A Review of Volume-Delay Functions Integrating Traditional Models and Emerging AI Technologies

Yuyan (Annie) Pan[a], Xianbiao Hu[a]*, George List[b], Xuesong (Simon) Zhou[c]*

a Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA 16802-1408, United States

b Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC, USA

c School of Sustainable Engineering and Built Environment, Arizona State University, Tempe, AZ 85281, United States

*Corresponding authors.

Email address: xbhu@psu.edu; xzhou74@asu.edu

Abstract

Link Performance Functions (LPFs) underpin the estimation of travel times, delays, and congestion in transportation network models. As multimodal systems increasingly incorporate CAVs, EVs, and complex queuing phenomena, traditional LPFs face limitations in accuracy, physical consistency, and adaptability. This review synthesizes the evolution of LPFs to address these modern challenges and bridge theory, data, and artificial intelligence (AI). We survey over 270 studies, covering (i) classical empirical forms (e.g., BPR, Davidson), (ii) theory-driven models (fundamental-diagram and fluid queue approaches), and (iii) AI-augmented frameworks (LSTM, GNN, Transformer, and physics-informed learning). Models are categorized by mathematical assumptions, data needs, integration with Dynamic Traffic Assignment tools, and multimodal flow support. Each class is evaluated against emerging criteria: queuing dynamics, CAV/EV impacts, data fusion, and computational tractability. Key findings include: (1) growing use of hybrid physics-machine learning models enforcing conservation laws while exploiting large-scale sensor data, and (2) gaps in AI evaluation, with reliance on point-error metrics rather than system-level Key Performance Indicators (e.g., delay, throughput, queue length). By linking theory, data, and AI, this review maps the LPF landscape, identifies open research directions, and offers a roadmap for next-generation multimodal transportation networks.

Keywords

Link Performance Functions (LPFs); Volume-Delay Functions (VDFs); Physical-Informed Neural Network (PINN); Emerging Transportation Applications; Multimodal Systems; Artificial Intelligence (AI)

1 Introduction

Link Performance Functions (LPFs) have long served as a cornerstone of traffic assignment, network design, and policy evaluation. Traditionally, LPFs such as the Bureau of Public Roads (BPR) function provided simple yet tractable formulations linking volume to delay. However, with the proliferation of connected and automated vehicles (CAVs), the emergence of high-resolution sensing, and the increasing availability of multimodal, multi-source traffic data, the limitations of conventional LPFs have become increasingly evident. Existing forms often fail to capture dynamic congestion phenomena, lack physical consistency with queuing processes, and exhibit poor generalizability across regimes and networks.

The rapid advancement of artificial intelligence (AI) and machine learning offers new opportunities to address these shortcomings. Models such as Long Short-Term Memory (LSTM) networks, Graph Neural Networks (GNNs), and Transformers can learn complex nonlinear patterns and exploit spatiotemporal dependencies in modern transportation systems. However, purely data-driven models remain limited: they risk violating conservation principles, offer little interpretability, and may not generalize under distribution shifts. This tension highlights the need for hybrid approaches that integrate domain knowledge from traffic flow theory with the predictive power of AI.

Against this backdrop, the unique contribution of this review lies not only in synthesizing prior research but, equally importantly, in proposing an original future framework and a unified pipeline for LPF modeling in the AI era. Specifically, we revisit LPFs through the lens of multi-resolution traffic flow theory, clarify technical misconceptions, and introduce reproducible calibration procedures such as demand-to-capacity (D/C) ratio estimation. Building on these insights, we develop a physics-informed, AI-driven pipeline and articulate a prioritized roadmap for future research. To our knowledge, this is the first attempt to consolidate the evolution of LPFs into a forward-looking, technically grounded framework that is both comprehensive and actionable.

In recent years, the rapid rise of large-language models and AI-enabled traffic forecasting has spawned numerous review articles in the machine-learning communities. Yet on the ground, transportation planners largely continue to rely on the classical four-step paradigm, which includes trip generation, distribution, mode choice, and assignment (Bliemer et al., 2017; Patriksson, 2015; Roughgarden and Tardos, 2002), even as many agencies experiment with activity-based, dynamic self-assignment, or simulation-based extensions. These methods often aim to improve link flow or density predictions via machine learning, but they lack robust, system-level key Performance Indicators (KPIs), such as network-wide delay, throughput, or multimodal performance, which are needed to evaluate large-scale investment and operational policies effectively.

At the heart of every assignment model lies the LPFs, also called a volume-delay, travel-time, or cost function, which relates flow (or demand) to link travel time (or delay). The purpose of a link performance function in traffic assignment is to model the relationship between traffic flow and travel time on a particular link in a transportation network. LPFs have many different name in literatures, they are also called a) link travel time functions (Beckmann et al., 1956; Fisk, 1991; Xiong and Davis, 2009), b) link capacity functions (Branston, 1976; Kosun et al., 2016), c) link performance functions (Patriksson, 2015; Ran et al., 1996; Sheffi, 1984), d) link cost functions (Small et al., 2007) with related marginal cost functions (Fosgerau and Small, 2012; Van Ommeren and Fosgerau, 2009), and e) edge latency function in computer science and game theory (Lianeas et al., 2018; Roughgarden and Tardos, 2002).

The early formulation of the LPFs can also be traced back to work by (Duffin, 1947) on electrical networks with a certain nonlinear resistance function. The classical work by Beckmann, McGuire and Winsten (1956) (BMW) produced mathematical programming models that formulate the static equilibrium traffic assignment problem and

adopt a generic nonlinear LPF. Charnes et al. (1958) and Charnes and Cooper (1959) independently proposed formulations for traffic assignment problems with fixed origin-destination (OD) flows and piecewise LPFs. While many studies on STAs with hard capacity constraints (Bliemer and Raadsen, 2020; Larsson and Patriksson, 1995; Nie et al., 2004a; Raadsen et al., 2020), representing and calibrating a LPF for oversaturated freeway facilities remains a significant challenge. This challenge has motivated another research line, dynamic traffic assignment, which aims to capture queue formation, propagation, and dissipation in congested networks (Peeta and Ziliaskopoulos, 2001; Brederode et al., 2019). Equally important for LPFs are examining the inherent theoretical connections between macroscopic LPFs and queue-theoretic models, and mesoscopic and microscopic simulation models, such as POLARIS (Auld et al., 2016); MATSim (Balmer et al., 2008); Dynus-T (Chiu et al., 2010); SUMO (Behrisch et al., 2011); TRANSIMS (Nagel et al., 1999); AIMSUN (Barceló and Casas, 2005); VISSIM (Fellendorf and Vortisch, 2010); TransModeler (Caliper, 2009); DynaMIT (Ben-Akiva et al., 1998). Insights from a cross-resolution perspective can contribute to achieving a higher level of consistency and accuracy in modeling the complex demand and supply interactions, and the importance and development of LPFs has been highlighted in surveys by Branston (1976) and Saric et al. (2019).

This review aims to fill that gap by guiding both planners and AI researchers through a concise yet systematic journey:

(1) **Core LPF concepts** (capacity, demand, queue dynamics);

(2) **Evolution of formulations** from simple linear and BPR curves to theory-driven and hybrid AI approaches;

(3) **Extensions to multimodal and cross-resolution applications**, and

(4) **Integration with AI frameworks** for real-time adaptation, physical consistency, and large-scale KPI design.

By highlighting how even a basic linear LPF connects to fundamental diagrams (FD), traffic-state estimation, and multimodal interactions, and showing where AI can be woven in, we aim to equip practitioners with the conceptual tools and evaluation metrics necessary to leverage AI without abandoning the rigor and transparency of established traffic-modeling practice.

This review progresses systematically from foundational concepts to emerging applications. Section 2 traces the theoretical foundations and historical evolution of LPFs across four phases (1950s-present). Section 3 introduces four core insights addressing key challenges: distinguishing true demand from measured flow, achieving multi-scale consistency, emphasizing congestion duration as a key delay factor, and enabling cross-resolution modeling for future technologies. Section 4 examines the shift toward physics-informed AI, identifying critical integration challenges and presenting a unified end-to-end AI-driven LPF framework that integrates data-driven approaches with physics-based constraints to balance theoretical consistency with cross-scenario generalization. Section 5 presents a case study that systematically compares empirical VDFs, theory-driven approaches, data-driven models, and physics-informed methods, covering the full process from case selection to model comparison and result analysis to provide a comprehensive and unified evaluation across paradigms. Section 6 outlines nine priority questions that guide future research, synthesizing findings and offering strategic recommendations for transitioning to next-generation physics-informed AI systems.

2 Fundamentals, Evolution and Contemporary Challenges of Link Performance Functions

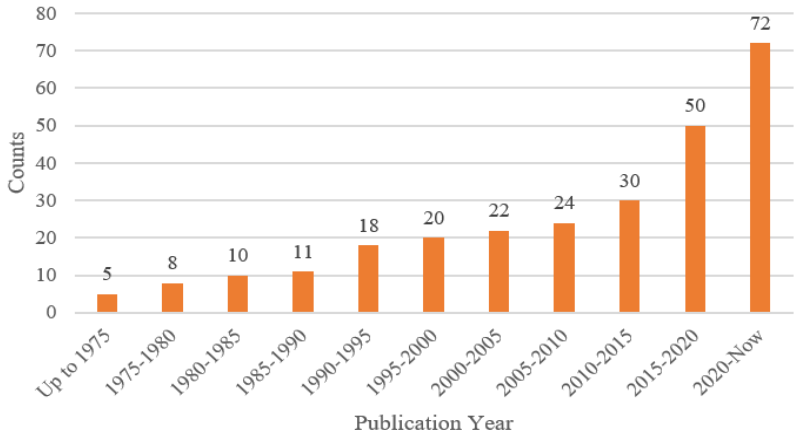
2.1 Bibliometric Analysis

We conducted a systematic literature search across Web of Science, Scopus, Transport Research International Documentation (TRID), and IEEE Xplore, using keywords such as "link performance function," "volume-delay function," "fundamental diagram," and "queue theory." Each study was then coded according to four key dimensions: (1) its theoretical foundation (FD-based, queue-based, simulation-derived, or data-driven), (2) its temporal scope (static, quasi-dynamic, or fully dynamic), (3) its spatial scale (link-level, corridor, or network-wide), and (4) its application domain (highway, urban arterial, multimodal, or emerging mobility).

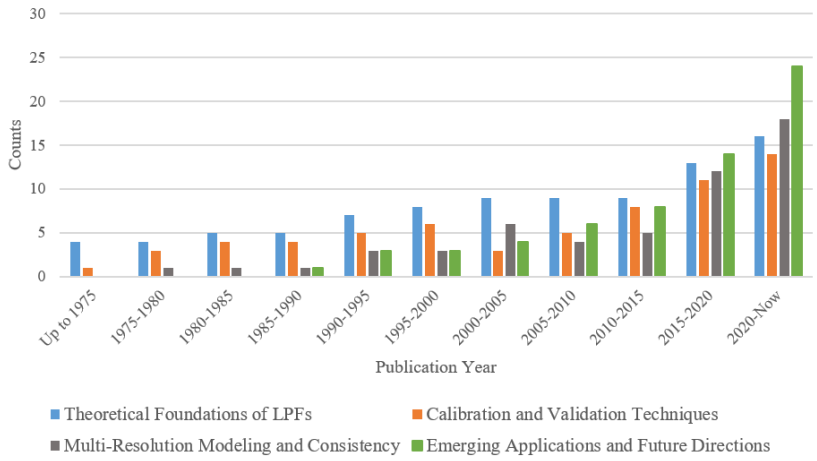
A bibliometric timeline highlights the field's evolution: during the 1950s-1970s, foundational piecewise-linear and BPR functions were established; the 1980s-1990s saw the integration of queue theory and advanced polynomial forms; the 2000s-2010s brought dynamic extensions and hybrid simulation frameworks; and from 2015 onward, machine-learning, AI, and physics-informed approaches have come to the fore.

These papers are from more than 50 journals, and all these journals belong to "transportation." Two of the top ones with the highest volume of publications are Transportation Research Part B and Transportation Research Record. Articles published in Transportation Research Part B account for 18.9% of the total number of publications, and Transportation Research Record accounts for 19.4% of publications.

We analyzed around 270 references, which are displayed in Fig. 1. We organized these references chronologically, and the graph in Fig. 1(a) provides a good indication of the publication trend in this field. Our analysis revealed an increase in related publications after 2010. LPFs remains a fundamental research topic in the transportation modeling field and are expected to evolve in the coming years. In addition, we also grouped the references into four thematic categories, as shown in Fig. 1(b): "Theoretical Foundations of LPFs," "Calibration and Validation Techniques," "Multi-Resolution Modeling and Consistency," and "Emerging Applications and Future Directions." This categorization highlights the interconnections between different approaches and their practical applications, providing readers with a more comprehensive understanding of the LPFs matter. In the subsequent sections, we provide detailed review about theoretical, practical and applications for each category, offering a more insightful analysis. Additionally, we observed that the standard form of LPF has been widely utilized in highly cited studies for numerous emerging applications, such as road pricing, tradable mobility credits, and autonomous vehicle management.



(a) Temporal distribution of the total publications



(b) Temporal distribution of publications by topic group

Figure 1 Analysis of literature publications related to LPFs.

We summarized the most frequently cited papers during the past few decades, which are shown in Table 1. It should be noted that the original report of the BPR function by the US Bureau of Public Roads has approximately 284 direct citations, while the majority of papers (approximately 3,184) only use the abbreviation "BPR function" without citing the original reference. The highly cited papers that directly cite the BPR function are shown in Table 2.

Table 1 The frequently cited papers on the topic of VDFs.

Rank	Journal	Total citations	Paper title	Author/s/Year
1	US Bureau of Public Roads	3,184 (284)	Traffic assignment manual for application with a large, high speed computer	US Bureau of Public Roads (1964)
2	Transportation Research Board	2,805	Traffic signal settings	Webster (1958)
3	Transportation Research Part B	485	Comparison of delay estimates at undersaturated and oversaturated pretimed signalized intersections	Dion et al. (2004)
4	Transportation Research Part A	472	Traffic assignment and signal control in saturated road networks	Yang and Yagar (1995)
5	Transportation Research	444	Link capacity functions: a review	Branston (1976)
6	Transportation Science	396	Conical volume-delay functions	Spiess (1990)
7	Australian Road Research	360	Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and alternative travel time function	Akcelik (1991)
8	Australian Road Research Board	360	A flow travel time relationship for use in transportation planning	Davidson (1966)
9	Transportation Science	356	Traffic queues and delays at road junctions	Kimber and Hollis (1979)
10	Journal of Urban Economics	349	The incidence of congestion tolls on urban highways	Small (1983)
11	ITE journal	245	The highway capacity manual delay formula for signalized intersections	Akcelik (1988)

12	Australian Road Research	205	Time-dependent expressions for delay, stop rate and queue length at traffic signals	Akcelik (1980)
13	Transportation Research Record	158	Signalized intersection delay models-a primer for the uninitiated	Hurdle (1984)
14	Transportation Research Part B	151	Estimation of delays at traffic signals for variable demand conditions	Akçelik and Rouphail (1993)
15	Highway Research Board	125	An iterative assignment approach to capacity restraint on arterial networks	Smock (1962)

Table 2 Highly cited papers that directly cite the original BPR paper (US Bureau of Public Roads, 1964).

Rank	Journal	Total citations	Paper title	Author/s/Year
1	Mathematics of Operations Research	574	Selfish routing in capacitated networks	Correa et al. (2004)
2	Transportation Science	427	The convergence of equilibrium algorithms with predetermined step sizes	Powell and Sheffi (1982)
3	Journal of transportation engineering	321	Pedestrian speed/flow relationships for walking facilities in Hong Kong	Lam and Cheung (2000)
3	Transportation science	185	Shelter location and evacuation route assignment under uncertainty: A benders decomposition approach	Bayram and Yaman (2018)
5	Transportation	183	A study of the bidirectional pedestrian flow characteristics at Hong Kong signalized crosswalk facilities	Lam et al. (2002)
6	International Transactions in Operational Research	123	Fragile networks: identifying vulnerabilities and synergies in an uncertain age	Nagurney and Qiang (2012)
7	Mathematics of Operations Research	121	On the existence of pure Nash equilibria in weighted congestion games	Harks and Klimm (2012)
8	Transportation Research Record	115	Flow breakdown and travel time reliability	Dong and Mahmassani (2009)
9	Transportation Science	108	Wardrop equilibria with risk-averse users	Ordóñez and Stier-Moses (2010)
10	Transportation Research Part B: Methodological	100	A bilevel model of the relationship between transport and residential location	Chang and Mackett (2006)
11	Journal of Intelligent Transportation Systems	95	An evaluation of environmental benefits of time-dependent green routing in the greater Buffalo-Niagara region	Guo et al. (2013)
12	Networks and Spatial Economics	94	Hybrid evolutionary metaheuristics for concurrent multiobjective design of urban road and public transit networks	Miandoabchi et al. (2012)

Link performance functions trace their origins to FD-based traffic flow theory, which relates flow, density, and speed to characterize free-flow, transitional, and congested regimes. By analytically mapping these regimes, FD models enable explicit delay formulations under varying demand and capacity conditions (e.g., Moses et al., 2013;

Kucharski and Drabicki, 2017). Subsequent applications have leveraged this approach to study bottleneck formation and breakdown phenomena in urban networks (Huntsinger and Rouphail, 2011; Geroliminis and Sun, 2011).

