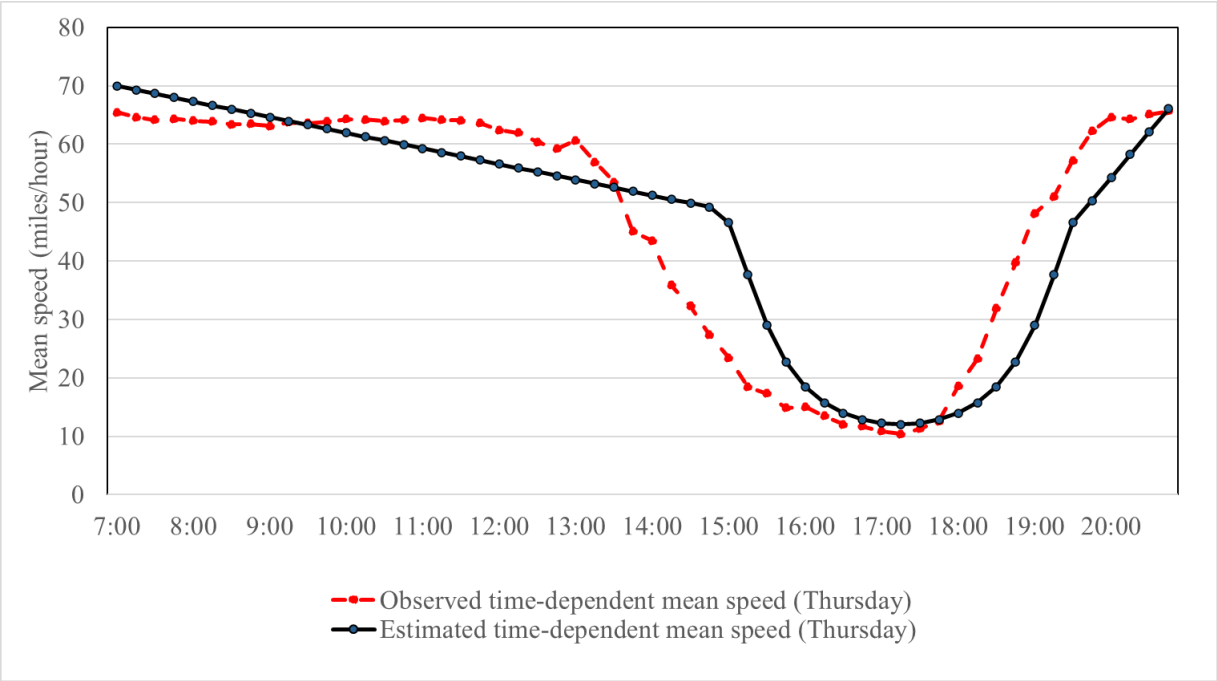
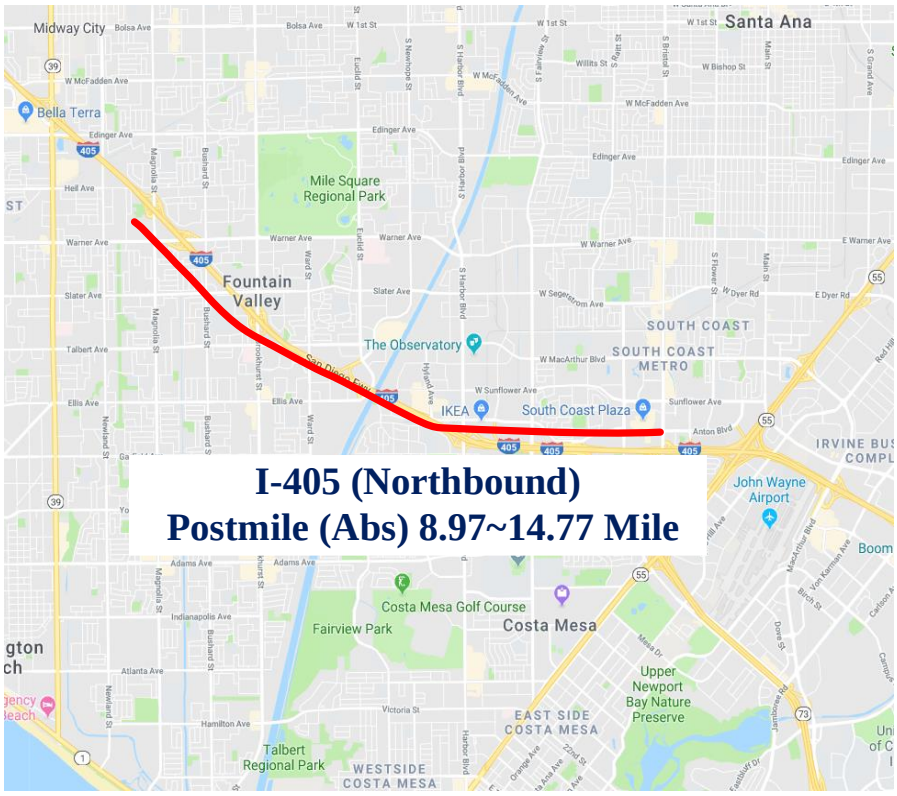


