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], Baloka Belezamo [b], Peiheng Li [c]*

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

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

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

** Corresponding author. Email: xzhou74@asu.com (X. Zhou), qcheng15@asu.edu (Q. Cheng), xinwu3@asu.edu (X. Wu), bbelezamo@azdot.gov (B. Belezamo), peiheng.li@nscorp.com (P. Li).*

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 the deficiencies of the static VDF in capturing traffic flow dynamics and queue evolution processes. Many queueing-based and simulation-based mesoscopic vehicular flow based dynamic traffic assignment (DTA) models are used to capture traffic system dynamics on different spatial scales; however, calibrating these DTA models involving many traffic flow parameters could be a challenging task in its own right, especially for real-world congested networks with complex traffic dynamics. By applying the fluid queue model with quadratic arrival rates proposed by Newell (1982) and cubic arrival rates by Cheng et al. (2022), we hope to explicitly establish a coherent connection between (a) the macroscopic average travel delay performance relationship in terms of inflow-demand-to-capacity ratio and (b) the deterministic dynamic queuing model during a single oversaturated period which is typically used in a mesoscopic simulation framework. The meso-to-macro derivation process includes the following steps: (1) assume a polynomial functional form for the inflow rate with a constant discharge rate; (2) derive a closed-form of time-dependent queue length based on the integration of 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 rates (and resulting congestion duration) and lowest speed as functions of macroscopic inflow-demand-to-capacity ratio. The proposed cross-resolution approach can provide numerically reliable and theoretically rigorous models to capture congested bottlenecks at both macro and meso scales.

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

1 Introduction

1.1 Necessity for cross-resolution modeling of traffic flow dynamics

Multi-Resolution Modeling (MRM) is a modeling technology that creates a family of models that represent the same phenomenon or a set of questions at more than two different resolutions. The fine-grained spatial scales could cover corridors, roads, and lane representations, and the temporal resolution refers to the time interval (or time stamps) at which the dynamic state of the model is updated, typically ranging from days or seconds. Each type of model (macroscopic, mesoscopic, or microscopic) has its own advantages and disadvantages, and represents a trade-off between scales and resolution levels. The ultimate goal of MRM is how to seamlessly integrate models with different temporal and spatial resolutions, while the focus of the cross-resolution approach is on how to bridge the gaps across macroscopic and microscopic levels, so as to provide strong theoretical support and deeper insights for both levels.

From a broader perspective of ensuring traffic flow modeling consistency across different resolutions, many important studies have been devoted to this research direction. The pioneering work by Gazis et al. (1959) first highlighted that the fundamental diagram, such as Greenshields' speed-density relationship, can be linked to the microscopic car-following model (Greenshields et al., 1935; Gazis et al., 1961). An important 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.

Network flow models in large-scale transportation networks are critically needed to assess at different spatial scales, e.g., corridor-level, segment-level and cell-level (e.g., seven meters in length for cellular automata models), and at various temporal resolutions (e.g., min-by-min, peak hours, the entire day, and multiple days). Link travel time functions (a.k.a. link performance functions or volume-delay functions) have been widely used in the static traffic assignment (STA) for transportation planning. Among a variety of link performance functions, the BPR function, created by the US Bureau of Public Roads in 1964, is recognized as an analytical building block for system-wide performance evaluation. Focusing on the relationship of traffic volume and delay, the BPR function as a simple polynomial equation is computationally efficient for transportation planning software implementations. On the other hand, the planning community has long recognized that the static BPR function cannot capture traffic flow dynamics and queue evolution process, particularly related to the queue formation, propagation, and dissipation. Besides, it has difficulties in using the average travel time measure to describe an oversaturated bottleneck with high density but low throughput.

Queueing models are widely used to describe system dynamics and queue evolution process with time-dependent arrival and discharge rates[1]. There are many extended queuing models in traffic flow modeling, and interested readers are referred

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

to the book by Newell (1982). 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 traffic flow 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 Daganzo's cell transmission model (Daganzo, 1994, 1995a) that adopts a "supply-demand" or "sending-receiving" framework to model flow dynamics between discretized cells. Besides, there are partial-differential-equation (PDE)-based numerical analysis approaches and customized simulation packages to capture microscopic interactions between the demand and supply (Behrisch et al., 2011; Marshall, 2018) at a very fine resolution.

Mesoscopic dynamic traffic assignment (DTA) and microscopic traffic simulation have been recognized as the key building blocks in describing traffic dynamics under congested conditions. Many agent-based simulation tools with mesoscopic traffic flow models, e.g., DynaMIT (Ben-Akiva et al., 1998), DYNASMART (Mahmassani et al., 1992), and DTALite (Zhou and Taylor, 2014), etc., with mesoscopic traffic flow models are developed to capture system dynamics and queue evolution process at a finer resolution. In general, the simulation-based dynamic system modeling approach could be computationally intensive, especially for real-world networks with complex traffic dynamics. A wide range of approaches is proposed for addressing such computational challenges. (1) A *coarse simulation approach* referred to as "Cellular Automata" (CA) can be used to keep up with the fast computational speed necessary to simulate a whole region. (2) A *consistent cross-resolution* traffic state representation based on Newell's simplified kinematic wave and linear car following models was also introduced by Zhou et al. (2015) for estimating emissions and fuel consumption efficiently. (3) *Parallel computing* is introduced to utilize multiple computer processors (CPUs) (e.g., in the TRANSIMS micro-simulator) to support a large number of travelers and a considerably sized transportation network. A recent effort by Qu and Zhou (2017) introduces a hybrid time-based and event-based data structure for parallel computing-based mesoscopic dynamic traffic simulation.

Nonetheless, there are multiple impediments associated with a loose linkage or potential inconsistency between macroscopic and mesoscopic traffic flow models. Although the low-resolution traffic flow models, for example, spatial queue or simplified kinematic wave models used in DTA can be viewed as a reasonable approximation to the complex real-world traffic flow dynamics. By simply mixing models with different resolutions together, microscopic traffic simulation results and the resulting link travel times could be significantly different from the link performance statistics at the macroscopic level, which leads to internal discrepancy between different levels of modeling resolution, and hinders tight interconnections and further iterations between different simulation/assignment components. A recent report on the state-of-the-practice and gap analysis by Zhou et al. (2021) provides a review of multi-resolution modeling terminology, tools, and literatures in traffic analysis applications.

Undoubtedly, addressing these critical concerns and barriers in full-scale implementation is an extremely important and challenging task before practitioners are ready to deploy consistent cross-resolution models for assessing the impacts of traveler information provision and management strategies in a wide range of applications. Analytical models are advocated to overcome the computational issues in simulation-based dynamic system modeling approaches due to their simplicity and tractability. If they are elaborately designed, analytical models can also capture system dynamics and describe the queue evolution process with time-dependent queue length, delay, and travel time. This opens a new window to model and analyze dynamic traffic systems more efficiently and effectively. In this research, we aim to address the following two technical challenges to enable internally consistent meso-to-macro cross-resolution traffic flow modeling at the level of corridor bottlenecks. (1) From the perspective of *macro-to-meso mapping*: how to ensure the simplified analytical volume-delay function (VDF) model, which is commonly used in macro-level traffic assignment tasks, can reasonably and accurately represent traffic flow dynamics under typical patterns of arrival flow and degrees of oversaturation? (2) In view of the *meso-to-macro mapping*: at the level of freeway bottlenecks, how to efficiently obtain and simulate meso-level vehicular trajectories of individual agents that are theoretically consistent with the macroscopic demand volume and the average delay in the path-level assignment? We select the term "meso" here and in the title because the microscopic simulation models (e.g., VISSIM, Paramics, Aimsun, TRANSIMS), which move traffic by capturing individual driver maneuvers such as car-following, overtaking, lane changing, and gap acceptance decisions, have been viewed as somewhat being distinct from the mesoscopic models, which provide a hybrid approach to modeling traffic propagation.

1.2 Literature review on volume-delay and link performance functions

In order to evaluate the link travel time performance, there are various VDFs proposed during the beginning of the transportation planning discipline. The seminal work was conducted by CATS (acronym of Chicago Area Transportation Study, 1960) in the early 1960s for the capacity restraint traffic assignment problem and the CATS function was explicitly stated in Muranyi (1963). This CATS function is to multiply the free-flow travel time by a factor of two raised to the power of volume-to-capacity 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, in which 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 the *Traffic Assignment Manual*, in which the classic BPR (1964) VDF 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 the capacity. Determining the values of BPR parameters α and β could be challenging as the proper values could significantly vary from region to region. Practitioners have also noted that the BPR function leads to an overestimation of speeds for *V/C* ratios of greater than one and an underestimation of volumes for *V/C* ratios of less than one. Spiess (1990) proposed a number of conditions for "well behaved" volume-delay functions to ensure that the original form and its first-order gradient are strictly increasing and change of congestion effects is reasonable when the capacity is reached. He also recommended a set of conical congestion functions and identified further research needs for directly estimating the parameters using observed speeds and volumes.