2.2 Historical Evolution of Link Performance Functions

(1) Early Linear and Piecewise-Linear Forms

In the 1950s, practitioners adopted simple linear approximations for tractability. The Chicago Area Transportation Study introduced a two-segment piecewise-linear LPF based on the volume-capacity ratio (V/C) (Campbell et al. 1968).

$$tt = \begin{cases} t_f, if \dfrac{V}{C} \leq 0.6 \\ t_f + \alpha\left(\dfrac{V}{C} - 0.6\right), otherwise \end{cases} \tag{1}$$

where t_f is free-flow travel time, tt is the average link travel time, V is hourly volume, C is hourly capacity, α is a parameter.

This linear form was also proposed earlier, though not based on the V/C ratio but instead expressed as the difference between volume and capacity ($V - C$). The formulation was later extended to include a third straight line that represents the over-saturated region, as illustrated below:

$$tt = \begin{cases} t_p + \alpha(V - C_p), if\ V < C_p \\ t_p + \beta(V - C_p), if\ C_p \leq V \leq C_u \\ t_u + \gamma(V - C_u), if\ V > C_u \end{cases} \tag{2}$$

where $t_u = t_p + \beta(C_u - C_p)$. t_p is base (or free-flow) travel time, corresponding to low-volume conditions. t_u is travel time at the end of the capacity interval. C_p is lower critical capacity (onset of congestion threshold), below this, travel time grows only slightly with volume. C_u is upper capacity threshold, beyond this, traffic enters the oversaturated region with severe queuing. α, β, γ are parameters.

While the linear form is simple and easy to apply, it is unfortunately insufficient to capture the curve-like patterns observed in field data, as travel time does not increase linearly with traffic volume.

(2) Exponential and Logarithmic Form

To capture nonlinear congestion onset, exponential and logarithmic LPFs emerged in the 1960s. Smock (1962) fitted a natural exponential of V/C in Detroit, and Soltman (1965) used a base-2 exponential for Pittsburgh, with a generalized exponential form as:

$$tt = t_f \cdot \alpha^{(V/C)^\beta} \tag{3}$$

A logarithmic form was also developed, for example by Mosher (1963), and expressed as $tt = t_f + \ln(\alpha) - \ln(\alpha - V)$, where α represents the saturation flow of a link.

$$tt = \begin{cases} t_f + \beta \ln(\alpha) - \beta \ln(\alpha - V)\ \ if\ V \leq C \\ t_f + \beta \ln(\alpha) - \beta \ln(\alpha - C) + \dfrac{\beta V}{\alpha - C}\ otherwise \end{cases} \tag{4}$$

(3) BPR Function and Polynomial-based Form

Bureau of Public Roads, in its 1964 Traffic Assignment Manual (BPR, 1964), introduced one of the most well-known and widely adopted VDFs:

$$tt(t) = t_f \cdot [1 + \alpha(\frac{q(t)}{C})^{\beta}] \quad (5)$$

where parameter α represents the ratio of travel time per unit distance at practical capacity to that at free flow, while β controls how sharply the curve rises from the free-flow travel time. For V/C ratios less than 1.0, the increase in travel time is gradual, but it accelerates rapidly once the V/C ratio exceeds 1. Higher values of β intensify the onset of congestion effects.

(4) Advanced Polynomial Forms

Several later studies identified shortcomings in the BPR function (Dowling et al., 1998; Skabardonis and Dowling, 1997; Spiess, 1990). The calibrated function may have a non-negligible error for urban streets due to their unique characteristics such as intersections, bus transit lanes, stop signs, and signals. To deal with those issues, several VDFs in polynomial form were also proposed. For instance, Spiess (1990) proposed the conical VDF, expressed as: $tt = t_f \cdot \left[2 + \sqrt{\beta^2(1-\frac{V}{C})^2 + \alpha^2} - \beta\left(1-\frac{V}{C}\right) - \alpha\right]$. In this formulation, α is directly related to β, defined as $\alpha = \frac{2\beta-1}{2\beta-2}$, with the condition that $\beta > 1$.

Additionally, Akçelik (1991) introduced another volume-delay function that employs a time-dependent form derived from a coordinate transformation technique. It is given by: $tt = 0.25t_f \cdot \left[\left(\frac{V}{C}-1\right) + \sqrt{(\frac{V}{C}-1)^2 + 8\frac{J_A V}{C^2 T_f}}\right]$ where T_f is the flow period (the analysis period, e.g., 15 minutes, 1 hour), and J_A is a delay parameter (accounts for queueing and coordination effects).

(5) System-Oriented Extensions

A truly system-oriented LPF framework must go beyond isolated volume-delay curves to account for interactions with land use, travel behavior, and environmental objectives. Empirical studies have demonstrated that surrounding land-use patterns fundamentally alter link sensitivities (Ma and Lo, 2012; Szeto et al., 2015), while heterogeneity in traveler value-of-time and route preferences can reshape the effective delay experienced on a link. Moreover, considerations of environmental factors such as emissions and fuel consumption are increasingly integrated into LPF design to promote sustainable network management (Bagherian et al., 2017; Benedek and Rilett, 1998).

In practice, agencies routinely recalibrate or extend the canonical BPR function to capture local signal timing, intersection density, and transit priority effects (Dowling and Skabardonis, 1993). For dynamically oversaturated corridors, Tisato (1991) modified Davidson's VDF to embed a congestion-duration term, while Boyce et al. (1981) warned that asymptotic growth in LPFs can produce counter-intuitive detours by over-penalizing highly congested links. More recently, Zhou et al. (2022) revisited Spiess's well-behaved congestion criteria and infused them with queuing-theoretic constraints, thereby improving LPF realism under sustained oversaturation.

(6) Link Performance Functions for Heterogeneous Traffic Flows

The presence of heterogeneous traffic flows composed of various vehicle classes such as cars, trucks, buses, and trains introduce critical complexity to LPFs by affecting average speeds, stopping behaviors, and capacity distributions (Montanino et al., 2021; Qian et al., 2017). Early LPFs that accounted for truck impacts include Kim and Mahmassani's (1987) truck-volume delay model and Müller and Schiller's (2015) German-motorway extension. Mesbah et al. (2011) adapted the Akçelik cost function for mixed car-transit priority, while Chen et al. (2011) introduced a time-dependent queuing model for stochastic truck service at gates. In rail applications, Petersen (1974)

and Prokopy and Richard (1975) developed multi-class train delay estimators; Crainic et al. (1990) proposed polynomial LPFs for yards and lines; Lai and Barkan (2009) fitted exponential delay-volume curves for single-track networks; and Sogin et al. (2016) quantified capacity gains of double-tracking via nonlinear delay functions. On highways, Zhang and Waller (2018) modified the BPR formula for High Occupancy Vehicle (HOV) lane versus General Purpose (GP) lanes, Hörcher et al. (2017) modeled crowding-dependent train times, and Anupriya et al. (2023) leveraged station-level fundamental diagrams to capture passenger boarding-alighting dynamics. Collectively, these studies show that explicitly embedding multivehicle-type considerations into LPFs produces more accurate, context-sensitive delay models essential for both road and rail network analysis.

2.3 Core Misconceptions and Technical Challenges in Link Performance Function Studies

Drawing on our comprehensive review, Fig. 2 synthesizes the full demand-capacity mapping across different regimes, and this end-to-end perspective dispels three persistent misconceptions about LPFs. It traces the demand-capacity (D/C) ratio through four linked regimes: Regime A (Uncongested): Speed declines gently as flow rises toward capacity; Regime B (Congested): Beyond the capacity peak, queue formation makes the speed-flow relationship non-monotonic; Regime C (Oversaturated Outflow): Observed outflow caps at the link's service rate μ even as demand grows.

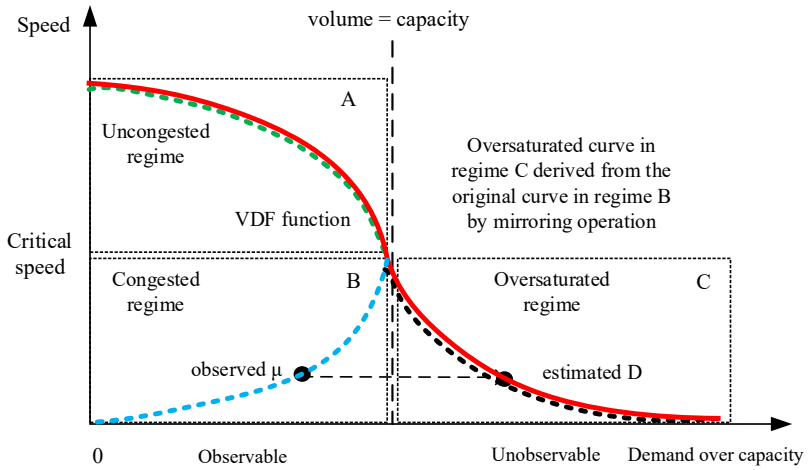


Figure 1 Fundamental-diagram-derived speed-flow relationship (Greenshields), illustrating Regime A (uncongested), Regime B (congested), and Regime C (oversaturated/unobservable demand). True demand D lies beyond the measured outflow and must be imputed.

Misconception 1: LPFs belong only in regional-scale, static traffic assignment.

(i) **Common belief**: As the final component of the four-step process, LPFs are often calibrated only once using annual or peak-hour traffic counts and are applicable solely to individual corridors.

(ii) **Why it's misleading and how to address it**: LPFs are grounded in fundamental diagram and queue-theoretic principles across micro- to macroscopic scales. By calibrating across shoulders, peak, and oversaturated regimes, the same functional (or non-functional) form links regimes B→C and infers demand D. This enables multimodal, regional analyses of delay, safety, and queuing, rather than isolated corridor costs.

Misconception 2: High-fidelity simulation is the only way to model congestion.

(i) **Common belief**: Detailed microsimulation or dynamic-assignment software is essential for capturing spillback effects, queuing behavior, and the impacts of traffic control, as analytical LPFs are unable to accurately represent real-world congestion dynamics.

(ii) **Why it's misleading and how to address it**: while simulators capture the regime A→B transition and regime C queues, they incur high computational cost and offer limited marginal-cost insight. An analytical LPF, extended by mirroring the congested branch into regime C, efficiently reproduces key KPIs (throughput, delay, congestion duration). A hybrid workflow, using LPFs for KPI-driven calibration and targeted simulation for local validation, balances speed with fidelity across all regimes.

Misconception 3: Black-box AI models eliminate the need for LPFs.

(i) **Common belief**: If you can forecast link speed or flow with LSTM, GNN, or Transformer models, we no longer need an LPF.

(ii) **Why it's misleading and how to address it**: Forecasting models (LSTM, GNN, Transformers) excel in regime A but typically ignore queue spillback (B→C) and unobserved demand (observed flow→inferred demand D). Without embedding the physical D/C mapping from Fig. 2, a model trained only on regime A will under-perform once capacity is exceeded. Rather than relying solely on predictive models, the integration of machine learning (ML) into a multi-stage LPF framework comprising fundamental diagram fitting, traffic-state estimation, and dynamic α and β parameter updates enable the extraction of corridor- and region-level insights. These include demand inference, capacity estimation, and real-time system adaptation, which are typically unattainable through standalone forecasting approaches

3 Key Insights from the Multiresolution Traffic Flow Modeling Perspectives

The development of multi-resolution methodologies (MRMs) for traffic analysis has gained significant attention, as highlighted in the comprehensive Department of Transportation (DOT) study by Zhou et al. (2021), Hadi et al. (2022). These methodologies, successfully applied in computational fluid dynamics, offer a framework for expressing short-range, mid-range, and long-range relationships consistently and explicitly. In the context of traffic flow, a well-defined VDF that satisfies MRM requirements can: (1) reduce computational complexity that would otherwise be required for detailed simulation, and (2) better capture long-range demand-supply interaction phenomena, improving numerical robustness. It should be noted that many researchers in the field of traffic flow theory also recognize that the traditional BPR formula has limitations in describing flow evolution when capacity is reached. As a result, the subject of macroscopic fundamental diagrams for urban traffic has recently received significant attention (Daganzo and Geroliminis, 2008; Gayah and Daganzo, 2011; Geroliminis and Sun, 2011).

3.1 Key Questions Driving the Field

Based on studies by Dowling et al. (1998) and Dowling et al. (2011), we identified critical questions that continue to shape VDF research:

(1) How can temporal resolutions and appropriate levels of service (LOS) be selected to define capacity consistently between design capacity, practical capacity, and ultimate capacity?

(2) How can the maximum possible value for capacity be defined while recognizing its stochastic distribution?

(3) How can inflow demand be defined and queue discharge rates be distinguished from design capacity, especially under heavy congestion?

(4) How can daily volume, peak hour volume, and demand duration congestion be mapped through widely used parameters?

(5) How can we obtain consistent average delay from the VDF with underlying queueing models?

These questions frame the four key insights and actionable takeaway messages that structure the following review.

3.2 Insight 1: Understanding True Demand Versus Measured Flow

3.2.1 The Measurement Problem: Evidence from Literature

The distinction between measured flow and true demand has been recognized by numerous researchers. Dowling and Skabardonis (1993), Akcelik (1996), and Fambro and Rouphail (1997) highlighted the importance of recognizing underlying traffic flow states when addressing demand over capacity conditions. As stated in Chiu et al. (2010), the dynamic traffic assignment (DTA) research community recognizes that the V/C ratio does not directly correlate with physical traffic state measurements describing congestion (e.g., speed, density, or queue). Small and Verhoef (2007) provide a particularly illuminating analysis, highlighting that a certain level of allowed flow can result from either a congested or a hyper-congested speed and density. They demonstrate that the BPR function does not account for how long traffic exceeds capacity and propose a piecewise-linear relationship model to estimate travel time patterns using congestion duration.

We summarize the different modeling definitions of demand D in published papers for congested conditions, as shown in Table 3.

Table 3 Various definitions of demand variables for congested conditions in the literature.

Categories	Author/s/Year	Detailed definition
Highest flow rates associated with traffic intensity	Kimber and Hollis (1979)	The definition of traffic demand in a stream is q and the capacity available to it μ (both expressed in average vehicles per unit time) at a given stage, traffic intensity is defined as $\rho = V/\mu$
Peak-period inflow Volume D	Small (1983)	Peak-period inflow volume D, uniform rate and delay result from queuing behind a single bottleneck with a constant capacity.
Approximated overflow rate	Moses et al. (2013)	$D = capacity + (capacity - volume)$
Queued demand	Huntsinger and Rouphail (2011)	$D = demand\ at\ capacity + queue$
Approximation through density	Kucharski and Drabicki (2017)	$D = \frac{k}{k_c} \cdot C$
Total volume during the congested period	Cheng et al. (2022); Newell (1982)	$D = \sum_{t_0}^{t_3} q(t), t\epsilon[t_0, t_3]$ Demand is the total volume during a congestion period

3.2.2 Theoretical Foundations from Queue Theory

Rakha and Zhang (2005) use shockwave and queuing theory procedures to analyze bottlenecks, proposing that the area between demand and capacity curves represents total delay. Building on this foundation, Huntsinger and Rouphail (2011) derive a demand-to-capacity function by analyzing bottlenecks and queues using freeway detector

data. Their key insight is using demand at capacity plus the queue as the demand under congestion conditions. Recent work by Bliemer and Raadsen (2020) reformulates the BPR function to propose a model with capacity and storage constraints for static traffic assignment models with possible residual queues and spillback.

3.2.3 Capacity Definitions from Traffic Bottleneck Identification Perspectives

The complexity of capacity definition has been extensively studied. Branston (1976) provides an early comprehensive analysis that highlights the critical importance of time interval selection for capacity estimation, while Horowitz (1991) and the Highway Capacity Manual (2000) establish widely used frameworks. The literature reveals multiple capacity concepts:

Practical Capacity: Defined by Branston (1976) as the maximum number of vehicles that can pass a given point during a specified time period. Dowling et al. (1998) state that practical capacity is 80% of actual capacity.

Ultimate Capacity: Horowitz (1991) emphasizes providing consistent procedures to find "design capacity" that yields reasonable delay estimates.

Stochastic Capacity: Lorenz and Elefteriadou (2000), Jia et al. (2011) and Zhang et al. (2025) demonstrate capacity's probabilistic nature.

A traffic bottleneck is a result of a specific physical condition, often related to the design of the road, a saturated traffic intersection, sharp curves, or incidents. Before the transition from uncongested to congested flow occurs, capacity drops in lanes that experience the most traffic. This capacity breakdown phenomenon has been investigated systematically in many studies (Banks, 1991; Hall and Agyemang-Duah, 1991; Cassidy and Bertini, 1999; Lorenz and Elefteriadou, 2000; Kerner, 2000). Hall and Agyemang-Duah (1991) propose that capacity is not the observed absolute maximum flow. Instead, it is the flow rate that can be repeatedly achieved, day in and day out. This is based on the premise that capacity has a distribution, not a single value achieved on all similar days. Cassidy and Bertini (1999) suggest that enhanced cumulative count curves constructed from counts and occupancies measured at neighboring locations along the roadway can serve as the diagnostics for a bottleneck. Zhang and Levinson (2004) suggest that the capacity at a bottleneck should be the weighted sum of the long-run average queue discharge flows and the long-run average pre-queue transition flows, both of which follow normal distributions. Yeon et al. (2009), Kim et al. (2010), and Mahmassani et al. (2013) discuss many types of flow rate measures related to capacity, including the pre-breakdown flow rate, the post-breakdown flow rate, breakdown flow, maximum queue discharge flow, etc. Based on our literature review findings, the following definitions were considered for data from July 5, 2017 (Wednesday) on a corridor along US I-405 in Los Angeles, California, based on data from the Performance Measurement System (PeMS) (2013) database, as shown in Fig. 3.