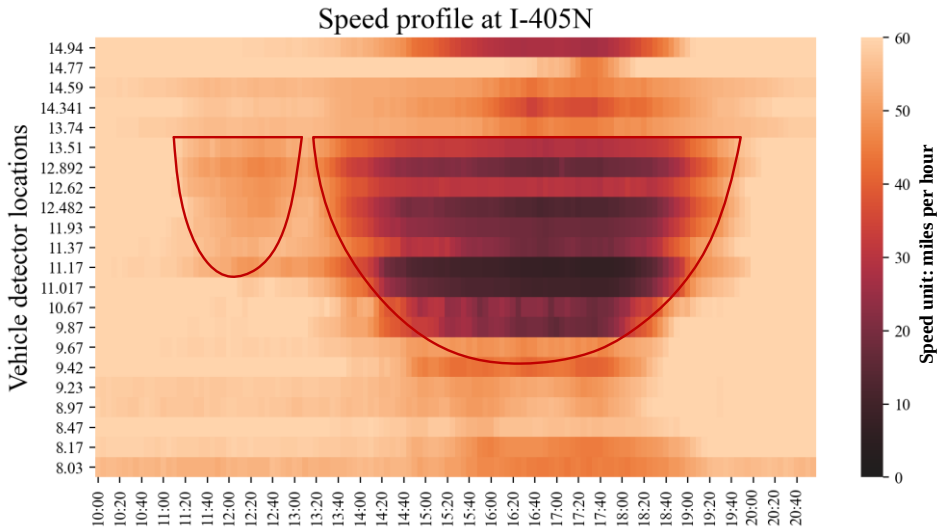
Fig. 18 Observed time-dependent mean speed and estimated speed at location with detector ID 87 on Thursday

5.2 Case study 2: A longitudinal analysis for a single bottleneck on a 6-mile freeway corridor in Los Angeles, California over 4 months

Traffic flow, speed, and occupancy data are collected every five minutes from 11:00 to 20:00 in April, May, June, and July 2019 from the 22 freeway detectors along the Northbound direction of I-405 freeway between the absolute milepost (Abs) 8.97 to 14.77 in Los Angeles (see **Fig. 19(a)**). The bottleneck is located at where Abs=13.51 mile. As shown in **Fig. 19(b)**, we are interested in one afternoon peak period from $t_0 = 13{:}10$ to $t_3 = 19{:}45$ at this single bottleneck.



(a) Areas of the traffic bottleneck



(b) Typical space-time extent of traffic congestion along this corridor

Fig. 19: Locations and speed profile of the corridor (Cheng et al. 2022).

The original data can be accessed at http://pems.dot.ca.gov (hosted by the California Department of Transportation). We choose a single recurrent bottleneck and calibrate congestion duration P, total inflow demand volume D, and constant discharge rate μ for each afternoon peak period on weekdays. P has a mean value of 6.75 hours while the average μ is 1339.4 vehicles per hour per lane. We can estimate elasticity coefficients n =1.046, and f_d=1.124, leading to mean absolute error, MAE = 0.761 hours and MAPE =

11.26% between observed and estimated P. As shown in **Fig. 20**, we express congestion duration P as a function of the observed D/C ratio using the calibrated coefficients.

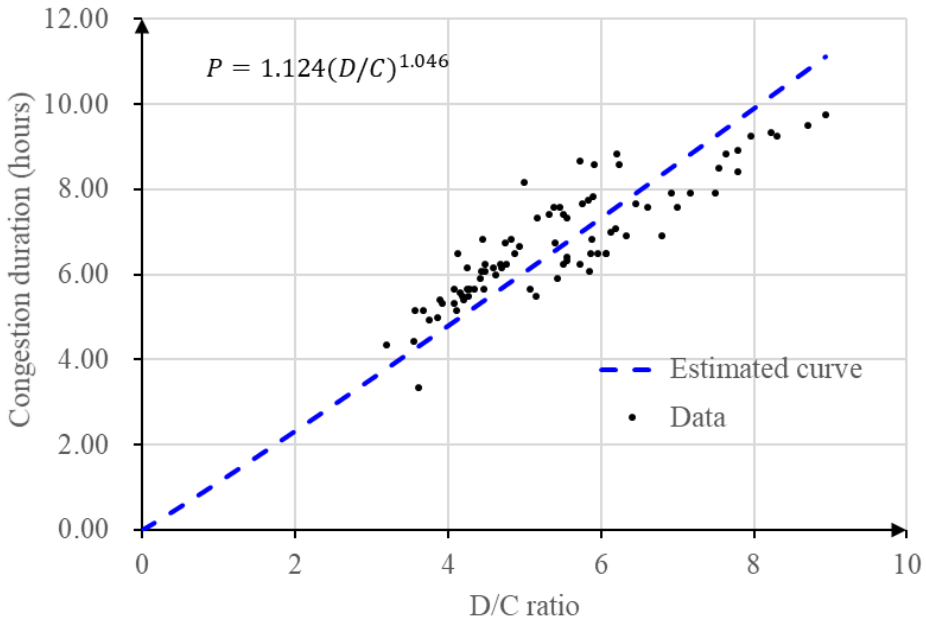


Fig. 20: Observed and estimated congestion duration for the inflow demand-to-capacity ratio.