Davidson (1966) derived a VDF based on stochastic queueing theory without a clear demonstration, while 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 traffic volume below the capacity since the travel time approaches infinity when the volume approaches the capacity in his formulation. In order to model the travel time when the volume is near or over the capacity, a variety of modified Davidson's form have been proposed. Akçelik (1978) modified Davidson's function by linearly extending the function's slope to yield a finite travel time under oversaturated conditions for both user equilibrium and system optimal traffic assignment problems. As claimed in Tisato (1991), Akçelik's (1978) modification behaves poorly when the 'quality of service' parameter m approaches its upper limit of one, and the modified function is still static rather than time-dependent. Tisato (1991) developed an alternative form on top of Akçelik's (1978) modified Davidson's function, which is time-dependent and influenced by the congestion duration under oversaturated conditions. As Tisato's modification is shown to overpredict the travel times for flows 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. However, all of these link travel time performance functions cannot describe the queue evolution process, and this paper will bridge this gap with the approximation of inflow rate by a polynomial function.

The link travel time performance functions, based on the number of vehicles on the link, the inflow and/or the outflow rate at time t, can be used for dynamic traffic assignment (DTA) problems as well. Ran et al. (1993), Friesz et al. (1993), Ran and Boyce (1996a, 1996b) 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 properties, including uniqueness, continuity, causality, consistency, and first-in-first-out (FIFO), are discussed on the link travel time functions (Carey, 2004). Zhang and his colleagues (Nie and Zhang, 2005a, 2005b; Nie et al., 2008) investigated various FIFO-consistent link travel time functions and revealed the following observations. The FIFO-consistent linear travel time function may overestimate the results, the improved 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 is FIFO-consistent only for some inflow profiles, and a polymorphic travel time function would be a good choice for the FIFO-consistent DTA. Recently, Carey et al. (2014) extended the link-travel-time-based DTA models to ensure the FIFO and capacity constraints and strengthen the realism and behavioral dynamics.

Many practitioners have the following interesting questions related to commonly used VDFs: (1) where are those coefficients of α and β in the BPR function coming from? (2) how to use observed traffic dynamics data to calibrate these coefficients? Aiming to help transportation planners to understand the above two questions, this study presents a new cross-resolution travel time performance model, namely the fluid-based polynomial arrival queue (PAQ) model, for system-wide performance evaluation. Based on continuous-time modeling and polynomial-based approximation, the proposed model explicitly establishes a coherent connection between the average performance and the deterministic dynamic queuing model during a single oversaturated period.

The organization of this paper is summarized as follows. Sections 2 and 3 first introduce the building block of the fluid queue model with quadratic arrival rates proposed by Newell (1968a, 1982) and cubic arrival rates by Cheng et al. (2022). Section 4 examines how to use a set of elasticity terms to approximate the overall queue evolution process, and establish the meso-to-macro relationship. Calibration results are presented in Section 5, which is followed by a range of discussions on the applications related to these cross-resolution models in Section 6. Finally, Section 7 concludes this study.

2 Early Efforts by Newell's Fluid Queueing Model

Although Newell's quadratic model (Newell, 1968a, 1982) is only applicable to mild traffic conditions (Cheng et al., 2022), it is a building block for our polynomial arrival queue (PAQ) model; therefore, we will first introduce Newell's model in this section. Table 1 summarizes the symbols used in this study and Fig. 1 illustrates the modeling procedures of Newell's model.

Table 1: Symbols and definitions used in this study.

Symbols	Definitions
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 demand during the whole peak period
t_f	free-flow travel time
ρ	shape parameter used in Newell's quadratic form of inflow rates
$\lambda(t)$	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

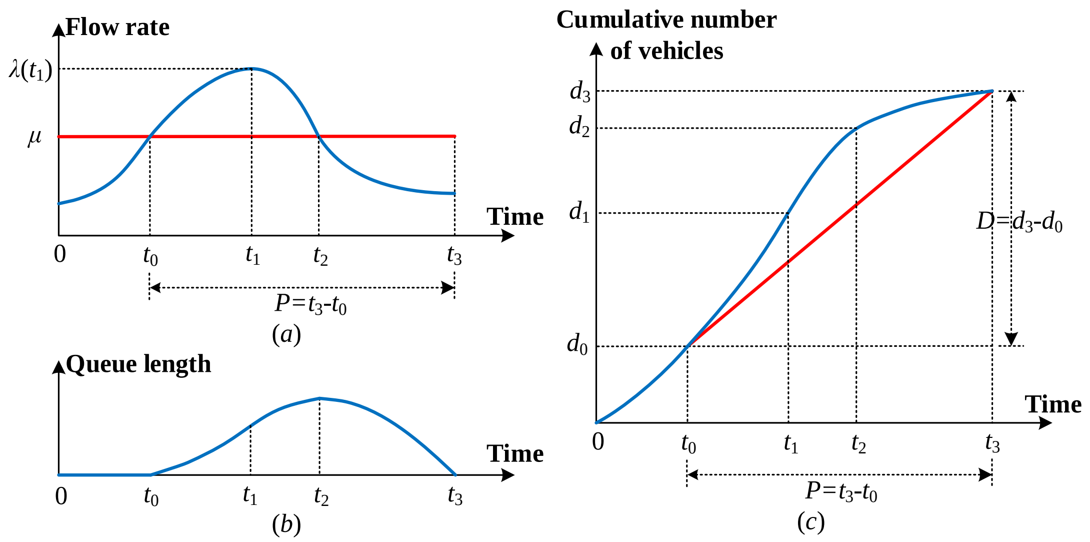


Fig. 1: General graphical illustration of queue evolution for a single oversaturated bottleneck (Newell, 1982).

Let t_1 be the time with maximum inflow rate as shown in Fig. 1(a). Newell assumes that the inflow rate near time t_1 can be approximated by the following quadratic function (Newell, 1982):

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

It is obvious that $\lambda'(t_1) = 0$. Let $\rho = -\frac{1}{2}\lambda''(t_1)$, then Eq. (1) can be transformed to

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

The discharge rate μ, (where $\mu = \lambda(t_0) = \lambda(t_2)$), can be estimated in terms of Eq. (2):

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

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

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

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

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

It is obvious that $\lambda(t)$ in Eq. (5) passes two points (t_0, μ) and (t_2, μ), and its second derivative with respect to t is -2ρ, which is consistent with Eq. (2).

The queue length at time t equals $A(t) - D(t)$, which can be obtained from Eq. (6):

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

After substituting Eq. (5) into Eq. (6), we can get the queue length at time t in terms of t_0, t_2, and ρ:

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

The maximum queue length at time t_2 is:

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

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:

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

The total delay between time t_0 and t_3 can also be calculated by the area between $A(t)$ and $D(t)$ in Fig. 1(c), which can be calculated by integrating Eq. (10):

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

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:

$$\bar{w} = \frac{W}{D} = \frac{\rho(t_3 - t_0)^4}{36D} \tag{12}$$

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:

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

Newell (1982) explicitly remarked that quadratic model is only applicable to analyze mild traffic conditions. Cheng et al. (2022) further pointed out that when the system is in an oversaturated state, the arrival rate might be a counterintuitive negative value using Newell's quadratic model. In addition, it is important to note that the classical BPR function is formulated via the demand-over-capacity ratio, while the fluid-based queueing model is established in terms of the demand-over-discharge-rate, in which the discharge rate μ is usually much lower than the maximum capacity C used in the BPR function. According to Cheng et al. (2022), the key parameter ρ in Eq. (14) can be highly sensitive to the underlying queueing behavior, as opposed to the typically assumed constant value of α in the BRP function.

3. Polynomial Arrival Queue Model with Cubic Forms

Recently, Cheng et al. (2022) revisited Newell's fluid-based queueing model (Newell, 1968a, 1968b, 1968c, 1982), and described queueing systems with analytical time-dependent arrival rates. To overcome the drawbacks in Newell's quadratic model, they extended it to a cubic model, i.e., the arrival rate is assumed to be a cubic function for one peak period, and the assumption was calibrated and validated using empirical data. This cubic model is ideal for efficient dynamic system modeling and management as it is able to analytically calculate the time-dependent queue length, delay, and travel time.

The general framework for the polynomial arrival queue model can be found in Cheng et al. (2022), and we only summarize its key points here. Considering a single bottleneck where the time-dependent arrival rate can be approximated by a polynomial function (which could be linear, quadratic, cubic, etc.) and the discharge rate is assumed to be constant during the peak period of interest. Without loss of generality, we use the cubic function form to illustrate the derivation of the system dynamics.

The net flow rate, i.e., the difference between the arrival rate $\lambda(t)$ and discharge rate μ, is expressed with a factor form as $\lambda(t) - \mu = \gamma(t - t_0)(t - t_2)(t - \bar{t})$, where t_0 and t_2 follow the same definitions in Newell's model, and $\bar{t}$ is an auxiliary time representing an additional root besides t_0 and t_2, and γ is a shape parameter to be calibrated. All other notations follow Newell's model. Fig. 2 illustrates the flow rate, queue length, and cumulative flow rate evolution process over time in the cubic model (Cheng et al., 2022).