Definition A (Free-flow rate): the flow rate corresponding to the free-flow speed.

Definition B (before breakdown rate): a per lane equivalent hourly rate, observed immediately before the onset of traffic breakdown.

Definition C (maximum pre-breakdown flow as ultimate capacity): the highest 5 min flow within 2 h. before the breakdown during undersaturated conditions. For example, suppose there is a previous breakdown within the 2 h limit. In that case, the maximum pre-breakdown flow is the 5 min flow per lane that occurred after the last congested conditions ended and before the subject breakdown flow started.

Definition D (Breakdown): the 5 min flow per lane for the interval before the breakdown event (i.e., before the abrupt speed drop).

Definition E (Lowest queue discharge flow): the lowest 5 min flow during the congestion period.

Definition F (Recovery flow): the 5 min flow per lane observed immediately after the traffic recovery.

In terms of the evolution of flow and speed measured along with the time series, we considered six states from A to F, accordingly, in one peak period of a bottleneck, as shown in Fig. 3.

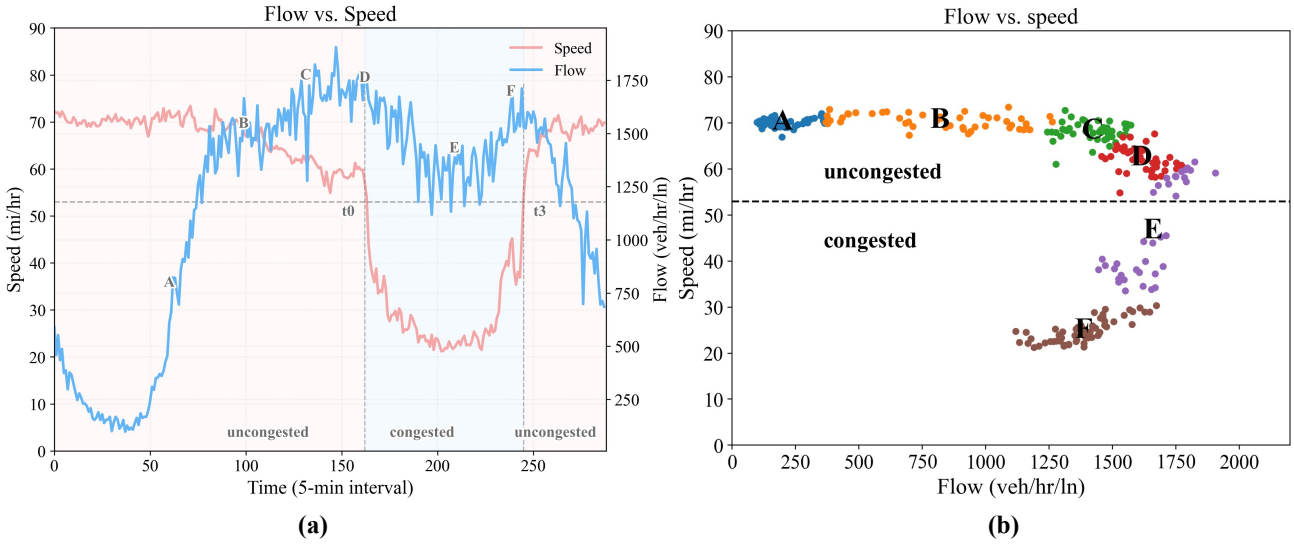


Figure 3 Different definitions of capacity-related flow rates measured in a bottleneck (5-min intervals): (a) flow and speed along with the time series, (b) flow and speed relationship. A. Free-flow rate; B. Before breakdown rate; C. Maximum pre-breakdown flow as "ultimate capacity"; D. Breakdown; E. Lowest queue discharge flow; F. Recovery flow. (adapted from Pan et al., 2025)

Takeaway: Extract cumulative arrival and departure curves to compute D/C ratios and use these in place of raw V/C ratio when calibrating delay functions. Substituting D/C ratio for V/C ratio ensures delay-function calibration reflects true demand, yielding more accurate travel-time predictions and network-performance assessments. Recent illustrative studies includes Belezamo (2020) introduced a polynomial point-queue model leveraging cumulative arrival/departure data to estimate congested travel times; Wu et al. (2021) developed a queue-theoretic extension of the BPR function capturing both free-flow and oversaturated regimes; Pan et al. (2022) formulated a dynamic volume-delay function calibrated on two loop-detector corridors using a rolling-horizon framework. Accurate determination of the D/C ratio, particularly under oversaturated or bottleneck conditions, is critical for reliable traffic assignment and planning. These advancements also create new opportunities for demand-driven calibration of LPF.

3.3 Insight 2: Fluid-Queue Models for Multi-Scale Consistency

Again, traditional link performance functions are primarily concerned with average travel time measurements. By examining how time-dependent performance measures are derived from queueing models and numerical methods for solving fluid approximation equations, we aim to shed light on the potential connection between time-varying delay and its aggregated form. The following section describes the evolution from queueing theory to fluid approximations.

3.3.1 Stochastic Queueing with Undersaturated Conditions

A wide range of link delay functions are adopted in related non-transportation disciplines, such as data communication networks (Bertsekas and Gallager, 2021), and many closed-form link performance functions can be seamlessly embedded in the upper layers of flow control and access control. It should be noted that these performance functions typically consider undersaturated conditions, and the V/C ratio is expressed in terms of utilization rates as

$\rho = \lambda/\mu < 1$ and the analytical form of expected waiting or response time can be derived based on specific queueing systems such as M/G/1 (Berman et al., 1987). In the traffic engineering domain, Davidson's function (Davidson, 1978) was developed based on the queueing model with random arrivals and an Erlang service distribution, and Akçelik (1991) proposes an alternative time-dependent queueing function to improve the estimation of intersection delays.

3.3.2 Bottleneck Models with Piecewise-constant Arrival and Departure Rates

Assuming deterministic, piecewise-constant arrival rates and fixed discharge capacities, the point-queue framework developed by Vickrey (1969) provides closed-form, piecewise-linear travel-time profiles for oversaturated bottlenecks without explicit storage constraints. Embedding early- and late-arrival penalties around a user's preferred arrival time further produces dynamic departure-time equilibrium solutions under a constant value of time. Extensions by Small and Verhoef (2007) introduce hyper-congestion effects, in which excessive density reduces flow, to account for queue spillback and upstream blockage. Meanwhile, Akcelik (1980) applies the same deterministic assumptions to signalized intersections, deriving analytical expressions for time-dependent queue length, delay, and stop rates based on effective green-time ratios (g/C) and saturation flows, as demonstrated in Fig. 4.

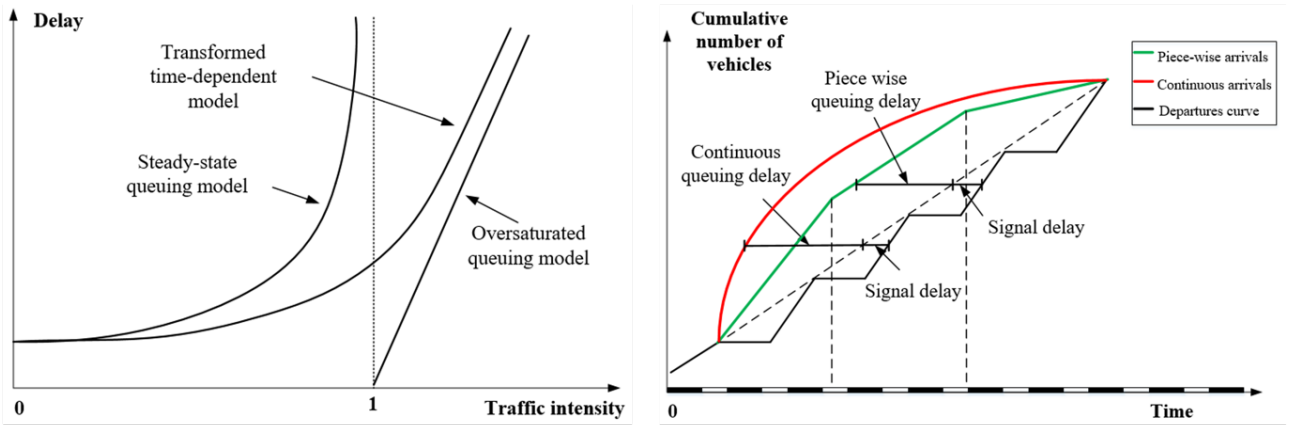


Figure 4 Deterministic queuing analysis in a signal intersection by Kimber and Hollis (1979): (a) steady-state delay, deterministic queuing delay, and transformed delay curves (b) Arrival and departure curves for a saturated link.

3.3.3 Fluid Approximation Queue with Polynomial Arrival Rates for Oversaturated Conditions

To analyze the queue evolution process, Newell (1968a, 1968b, 1982) proposes fluid-based queueing models with time-dependent arrival rates, characterized by quadratic functions, in oversaturated conditions. As noted by Hall (1991), this type of deterministic fluid approximation model has also been used as an essential tool to represent non-stationary queueing systems, especially with "predictable fluctuations", which tend to overwhelm random changes in arrival and service processes. Schwarz et al. (2016) systematically review the performance of time-dependent queueing systems, and they classify them into three groups: numerical and analytical, piecewise constant, and modified system models. Recently, Cheng et al. (2022) extend Newell's queue model to consider cubic arrival rates to analytically capture the evolution of queue saturation with closed-form delay functions.

It is crucial to understand how polynomial arrival queue (PAQ) models represent different system performance characteristics based on a parsimonious set of critical parameters. Such an understanding can help planners establish

a consistent definition and unified modeling framework to map time-varying traffic performance observations to different modeling elements in traditional VDFs. As indicated in Table 4 and Fig. 5, we offer the following four observations. (1) In a given analysis period $[t_s, t_e]$, inflow demand $\lambda(t)$ can be represented by the congestion duration P and the queue discharge rate μ during the congestion period $[t_0, t_3]$. (2) The PAQ is able to provide a closed-form of average delay $\bar{w}$ consistent with the time-dependent queueing delay $w(t)$. (3) There are subtle but important differences between the macroscopic parameters of V/C in a VDF and parameters D/μ. (4) The time-dependent travel time $tt(t)$ in the underlying fluid approximation queue model should satisfy the First in First out (FIFO) property and capacity constraints.

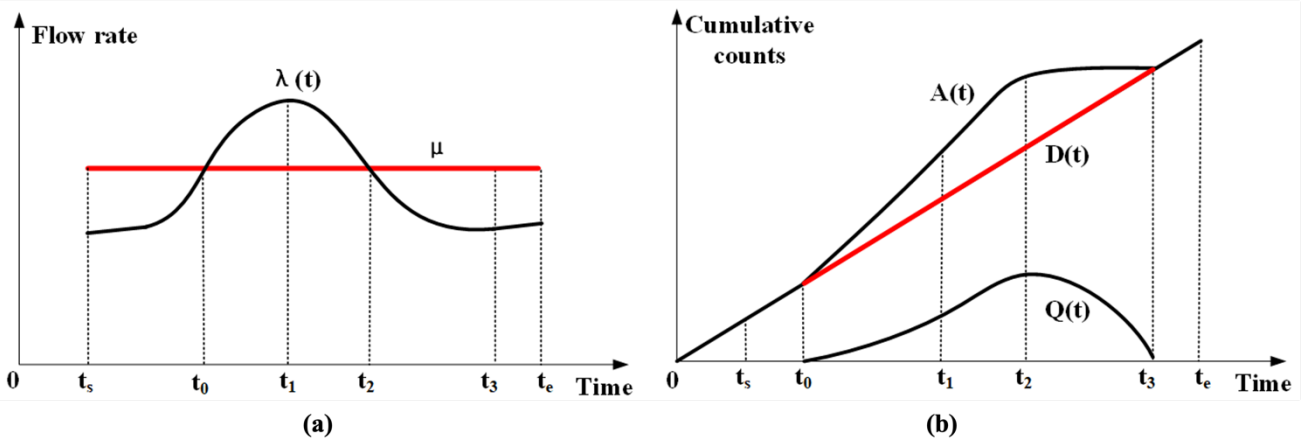


Figure 5 Deterministic queuing analysis in a bottleneck: (a) polynomial arrival and constant discharge rate, (b) cumulative arrival and departure rate.

Balance of Fidelity and Parsimony: While Newell's closed-form relationship offers a lightweight, direct mapping, the more sophisticated fluid-queue and kinematic-wave schemes (e.g. cell transmission model by Daganzo (1994), link transmission model by Yperman et al. (2005), modified input-output diagrams), provide richer spatio-temporal detail. However, these methods require greater computational effort and more extensive data inputs for dynamic traffic assignment (Ran et al., 1996; Ran and Boyce, 1996). Lawson et al. (1997) use a modified input-output diagram to connect the space-time extents of a queue and the delay of a bottleneck to simplify evaluating the waiting time, queue length, and travel time in a jam. As highlighted by Carey (2004) and Carey et al. (2014), the link travel time function should satisfy certain properties, in particular, uniqueness and continuity, FIFO, causality, and time-flow consistency, there are consistent research efforts along this direction (Long et al., 2011; Raadsen and Bliemer, 2019).

Table 4 Key parameters and performance measurements among the different queueing models.

Models	Loading ratio	Arrival process	Departure process	Travel time output	Modeling of interaction
Stochastic queue	λ/μ for undersaturated conditions	Random arrival processes such as Poisson distribution	Determined by service time distribution	Average waiting time or response time	Delay is caused by interference among vehicles/customers
Standard point queue in bottleneck model	A queue will form if the outflow rate exceeds the maximum	Piecewise constant time-varying arrival rates	Constant maximum discharge rate	Closed-form solutions for time-varying travel time	Outflow depends on its own inflow, bottleneck capacity; vehicles stack vertically, and the

	discharge rate (capacity flow).				queue occupies no space and does not affect upstream links.
Fluid approximation queue	D/μ for oversaturated conditions.	Polynomial arrival rates with quadratic or cubic form, specifically for the congested duration only. The arrival rate outside congestion duration is not modeled due to possible approximation errors.	Constant discharge rate.	Closed-form solutions for time-varying travel time.	During congestion duration, the vehicle's space-time trajectories are mapped from the fluid approximation dynamic point queueing process to spatial queues with constant congestion speed.
Link travel time function	Time-dependent flow rate.	Allow form-free time-dependent arrival rates and capacity-based inflow restrictions.	Link exit flow function with link capacities as an upper bound.	The explicit function forms for time-dependent travel time; link travel times reach an upper bound after spillback.	Capture travel time by modeling delays without capturing queues explicitly, a special type of FD lacking the capability of withholding traffic flows when exceeding the physical link capacity.
Kinematic wave related models	Determined by initial and boundary conditions in terms of cumulative vehicle numbers.	Allow form-free time-dependent arrival rates during discretized time intervals.	Outflow depends on its own inflow, bottleneck capacity, and downstream spatial capacity supplies.	Indirectly derived from speed in numerical solutions, link travel time could reach an upper bound after spillback.	Simplified kinematic wave propagation principle for modeling shock waves.

Takeaway: Use Newell's closed-form to translate time-dependent queue length $Q(t)$ into time-dependent travel times rather than relying solely on empirical VDFs. Given many underlying queuing representations as shown in Table 4, modelers must therefore strike an application-driven balance between representational fidelity and parsimonious model requirements, choosing the simplest formulation that still captures the key dynamics needed for their real-world context.

3.4 Insight 3: Congestion Duration as the Primary Delay Driver for Traffic Bottleneck

3.4.1 Beyond Instantaneous Metrics: Literature Evidence

Travel planning models need to forecast future travel demand and load that demand onto a roadway network. Typically, in engineering practice, a peak hour factor (PHF) is used to convert the peak hour traffic volume into the flow rate representing the busiest 15 minutes of the rush hour in the assignment period (e.g., AM or PM). Previous

research indicates that the PHF can strongly impact traffic analysis results (Tarko and Perez-Cartagena, 2005). It can be used to quantify the effects of short-term traffic peaking, leading to congestion.

In traffic assignment applications, to obtain a more temporally comprehensive understanding of congestion impacts, planners can use a peak spreading method (Horowitz et al., 2014) to expand the peak period of traffic from the traditional one-hour peak to a multi-hour peak period. In a recent study by Wu et al. (2021), the threshold speed is selected first so that congestion durations and queued demand volume for each link are computed more precisely. The relationship between PHF, peak spreading characteristic, and assignment period is shown in Fig. 6. Based on speed data, one can determine which hours of traffic surpass capacity, and this hour is typically defined as the peak hour. There are essential differences between congestion duration and heaviest volume hour: all the hours in which inflow demand volumes exceed capacity have been identified in congestion duration.