As shown in **Fig. 21**, our method provides reasonable estimates on P via the elasticity-form power function of the D/C ratio. If we set the intersect as zero, the regression function will be $y = 0.9842x$.

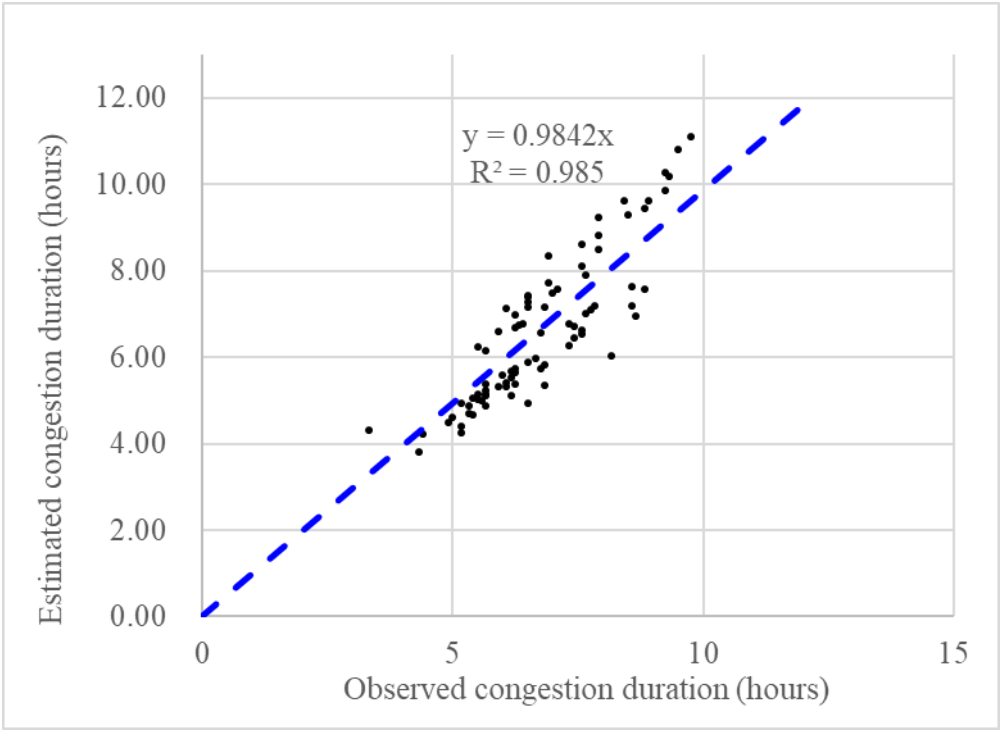


Fig. 21: Estimated vs. observed congestion duration P based on elasticity form.

Since $\frac{v_{co}}{v_{t_2}} - 1 = f_p(P)^s$, we have

$$v_{t_2} = \frac{v_{co}}{f_p \times P^s + 1} = \frac{v_{co}}{f_p {f_d}^s \left(\frac{D}{C}\right)^{ns} + 1} \tag{35}$$

Through calibration using **Eq. (35),** we can estimate elasticity coefficients s = 0.939, and f_p= 0.219, leading to MAE = 6.23 hours and MAPE = 31% between observed and estimated v_{t_2}. As displayed in **Fig. 22**, we further have $f_p {f_d}^s = 0.245$, $ns = 0.982$. Overall, the lowest speed could decrease if the congestion lasts longer. There are outliers that the curve cannot capture. They might be due to irregular traffic states as a result of incidents or bad weather conditions.