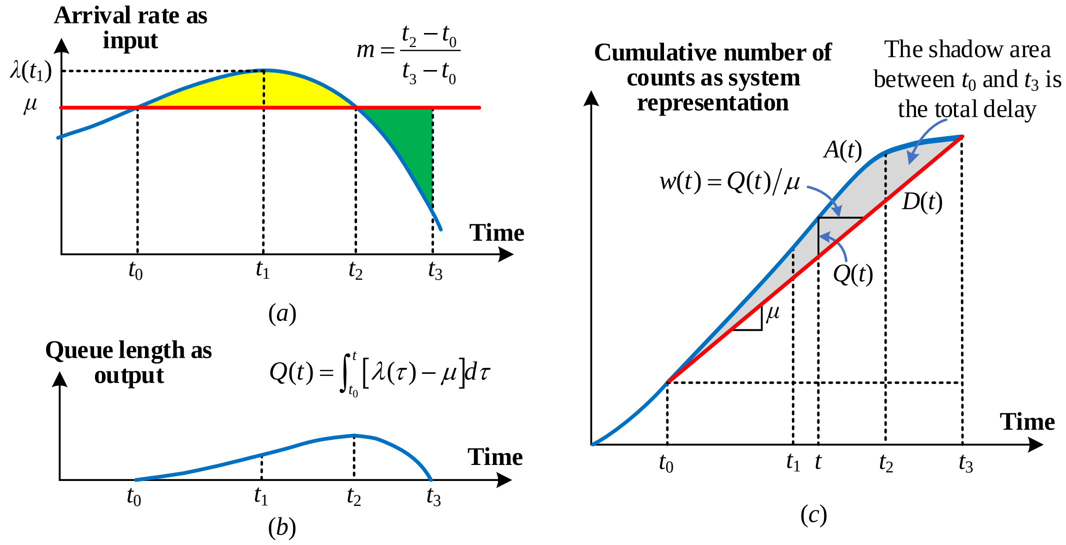

Arrival rate as input
$\lambda(t_1)$
μ
$m = \frac{t_2 - t_0}{t_3 - t_0}$
Time
t_0
t_1
t_2
t_3
(a)
Queue length as output
$Q(t) = \int_{t_0}^{t} [\lambda(\tau) - \mu] d\tau$
Time
t_0
t_1
t_2
t_3
(b)
Cumulative number of counts as system representation
The shadow area between t_0 and t_3 is the total delay
$A(t)$
$w(t) = Q(t)/\mu$
$D(t)$
$Q(t)$
μ
Time
t_0
t_1
t
t_2
t_3
(c)

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

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

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

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

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

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

where,

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

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

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

Note that $\gamma < 0$ is only applicable to mild traffic conditions and the model reduces to Newell's quadratic model when $\gamma = 0$. Therefore, we only recommend $\gamma > 0$, which applies to both mild and oversaturated traffic conditions. The detailed derivation can be found in Cheng et al. (2022).

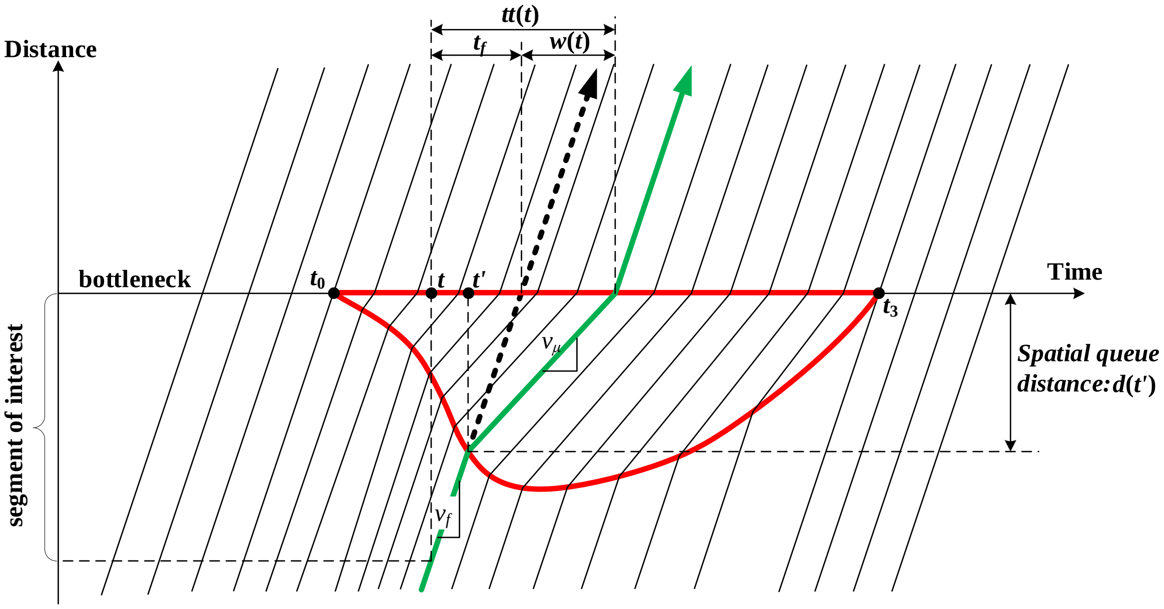


Fig. 3: Illustration of the mesoscopic physical vehicle trajectory and virtual queue length (Lawson et al.,1997, Cheng et al., 2022).

In quadratic and cubic PAQ models, a point queue model is presumed where vehicles are assumed to be zero-length. Thus, when the traffic system is oversaturated, vehicles will queue at the bottleneck until there is enough receiving capacity downstream of the bottleneck. The number of those with zero-length waiting at the bottleneck is regarded as the virtual queue length. When calibrating the PAQ model, one may need to map the virtual queue length to the physical queue length, which is the number of vehicles upstream of the bottleneck. In addition, the spatial queue distance is the queue extent spatially upstream of the bottleneck (as shown in Fig. 3).

According to the input-output diagram proposed by Lawson et al. (1997), the physical mesoscopic queue length can be obtained in terms of the virtual queue length calculated by Eq. (15) (Cheng et al., 2022):

$$Q^p(t') = \frac{Q(t)}{1 - \frac{v_\mu}{v_f}} \tag{22}$$

where t' is the time when a vehicle encounters the tail of the queue (see Fig. 3 for illustration), and v_f and v_μ are the free-flow speed and the speed at capacity, respectively. Thus, one can obtain the physical queue length from virtual queue length and map the mesoscopic vehicle trajectories to the macroscopic system performance in terms of Eq. (22).

4 Linking Polynomial Arrival Queue with Macroscopic Volume-Delay Function

4.1 Definition of inflow demand volume-to-capacity ratio

From the perspective of macroscopic stationary-state analysis, traffic demand is usually defined as vehicles planning to pass a roadway section, possibly endogenously, in a given time period. If the traffic 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 traffic demand rates exceed the capacity, there are vehicles being held and the measured flow rate only represents the discharge rate that can be handled by the transportation system but not the demand rate. More importantly, the congestion duration P is dependent on two factors, i.e., the overall inflow demand pattern and the discharge rate. Understanding the inflow demand significantly helps us to derive the link performance 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. Some researchers, such as Huntsinger and Rouphail (2011), stated that the volume-over-capacity ratio should be demand-over-capacity in the BPR function. Huntsinger and Rouphail (2011) showed that the demand beyond capacity can be estimated by performing bottleneck and queue analysis according to the following equation: *demand = capacity + queue length.* Their form is more suitable for time-dependent delay calculation with clear information in the queue influence area. Mtoi and Moses (2014), on the other hand, proposed a straightforward method of computing the term of "demand above capacity" as *demand = capacity + (capacity – measured flow).* This method simply means that the demand is equal to the capacity plus the drop in the measured flow after reaching the capacity. Dowling et al. (2016) referred to the extra demand as carry-over demand, which indicates the presence of a queue at a specific location. Teknomo and Gardon (2018) proposed that traffic flow is modeled as a function of the proportional capacity of inflow and outflow. A recent dissertation by Belezamo (2020) further examined the trade-offs between the aforementioned demand definitions.

In our proposed approach for the macroscopic VDF, D is defined as the total inflow demand volume during the peak period from t_0 to t_3. The congestion period is defined as P (i.e., $P = t_3 - t_0$). The inflow-demand-to-capacity D/C ratio can offer a more precisely defined and queue-theoretic measure consistent with the congestion dynamic than the volume over capacity (V/C) ratio. A recent study by Wu et al. (2020) examined different ways to calibrate critical density or critical speed based on a well-defined fundamental diagram and further defined the congestion duration P and total demand for different types of facility and area.

Focusing on the inability of the standard BPR function to accommodate the duration of traffic exceeding capacity, Small (1983) developed a duration-dependent function to express the average travel delay over P as a piecewise-linear function $\bar{w} = \frac{1}{2}P\left(\frac{D}{C} - 1\right)$ where D follows a uniform rate and delay results from the queue behind a single bottleneck with a constant capacity C. Small (1983) found that above equation well approximates the travel delay pattern during the afternoon peak period on an 11-mile freeway segment in the San Francisco Bay area. Compared to the fluid based model, the formula by Small (1983) gives an intuitive value of $\bar{w}$=0 when *D/C*=1, but it is difficult to directly measure P using the real-world data if a constant flow is assumed over a period, as further discussed by Small and Verhoef (2007). Furthermore, P and inflow rate are typically assumed to be exogenous parameters, and the rational demand behavior are presumed by an endogenous departure time choice model.

4.2 Requirements for a well-behaved queueing-theoretical volume-delay function

There are a number of issues to address to construct the VDF $f(x)$, where $x = D/C$, typically expressed in terms of volume-over-capacity. Let us revisit and enhance the requirements for a well-behaved congestion function proposed by Spiess (1990), with the highlighted features related to our queue-theoretical delay function under consideration.