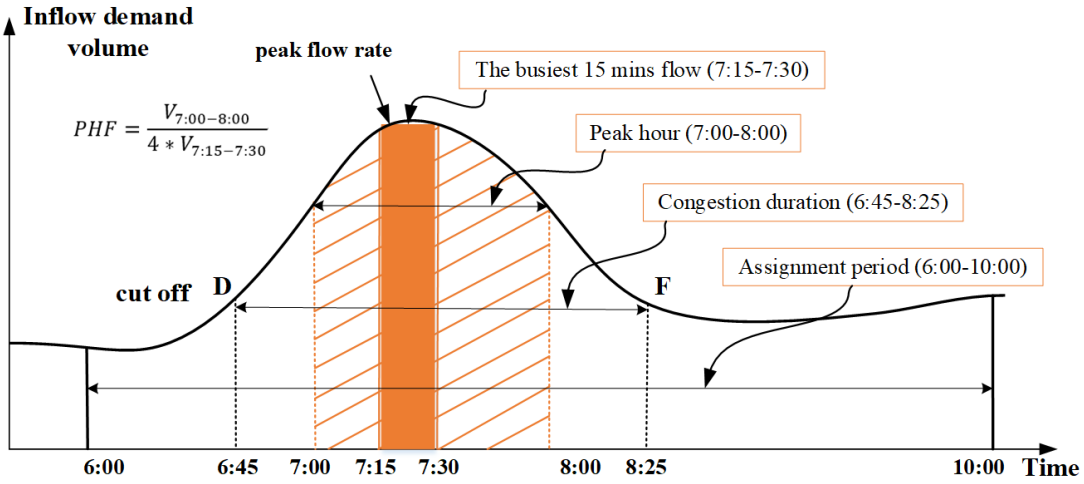


Figure 6 The relationship between peak flow rate, congestion duration, and assignment period; points D and F correspond to states D and F in Fig. 3. The y-axis is the inflow demand volume, as the observed flow count between states D and F is lower than the ultimate hourly capacity, as illustrated in Fig. 3.

It should be noted that in the literature on traffic delay functions for signalized intersections, the degree of oversaturation is expressed as $x = traffic\ peaking\ intensity$ (Rouphail et al. 1992), which is in turn represented by the ratio of highest volume and capacity. From the perspective of fluid approximation queues with quadratic arrival rates, Newell (1982) provides a closed-form solution as follows in Eq. (6).

$$\frac{D}{\mu} = P = 3 \times \left[\frac{\lambda(t_1) - \mu}{\gamma}\right]^{\frac{1}{2}} \tag{6}$$

where $\lambda(t_1)$ is the highest volume at time t_1, as shown in Fig. 5(a), and γ is the inflow demand curvature parameter, P is congestion duration.

As discussed in the trip scheduling literature (Small and Verhoef, 2007), the key variables of congestion duration P and longest waiting time in the Queueing-based Volume-Delay Function (QVDF) (Zhou et al. 2022) should be integrated into the departure time choice behavioral model with considerations of preferred arrival time and schedule delay in bottleneck models. Recently, Wu et al. (2021) and Zhou et al. (2022) propose the use of the total inflow demand volume during the congestion duration as the D variable in the D/C ratio. By introducing two types of elasticity functional forms, Zhou et al. (2022) further demonstrate connections from the macroscopic inflow D/C ratio to the congestion duration of a bottleneck, from the congestion duration to the magnitude of speed reduction.

Takeaway: Traditional metrics such as speed, volume, and peak V/C ratio often overlook the fact that congestion duration ($P = t_3 - t_0$) explains far more delay variation than peak V/C ratio alone. We therefore recommend fitting a fluid-queue model to identify $[t_0, t_3]$ and adopting Δt (or P) as a core KPI in both VDF calibration and signal-timing analyses. As illustrated in Fig. 7 starting from the inflow rate $\lambda(t)$, the diagram contrasts two regimes:

(i) Uncongested conditions: A higher V/C ratio leads to a nonlinear increase in travel time (TT). Enhancing link capacity c reduces the V/Cratio, thereby decreasing travel time TT.

(ii) Congested conditions: The demand-to-capacity ratio D/C governs key traffic state variables, including queue length $Q(t)$, discharge rate μ, and congestion duration P. As the D/C ratio increases, $Q(t)$ becomes longer, μ decreases, and P extends, which amplifies overall delay and increasing time-dependent travel time $TT(t)$.

This causal map encapsulates how duration, not just intensity, underpins bottleneck delay and highlights where parsimonious fluid-queue models can improve LPF accuracy.

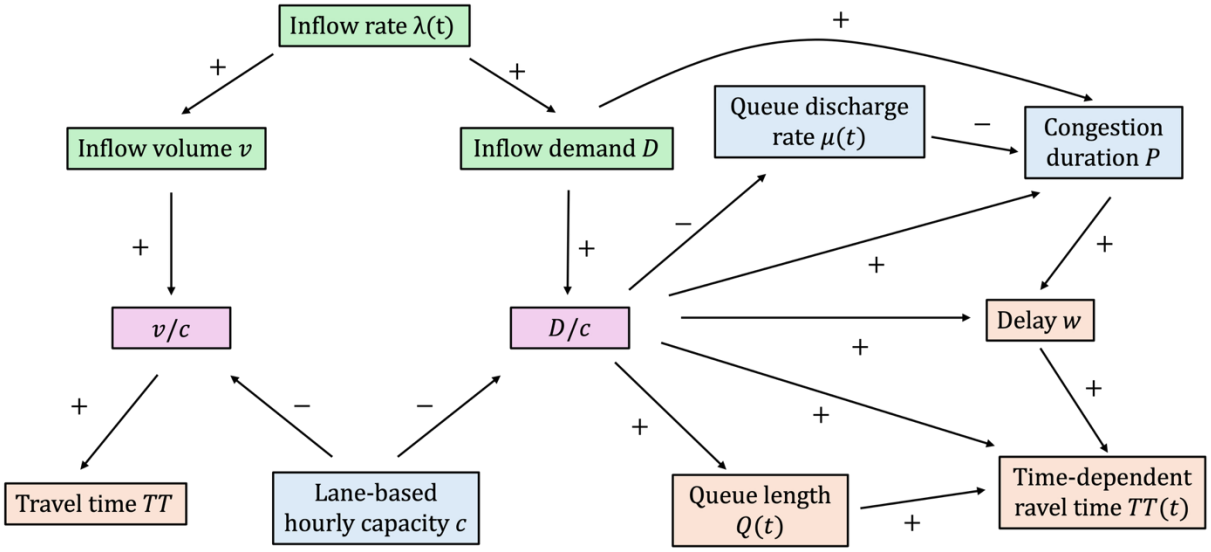


Figure 7 Causal interactions of key variables in the LPF framework.

3.4.2 Operational Framework for D/C Ratio Calibration and Implementation

With the advancement of transportation sensing techniques, it has become increasingly important to systematically utilize available data to select appropriate VDF forms and calibrate their associated parameters. The general workflow for developing and calibrating VDF models can be summarized as follows, as illustrated in Fig. 8.

(1) Fundamental diagram calibration: The process of calibrating the fundamental diagram involves determining and adjusting coefficients based on traffic flow models, including road capacity, critical speed, critical density, and free-flow speed. A multitude of empirical studies (Cheng et al., 2021; Wang et al., 2011; Ni et al., 2016) have confirmed that the FD provides a clearer representation of traffic flow characteristics and can be used to distinguish between various traffic states based on important parameters.

(2) Computation of V/C or D/C ratios: Determining the V/C or D/C ratio involves analyzing existing congestion bottlenecks by determining the congested period and average waiting time, and computing the demand during peak periods. The key parameters, the V/C ratio and the D/C ratio can be calculated for different facility type.

(3) Calibrating volume-delay functions: Calibrating Volume-Delay Functions: Develop and calibrate one or multiple functional forms of volume-delay functions (e.g., α, β in BPR function), taking into consideration the probabilistic distributions of capacity, demand, and travel time parameters.

(4) Traffic assignment: The calibrated VDFs are embedded into traffic assignment models for network design or policy analysis. The traffic assignment results are compared against baseline data using a specified origin-destination matrix and model outputs are validated against observed traffic conditions in the base year.

(5) Evaluating measures of effectiveness (MOE): Based on the assignment outputs, additional MOEs such as queue length, delay, and emissions can be derived for both short-term and long-term planning tasks.

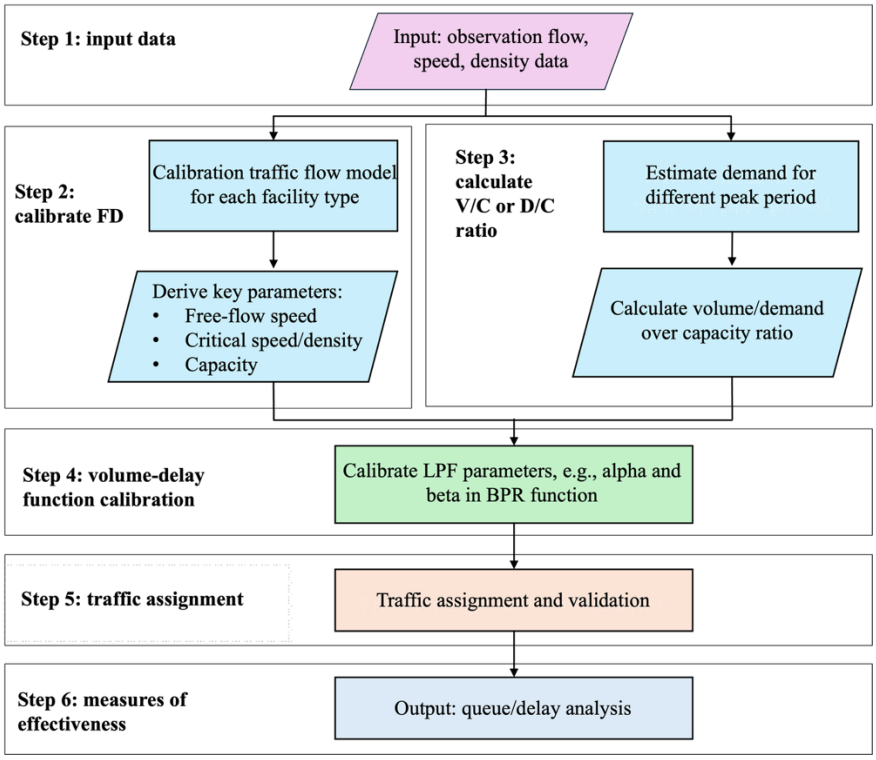


Figure 8 Modeling calibration frameworks for characterizing VDFs.

3.5 Insight 4: Micro-to-Macro Cross-Resolution Modeling—Extending LPFs for Connected and Automated Vehicle Evaluation and Network Design

Connected and automated vehicles introduce unique microscopic traffic flow characteristics that challenge traditional assumptions embedded in LPFs. Understanding how these microscopic behaviors scale to macroscopic traffic dynamics is crucial for designing efficient and sustainable transportation systems. We first synthesize the key literature, then demonstrate how microscopic CAV behaviors map to macroscopic LPFs via concrete examples.

3.5.1 Literature based on Microscopic Dynamics and Car Following Model Integration

An early study by Chen et al. (2011) applied the describing function method, substituting a nonlinear car-following model with a linear one. This process begins with Newell's parsimonious car-following model, which correlates reaction time and minimum safe distance. Li (2022) emphasized the crucial role of microscopic car-following models in balancing safety, mobility, and stability for commercial AVs. Nonlinear effects within AV control

may potentially amplify traffic oscillations, posing challenges for modeling such dynamics. Levin and Boyles (2015), Mahmassani (2016), and Huang et al. (2023) examined how the minimum safety distance can be reduced in CAV scenarios. More recent work has built on these foundations: Vehicle-to-Infrastructure (V2I)-enabled automation can cut intersection saturation headways by more than 20 % (Hajbabaie et al. 2024), while mixed-traffic throughput gains only emerge past certain CAV thresholds (Li et al. 2025). Moderate AV penetration improves urban travel-time reliability (Samaranayake et al. 2024), though intersection capacity still hinges on market share and signal logic (Song and Fan, 2023), and platooning models extend turbo-roundabout capacity forecasts (Guerrieri 2024).

3.5.2 Literature based on Future Transportation Network Design

Transportation network design is an essential planning task in improving transportation networks' speed and travel times by selecting the optimal projects among alternatives. Through a case study in Stockholm, Engelson et al. (2015) highlighted the importance of refining the VDF for the forecasted year in evaluating network design to avoid overestimation or underestimation of the flow/travel time changes. In addition, the design of a new road network or the upgrading of existing roadways would require a good understanding of the uncertainty involved and the impact on the system-wide performance (Chen et al., 2011). Many studies, for example by Yang and Wang (2011), Correa et al. (2004), Zhang and Levinson (2008), Nagurney and Qiang (2012), Wong and Wong (2016), and Wei et al. (2019), also focused on evaluating various problems related to transportation network design problems.

3.5.3 Some Key Questions Need to be Considered in Future Studies

(a) How can consistent definitions and a unified modeling framework be developed for multiresolution modeling? Future studies should use an integrated supply-side parameter calibration package with consistent definitions of traffic flow models and macroscopic VDFs. They could capture the relationship between discharge rates, demand-to-capacity ratio, during the congestion period, and queue spillback.

(b) How can the VDF function be developed to analyze performance on both the demand and supply sides? It is important to determine the cost equilibrium point where demand equals supply. While the traditional static models could significantly underestimate network congestion levels in traffic networks (Boyles et al., 2008), how to connect VDFs with available DTA models to account for variable demand and traffic dynamics can be a critical future research direction, especially under a policy of time-dependent congestion pricing.

Using Fig. 9. and Table 5, we examine three distinct scenarios: (0) baseline with zero AV penetration (blue dotted line), (1) high AV penetration (solid red line), and (2) managed demand scenario (solid green line).

(1) **Microscopic Car-Following Dynamics.** Our analysis begins with Newell's parsimonious car-following model, correlating reaction time and minimum safe distance. Subfigure (a) shows trajectory transitions from free-flow to congested speeds, while subfigure (b) displays the resulting piecewise linear approximation trajectories. Consistent with previous research (Levin and Boyles, 2015; Mahmassani, 2016; Huang et al., 2023), minimum safety distance reduces from 7 meters in the baseline to 5 meters under high AV penetration, directly influencing backward wave speed. Subfigure (c) demonstrates the linear relationship between distance headway and speed in Newell's model, where the slope represents reaction time and the intercept signifies minimum distance headway. This micro-to-macro linkage is essential for understanding how individual vehicle interactions influence broader traffic dynamics.

(2) **Macroscopic Flow and Capacity.** Subfigures (d) and (e) show macroscopic traffic flow and capacity characteristics approximated using a triangular fundamental diagram (TFD). The backward wave speed in

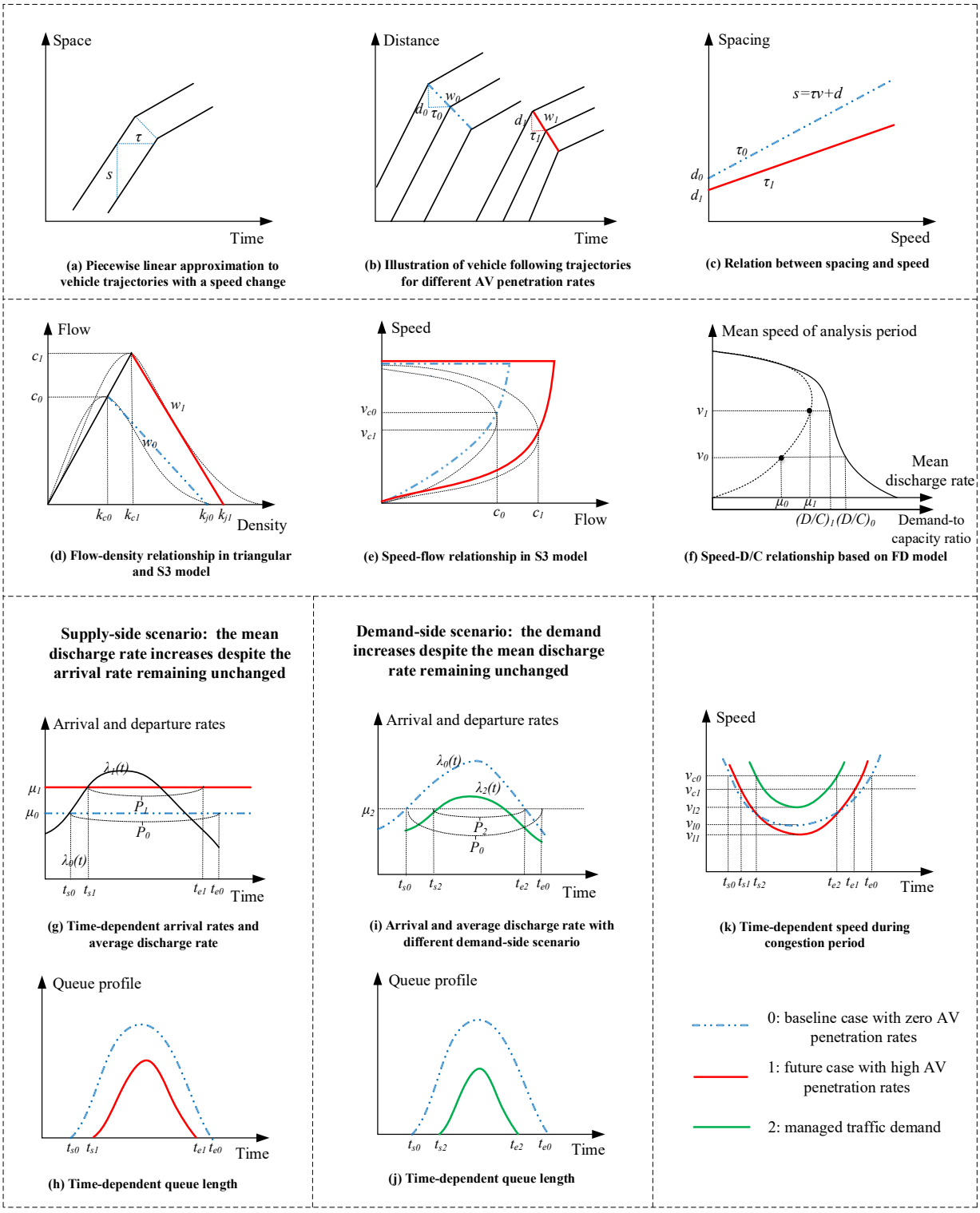
the TFD aligns significantly with the microscopic car-following model's wave speed. While the TFD provides straightforward approximation, other fundamental diagram shapes like the S3 model (Cheng et al., 2021) could be represented (dotted line). The abrupt transitions from free-flow to congested states in the triangular flow-density curve (subfigure (e)) is generally undesirable for calibrating traffic volume-delay functions. Traditional highway capacity studies demonstrate smoother transitions through LOS A-D, mirroring the S3 model's speed-flow relationship. Subfigure (f) illustrates the "mirroring" perspective linking fundamental diagrams and link performance functions by mapping congested states to regimes with demand-to-capacity ratios greater than 1.