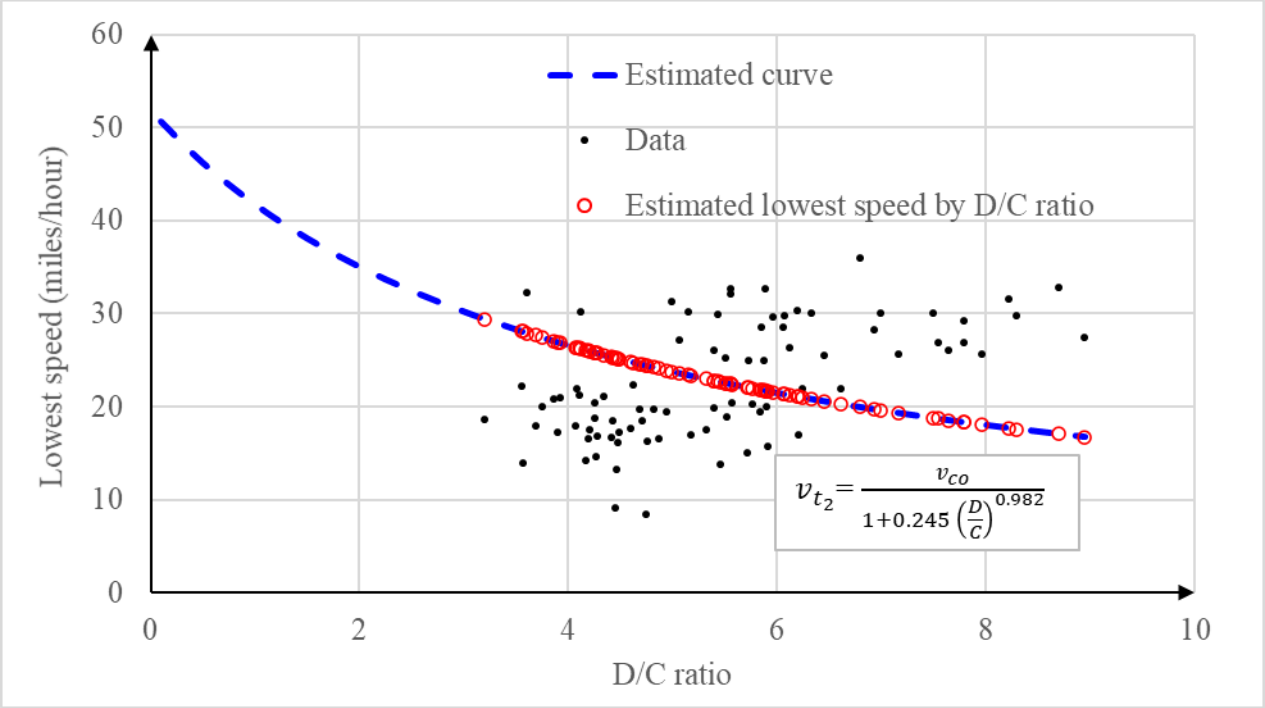


Fig. 22: Lowest speed v_{t_2} vs. observed D/C ratio.

In **Fig.23**, we connect the D/C ratio and the mean speed, i.e., $\bar{v}$ using **Eq. (34)**. The results are illustrated in Fig. **23**, where $\alpha = \theta f_p {f_d}^s = 0.1305$, $\beta = ns = 0.982$. The MAPE of the model is 27%, while MAE = 6.66 mph. Overall, the observed (short-run) demand-to-congestion elasticity show the response.

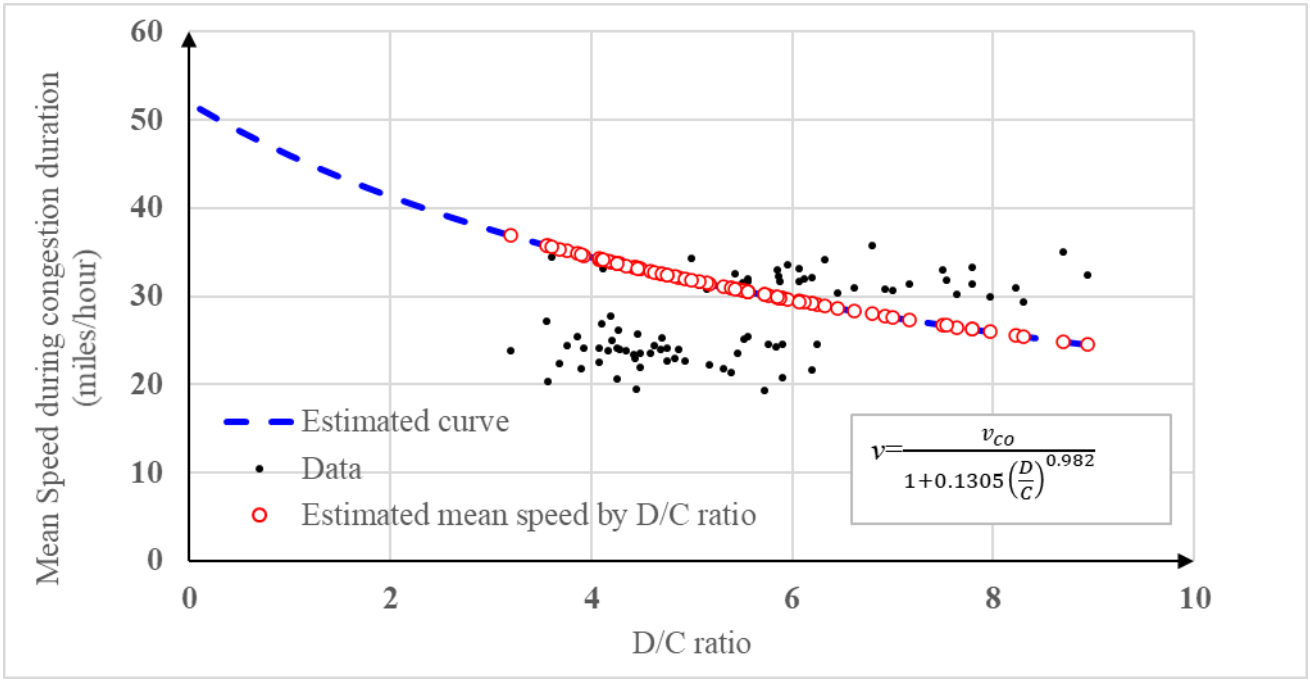


Fig. 23: Mean speed $\bar{v}$ vs. observed D/C ratio.