(1) $f(x)$ *should be strictly increasing.* This is the necessary condition for the assignment to converge to a unique solution.

(2) $f(x = 0) = 0$ *and* $f(x = 1) = constant$ These conditions ensure compatibility with the well-known BPR type functions.

(3) $f'(x)$ *exists and is strictly increasing.* This ensures convexity of the congestion function and this is important when calculating marginal costs in system optimum assignment.

(4) $f'(x) < M$*, where* M *is a positive constant.* The steepness of the congestion curve is limited so that the derived average delay is reasonable when considering a relatively large value of D/C ratio especially during the first few iterations of an equilibrium assignment, e.g., values of D/C could reach values of 3, 5 or even larger.

(5) **(Consistency of demand definition)** Definitions of demand *D* and analysis duration *P* for time-averaged delay should be consistent to the queueing dynamics.

(6) **(Consistency of time-averaged delay)** The time-averaged delay in both macroscopic model and mesoscopic vehicular fluid model should be consistent.

(7) **(FIFO)** The time-dependent travel time in the underling mesoscopic vehicular

fluid model should satisfy the first in and first out property and capacity constraints.

Table 2 summarizes different functional forms and key parameters of travel delay function under different orders of arrival rate pattern (Cheng et al., 2022). A common feature is that they are functions of inflow demand volume *D*, discharge rates μ, and shape parameters. Shape parameters vary over different orders of polynomial arrival rate functions, namely, π_1 and π_2 (i.e., the over-capacity and under-capacity flow rates) for the constant form, κ for the linear form, ρ for quadratic form, and γ for the cubic form.

Table 2: The arrival rates of mesoscopic vehicular flow and corresponding macroscopic travel delay functions (details on the figure explanations and derivations are referred to Cheng et al., 2022).

Order of polynomial form	Arrival rate function	Average travel delay function
Constant form	$\lambda(t)=\begin{cases}\pi_1>\mu,\ t_0\le t<t_2\\ \pi_2<\mu,\ t_2\le t\le t_3\end{cases}$	$\bar{w}=\frac{(\pi_1-\mu)(\mu-\pi_2)}{2\mu(\pi_1-\pi_2)}\cdot\left(\frac{D}{\mu}\right)$
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	$\lambda(t)=-\rho(t-t_0)(t-t_2)+\mu,\ \rho>0$	$\bar{w}=\frac{\rho}{36\mu}\cdot\left(\frac{D}{\mu}\right)^3$
Cubic form	$\lambda(t)=\gamma(t-t_0)(t-t_2)(t-\bar{t})+\mu$	$\bar{w}=\frac{\gamma\cdot g(m)}{\mu}\cdot\left(\frac{D}{\mu}\right)^4$

4.3 Derive the average speed based on two assumptions

In this study, we select the PAQ with cubic arrival rates as an example for deriving a queue-theoretic VDF, and construct the following two mapping functions, *D/C* to *P* and *P* to the lowest speed, so that we can express discharge rate μ and shape parameter γ as a function of *D/C*. One can apply similar derivation for different orders of PAQ in Table 2.

(1) Mapping function to express discharge rate μ as a function of *D/C* In general, the elasticity term of a function (e.g., demand as a function of ticket price) shows the relative percentage change of a given variable due to a relative percentage change from another variable. Specifically, the elasticity of a function is a constant p if the function has the form $f(x)=x^p$. The point elasticity of a positive differentiable function at point x is defined as $Ef(x)=\frac{d\log f(x)}{d\log x}$. Similarly, we define n as the oversaturation-to-duration elasticity factor, and consider the congestion duration as a function of D/C as below,

$$P=\left(\frac{D}{C}\right)^n \tag{23}$$

According to Eq. (23), we have $P^{(n-1)/n}=(D/C)^{n-1}$. As $P=D/\mu$, one can rewrite Eq. (23) as the following mapping function with respect to *D/C*:

$$\mu=\frac{D}{P}=\frac{D}{(D/C)^n}=\frac{C}{(D/C)^{n-1}}=\frac{C}{P^{(n-1)/n}} \tag{24}$$

(2a) Mapping function to express shape parameter γ as a function of *D/C*

Let us consider v_c/v_{t_2} as the magnitude of speed reduction between the speed at capacity and the lowest speed. We further define s as the duration-to-speed reduction elasticity factor and construct Eq. (25) as an elasticity function of congestion duration P, where both variables P and v_{t_2} are directly observable

$$\frac{v_c}{v_{t_2}} = (P)^s = \left(\frac{D}{C}\right)^{n \cdot s} \tag{25}$$

The above formula implies that $P = 1$ when $v_c/v_{t_2} = 1$. When $P = 1$ (which is the boundary condition), v_{t_2} in reality is already lower than v_c. To mitigate this approximation issue, one can introduce a base rate ($v_c/v_{t_2} = r_0(P)^s$) and then calibrate the model through the boundary condition. Without loss of generality, we assume $r_0 = 1$ in this study. A more systematic evaluation of queue evolution is needed for different horizon lengths of P in the future.

By assuming $\gamma > 0$ and $m = 1/2$, we can obtain the time-dependent delay in t_2 according to Cheng et al. (2022) as shown in Eq. (26) and the average delay in Eq. (27).

$$w_{t_2} = \frac{L}{v_{t_2}} - \frac{L}{v_c} = \frac{\gamma}{4\mu} \cdot \left(\frac{P}{2}\right)^4 = \frac{\gamma}{64\mu} \cdot P^4 \tag{26}$$

$$\bar{w} = \frac{\gamma}{120\mu} \cdot P^4 \tag{27}$$

This leads to the following property which is important for mapping the easy measured lowest speed back to the average waiting time in the queuing model.

Property 1: The ratio of average waiting time and longest waiting time in the fluid queue, $\bar{w}/w_{t_2}$, has a constant value α.

Take the PAQ model with the cubic arrival rate with *m*=0.5 for instance, $\alpha = \frac{\bar{w}}{w_{t_2}} = \frac{64}{120} = \frac{8}{15}$ (as a result of dividing Eq. (27) by Eq. (26)). In the case of *m*=3/4, $\bar{w} = \frac{\gamma}{120\mu} \cdot P^4$, and this will generate a different value of α.

Let $\beta = ns$ and $\beta > 1$. One example would be $\beta = 1.100$ when $n = 1.28$ and $s = 0.859$. The average speed during congestion can be transformed in terms of speed reduction factor v_c/v_{t_2}, which can be in turn expressed as a function of *D/C* below.

$$v = \frac{L}{tt} = \frac{L}{\frac{L}{v_c} + \bar{w}} = \frac{L}{\frac{L}{v_c} + \alpha w_{t_2}} = \frac{L}{\frac{L}{v_c} + \alpha\left(\frac{L}{v_{t_2}} - \frac{L}{v_c}\right)} \tag{28}$$

$$= \frac{v_c}{\frac{v_c}{v_c} + \alpha\left(\frac{v_c}{v_{t_2}} - \frac{v_c}{v_c}\right)} = \frac{v_c}{1 + \alpha\left[\left(\frac{D}{C}\right)^{\beta} - 1\right]}$$

Then from Eq. (26), we can derive the parameter γ as a composite function of D/C, as P, μ and w_{t_2} can be also expressed as a function of *D/C* based on Eqs. (23),

(24) and (27).

$$\gamma = \frac{120\mu \cdot w_{t_2}}{(P)^4} \tag{29}$$

(2b) Mapping function to express shape parameter γ as a function of *D/C*

Alternatively, we can consider $v_c/\bar{v}$ as the magnitude of speed reduction between the speed at capacity and the mean speed. Accordingly, we can consider the following formula,

$$\frac{v_c}{\bar{v}} = (P)^s = \left(\frac{D}{C}\right)^{n\cdot s} \tag{30}$$

which leads to

$$\bar{v} = \frac{v_c}{\left(\frac{D}{C}\right)^{n\cdot s}} = \frac{v_c}{\left(\frac{D}{C}\right)^{\beta}} \tag{31}$$

The above formula implies that $P = 1$ when $\bar{v} = v_c$. The value of γ and resulting time-dependent delay can be derived as $\gamma = \frac{120\mu w}{P^4}$.