(3) **System-Level Implications.** At the macroscopic level, we explore supply-side and demand-side scenarios for regional policy decision support. Supply-side improvements through higher CAV penetration enhance capacity, resulting in shorter congestion durations despite constant arrival rates (subfigures (g) and (h)). In demand-side scenarios employing active traffic management strategies, including time-varying tolling and advanced reservation systems, arrival rates are reduced, and queue lengths are shortened, while discharge rates remain consistent, as illustrated in subfigures (i) and (j). Subfigure (k) illustrates time-dependent speed variations across all three scenarios during congestion periods, with corresponding parameters detailed in Tables 6 and 7. These visualizations collectively demonstrate the intricate relationships between flow rates and system characteristics across various temporal and spatial scales.

To operationalize the proposed multi-resolution framework, microscopic CAV parameters must be systematically transitioned into mesoscopic delay representations. Reductions in reaction time τ_0 and minimum gap d_0 at the microscopic level translate into higher critical density k_c and modified backward wave speed w in the fundamental diagram. These macroscopic parameters directly influence mesoscopic delay estimation through changes in discharge rate μ and congestion duration P. Calibration is performed hierarchically: (i) microscopic parameters are estimated from CAV trajectory data or simulation; (ii) macroscopic parameters are obtained by fitting the fundamental diagram; and (iii) mesoscopic parameters such as discharge rate and D/C ratios are extracted from cumulative arrival-departure curves. Data requirements thus vary across scales, ranging from high-resolution trajectory datasets at the micro-level, to loop-detector or probe-based aggregate flows at the macro-level, to OD demand and bottleneck-specific measurements at the meso-level. A summary of these parameter transitions and calibration procedures is provided in Table 5.

Table 5 Parameter transitions and data requirements across micro-meso-macro scales.

Scale	Key parameters	Parameter transition	Data requirements
Micro-level	Reaction time τ_0, minimum gap d_0, acceleration distribution	Driving behavior affects car-following and lane-changing dynamics	High-resolution CAV trajectory data, simulation outputs
Macro-level	Critical density k_c, capacity C, backward wave speed w	τ_0 and d_0, translate into k_c and w; fundamental diagram fitting yields C	Loop detector flow and speed data, probe vehicle measurements
Meso-level	Discharge rate μ, D/C ratio, flow rate q and congestion duration P	k_c, q and w result in μ, D/C, P derived from cumulative curves.	Cumulative arrival-departure curves, OD demand, and bottleneck data



(a) Piecewise linear approximation to vehicle trajectories with a speed change

(b) Illustration of vehicle following trajectories for different AV penetration rates

(c) Relation between spacing and speed

(d) Flow-density relationship in triangular and S3 model

(e) Speed-flow relationship in S3 model

(f) Speed-D/C relationship based on FD model

(g) Time-dependent arrival rates and average discharge rate

(h) Time-dependent queue length

(i) Arrival and average discharge rate with different demand-side scenario

(j) Time-dependent queue length

(k) Time-dependent speed during congestion period

Figure 9 Comprehensive LPF modeling framework: A micro-to-macro level perspective across diverse scenarios.

Table 6 Comparing microscopic and macroscopic parameter differences between the baseline and future scenarios.

Parameters	0: baseline case with zero AV penetration rates	1: future-year case with high AV penetration rates	References
Minimum distance between the leading and following vehicle d (m)	$d_0 = 7$	$d_1 = 5$	Huang et al. (2023); Levin and Boyles (2015); Mahmassani (2016);
Reaction time τ (s)	$\tau_0 = 1.5$	$\tau_1 = 1$	Levin and Boyles (2015); Shi and Li (2021); Talebpour and Mahmassani (2016); Wei et al. (2017)
Backward wave speed w (km/h)	$w_0 = -16.8$	$w_1 = -18$	$w = \frac{d}{\tau}$
Free-flow speed v_f (km/h)	$v_{f0} = 112$	$v_{f1} = 112$	Shi and Li (2021); Zhou et al. (2022)
Capacity C (veh/hr/ln)	$c_0 = 1800$	$c_1 = 2160$	
Jam density k_j (veh/km/ln)	$k_{j0} = 143$	$k_{j1} = 200$	$k_j = \frac{1000}{d}$
Critical density k_c (veh/km/ln)	$k_{c0} = 36$	$k_{c1} = 80$	$k_c = k_j - \frac{C}{w}$
Critical speed v_c (km/h)	$v_{c0} = 50$	$v_{c1} = 27$	$v_c = \frac{C}{k_c}$
Reference queue-based VDF parameters	$\alpha = 0.21$ $\beta = 1.87$ $f_p = 0.23$ $f_d = 1.37$ $s = 1.64$ $n = 1.14$		Zhou et al. (2022)

Table 7 Comparative analysis of demand and supply-side parameters across three scenarios.

Typical values	0: baseline case with zero AV penetration rates	1: future-year case with high AV penetration rates	2: manage active traffic demand	References
Demand D (veh/ln)	$D_0 = 3600$	$D_1 = 3600$	$D_2 = 2880$	Assumed for illustration with 20% congested demand decrease
Demand over capacity ratio D/C	$\frac{D_0}{C_0} = 2$	$\frac{D_1}{C_1} = 1.67$	$\frac{D_0}{C_0} = 1.6$	Zhou et al. (2022)
Congestion duration P (h)	$P_0 = 3.02$	$P_1 = 2.45$	$P_2 = 2.34$	$P = f_d \cdot (\frac{D}{C})^n$
Mean speed during congestion duration $\bar{v}$ (km/h)	$\bar{v}_0 = 28.51$	$\bar{v}_1 = 17.47$	$\bar{v}_2 = 33.47$	$\bar{v} = \frac{v_c}{1 + \alpha(\frac{D}{C})^\beta}$
Queue discharge rate μ (veh/hr/ln)	$\mu_0 = 1192$	$\mu_1 = 1468$	$\mu_2 = 1230$	$\mu = \min\left[\frac{C}{f_d\left(\frac{D}{C}\right)^{n-1}}, C\right]$
Lowest speed v_l (km/h)	$v_{l0} = 20.93$	$v_{l1} = 13.49$	$v_{l2} = 26.14$	$v_l = \frac{v_c}{f_p f_d^s (\frac{D}{C})^{ns} + 1}$
Start congestion time t_s	t_{s0} =7:00	t_{s1} =7:10	t_{s2} = 7:20	$P = t_e - t_s$

End congestion time t_e	t_{e0} =10:01	t_{e1} =9:37	t_{e2} =9:42

Takeaway: The cross-resolution methodology enables comprehensive evaluation of emerging transportation technologies by integrating microscopic car-following models, mesoscopic link performance functions, and macroscopic fundamental diagrams. This approach captures critical operational metrics beyond travel time, including congestion duration, network resilience, and capacity variations, while incorporating real-world constraints like technology penetration rates and infrastructure limitations. For practitioners, this framework provides evidence-based insights for strategic deployment (planners), infrastructure design optimization (engineers), and robust analytical tools that bridge behavioral dynamics with system performance (modelers), ensuring technology assessments reflect operational reality rather than idealized scenarios.

The four insights collectively highlight the need for physically informed and scalable delay models, which motivates the AI-driven framework presented in Section 4.

4 Physics-Informed AI for Delay Modeling: Research Challenges in an Integrated Demand-Supply Framework

This section presents an original, unified end-to-end AI-driven LPF framework, which integrates data-driven approaches with physics-based constraints in VDF modeling. Whereas the preceding sections primarily synthesize existing literature, the framework proposed here is constitutes an original contribution, designed to address shortcomings in current practices, which include the lack of physical consistency in purely data-driven models and the limited flexibility of traditional formulations. Our framework combines the strengths of both paradigms: embedding traffic flow conservation constraints into learning architectures, while also introducing regime-aware benchmarking and a standardized evaluation protocol to systematically compare model performance. In addition, the framework leverages feature engineering, graph neural networks, and attention mechanisms to integrate roadway attributes with external factors, enabling the prediction of link capacity and the updating of LPF parameters. In doing so, it balances theoretical consistency with cross-scenario generalization.

4.1 From Empirical LPFs to AI-Driven Delay Models

Traditional LPFs are typically based on empirical observations and theoretical assumptions about traffic flow characteristics, such as steady-state conditions and uniform congestion propagation. However, these assumptions may limit their accuracy in complex urban environments. With advances in ML and AI, researchers have begun exploring data-driven approaches to improve the predictive accuracy of LPFs and enhance their adaptability to dynamic and complex traffic conditions. ML techniques can capture nonlinear traffic flow relationships and integrate heterogeneous data from multiple sources, such as sensor-based traffic counts, GPS trajectories, and real-time incident reports. Studies have shown that machine learning models such as neural networks (Wang et al., 2018), support vector regression (Wu et al., 2004), random forests (Cheng et al., 2019) and long short-term memory (Duan et al., 2016) can reduce prediction errors and better adapt to different congestion patterns, outperforming traditional LPFs.

Beyond improving predictive accuracy, AI-enhanced LPF studies have also made progress in traffic management applications, such as congestion pricing optimization (Genser and Kouvelas, 2022), dynamic route choice (Zhao and Liang, 2023), and adaptive signal control (Srinivasan et al., 2006). These studies indicate that integrating AI with

LPFs not only overcomes the constraints of static empirical formulas but also enables the development of adaptive, real-time traffic models. Despite these advancements, AI-driven LPFs still face challenges in practical applications, including data sparsity, the interpretability of black-box models, and computational efficiency issues.

Fig. 10 illustrates the positions of three LPF modeling approaches within a three-dimensional space defined by reliance on traffic-flow theory (vertical axis), utilization of empirical data (depth axis), and ability to generalize across networks (horizontal axis). Classical physics-driven LPFs are located in the high-theory, low-data, high-generalization region. They are firmly grounded in fundamental diagram theory but show limitations in adapting to real-time and complex traffic patterns. Purely data-driven LPFs are situated in the low-theory, high-data, low-generalization region. They rely on sensor and GPS data to capture nonlinear relationships but often lack interpretability and transferability.

Physics-Informed Neural Network (PINN) is a class of deep learning algorithms that can seamlessly integrate data and abstract mathematical operators, including partial differential equations PDEs with or without missing physics (Karniadakis et al., 2021). PINN-based LPFs appear at the top-right corner of the cube as a more comprehensive and enriched framework that integrates traffic-theory constraints with large-scale data learning. The arrows pointing from the green and blue regions toward the red region indicate that PINNs draw on the strengths of both physics-based and data-driven approaches, enabling them to balance interpretability, adaptability, and cross-scenario performance, thereby forming an ideal "sweet spot." It should be noted that Fig. 10 does not present a strict comparison of the absolute level of theory contained in each model. Instead, it highlights how traffic theory is incorporated into different modeling paradigms: traditional physical models rely directly on traffic theory, machine learning models typically do not explicitly include it, while PINNs embed theoretical constraints more deeply into the learning process, which explains their slightly higher placement in the figure.

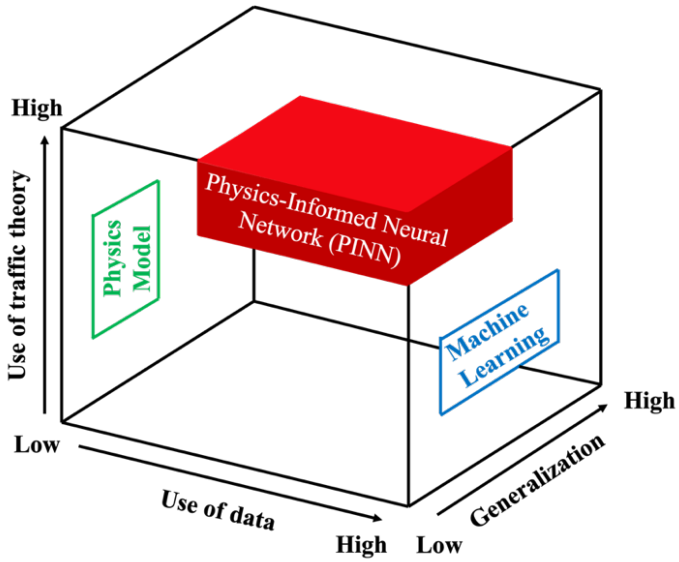


Figure 10 Modeling paradigms for link performance functions: Balancing theory, data, and generalization.

In practice, PINNs are not merely positioned conceptually at the intersection of theory and data, but rather operationalize this integration by embedding physical residuals directly into the training loss. For traffic applications, residual terms derived from the Lighthill-Whitham-Richards (LWR) continuity equation, queue balance relationships, and FD consistency are added as penalty components to the loss function. This explicit formulation ensures that predicted density, flow, and delay trajectories evolve in compliance with conservation laws and remain physically

interpretable, thereby distinguishing PINNs from conventional machine learning approaches that optimize purely against data-fitting error.

4.2 Physics-Informed Hybrid Architectures and Open Challenges

Various ML methods have been adopted in transportation studies (Rowan et al., 2025) to capture complex spatiotemporal patterns in traffic data. Representative examples include k-nearest neighbor algorithms (Tak et al., 2014), convolutional neural networks (Han et al., 2022), long short-term memory networks (Ma et al., 2015), graph neural networks (Cui et al., 2019; Liu and Meidani, 2024), and reinforcement learning approaches (Zhou and Gayah, 2023). In contrast to traditional physics-based models, ML methods are capable of learning subtle and nonlinear relationships among traffic state variables while accommodating data variability.

Several studies have laid the foundation for deep learning and hybrid modeling in transportation, progressively advancing temporal, spatial-temporal, and physics-informed approaches to LPFs. The first application of LSTM to traffic speed prediction was by Ma et al. (2015). Ke et al. (2017) extended LSTM with convolutional layers in their FCL-Net to model short-term passenger demand, effectively fusing spatial and temporal dependencies. Wu et al. (2018) introduced a Hierarchical Flow Network (HFN) that uses computation graphs to represent multi-level travel demand variables, including trip generation, OD matrices, path/link flows, and behavior parameters, thereby enabling structured propagation of errors across planning layers.

In the broader context of delay modeling, Shi et al. (2022) developed "PIDL+FDL", integrating fundamental diagram estimation, state estimation, and parameter identification within a unified neural framework for highway traffic. Recent studies have extended PINNs by integrating them with various traffic flow theories and models to enhance physical consistency and interpretability. Notable efforts include the incorporation of car-following models (Liu et al., 2023), first-order partial differential equations (Huang and Agarwal, 2023), fundamental diagrams (Pan et al., 2024), macroscopic fundamental diagrams (Tang et al., 2024), fluid queue theory (Doll et al., 2024; Xu et al., 2024), deterministic queuing models (Lu et al., 2023; Pan et al., 2025), the bathtub model (Ka et al., 2024), Newell's theory (Sengupta et al., 2025), Kalman filter-based frameworks (Deshpande and Park, 2025), and vehicle trajectory prediction (Long et al., 2024, 2025).

We present the following open challenges to the machine learning community, encouraging contributions that integrate VDF-based D/C analysis with multi-source data and AI methods for transportation network analysis:

(1) **Data and AI Integration: Beyond Black-Box Approaches.** Traditional machine learning approaches in traffic state estimation bifurcate into model-driven and data-driven categories, each suffering from either deficient physics or insufficient data. The emergence of PINN represents a paradigmatic shift, offering a deep learning framework that embeds knowledge of physical laws governing traffic flow datasets in the learning process. The computational complexity of transportation networks demands sophisticated AI architectures that can simultaneously handle temporal patterns and spatial correlations. Physics-aware neural networks using VDF embedded in network architectures could yield traffic state estimations physically constrained by macroscopic traffic flow models, addressing the fundamental limitation of black-box approaches that fail to capture the underlying physics of traffic behavior.