As an exploration, in this experiment, we estimate the average discharge rate μ using **Eq. (26a).** The MAPE between observed average discharge rate within the congestion duration and estimated μ is 11.86%. **Fig. 24** illustrates the scatters of observed P and observed μ/C. We can also find that the elasticity form is only able to produce a steady mean estimate on discharge rate μ, which is unable to cover all possible variations of capacity discount or throughput of bottlenecks. To capture the day-to-day dynamics of queue discharge rate, further research is needed to incorporate other factors, such as weather and different queue influence areas.

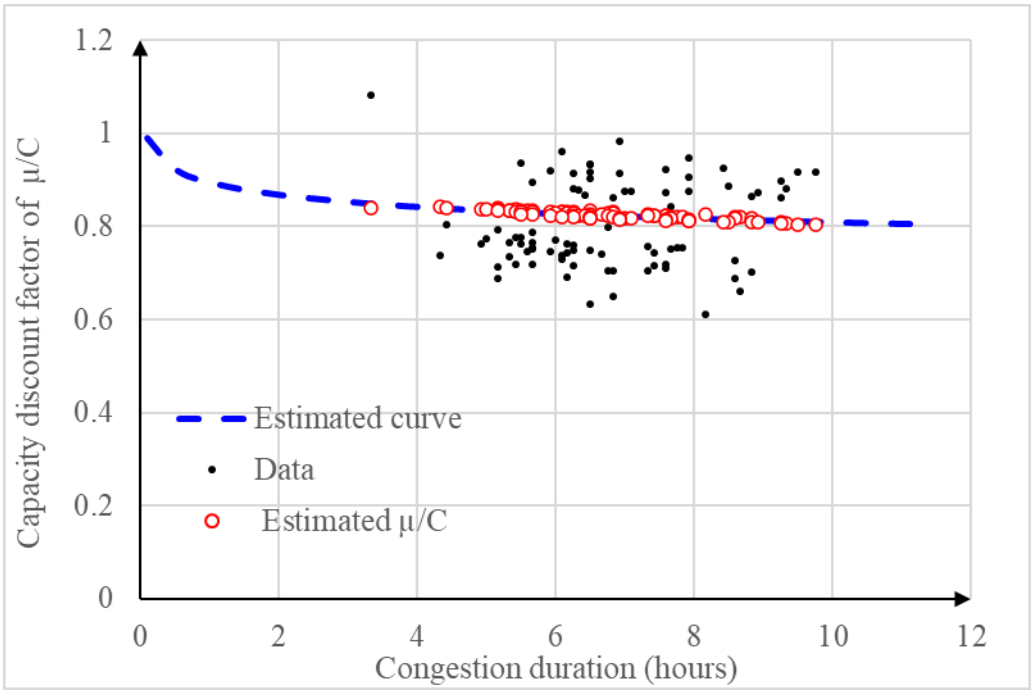


Fig. 24: Capacity discount factor vs. observed congestion duration.

6 Discussion and Conclusions

6.1 System observability and controllability

One of the core purposes to conduct cross-resolution modeling evaluation is to increase the observability and controllability of the system at different levels of fidelity. From the system identification perspective, the time-dependent arrival rate at the bottleneck is not directly observable and must be estimated through other related measurements and controlled by a combination of information and management measures, such as traffic signal controls, road pricing, and smart route guidance. To provide effective congestion mitigation strategies for oversaturated queueing systems, decision-makers need to know: (1) how reliably we can estimate and predict the underlying demand $\lambda(t)$ and supply μ as part of observability quantification tasks; and (2) to what extent we can proactively control the demand inflow curves $\lambda(t)$ and supplied capacity μ at different scales as part of controllability quantification tasks. Strong system observability and controllability could help agencies inform and divert traveling agents around the bottlenecks to avoid recurring and nonrecurring congestions. The proposed cross-resolution model, which can be calibrated at the macroscopic level using a few measurements as inputs, provides good boundary conditions for other models to tackle extremely complex and highly stochastic and dynamic systems at a finer resolution. Specifically, the control measures for an oversaturated dynamic queueing system can be categorized as follows: (1) decreasing the arrival rate from the demand side, so that the congestion occurs later and dissipates earlier, i.e., t_0 shifts right and t_3 moves left; (2) increasing the discharge rate μ from the supply side, which also makes the congestion occur later and dissipate earlier; (3) coordinating traveling agents in the queueing system through pricing, incentives or slot reservation to enable peak shifting in terms of changing the inflow curvature parameter γ; and (4) designing appropriate capacity management strategies, e.g., traffic signal timing, transit scheduling and freeway ramp metering for the traffic system, to effectively enhance the system-level discharge rate μ.

6.2 Conclusion

Based on a family of fluid queue models with different orders of polynomial arrival rates, this study introduces a general meso-to-macro framework with analytical formulations. Specifically, a coherent connection between the macroscopic average travel delay function and the mesoscopic queueing-based vehicular flow model is established. The meso-to-macro derivation process includes the following four steps: (1) assume a polynomial functional form for the inflow rate along with a constant discharge rate; (2) derive a closed-form time-dependent queue length through integrating the difference between the inflow rate and discharge rate; (3) obtain the analytical form of the average delay function in terms of oversaturation period; and then (4) introduce elasticity terms to approximate the overall queue evolution process, that is, express the relative changes of discharge rate (and resulting congestion duration) and lowest speed as functions of macroscopic inflow demand-to-capacity ratio. The proposed cross-resolution approach provides numerically reliable and theoretically rigorous models to capture congested bottlenecks at both macro and meso scales.