4.4 Travel time function for uncongested state

According to a recently proposed traffic flow model S3 (Cheng et al., 2021), we can use the following Eq. (32) to capture the speed-density relationship over a wide range of possible densities, where m is the flow maximization factor in this section and is different from that in $g(m)$ under the cubic model with the PAQ.

$$v = \frac{v_f}{\left[1 + \left(\frac{k}{k_c}\right)^m\right]^{\frac{2}{m}}} \tag{32}$$

Then we can derive the density as follows:

$$k = k_c \cdot \left[\left(\frac{v_f}{v}\right)^{\frac{m}{2}} - 1\right]^{\frac{1}{m}} \tag{33}$$

The unique characteristic of the S3 model is that the speed-flow relationship and flow-density relationship can be derived analytically. According to the conservative law $q = kv$, we have the following equation.

$$q = v \cdot k_c \cdot \left[\left(\frac{v_f}{v}\right)^{\frac{m}{2}} - 1\right]^{\frac{1}{m}} \tag{34}$$

Then we can derive the following equation with one unknown variable of speed.

$$v^m \cdot k_c^m - k_c^m \cdot v^{\frac{m}{2}} \cdot v_f^{\frac{m}{2}} + q^m = 0 \tag{35}$$

Let $x = v^{\frac{m}{2}}$, $a = k_c^m$, $b = -k_c^m v_f^{\frac{m}{2}}$, $c = q^m$, then we can derive the speed in the uncongested state, leading to the average speed function that includes both oversaturated ($P > 1$) and under-saturated ($P \leq 1$) states in Eq. (36).

$$v = \begin{cases} \left[\dfrac{\sqrt{k_c^{2m} v_f^m - 4k_c^m q^m} + k_c^m \sqrt{v_f^m}}{2k_c^m}\right]^{\frac{2}{m}}, & if\ P \leq 1 \\ \dfrac{v_c}{\frac{7}{15} + \frac{8}{15}\left(\frac{D}{C}\right)^{\beta}}, & if\ P > 1 \end{cases} \tag{36}$$

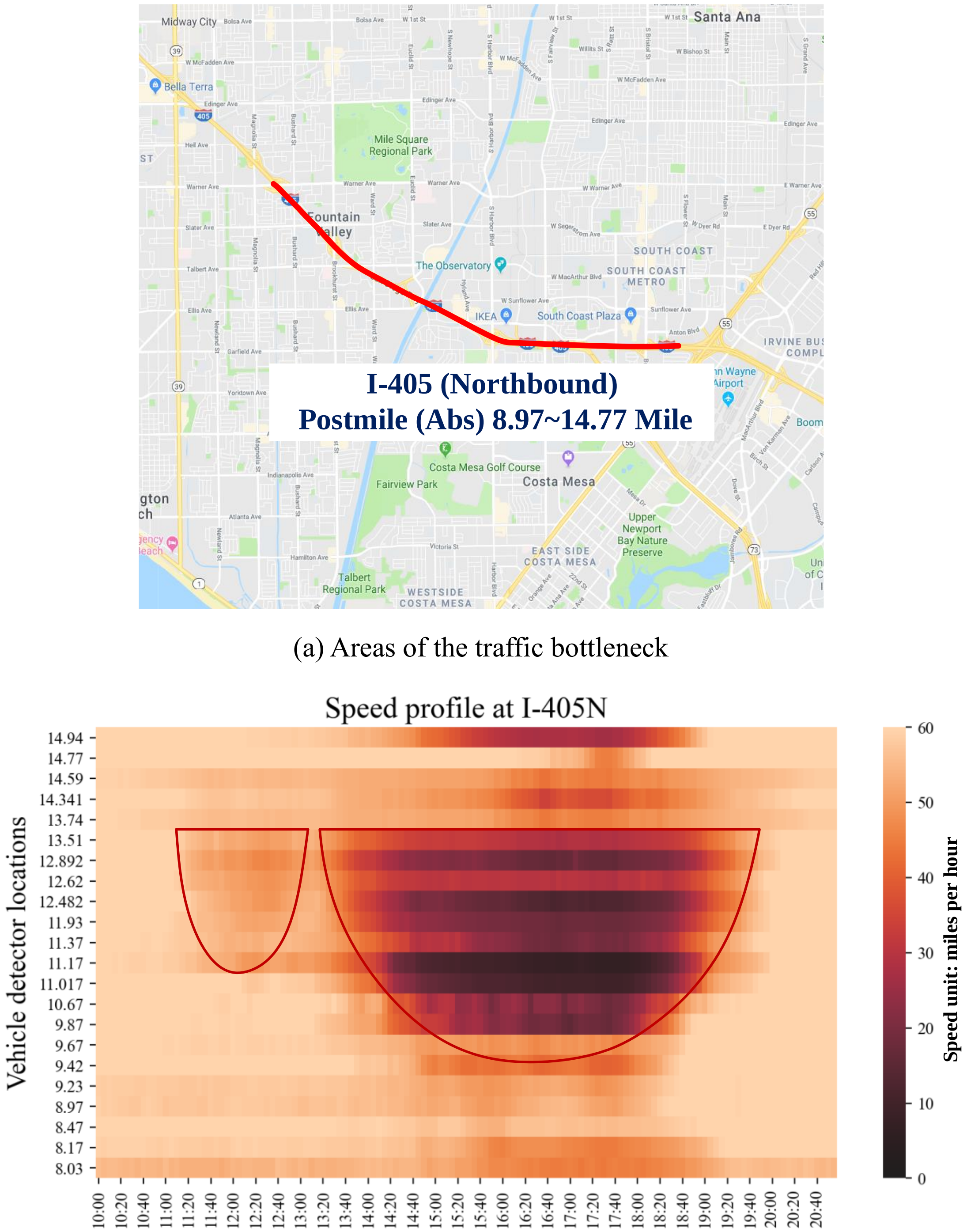
For each link, we have the ordinary differential equation within congested space time regimes. Assuming dynamic arrival rates, departure rates, queue length process, and introducing elasticity parameters n and s, we can capture the relationship between the congestion duration and the average discharge rate. We correlate congestion duration and maximum virtual queues (at the lowest speed) through regression analysis as well. Furthermore, a detailed modeling on traffic congestion is made possible by the proposed model, which uses a more systematical way to model queue spillback and physical bottlenecks, such as merge, diverge, reduced capacity, capacity drops, etc.

5 Calibration and Results

We conduct 2 case studies to evaluate the effectiveness of proposed methodology. The first one considers a single bottleneck as a whole modeling element or “link”, on a 6-mile corridor with 4 months of sensor data. The focus is on the relationship between the inflow demand-to-capacity ratio and key parameters of macroscopic VDF, to name a few, congestion duration, lowest speed, and mean speed during congested period. The second data set treats 4 different locations as 4 modeling “links” in a relatively complex freeway corridor, and we hope to demonstrate how the time-dependent queuing dynamics and travel time can be modelled in a consistent fashion from the fundamental diagram calibration, VDF calibration and time-dependent speed profile estimation.

5.1 Case study 1 for a single bottleneck on 6-mile freeway corridor in Los Angeles, California

Traffic flow, speed, and occupancy data are collected every five minutes from 11:00 a.m. to 20:00 p.m. from the 22 freeway detectors along the Northbound direction of I-405 freeway between the absolute postmiles 8.97 to 14.77 in Los Angeles, shown in Fig. 4(a), in April, May, June and July 2019. It is clear that the bottleneck is located at Abs=13.51 mile. As shown in Fig. 4(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. 4: Locations and speed profile of the corridor (Cheng et al. 2022).

The original data can be accessed from http://pems.dot.ca.gov (hosted by the California Department of Transportation). We choose a single recurrent bottleneck. For each afternoon peak period on weekdays, we calibrate congestion duration P, total inflow demand volume D, and constant discharge rate μ. P has a mean value of 6.75 hours while the average μ is 1339.4 vehicles per hour per lane.

With Eqs. (23) and (25), we can calculate elasticity coefficients n and s by $n = \frac{\ln P}{\ln(D/C)}$ and $s = \frac{\ln(v_c/v_{t_2})}{\ln P}$. The mean values of n and s over 83 days are 1.0804 and

0.4743, leading to $\beta = 0.5123$ in Eqs. (28) and (29). As shown in Fig. 5, we can plot the observed values *vs.* estimated values of congestion duration *P* as a function of the observed *D/C* ratio by using the mean value of *n* as the calibrated value. It shows the proposed formulas can well approximate the ground truth with a correlation coefficient of 0.874.

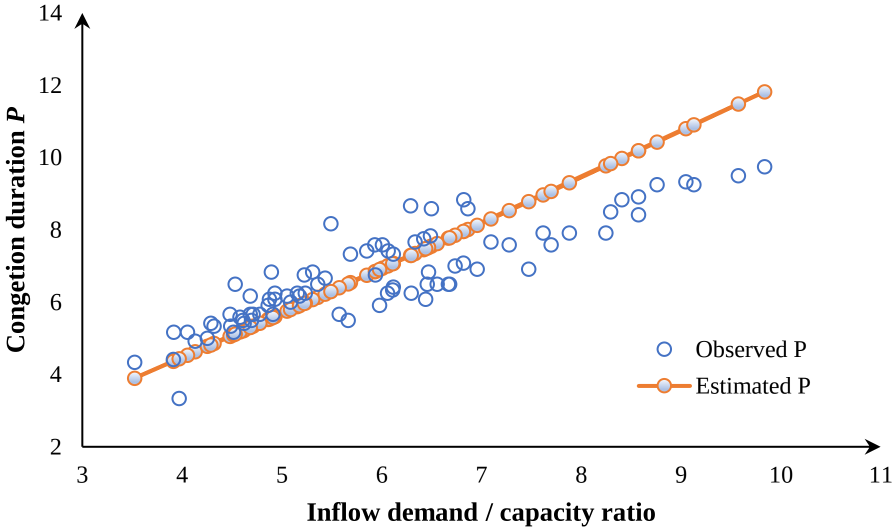


Fig. 5: Observed and estimated congestion duration with respect to the inflow-demand-to-capacity ratio.