(2) **Sensor Coverage, Link Proportion Matrices and VDF Calibration.** Real networks rely on sparse loop detectors and intermittent GPS probes, yet accurate VDFs demand fine-grained link proportion matrices for OD mapping and modular network design. Reliable calibration of time-varying VDF parameters (α, β, $\mu(t)$)

requires robust denoising, synthetic data augmentation, and transfer learning across corridors to ensure dynamic D/C profiles reflect true area-level flows rather than isolated bottlenecks.

(3) **Differentiable VDF Formulations with Area-Scale D/C and Queue Spillback.** Classical VDFs treat V/C at a single link; hybrid architectures must generalize to area-scale D/C ratios and model queue spillback across adjacent links. Smooth, gradient-friendly representations of congestion duration P (oversaturated cycles) and macroscopic flow-density-speed relationships enable LSTM and other models to learn delay functions adhering to flow conservation and interpretability.

(4) **Online, Multi-Day, Multi-Resolution Inference.** Delay dynamics span signal-cycle (seconds), queue buildup (minutes), and demand fluctuation (days). Hybrid VDFs should bridge multiple spatial and temporal scales by employing tensorized flow representations or hierarchical time grids. This enables support for real-time and multi-day inference in traffic simulators and operational platforms, allowing continuous adaptation to varying inflow rates and control actions.

(5) **Multi-Modal Data Fusion and Physical Spatial Behavior.** Beyond vehicle counts, VDFs should incorporate pedestrian flows, transit schedules, EV charging profiles, CAV trajectories, weather conditions, and incident data, integrating them within a unified computational graph. This fusion allows AI modules to capture how diverse demand and supply streams jointly influence delay, reflecting spatial correlations that black-box methods often overlook.

(6) **Parsimonious AI and Control-Oriented Design**. Complex models risk overfitting and slow inference. A parsimonious approach begins with simple machine learning models, such as LSTM integrated with physics-based priors, to estimate delay functions. It then incrementally incorporates hybrid components including noisy data filtering, event embedding, and flow-through tensor representations to support lightweight control policies such as signal timing and dynamic pricing under system-optimal objectives.

A key advantage of PINNs lies in their ability to incorporate domain-specific knowledge into the training objective, thereby ensuring that the predicted traffic states adhere to first principles. In the context of traffic flow modeling, several classes of physical constraints can be systematically embedded:

(A) Conservation constraints. Vehicle conservation can be enforced by incorporating the residuals of the Lighthill-Whitham-Richards (LWR) continuity equation:

$$\frac{\partial k(x,t)}{\partial t} + \frac{\partial q(x,t)}{\partial x} = 0 \tag{7}$$

This penalization ensures that the predicted density $k(x,t)$ and flow $q(x,t)$ evolve consistently with conservation laws, thereby preventing unrealistic outcomes such as vehicles appearing or disappearing within the network.

(B) Queue balance constraints. For bottleneck or fluid-queue models, the consistency between cumulative arrivals and departures:

$$Q(t) = A(t) - D(t) \geq 0 \tag{8}$$

where $A(t)$ is cumulative arrival counts, $D(t)$ is cumulative departure counts. The queue length also could represent the cumulative imbalance between inflow and outflow rate from the onset of congestion t_0 to time t.

$$Q(t) = \int_{t_0}^{t} [\lambda(\tau) - \mu(\tau)] d\tau \tag{9}$$

where $\lambda(t)$ is time-dependent inflow demand rate, $\mu(t)$ is time-dependent outflow rate. This equation is introduced as an additional training constraint. This formulation guarantees that queue length $Q(t)$ reflects the cumulative imbalance between arrivals and departures, ensuring physically meaningful and bounded queue evoluation.

(C) Fundamental diagram consistency. The relationship between flow and density can be explicitly enforced by penalizing deviations from a calibrated FD, e.g., the triangular FD:

$$q(t) = \begin{cases} v_f \cdot k(t), & for\ k(t) \le k_c \\ w \cdot [k_j - k(t)], & for\ k(t) > k_c \end{cases} \tag{10}$$

which restricts predictions to physically feasible traffic regimes. w is backward wave speed, k_c is critical density, k_j is jam density, $q(t)$ is flow rate at time t, $k(t)$ is density at time t. This restriction ensures that predicted states remain within physically feasible regimes, preventing violations, such as flows exceeding capacity.

(D) Feasibility parameterization. Non-negativity and monotonicity can be preserved through parameterization strategies in the neural network outputs. For example, applying exponential or softplus activation functions ensures that flow and delay remain non-negative, while parameter bounds maintain monotonic behavior near capacity.

Following the classical review by Karniadakis et al. (2021), three major pathways for embedding physics into neural networks are identified: (i) through the data (observational bias), (ii) through the training process (learning bias), and (iii) through the model structure (inductive bias). The four physics-informed constraints summarized in this study are not homogeneous but primarily belong to the second and third categories. Specifically, Types (A-C) correspond to the "training-level" integration, where physical residuals are incorporated into the loss function to guide optimization, whereas Type (D) aligns with the "model-structure" integration, embedding physics directly into activation functions or network architectures.

Moreover, physical constraints can be divided into two subcategories: hard physical laws (e.g., conservation or continuity equations), which act as strict regularization terms ensuring physically consistent outputs; and soft empirical relations (e.g., fundamental diagrams or car-following relations), which serve as guiding priors that may introduce bias when the underlying empirical model is imperfect. This distinction echoes observations from Mo et al. (2021) and Long et al. (2024), who demonstrated that iterative calibration or refinement of empirical models is crucial when soft physical priors are integrated into PINNs. Table 8 summarizes how each constraint type corresponds to its embedding level (data, training, or structure) and physical category (hard or soft), enhancing conceptual clarity and theoretical consistency. Furthermore, by explicitly distinguishing between hard physical laws and soft empirical relations, this classification clarifies how different constraint types correspond to varying integration levels between physics and neural networks.

Table 8 Integration of physics-informed constraints for delay modeling

Constraint type	Integration level	Nature of physics constraint	Description and role in model
Conservation constraint	Training-level	Hard law	Penalizes violation of vehicle conservation; prevents mass-gain/loss artifacts in predicted density and flow fields.
Queue balance constraint	Training-level	Hard law and soft behavioral relation	Ensures inflow-outflow consistency and physically meaningful queue evolution; anchors temporal dynamics near observed arrival/departure patterns.

Fundamental diagram consistency	Training-level	Soft empirical relation	Constrains predictions within feasible flow-density regimes; introduces domain-specific bias reflecting traffic equilibrium relations.
Structural/activation constraint	Model-structure	Hard/Soft	Embeds physics into network design itself to enforce non-negativity and capacity-bounded behavior without explicit loss terms.

Compared with conventional ML approaches such as LSTM, GNN or Transformer models, which can effectively reduce error metrics like RMSE in delay estimation or LPF fitting, PINNs exhibit a distinct advantage in physical consistency. Purely data-driven models often violate conservation laws or generate unrealistic queue dynamics, limiting their transferability across networks and traffic regimes. By embedding residual terms derived from the LWR continuity equation, queue balance relationships, and FD consistency directly into the loss function, PINNs ensure that predicted delay and capacity patterns remain physically interpretable.

Although PINN method comes at the cost of higher training complexity and computational overhead, the resulting models demonstrate more robust and generalizable performance in delay estimation and LPF calibration tasks. This integration provides a principled mechanism to combine data-driven learning with traffic flow theory, retaining predictive flexibility while reducing the risk of physically inconsistent outcomes, such as negative delays or violations of capacity constraints.

4.3 Limitations of PINNs

Despite their potential, the application of PINNs to traffic flow modeling remains subject to several limitations.

First, computational cost is substantially higher than that of conventional machine learning models, as the evaluation of additional physics-based residuals (e.g., conservation and queue balance equations) introduces considerable overhead during training, particularly for large-scale networks. Second, the design of physics terms requires strong prior knowledge from the domain expert, and alternative formulations of the same constraint may lead to divergent performance outcomes. Third, PINNs exhibit a high degree of sensitivity to hyperparameters, such as the relative weighting between data-fitting and physics-based losses, learning rate schedules, and collocation point sampling strategies; improper configurations may result in unstable convergence or suboptimal solutions. Finally, PINNs often suffer from performance degradation under sparse or noisy data, which are prevalent in real-world sensing environments. These issues collectively highlight the need for robust training algorithms, adaptive loss-weighting mechanisms, and scalable implementations to make PINNs a practical tool for transportation applications.

4.4 A Unified AI-Driven LPF Framework

Rather than treating LPFs as fixed algebraic forms, future research will emphasize hybrid architectures that fuse physics-based traffic flow models with AI-driven learning. This fusion aims to produce formulations that are both interpretable and broadly applicable across network contexts. Building on these insights, we modified the framework proposed by Huo et al. (2022) and propose an original unified end-to-end AI-driven LPF pipeline, as shown in Fig. 11. The pipeline proceeds as follows:

(1) **Data preparation:** split raw observations into 80 % training and 20 % test sets; apply normalization, scaling, and feature engineering; map categorical or high-dimensional inputs through embedding layers.

(2) **Feature extraction**: derive core variables such as observed mean travel time, free-flow travel time, and nominal link capacity.

(3) **Interpretability:** Interpretability is strengthened through SHAP analysis and attention weight visualization to identify the relative importance of link attributes, exogenous factors (e.g., weather, incidents), and network topology. This allows the framework to provide not only accurate predictions but also actionable insights for planners.

(4) **Capacity estimation**: employ graph neural networks, Transformers, and attention mechanisms to ingest link attributes plus exogenous factors (e.g., weather, incidents) and predict a dynamic link capacity profile.

(5) **Iterative loop:** Rather than treating capacity estimation and travel time prediction as isolated tasks, the pipeline adopts an iterative, physically-informed modeling process. Predicted delays are fed back to update flow distributions, which in turn refine capacity estimation. This feedback loop operationalizes the dynamic coupling between demand, supply, and delay in real-world networks.

(6) **Model training and physical consistency**: To ensure theoretical validity, we incorporate physics-informed regularizers into the training objective. Specifically, conservation-based residuals enforce that inflows equal outflows plus changes in queue length, monotonicity constraints guarantee that volume-delay curves remain non-decreasing, and non-negativity constraints prevent unrealistic negative delays or capacities. These conditions are embedded into the loss function alongside conventional error metrics, thereby preserving physical consistency across predictions.

(7) **Travel-time prediction**: Combine predicted capacity, traffic volume, and free-flow time into a generalized LPF to compute time-dependent travel times.

(8) **Evaluation and robustness testing**: Beyond standard accuracy metrics (MAE, RMSE, MAPE), the framework incorporates uncertainty quantification through prediction intervals and conformal prediction methods. Robustness is further evaluated under distribution shifts, including cross-temporal and cross-network validation, to ensure generalizability across diverse traffic regimes.

This unified AI-driven LPF framework is specifically tailored to the unique challenges of LPF modeling. At its core, it enforces physical consistency, embeds capacity estimation and delay prediction into an iterative loop, and integrates uncertainty quantification with interpretability modules. As such, the framework moves beyond black-box learning to provide a physically grounded, reproducible, and extensible methodology. By embedding machine learning and artificial intelligence into each stage of the process, from parameter calibration to travel time estimation, it enables LPFs to adapt to evolving traffic patterns, enhances predictive accuracy, and supports more effective congestion management and strategic transportation planning.

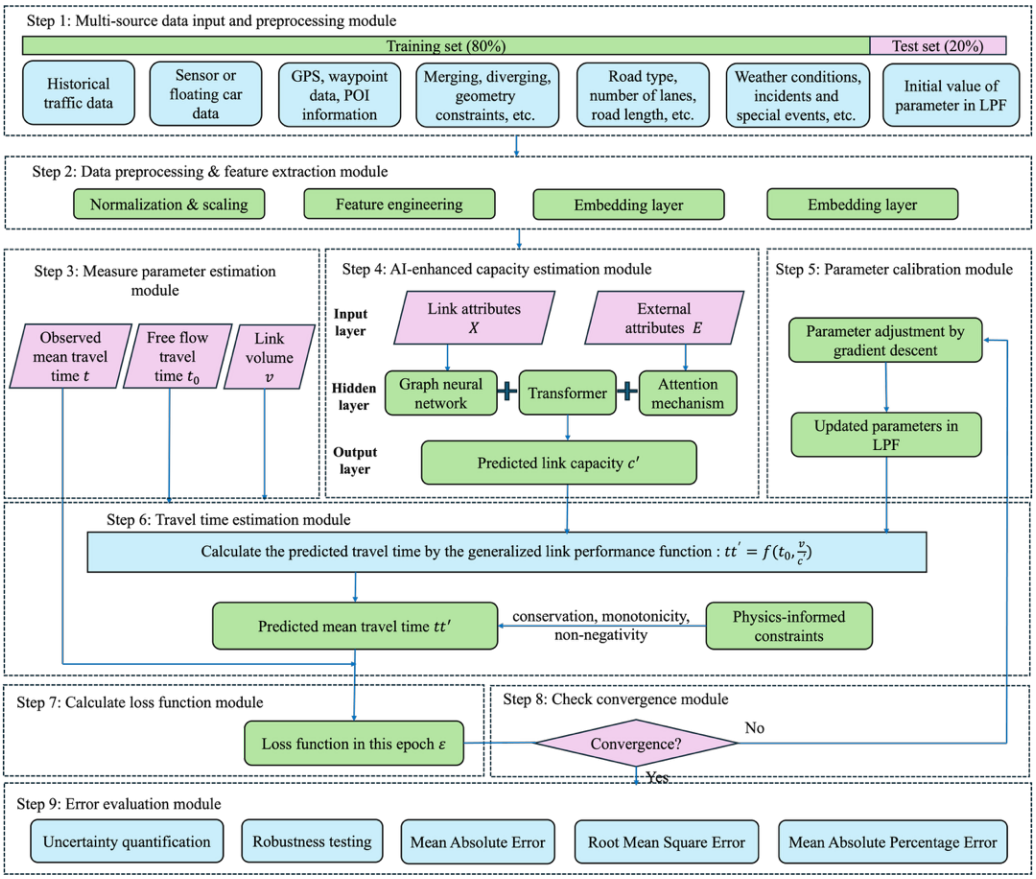


Figure 11 Physics-informed neural network and LPF-based travel time prediction framework.

5 Case Study

5.1 Data and Study Site

In this section, we select traffic data from the Los Angeles I-405 corridor, which exhibits both recurring and non-recurring congestion patterns across freeway segments, providing a solid basis for evaluating the generalizability of the model. The dataset, obtained from the PeMS system, contains flow, speed, and occupancy measurements collected at 5-minute intervals via loop detectors from April 1, 2017, to July 31, 2017, covering a total of 81 weekdays and excluding holidays and weekends. The spatial extent of the study corridor is shown in Fig. 12(a), while the corresponding bottleneck speed heatmap is presented in Fig. 12(b).

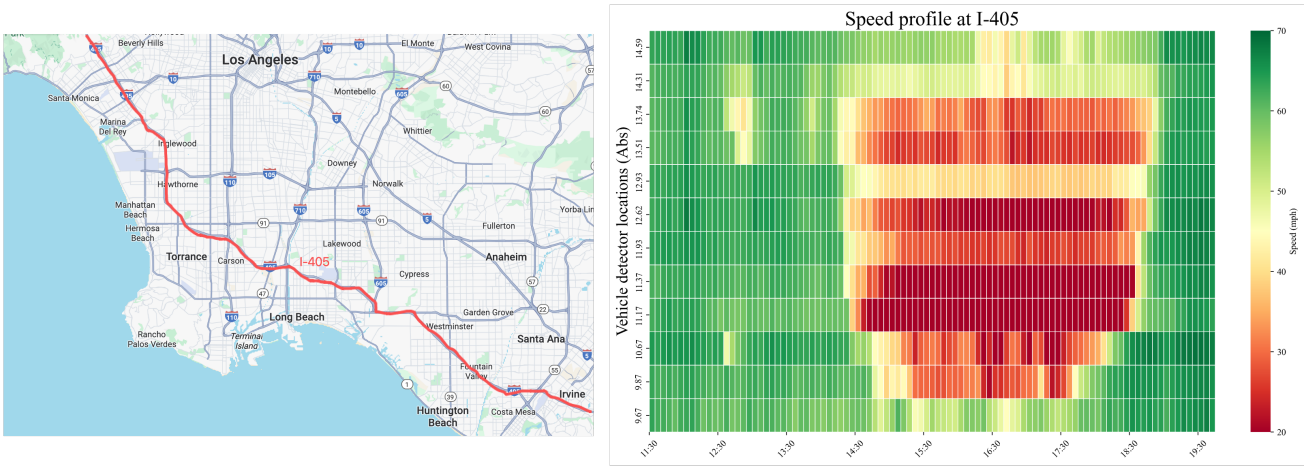


Figure 12 Areas of Los Angeles I-405 corridors: (a) corridor location; (b) speed heatmap of I-405 corridor.