We should point out some simplifications in the proposed mesoscopic vehicle flow models based on fluid queues. (1) The origin-destination matrix cannot be directly generated even in a subarea network. (2) The impact of signal controllers is not considered and modeling on queue spillback over complex freeway corridors might not be detailed enough. Future work can also be devoted to extending the proposed analytical cross-resolution formulas in a tighter integration through a feedback loop between mesoscopic DTA models and travel demand models, especially when there are multidimensional travel choice adjustments, such as departure time and/or mode choice. The resulting congestion dynamics can be rapidly evaluated with a consistent queueing dynamic representation.

To manage the modeling complexity in preparing and calibrating system parameters in region-wide or subarea-wide dynamic traffic assignment applications, we recommend a two-stage macro-to-meso process. First, for the base-year network, parsimonious traffic flow models and link-performance models such as QVDF can be calibrated using time-dependent speed and volume measurements, so that macroscopic queueing variables, such as link-level queue discharge rates, congestion duration, and lowest speed, can be estimated in the DTA model. Careful validation efforts are also needed to examine how the observed queue evolution process can be reproduced by the traffic assignment and network loading models using estimated OD matrices. Second, with the analytically derived queue discharge rate and time-dependent in-flow demand flow curve from QVDF as the starting point, planners can further calibrate the dynamic origin-destination demand patterns, spatial queue spillback patterns with additional input data at mesoscopic network levels such as detailed signal timing, time-varying queue discharge rates, and outflow capacity distribution at merges and diverge points. For the future year analysis, the changes of macroscopic traffic flow parameters should be also first considered systematically at the facility type, area type, or corridor levels, e.g., by considering enhanced capacity due to connected automated vehicles and smoothed inflow curvatures due to smart road system reservations. Then a well-estimated macroscopic representation can be followed by detailed mesoscopic modeling efforts for various congestion mitigation strategies and operational improvements.

This paper aims to offer new theoretical insights on cross-resolution modeling of traffic demand-to-congestion elasticity, and we hope the proposed QVDF functional form on the supply side can be tightly integrated with demand-side analysis methods, to name a few, bottleneck with elastic demand (Arnott et al., 1993), road pricing (Yang and Huang, 2005), integration of transportation demand, supply and land-use models (Pinjari et al., 2011), transportation demand elasticity modeling (Litman, 2017; Verbas et al., 2015). It would be interesting to see how such meso-to-macro forms can be adapted to other broader areas such as the modeling of congestion and emission externality (Small amd Verhoef, 2007, Yin and Lawphongpanich, 2016) and economic analysis of ride-sourcing markets (Zha et al., 2016), trip distribution (Yan et al., 2017), and the design of public infrastructure systems with elastic demand (Daganzo and Ouyang, 2019). Equally important, consistent meso-to-macro multimodal volume-delay function are critically needed for transportation planning, especially under mixed traffic conditions with pedestrians and bikes. It would be interesting to establish connections from fluid queue-oriented VDF to many recent multimodal traffic stream models, including the multi-modal MFD proposed by Loder et. (2019) and the analysis framework on bicycle jam density and discharge rate by Wierbos et al. (2021).

Acknowledgments

The first author would like to thank the insightful discussions from Dr. Mohammed Hadi (Florida International University) and Dr. David Hale (Leidos, Inc.) about Congestion and Bottleneck Identification (CBI) Tool Software and multiresolution modeling in transportation systems analysis. The encouragements of Norman L Marshall (Smart Mobility Inc.) and Dr. Nagui M. Rouphail (NC State University) are gratefully acknowledged. We also appreciate the efforts and help from Dr.Arup Dutta, Wang Zhang, Haidong Zhu, and Dr.Vladimir Livshits from Maricopa Association of Governments in an early stage of this research direction.

References

Akçelik, R., 1991. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and an alternative travel time function. Australian Road Research, 21, 49–59.

Akçelik, R., 1978. A new look at Davidson's travel time function. Traffic Engineering and Control, 19(10), 459–463.

Arnott, R., De Palma, A., & Lindsey, R. (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, 161-179.

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

Belezamo, B., 2020. Data-driven methods for characterizing transportation system performances under congested conditions: A Phoenix study. Arizona State University.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., Mishalani, R., 1998. DynaMIT: A simulation-based system for traffic prediction, in: DACCORD Short Term Forecasting Workshop. pp. 1–12.

Boyce, D., Williams, H., 2015. Forecasting urban travel: past, present and future. Edward Elgar Publishing Limited.

Branston, D. (1976). Link capacity functions: A review. Transportation research, 10(4), 223-236.

BPR, 1964. Traffic Assignment Manual.

Carey, M., 2004. Link travel times I: desirable properties. Networks and Spatial Economics, 4, 257–268.

Carey, M., Ge, Y.E., McCartney, M., 2003. A whole-link travel-time model with desirable properties. Transportation Science, 37(1), 83–96.