As shown in Fig. 6, Eq. (23) gives reasonably good estimates on *P* as a function of *D/C*, where the R^2 is 0.765 and the resulting Mean Absolute Percentage Error (MAPE) is 11.470%. If we set the intersect of the slope as zero, the regression function will be $y = 1.0313x$, which is an unbiased estimate.

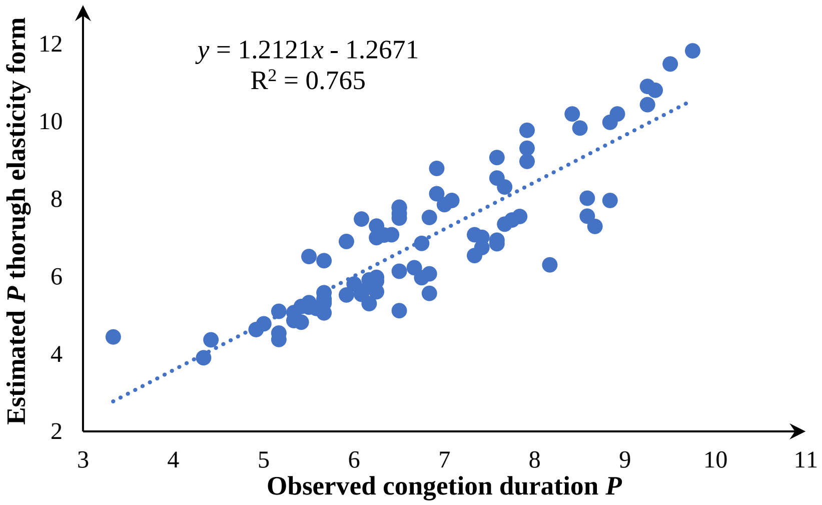


Fig. 6: Estimated vs. observed congestion duration *P* with elasticity form.

Using Eq. (24), we can also obtain reliable estimates on discharge rate as a function of *D/C* with MAPE = 11.06%. However, Fig. 7, which uses μ/C as the capacity discount factor, demonstrates the simple elasticity form is only able to produce a steady mean estimate of the discharge rate. It is unable to capture all possible variations of capacity discount. To capture the day-to-day dynamics of queue discharge rate, a further research is needed to incorporate other factors such as weather and different queue influence areas.

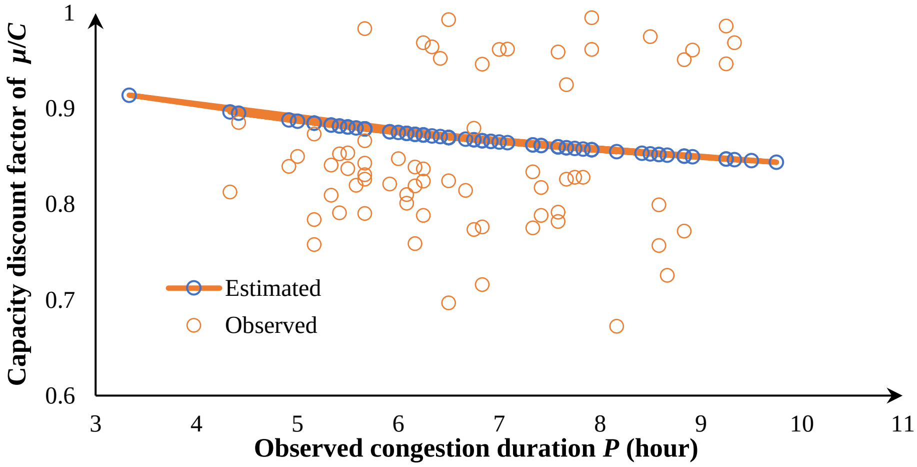


Fig. 7: Capacity discount factor vs. observed congestion duration.

We further calibrate the mean speed (i.e., 52.42 miles per hour) at capacity and use Eq. (35) to produce the comparison plot between the observed and estimated speed reduction factor v_c/v_{t_2} as a function of congestion duration as shown in Fig. 8. Overall, the lowest speed could decrease further if the congestion lasts longer, The proposed functional form performs reasonably well with a correlation coefficient of 0.8807. There are obvious outliers that the near linear trend cannot capture. They might be due to irregular traffic events such as incidents or weather condition.

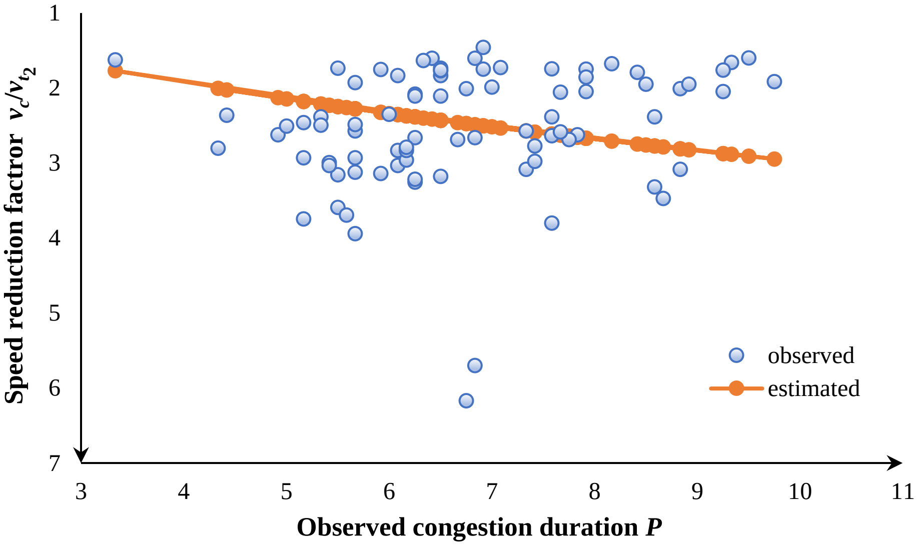


Fig. 8: Speed reduction factor vs. observed congestion duration.

Fig. 9 shows the final mean speed estimation results from two proposed models, with parameter β only (Eq. (31)) and with both parameters of α and β (Eq. (28)). The MAPE of two models are 22.14% and 19.98%, respectively. The mean absolute errors (MAE) of two models are 5.85 and 5.72, respectively. Root Mean Square Error (RMSE) of two models are 6.291 and 6.897, respectively. Overall, two alternative models reach very similar estimation performance.

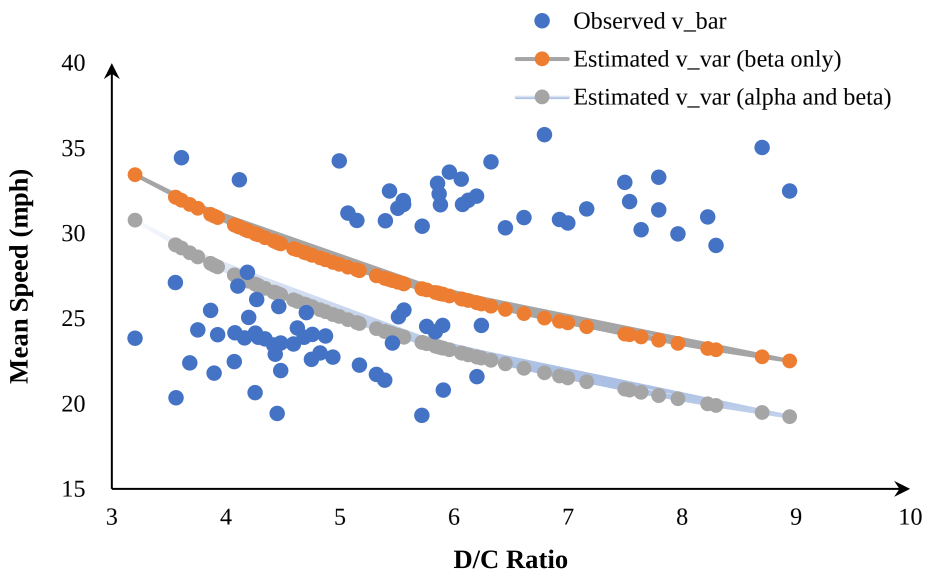


Fig. 9: Observed and estimated data points based on Eqs. (28) and (31).

5.2 Case study 2 for 4 individual locations along a freeway corridor in Phoenix, Arizona

This section implements a case study based on a freeway bottleneck on Phoenix I-10 westbound corridor equipped with four loop detectors that continuously collect traffic data. Figures 10 illustrates the location of the loop detectors. The detectors, installed by the Arizona Department of Transportation (ADOT) in this Phoenix downtown corridor, recorded both traffic counts and speeds every five minutes, each weekday from January 1st. 2016 to Match 1st 2016.

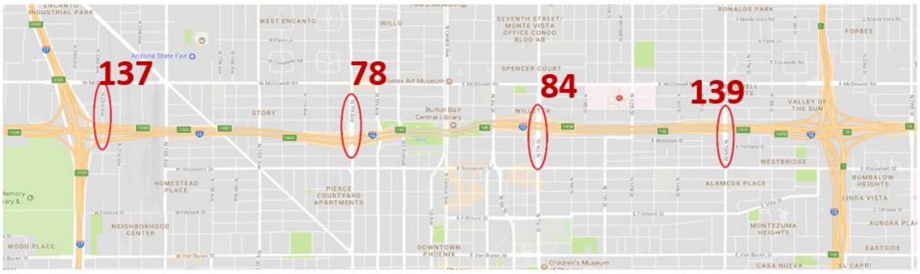


Fig.10. Phoenix I-10 freeway corridor location

Table 3 summarizes the detector identification numbers and types of collected data, and lane configurations of the selected freeway corridor. Belezamo (2020) describes detailed data processing process.