5.2 Performance metrics

To evaluate the performance of various traffic flow prediction models, it is crucial to define clear criteria that both measure prediction accuracy and guide methodological improvements. The evaluation process primarily consists of comparing the predicted traffic states with the observed real-time conditions. In this study, we focus on three widely used error metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The MAE quantifies the average magnitude of errors by computing the mean of the absolute differences between observed and predicted values. Formally, it can be expressed as:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{11}$$

The RMSE is defined as the square root of the average of the squared differences between the predicted values produced by the model and the corresponding observed values. Mathematically, it can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N}} \tag{12}$$

The MAPE is frequently used in traffic forecasting as it helps reduce the disproportionate influence of instances with very low traffic flow on overall error measurement. Unlike absolute error metrics, MAPE incorporates both the magnitude of the error and its relative size by expressing the discrepancy between predicted and observed values as a proportion of the observed value. Formally, it is defined as:

$$MAPE = \frac{100\%}{N}\sum_{n=1}^{N}\left|\frac{y_n - \hat{y}_n}{y_n}\right| \tag{13}$$

where y_n is the measured traffic flow value for observation m, $\hat{y}_n$ is the estimated traffic flow value for observation n, N is the total number of flow counts.

5.3 Representative Models for Comparison

This case study is based on PeMS data from the I-405 corridor and compares different types of VDF models uniformly on the travel time dimension. The outputs of all models are converted into travel time or delay through speed-time or flow-density relationships, and then compared against the observed travel time series.

The study considers six representative models as illustrated in Table 9, covering empirical, theory-driven, data-driven, and physics-informed approaches. The BPR function is simple and widely applied, but its accuracy is limited and it lacks physical consistency. FD-based models (e.g., the Greenshields model) offer better interpretability and remain consistent with fundamental traffic flow principles, while fluid-queue models can further capture oversaturation and congestion duration with a strong physical foundation.

(1) The classical BPR function is formulated as Eq. (5). We could modify and use the following time-dependent flow-based model to estimate travel time with the observed data.

$$tt(t) = t_f \cdot [1 + \alpha(\frac{q(t)}{C})^{\beta}] \tag{14}$$

We also test the following time-dependent density-based modified BPR function (Kucharski and Drabicki , 2017), as follows:

$$tt(t) = t_f \cdot [1 + \alpha(\frac{k(t)}{k_c})^{\beta}] \tag{15}$$

where $q(t)$ is the time-series flow, $k(t)$ is the time-series density and k_c is critical density.

(2) The mathematic relationships in the Greenshield's FD model are the speed-density relationship, it could derive the time-dependent travel time as follows:

$$tt(t) = \frac{L}{v(t)} = \frac{L}{v_f \cdot (1 - \frac{k(t)}{k_j})} = \frac{t_f}{1 - \frac{k(t)}{k_j}} \tag{16}$$

where L is the link length, $v(t)$ is the time-series speed and t_f is free-flow travel time.

(3) The time-dependent speed of the fluid queue-based volume-delay function could be represented as follows:

$$v(t) = \frac{L}{\frac{L}{v_{co}} + w(t)} \tag{17}$$

If we use the quadratic PAQ model, then we have the time-dependent waiting time:

$$w(t) = \frac{\gamma}{3\mu} \cdot (t - t_0)^2 \cdot (t_3 - t) \tag{18}$$

The travel time can be derived as follows:

$$tt(t) = \frac{L}{v_{co}} + w(t) = \frac{L}{v_{co}} + \frac{\gamma}{3\mu} \cdot (t - t_0)^2 \cdot (t_3 - t) \tag{19}$$

The travel time of the fluid queue-based volume-delay function could be represented by the sum of the free-flow travel time and the waiting time, as follows:

(4) In the LSTM and Transformer method, we could convert the time-series observed speed into a time-series of travel time, which then serves as the input for the data-driven model.

$$tt(t) = \sum_i \frac{L_i}{v_i(t)} \tag{20}$$

where L_i is length of link i, $v_i(t)$ is the observed speed on link i at time t.

(5) Finally, to unify all models under the "travel time/delay" dimension, we adopt a time-dependent PINN. In addition to conventional data features (flow, speed, density, and their historical sequences), this method integrates traffic flow theory-based time-varying variables such as free-flow travel time, segment flow, estimated capacity, congestion duration, and queue length as additional inputs or feature-engineering variables into the neural network.

A key feature of the proposed PINN is the explicit incorporation of physical constraints into the learning process. Specifically, the travel time is formulated as $tt(t) \geq t_f$, ensuring that the predicted travel time at any time t never falls below the physically feasible lower bound. The additional delay component $\varphi_t(D, P, Q(t), C, X, E)$ is constrained to be nonnegative and to reflect dynamic traffic-flow characteristic, such as the time-dependent queue length. While the neural network function $\varphi_t(\cdot)$ captures complex nonlinear dependencies among demand, congestion duration, queue dynamics, capacity, and exogenous factors, the overall structure remains consistent with queueing theory, thereby improving both predictive accuracy and physical interpretability.

$$tt = t_f \cdot [1 + \varphi_t(D, P, Q(t), C, X, E)] \tag{21}$$

where D is demand, P is congestion duration, $Q(t)$ is queue length, C is capacity, X is exogenous factors, E is environmental or error terms, D, P, $Q(t)$, C are positive.

Data-driven methods predict travel time by inputting observed temporal features such as flow, speed, density, time of day, and day of week, and then transforming point-based speed/flow data into link-level travel time through spatial aggregation. LSTM is well-suited for capturing short-term temporal dependencies and can effectively predict dynamic fluctuations in travel time, whereas Transformer excels at modeling long sequences, accounting for spatiotemporal dependencies across multiple road segments, and is suitable for large-scale prediction. These methods are capable of capturing complex nonlinear relationships and handling high-dimensional inputs (e.g., speed, flow, weather). However, their limitation lies in the lack of physical constraints, which may lead to unrealistic results (such as negative delays or excessively high speeds). In contrast, the core idea of the PINN method is to embed traffic flow physical principles (e.g., fundamental diagram relations, queueing delay constraints) directly into the loss function of the neural network, enabling the model to learn the mapping between speed/flow and travel time while ensuring physical consistency in predictions. This approach combines the high predictive accuracy of data-driven models with the interpretability and rationality of theoretical models, thereby achieving more robust travel time prediction under limited data and complex traffic conditions.

Table 9 Performance comparison of empirical, theory-driven, data-driven, and physics-informed models.

Model category	Representative methods	Accuracy	Physical consistency	Interpretability	Remarks
Empirical VDFs	BPR	Low (fit simple trends, inaccurate under oversaturation)	Low (rely only on V/C, lack conservation and queue dynamics)	Medium (parameters α, β partially interpretable)	Simple and widely applied in planning practice
Theory-driven	FD-based (e.g., Greenshields model)	Medium (capture free-flow to critical flow)	High (consistent with traffic flow theory)	High (parameters v_f,k_j,w have clear physical meaning)	Effective for capacity estimation and regime distinction

	Fluid-queue based (e.g., Newell model)	High (accurate for oversaturated travel time curves)	Very high (explicitly models queues, congestion duration P and μ)	High (parameters μ, P, D are directly interpretable)	Emphasizes congestion duration as the main driver of delay
Data-driven	LSTM	High (strong short-term prediction)	Medium (no explicit physical constraints)	Medium (black-box with limited interpretability)	Sensitive to data distribution shifts
	Transformer	Very high (long-sequence modeling is strong)	Low-medium (may violate physical rules without constraints)	Medium (attention weights provide partial interpretability)	Data-hungry, computationally costly
Physics-informed	PINNs	High (balance between prediction and constraints)	Very high (conservation, capacity and non-negativity embedded)	High (physical terms interpretable, flexible prediction)	High training cost, prior knowledge required, hyperparameter-sensitive

5.4 Results

The dataset was divided chronologically into a training and a testing set. For the BPR volume-based, density-based, and fluid-queue-based models, the unknown parameters α and β were calibrated on a daily basis using observed data from the training period. The daily parameter estimates were then aggregated to obtain stable link-specific parameters that account for observational variability. These aggregated parameters were applied to the testing set to generate travel times, which were compared against observations using MAE, RMSE, and MAPE.

For the FD-based model, the free-flow speed v_f and jam density k_j were estimated daily from density-speed data in the training period. The estimated parameters were aggregated to yields stable values of v_f and k_j, which were applied to the testing set to compute travel time estimates and evaluate prediction errors.

For data-driven sequence models such as LSTM and Transformer, the training samples are generated using a sliding-window approach, where a sequence of past observations with length seq_len is mapped to the next time step. Normalization is performed based solely on training set statistics. After training, rolling prediction is carried out on the testing set to generate travel time trajectories, which are compared with ground truth. For the PINN framework, model training is conducted on the training set by minimizing a composite loss that integrates data fidelity, physics-based residuals, and boundary/initial conditions. Out-of-sample evaluation on the testing set is performed using the selected error metrics. Table 10 summarizes the predictive performance of different models in terms of the MAE, RMSE, MAPE in the test dataset.

Table 10 Comparative prediction performance of different models.

Models	MAE	RMSE	MAPE (%)
Time-dependent, volume-based BPR function using Eq. (14)	0.124	0.239	21.076
Time-dependent, density-based BPR function using Eq. (15)	0.024	0.050	5.154
Fluid-queue based using Eq. (19)	0.069	0.123	15.213
Greenshield's FD-based using Eq. (16)	0.464	3.246	40.241
LSTM neural network	0.021	0.047	4.090
Transformer	0.012	0.024	2.462
PINN (proposed) using Eq. (21)	**0.002**	**0.004**	**0.312**

The results in Table 10 are derived from the 16 test-day datasets, and the performance metrics represent the average values across all test days. Table 10 compares the predictive performance of different models using three metrics: MAE, RMSE, and MAPE. The classical BPR volume-based model, with an MAE of 0.124, applies a single V/C ratio to produce a constant speed across the entire time interval, leading to poor performance when compared with time-varying observed travel times.

The fluid-queue model, with an MAE of 0.069, demonstrates the theoretical advantage of embedding congestion duration and demand-to-capacity ratios, generating structurally consistent and time-dependent travel times through its analytical framework. Although its numerical accuracy appears lower, this model provides physically meaningful insights into congestion dynamics that purely empirical methods cannot capture.

The density-based BPR variant, with an MAE of 0.024, and the Greenshield's FD-based model, with an MAE of 0.464, remain time-series generators that map demand to speed based on instantaneous observations.

Deep learning models such as the LSTM neural network, with an MAE of 0.021, and the Transformer, with an MAE of 0.012, achieve higher accuracy by learning temporal dependencies and short-term variations, though their generalizability across networks and demand conditions remains a concern.

The proposed PINN framework achieves the best performance, with an MAE of 0.002 and RMSE of 0.004, highlighting the benefit of integrating physical constraints with data-driven learning. Further sensitivity analyses are needed to evaluate the reasonableness of its demand-to-capacity and queue-to-speed relationships. Given the complex interdependence of demand-to-capacity ratios, queue dynamics, and external factors, future research should explore more customized and time-varying functional forms and improve the incorporation of external variables and short-term sequences to better represent demand fluctuations.

For the single test day on July 25, the model comparison results in Fig.13 (a) shows a consistent pattern with the overall findings. The proposed PINN model again achieves the best performance. The queue-based model performs moderately well, followed by the density-based and Transformer models, while the FD-based model records the highest errors, reflecting its limited ability to capture the nonlinear dynamics of congestion evolution. The LSTM model maintains relatively stable accuracy but still falls behind the PINN approach. These results confirm that the physics-informed learning mechanism of the PINN model enhances its ability to represent the temporal evolution of travel times even under single-day fluctuations.

Consequently, Fig. 13(b) does not include the classical BPR volume-based results, as they would appear as a flat horizontal line that cannot capture the dynamic variations shown in the time-dependent travel time profiles. Fig. 13(b) illustrates these paradigmatic differences. The fluid queue-based method produces smooth, physically consistent travel time profiles critical for understanding system behavior and policy evaluation. Data-driven methods excel at reproducing observed patterns but could struggle with scenario analysis where demand patterns differ from historical observations. The PINN framework offers a promising direction, though further development is needed to fully realize its potential for unified modeling.

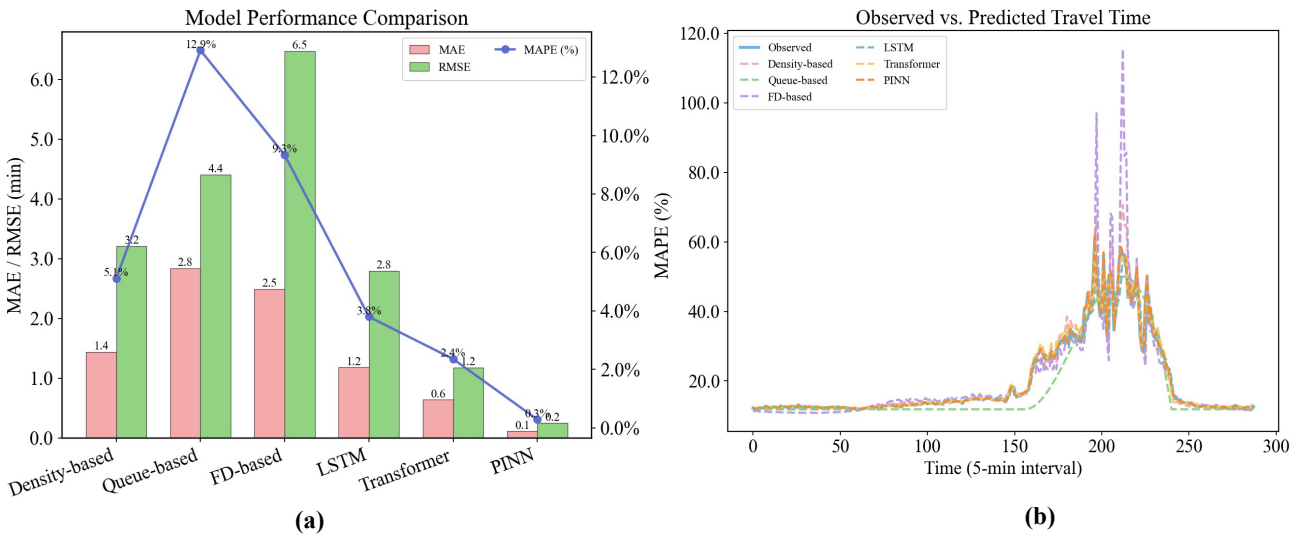


Figure 13 Comparison of model performance and travel time predictions: (a) models performance comparison (b) comparison between observed and predicted travel times in the time series.

6 Future Research

The role of VDFs in transportation research and practice is critical, as highlighted by this review. This paper has offered a comprehensive analysis of the theoretical fundamentals, practical deployment, and emerging applications of VDFs. While these findings highlight significant advancements, they also reveal persistent gaps that limit the full potential of VDFs in addressing evolving transportation challenges. To provide actionable guidance, we structure the future research agenda into three layers of priority: (A) short-term priorities in Calibration and congestion identification with multisource data, to improve operational reliability and reproducibility, (B) Medium-term priorities in incorporating reliability and stochasticity into VDFs, to ensure robustness across varying traffic regimes, and (C) Long-term priorities in leveraging AI-driven and machine learning-enhanced VDFs, to enable adaptive, scalable, and real-time applications in intelligent transportation systems.

Each layer is detailed below, with specific research questions and methodological pathways.

A. VDFs calibration with congestion identification and multisource data: The development of reliable VDFs for operational planning demands a deeper understanding of bottleneck congestion dynamics, queue formation, and the variability of traffic flow. We pose the following future research questions (FRQ) to be explored in future research.

FRQ 1: How should current VDFs be enhanced to incorporate queue spillback effects and downstream constraints?

Pathway: integrate cumulative arrival–departure curves and shockwave-based models into LPF calibration, embedding spillback as boundary conditions.

FRQ 2: In what ways can the breakdown probability be formulated as a function of the D/c ratio and inflow demand curvature to support reliability-aware VDFs?

Pathway: use stochastic capacity formulations and survival analysis techniques to link D/C ratios with breakdown probability distributions.

FRQ 3: Can advanced data-driven methods, such as deep learning or Bayesian inference, improve real-time VDF calibration while quantifying uncertainty from heterogeneous data inputs?

Pathway: employ Bayesian neural networks and variational inference to capture parameter uncertainty, combined with online learning for real-time updates.

B. VDFs with reliability considerations: Building upon the calibration foundation, it is essential to integrate reliability considerations into VDFs to ensure robustness across varying traffic conditions and demand scenarios. VDFs that reflect temporal and spatial variability in travel time can support planning tools such as congestion pricing, adaptive signal control, and active demand management. Additionally, incorporating reliability into VDFs enables system-level performance evaluation by linking link-level variability with path- or OD-level indicators.

We pose the following FRQ to be explored in future research:

FRQ 4: How can traffic flow variability and reliability metrics (e.g., travel time standard deviation, trip rate distribution) be consistently integrated into VDF formulations for both recurring and non-recurring conditions?

Pathway: extend VDFs with stochastic terms calibrated from empirical reliability metrics, and apply Monte Carlo simulation to capture non-recurring disruptions.