Carey, M., Humphreys, P., McHugh, M., McIvor, R., 2014. Extending travel-time based models for dynamic network loading and assignment, to achieve adherence to first-in-first-out and link capacities. Transportation Research Part B, 65, 90–104.

Carey, M., McCartney, M., 2002. Behavior of a whole-link travel time model used in dynamic traffic assignment. Transportation Research Part B, 36, 85–93.

CATS, 1960. Data Projections. Chicago.

Cheng, Q., Liu, Z., Guo, J., Wu, X., Pendyala, R., Belezamo, B., Zhou, X., 2022. Estimating key traffic state parameters through parsimonious spatial queue models. Transportation Research Part C, 137, 103596.

Cheng, Q., Liu, Z., Lin, Y., Zhou, X., 2021. An s-shaped three-parameter (S3) traffic stream model with consistent car following relationship. Transportation Research Part B, 153, 246–271.

Daganzo, C.F., 2006. In traffic flow, cellular automata = kinematic waves. Transportation Research Part B, 40, 396–403.

Daganzo, C.F., 1995a. The cell transmission model, part II: Network traffic. Transportation Research Part B, 29(2), 79–93.

Daganzo, C.F., 1995b. Properties of link travel times under dynamic loads. Transportation Research Part B, 29, 95–98.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B, 28(4), 269–287.

Daganzo, C. F. (1995). Properties of link travel time functions under dynamic loads. Transportation Research Part B: Methodological, 29(2), 95-98.

Daganzo, C. F., & Ouyang, Y. (2019). A general model of demand-responsive transportation services: From taxi to ridesharing to dial-a-ride. *Transportation Research Part B: Methodological*, *126*, 213-224.

Davidson, K.B., 1978. The theoretical basis of a flow travel-time relationship for use in transportation planning. Australian Road Research, 8(1), 32–35.

Davidson, K.B., 1966. A flow–travel time relationship for use in transportation planning. Proceedings of th[e] 3rd Australian Road Research Board (ARRB) Conference, 3(1), 183–194.

Dowling, R., Nevers, B., Jia, A., Skabardonis, A., Krause, C., Vasudevan, M., 2016. Performance benefits of connected vehicles for implementing speed harmonization. Transportation Research Procedia, 15, 459–470.

Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L., Wie, B.W., 1993. A variational inequality formulation of the dynamic network user equilibrium problem. Operations Research, 41(1), 179–191.

Gazis, D.C., Herman, R., Potts, R.B., 1959. Car-following theory of steady-state traffic flow. Operations Research, 7(4), 499–505.

Gazis, D.C., Herman, R., Rothery, R.W., 1961. Nonlinear follow-the-leader models of traffic flow. Operations Research, 9(4), 545–567.

Greenshields, B.D., Channing, W., Miller, H., others, 1935. A study of traffic capacity, in: Highway Research Board Proceedings.

Highway capacity manual. (2010). Highway capacity manual. *Washington, DC*

Hadi, M., Zhou, X. and Hale, D., 2022. Multiresolution Modeling for Traffic Analysis: Guidebook (No. FHWA-HRT-22-055). United States. Federal Highway Administration.

Hale, D., Jagannathan, R., Xyntarakis, M., Su, P., Jiang, X., Ma, J., Hu, J. and Krause, C., 2016. Traffic bottlenecks: identification and Solutions (No. FHWA-HRT-16-064). United States. Federal Highway Administration. Office of Operations Research and Development.

Huntsinger, L.F., Rouphail, N.M., 2011. Bottleneck and queuing analysis: calibrating volume–delay functions of travel demand models. Transportation Research Record, 2255(1), 117–124.

Litman, T. (2017). *Understanding transport demands and elasticities*. Victoria, BC, Canada: Victoria Transport Policy Institute.

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. Transportation Research Record, 1572(1), 140–147.

Loder, A., Bressan, L., Wierbos, M. J., Becker, H., Emmonds, A., Obee, M., Knoop, V. Menendez, M., & Axhausen, K. W. (2019). A general framework for multi-modal macroscopic fundamental diagrams (MFD). *Arbeitsberichte Verkehrs- und Raumplanung*, *1444*.

Mahmassani, H.S., Hu, T.Y., Jayakrishnan, R., 1992. Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART), in: Proceedings of the 2nd International Capri Seminar on Urban Traffic Networks. Capri, Italy.

Mahmassani, H.S., Saberi, M. and Zockaie, A., 2013. Urban network gridlock: Theory, characteristics, and dynamics. Procedia-Social and Behavioral Sciences, 80, pp.79-98.

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. Research in Transportation Business and Management, 29, 85–92.

Mtoi, E.T., Moses, R., 2014. Calibration and evaluation of link congestion functions: Applying intrinsic sensitivity of link speed as a practical consideration to heterogeneous facility types within urban network. Journal of Transportation Technologies, 4(2), 141–

149.

Muranyi, T.C., 1963. Trip distribution and traffic assignment, in: Traffic Assignment Conference. Chicago Area Transportation Study, Chicago.

Nagel, K., 1996. Particle hopping models and traffic flow theory. Physical Review E, 53(5), 4655–4672.

Newell, G.F., 2002. A simplified car-following theory: A lower order model. Transportation Research Part B, 36(3), 195–205.