Table 3. Detectors and lane configuration data.

Detector	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

Now, to model the queueing at the bottleneck, we should, to begin with, calibrate the parameters of the underlying traffic flow model and we select a recent 3-paramter model by Cheng et al., (2021). In Fig.11, volume-speed scatters of the four detectors of January and February data are provided, and we utilize the data to calibrate the ultimate capacity and other parameters for each detector, respectively. As shown in the Fig.11, it is evident that the four fundamental diagrams have different volume-speed patterns. For simplicity, we postulate average values for the different parametersL

1. free-flow speed: $v_f = 70$mile/hour,
2. critical speed (speed at capacity): $v_c = 51$mile/hour,
3. curve flatness: m =4.5,
4. ultimate capacity, C =1,750 veh/(lane*hour).

It is important to denote that

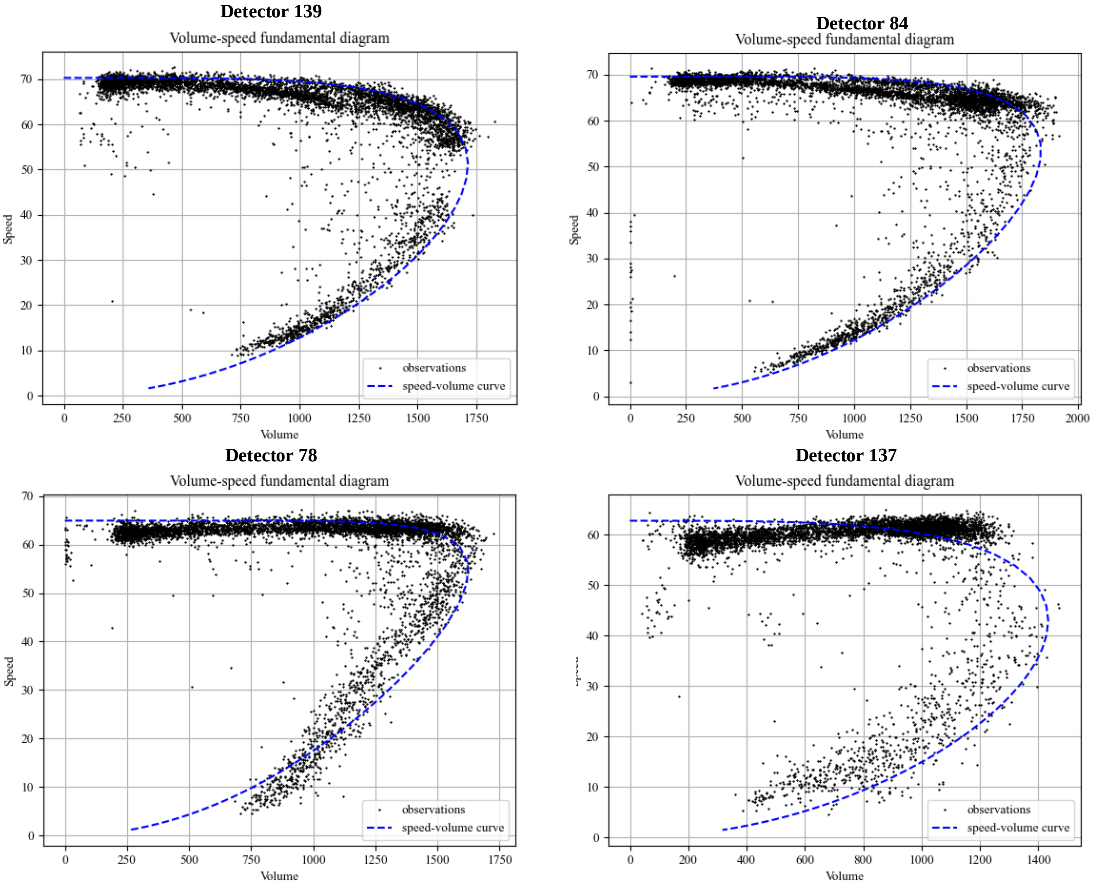


Fig.11 volume-speed scatters of January and February data and calibrated traffic flow model

Table 4 reports the calibrated results of different locations. The calibrated critical speed is applied to determine the congestion duration P, starting time and ending time of queue, as well as the demand (Wu et al., 2020). The statistics indicate this portion of I-10 corridor is extremely congested with congestion duration longer than 4 hours. The table also reports the mean speed during the congestion duration where time-dependent speed is lower than the critical speed 51 mile/hour, as shown in Fig.12. Fig.13 displays the queue-based VDF calibrated using the proposed equation (36) where $\beta = ns = 1.07$. Table 4 and Fig.16 also compare the estimated speed within the congestion duration with the corresponding mean speed values calculated based on the observed data.

Table 4. Calibrated parameters of each detector for macroscopic and mesoscopic models

	Detectors ID			
	137	78	84	139
Inflow demand	3359.86	3894.74	5842.72	5869.39
D/C ratio	1.92	2.23	3.34	3.35
P (unit: hour)	4.00	4.00	4.42	4.67
γ	105.81	64.62	10.90	13.35
μ (unit: veh/hour)	1573.83	1505.51	1301.22	1298.88
$v_{\min}$	25.37	12.46	16.24	17.16
$\bar{v}$ (estimated)	33.27	29.74	21.33	21.25
$\bar{v}$ (observed)	34.02	27.86	27.16	27.43
Congestion duration elasticity factor n	1.42			
Speed reduction elasticity factor s	0.75			

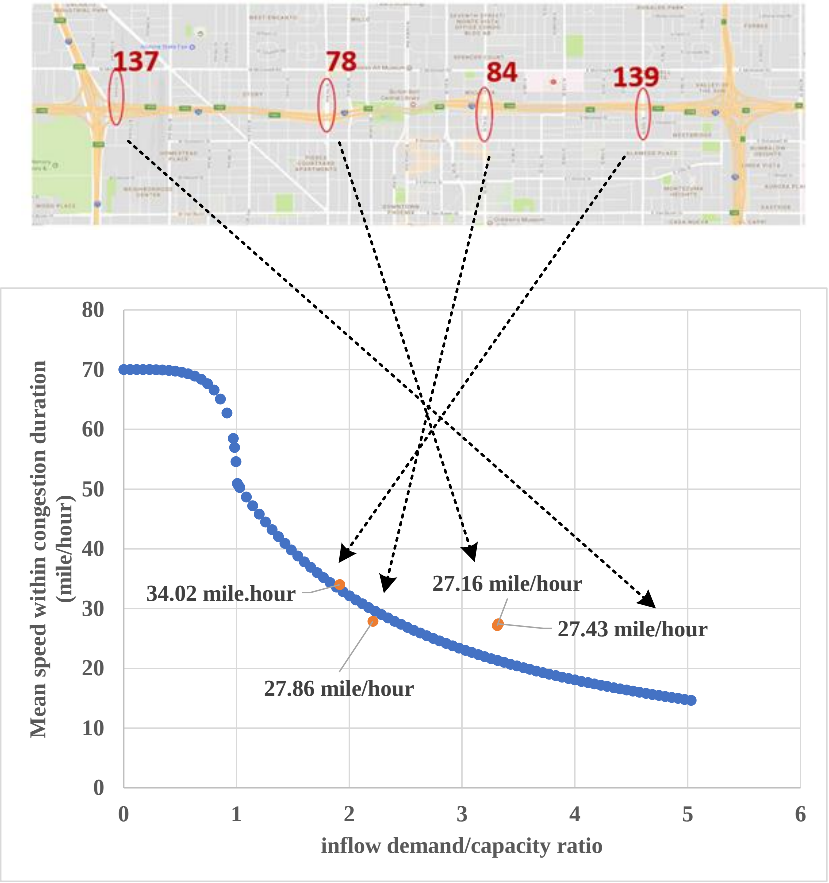


Fig.12 Calibrated queue-based VDF with data points from 4 locations based on models with both alpha and beta parameters.

Using the polynomial queuing model, we derive the time-dependent speed for different locations based on PAQ with the calibrated parameters in Table 4. The derived time-dependent speed profiles are displayed in Fig.13(a)-Fig.13(c).

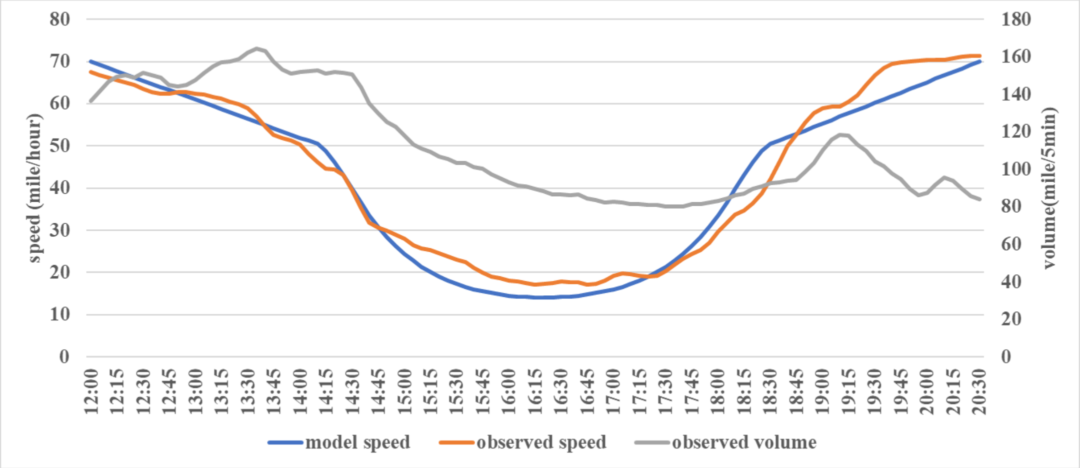


Fig.13 (a) Observed time-dependent mean speed and modeled time-dependent speed at location 139

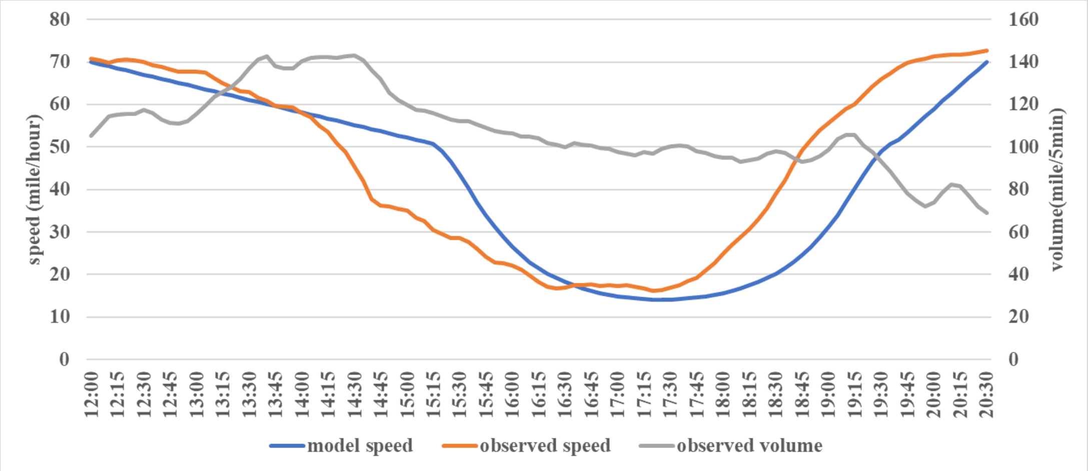


Fig.13(b) Observed time-dependent mean speed and modeled time-dependent speed at location 84

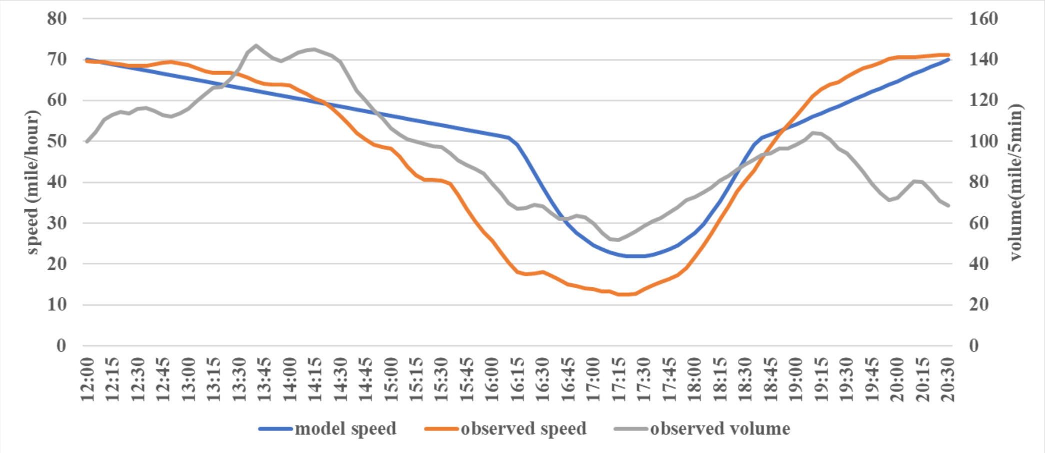


Fig.13(c) Observed time-dependent mean speed and modeled time-dependent speed at location 78

6 Discussion

The above cross resolution approach can be also extended in the macroscopic activity origin distributions. The related system identification and estimation questions are important to discuss as well.

6.1 Mapping inflow demand from space-varying distribution of activity locations

From the perspectives of activity-travel and land-use (Yan et al., 2017) that focus on the sociodemographic aspects of traffic congestion, the spatial distribution of travel demand at different activity locations (e.g., home to work, home to shopping, and work to home) can be viewed as the driving forces for the incoming demand flow toward a bottleneck. As shown in Fig. 21, l_0 is the activity location of a traveler who would be the first one encountering the congestion at the bottleneck, l_1 is the activity location with the maximum traffic demand, and l_3 is the activity location where a traveler would not encounter the congestion at the bottleneck. With a systematic understanding of system variables (e.g., inflow demand volume and arrival flow shape parameters) at different orders, we can analyze the root causes of congestion through a better activity scheduling with different behavioral constraints at the household level, such as alternative work arrangements with a flexible work-time system.

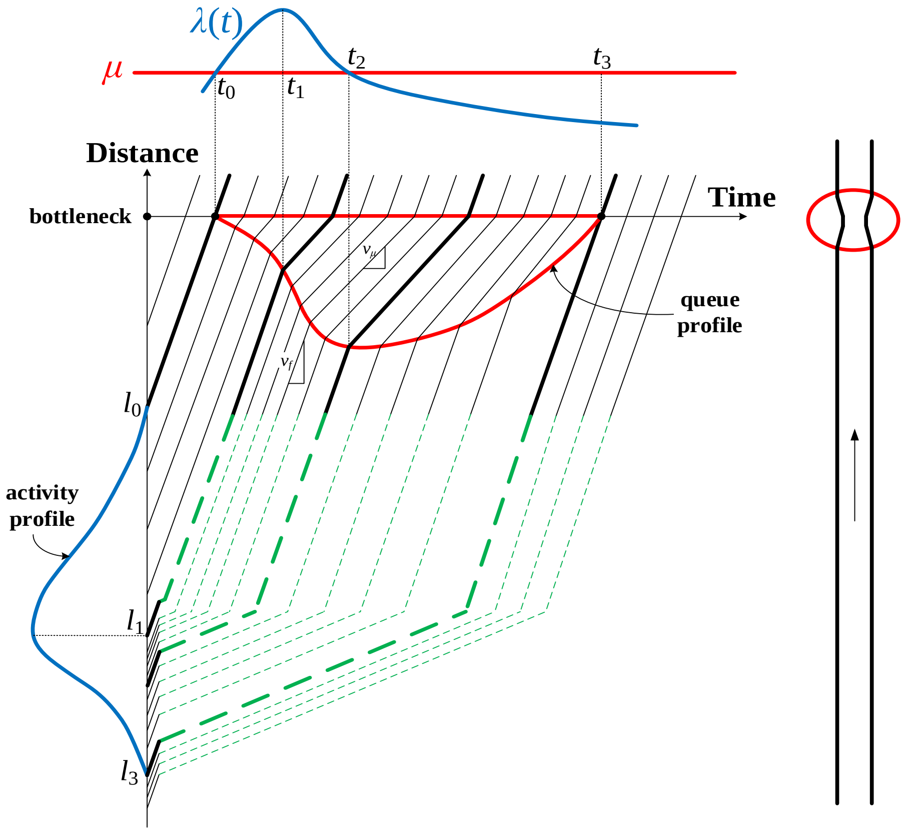


Fig. 21: Illustration of mapping between the spatial distribution of activity locations, travel demand, and inflow pattern in an oversaturated queueing system.

Due to space limitations, part of the vehicle trajectories in Fig. 21 is in green dash lines, corresponding to the solid straight lines of free-flow travel before they encounter the tail of the queue. Through this mapping in the space-time plane, we can further analyze how the time-varying demand of $\lambda(t)$ is contributed from the spatial distribution of activity locations (e.g., home or office) upstream of the bottleneck. This sheds some light on the root land-use causes of traffic congestion in many metropolitan areas, namely, urban sprawl, housing affordability, and home-work separation.

6.2 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 should 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 with 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 oversaturated dynamic queueing systems 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 shape 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.

7 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 with a constant discharge rate; (2) derive a closed-form of time-dependent queue length based on the integral 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 rates (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 also point out a number of 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.

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.

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.

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. Under review with Transportation Research Part C, 1–36.

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.

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 the 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.

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.

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.

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.

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: III. A mild rush hour. Journal of Applied Probability, 5(3), 591–606.

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

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.

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., 1996. Modeling dynamic transportation networks: An intelligent transportation system oriented approach. Springer-Verlag, Heidelberg.

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.

Ran, B., Hall, R.W., Boyce, D.E., 1996. A link-based variational inequality model for dynamic departure time/route choice. Transportation Research Part B, 30(1), 31–46.

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.

Teknomo, K., Gardon, R.W., 2018. Intersection analysis using the ideal flow model, in: Proceedings of the 20th IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 1–6.

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

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

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.

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.

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.