Newell, G.F., 1993a. A simplified theory of kinematic waves in highway traffic, part I: General theory. Transportation Research Part B, 27(4), 281–287.

Newell, G.F., 1993b. A simplified theory of kinematic waves in highway traffic, part II: Queueing at freeway bottlenecks. Transportation Research Part B, 27(4), 289–303.

Newell, G.F., 1993c. A simplified theory of kinematic waves in highway traffic, part III: Multi-destination flows. Transportation Research Part B, 27(4), 305–313.

Newell, G.F., 1982. Applications of queueing theory, 2nd ed. Chapman and Hall Ltd, New York.

Newell, G.F., 1968a. Queues with time-dependent arrival rates I—the transition through saturation. Journal of Applied Probability, 5(2), 436–451.

Newell, G.F., 1968b. Queues with time-dependent arrival rates: III. A mild rush hour. Journal of Applied Probability, 5(3), 591–606.

Newell, G.F., 1968c. Queues with time-dependent arrival rates: II. The maximum queue and the return to equilibrium. Journal of Applied Probability, 5(3), 579–590.

Nie, X., Zhang, H.M., 2005a. Delay-function-based link models: their properties and computational issues. Transportation Research Part B, 39, 729–751.

Nie, X., Zhang, H.M., 2005b. A comparative study of some macroscopic link models used in dynamic traffic assignment. Networks and Spatial Economics, 5, 89–115.

Nie, Y., Ma, J., Zhang, H.M., 2008. A polymorphic dynamic network loading model. Computer-Aided Civil and Infrastructure Engineering, 23, 86–103.

Pinjari, A. R., Pendyala, R. M., Bhat, C. R., & Waddell, P. A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, *38*(6), 933-958.

Qu, Y., Zhou, X., 2017. Large-scale dynamic transportation network simulation: A space-time-event parallel computing approach. Transportation Research Part C, 75, 1–16.

Ran, B., Boyce, D.E., 1997. Toward a class of link travel time functions for dynamic assignment models on signalized networks. Transportation Research Part B, 31(4), 277–290.

Ran, B., Boyce, D.E., LeBlanc, L.J., 1993. A new class of instantaneous dynamic user-optimal traffic assignment models. Operations Research, 41(1), 192–202.

Small, K.A., 1983. The incidence of congestion tolls on urban highways. Journal of Urban Economics, 13(1), 90–111.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. Routledge.

Smock, R., 1962. An iterative assignment approach to capacity restraint on arterial networks. Highway Research Board Bulletin, 347, 60–66.

Smock, R.B., 1963. A comparative description of a capacity-restrained traffic assignment. Highway Research Record, 6, 12–40.

Spiess, H., 1990. Conical volume-delay functions. Transportation Science, 24(2), 153–158.

Tisato, P., 1991. Suggestions for an improved Davidson travel time function. Australian Road Research, 21(2), 85–100.

Verbas, İ. Ö., Frei, C., Mahmassani, H. S., & Chan, R. (2015). Stretching resources: sensitivity of optimal bus frequency allocation to stop-level demand elasticities. *Public Transport*, *7*(1), 1-20.

Vickrey, W., 1969. Congestion theory and transport investment. The American Economic Review, 59, 251–260.

Wierbos, M. J., Knoop, V. L., Bertini, R. L., & Hoogendoorn, S. P. (2021). Influencing the queue configuration to increase bicycle jam density and discharge rate: An experimental study on a single path. *Transportation research part C: emerging technologies*, *122*, 102884.

Wu, X., Dutta, A., Zhang, W., Zhu, H., Livshits, V., Zhou, X., 2020. Characterization and calibration of volume-to-capacity ratio in volume- delay functions on freeways based on a queue analysis approach (TRBAM-21-04304), in: Proceedings of the 100th Annual Meeting of Transportation Research Board.

Yin, Y., & Lawphongpanich, S. (2006). Internalizing emission externality on road networks. *Transportation Research Part D: Transport and Environment*, *11*(4), 292-301.

Yan, X.Y., Wang, W.X., Gao, Z.Y., Lai, Y.C., 2017. Universal model of individual and population mobility on diverse spatial scales. Nature Communications, 8, 1639.

Yang, H., & Huang, H. J. (2005). *Mathematical and economic theory of road pricing. Elsevier*

Zha, L., Yin, Y., & Yang, H. (2016). Economic analysis of ride-sourcing markets. Transportation Research Part C: Emerging Technologies, 71, 249-266.

Zhou, X., Hadi, M., Hale, D., 2021. Multiresolution modeling for traffic analysis: State-of-practice and gap analysis report (FHWA-HRT-21-082).

Zhou, X., Tanvir, S., Lei, H., Taylor, J., Liu, B., Rouphail, N.M., Frey, H.C., 2015. Integrating a simplified emission estimation model and mesoscopic dynamic traffic simulator to efficiently evaluate emission impacts of traffic management strategies. Transportation Research Part D, 37, 123–136.

Zhou, X., Taylor, J., 2014. DTAlite: A queue-based mesoscopic traffic simulator for fast model evaluation and calibration. Cogent Engineering, 1(1), 961345.