FRQ 5: How can reliability-aware VDFs be scaled up to OD and route levels to assess system-wide performance, and what are the implications for equilibrium modeling under stochastic demand?

Pathway: Embed uncertainty assessment methods into the calibration process and integrate online updating to enable rapid response to dynamic traffic conditions.

FRQ 6: How do within-day and day-to-day demand fluctuations impact travel time reliability, and how can VDFs be adapted to account for these temporal dynamics?

Pathway: employ time-varying stochastic processes and day-to-day dynamic traffic assignment frameworks to explicitly capture demand variability.

C. Machine learning-enhanced VDFs modeling for intelligent transportation systems: The emergence of high-resolution traffic data and ML tools presents new opportunities for dynamic and scalable VDF modeling. However, challenges remain in data heterogeneity, anomaly handling, computational scalability, and real-time system integration. Efficiently leveraging ML, deep learning, and connected vehicle data can significantly enhance VDF accuracy and responsiveness, especially under dynamic traffic conditions.

We pose the following FRQ to be explored in future research:

FRQ 7: How can data fusion algorithms be developed to effectively integrate heterogeneous data sources (e.g., probe data, sensors, CAVs) for VDF training and calibration?

Pathway: develop graph-based data fusion and tensor decomposition methods to align temporal and spatial dimensions across diverse sources.

FRQ 8: How can graph-based learning methods (e.g., GNNs) and distributed computing architectures be employed to scale ML-based VDF models to large urban networks while maintaining accuracy and computational efficiency?

Pathway: adopt distributed GNN frameworks and parallel computing platforms to achieve scalability in metropolitan-scale networks.

FRQ 9: How can VDFs be dynamically adapted to reflect the real-time impact of emerging technologies such as CAVs, and how can reinforcement learning be leveraged to integrate VDF modeling with real-time traffic control decisions?

Pathway: implement reinforcement learning agents that adjust VDF parameters online, informed by streaming CAV and IoT data.

As transportation systems continue to evolve in the digital era, VDF research must move beyond generic problem statements toward prioritized and technically grounded solutions. Our proposed roadmap emphasizes (i) robust calibration using multisource data, (ii) integration of reliability metrics and stochasticity for scalable planning, and (iii) AI-driven adaptivity for real-time applications. Collectively, these pathways define a forward-looking, physics-informed, and data-enhanced research agenda that is both actionable and innovative.

Appendices

Appendix A: Notation List

Table A1 Summarizes the notations used in this paper.

Symbols	Definitions
v	Speed
$\bar{v}$	Mean speed
v_f	Free-flow speed
v_{co}	Cut-off speed
v_c	Critical speed
k	Density
k_c	Critical density
k_j	Jam density
w	Backward wave speed
V	Hourly volume
C	Roadway hourly capacity/ cycle time
C_p	Lower critical capacity (onset of congestion threshold), below this, travel time grows only slightly with volume,
C_u	Upper capacity threshold, beyond this, traffic enters the oversaturated region with severe queuing
μ	Constant discharge rate/ link's service rate
D	Demand volume (during congestion)
V/C	Volume-to-capacity ratio
D/C	Demand-to-capacity ratio
V^H	Human-driven hourly volume
V^A	Autonomous vehicle hourly volume
P	Congestion duration
T	Analysis flow period
T_f	Flow period (the analysis period, e.g., 15 minutes, 1 hour),
tt	Average travel time per unit distance
t_f	Free-flow travel time
t_p	Base (or free-flow) travel time, corresponding to low-volume conditions
t_u	Travel time at the end of the capacity interval
t_0	Start time of congestion period
t_1	The time of the highest inflow rate
t_2	The time of equal inflow and outflow rates during the congestion period
t_3	End time of congestion period
t_s	Start time of analysis period
t_e	End time of analysis period
$q(t)$	Time-dependent flow rate
$k(t)$	Time-dependent density
$Q(t)$	Time-dependent queue length
$w(t)$	Time-dependent waiting time
$tt(t)$	Time-dependent travel time
$\lambda(t)$	Time-dependent inflow rate
$\mu(t)$	Time-dependent outflow rate

$A(t)$	Cumulative arrival counts
$D(t)$	Cumulative departure counts
$\lambda(t_1)$	The highest inflow rates during the congestion period
$\bar{w}$	Average delay during the congestion period
g	Effective green time
g/C	As the proportion of the cycle which is effectively green for the phase under consideration
x	The degree of saturation, traffic intensity
ρ	Utilization rate as a queueing system
γ	Inflow demand curvature parameter
α, β, J, J_A	Parameters in VDFs
$g(m)$	Conversion factor in polynomial arrival queue models
d	Minimum gap
τ	Reaction time
f_p	Constant in elasticity function for mapping congestion duration to the magnitude of speed reduction
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 charges, i.e., oversaturation-to-duration elasticity
s	Elasticity coefficient of speed reduction magnitude in response to congestion duration changes, i.e., duration-to-speed reduction elasticity

Appendix B: Summary of the Key VDFs and Related Publications

Table B1 The most representative leading VDFs with the publication papers.

Author/s/Year	Paper title	Functions	Remarks
Webster (1958)	Traffic signal settings	$d = \frac{C(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2V(1-x)} - 0.65(\frac{C}{V^2})^{\frac{1}{3}} x^{(2+5\lambda)}$	Queueing model-based, $\lambda = g/C$ ratio
CATS (1960)	Data projections	$tt = t_f \cdot 2^{(\frac{V}{C})}$	
Smock (1962)	An iterative assignment approach to capacity restraint on arterial networks	$tt = t_f \cdot e^{(\frac{V}{C}-1)}$	
BPR (1964)	Traffic Assignment Manual	$tt = t_f \left[1 + \alpha \left(\frac{V}{C}\right)^{\beta}\right]$	
Davidson (1966)	A flow-travel time relationship for use in transportation planning	$tt = t_f \left[1 + j \cdot \frac{V}{C-V}\right]$	Single queueing with random arrival rates and exponentially distributed service rates
Davidson (1978)	The theoretical basis of a flow-travel time relationship for use in transportation planning	$tt = t_f \left[1 + J \cdot \frac{V}{C-V}\right], J = \frac{k+1}{2k}$	Queue system with random arrivals and an Erlang service distribution
Newell (1982)	Applications of queueing theory	$tt = t_f \left[1 + \frac{\gamma}{36\mu \cdot t_0} \cdot \left(\frac{D}{\mu}\right)^3\right]$	Fluid-based queue with quadratic arrival rates and fixed

			inflow curvature parameter γ
Small (1983)	The incidence of congestion tolls on urban highways	$tt = \begin{cases} t_f, if\ V \le C \\ t_f + \frac{1}{2}P \cdot \left(\frac{V}{C} - 1\right), if\ V > C \end{cases}$	Peak hour-based
Spiess (1990)	Conical volume-delay functions	$tt = t_f \cdot [2 + \sqrt{\alpha^2(1 - \frac{V}{C})^2 + \beta^2} - \alpha\left(1 - \frac{V}{C}\right) - \beta]$	Well-defined VDF properties proposed
Akçelik (1991)	Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function	$tt = t_f + 0.25T \cdot [\left(\frac{V}{C} - 1\right) + \sqrt{(\frac{V}{C} - 1)^2 + \frac{8JV}{TC^2}}]$	Time-dependent form, derived by using the coordinate transformation technique
Huntsinger and Rouphail (2011)	Bottleneck and queuing analysis: calibrating volume-delay functions of travel demand models	$tt = \begin{cases} t_f \left[1 + \alpha\left(\frac{V}{C}\right)^{\beta}\right], if\ V \le C \\ t_f \left[1 + \alpha\left(\frac{D}{C}\right)^{\beta}\right], if\ V > C \end{cases}$ $D = demand\ at\ capacity + queue$	Queued demand-based
Moses et al. (2013)	Development of speed models for improving travel forecasting and highway performance evaluation	$tt = \begin{cases} t_f[1 + \alpha(C)^{\beta}], if\ V \le C \\ t_f \left[1 + \alpha\left(\frac{D}{C}\right)^{\beta}\right], if\ V > C \end{cases}$ $D = C + (C - V)$	
Bliemer et al. (2014)	Quasi-dynamic traffic assignment with residual point queues incorporating a first order node model	$tt = \sum_{a \epsilon A} \frac{L_a}{v_f^a} \cdot \delta_{ap} + \tau_p^{queue}, \forall a \epsilon A$	Focus on incorporating hard capacity constraints and residual queues (including point queues) to effectively simulate congestion.
Kucharski and Drabicki (2017)	Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows	$tt = t_f \left[1 + \alpha\left(\frac{k}{k_c}\right)^{\beta}\right]$	Density as a proxy of V/C
Lazar et al. (2020)	Routing for traffic networks with mixed autonomy	$tt = t_f \left[1 + \alpha\left(\frac{V^H + V^A}{C}\right)^{\beta}\right]$	Mixed flow-based
Cheng et al. (2022)	Estimating key traffic state parameters through	$tt = t_f \left[1 + \frac{\gamma \cdot g(m)}{\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^4\right]$	Fluid-based queue with cubic arrival rates and fixed

	parsimonious spatial queue models		inflow curvature parameter γ
Zhou et al. (2022)	A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio	$tt = t_f\left[1 + \alpha\left(\frac{D}{C}\right)^{\beta}\right]$ $\alpha = \theta f_p \cdot {f_d}^s$ $\beta = ns$ $\theta = \bar{w}/w_{t_2}$	Fluid based queue with a family of inflow curvature parameter values, derived from two elasticity forms and D/C ratio.

CRediT Authorship Contribution Statement

Yuyan Pan: Writing– review & editing, Writing– original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Xianbiao Hu:** Writing– review & editing, Investigation, Supervision, Methodology. **George List:** Writing– original draft, Writing–review & editing, Methodology, Conceptualization. **Xuesong (Simon) Zhou:** Writing–review & editing, Writing– original draft, Supervision, Methodology, Investigation, Conceptualization.

Replication and Data Sharing

The dataset used in this research is partially available upon request by emailing the corresponding author or can be download from https://github.com/panyuyan/VDF-review-code.

Declaration of Competing Interest

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

Acknowledgments

The last author would like to extend sincere thanks to Dr. David Levinson at the University of Sydney for many insightful comments. The last author gratefully acknowledges the guidance from Prof. David Boyce, the feedback from Dr. Baloka Belezamo at Arizona Department of Transportation, Dr. Livshits, Dr. Arup Dutta, Dr. Wang Zhang, Haidong Zhu from the Maricopa Association of Governments (MAG).

References

Akcelik, R., 1996. Relating flow, density, speed and travel time models for uninterrupted and interrupted traffic. Traffic Eng. Control 37(9), 511–516.

Akçelik, R., 1991. Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. Presented at the Australian Road Research.

Akcelik, R., 1988. The Highway Capacity Manual delay formula for signalized intersections. ITE J. 58(3), 23–27.

Akcelik, R., 1980. Time-dependent expressions for delay, stop rate and queue length at traffic signals.

Akçelik, R., Rouphail, N.M., 1993. Estimation of delays at traffic signals for variable demand conditions. Transp. Res. Part B Methodol. 27(2), 109–131.

Anupriya, Graham, D.J., Bansal, P., Hörcher, D., Anderson, R., 2023. Optimal congestion control strategies for near-capacity urban metros: Informing intervention via fundamental diagrams. Phys. Stat. Mech. Its Appl. 609, 128390.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. Transp. Res. Part C Emerg. Technol. 64, 101–116.

Bagherian, M., Hickman, M., Tavassoli, A., 2017. Considering the impact of precipitation on the accuracy of delay-function parameters, in: Proceedings from the 39th Australasian Transport Research Forum.

Bahrami, S., Roorda, M.J., 2020. Optimal traffic management policies for mixed human and automated traffic flows. Transp. Res. Part A Policy Pract. 135, 130–143.

Balmer, M., Meister, K., Rieser, M., Nagel, K., Axhausen, K.W., 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. Arbeitsberichte Verk.-Raumplan. 37.

Banks, J.H., 1991. Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering. Transp. Res. Rec. 1320.

Barceló, J., Casas, J., 2005. Dynamic network simulation with AIMSUN. In: Simulation Approaches in Transportation Analysis: Recent Advances and Challenges, pp. 57–98. Springer, Boston, MA.

Bayram, V., Yaman, H., 2018. Shelter location and evacuation route assignment under uncertainty: A Benders decomposition approach. Transp. Sci. 52(2), 416–436.

Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. Studies in the Economics of Transportation. Yale University Press.

Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D., 2011. SUMO – Simulation of Urban Mobility. Presented at The Third International Conference on Advances in System Simulation.

Belezamo, B., 2020. Under congested conditions: A Phoenix study (Doctoral dissertation). Arizona State University.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., Mishalani, R., 1998. DynaMIT: A simulation-based system for traffic prediction. Presented at the DACCORD Short-Term Forecasting Workshop, Delft, The Netherlands.

Benedek, C.M., Rilett, L.R., 1998. Equitable traffic assignment with environmental cost functions. J. Transp. Eng. 124, 16–22.

Berman, O., Larson, R.C., Parkan, C., 1987. The stochastic queue p-median problem. Transp. Sci. 21, 207–216.

Bertsekas, D.P., Gallager, R.G., 2021. Data networks, in: Data Networks. Athena Scientific.

Bliemer, M.C.J., Raadsen, M.P.H., 2020. Static traffic assignment with residual queues and spillback. Transp. Res. Part B Methodol. 132, 303–319.

Bliemer, M.C.J., Raadsen, M.P.H., Brederode, L.J.N., Bell, M.G.H., Wismans, L.J.J., Smith, M.J., 2017. Genetics of traffic assignment models for strategic transport planning. Transp. Rev. 37, 56–78.

Bliemer, M.C.J., Raadsen, M.P.H., Smits, E.-S., Zhou, B., Bell, M.G.H., 2014. Quasi-dynamic traffic assignment with residual point queues incorporating a first order node model. Transp. Res. Part B Methodol. 68, 363–384.

Boyce, D.E., Janson, B.N., Eash, R.W., 1981. The effect on equilibrium trip assignment of different link congestion functions. Transp. Res. Part A Gen. 15, 223–232.

Boyles, S., Ukkusuri, S.V., Waller, S.T., Kockelman, K.M., 2008. A comparison of static and dynamic traffic assignment under tolls in the Dallas-Fort Worth region. Presented at the Transportation Research Board Conference Proceedings, pp. 114–117.

BPR, 1964. Traffic assignment manual. Planning Division, U.S. Department of Commerce, Washington, DC.

Branston, D., 1976. Link capacity functions: A review. Transp. Res. 10, 223–236.

Brederode, L., Pel, A., Wismans, L., de Romph, E., Hoogendoorn, S., 2019. Static traffic assignment with queuing: Model properties and applications. Transp. Sci. 15, 179–214.

Caliper, 2009. TransModeler traffic simulation software – version 2.5 user's guide.

Campbell, E.W., 1968. The Transportation System: An Evaluation of Alternative Land Use and Transportation Systems in the Chicago Area (No. 234).

Carey, M., 2004. Link travel times I: Desirable properties. Netw. Spat. Econ. 4, 257–268.

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. Transp. Res. Part B Methodol. 65, 90–104.

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

CATS, 1960. Data projections. Chicago.

Chang, J.S., Mackett, R.L., 2006. A bi-level model of the relationship between transport and residential location. Transp. Res. Part B Methodol. 40(2), 123–146.

Charnes, A., Cooper, W.W., 1959. Chance-constrained programming. Manag. Sci. 6, 73–79.

Charnes, A., Cooper, W.W., Symonds, G.H., 1958. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. Manag. Sci. 4, 235–263.

Chen, A., Zhou, Z., Chootinan, P., Ryu, S., Yang, C., Wong, S.C., 2011. Transport network design problem under uncertainty: A review and new developments. Transp. Rev. 31, 743–768.

Chen, Q., Li, X., Ouyang, Y., 2011. Joint inventory–location problem under the risk of probabilistic facility disruptions. Transp. Res. Part B Methodol. 45(7), 991–1003.

Cheng, Q., Liu, Z., Guo, J., Wu, X., Pendyala, R., Belezamo, B., Zhou, X. (Simon), 2022. Estimating key traffic state parameters through parsimonious spatial queue models. Transp. Res. Part C Emerg. Technol. 137, 103596.

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

Cheng, J., Li, G., Chen, X., 2019. Developing a travel time estimation method of freeway based on floating car using random forests. J. Adv. Transp. 2019(1), 8582761.

Chiu, Y.-C., Zhou, L., Song, H., 2010. Development and calibration of the anisotropic mesoscopic simulation model for uninterrupted flow facilities. Transp. Res. Part B Methodol. 44, 152–174.

Chu, T., Wang, J., Codecà, L., Li, Z., 2019. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans. Intell. Transp. Syst. 21(3), 1086–1095.

Correa, J.R., Schulz, A.S., Stier-Moses, N.E., 2004. Selfish routing in capacitated networks. Math. Oper. Res. 29, 961–976.

Crainic, T.G., Florian, M., Léal, J.-E., 1990. A model for the strategic planning of national freight transportation by rail. Transp. Sci. 24, 1–24.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp. Res. Part B Methodol. 28, 269–287.

Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transp. Res. Part B Methodol. 42, 771–781.

Davidson, K.B., 1978. The theoretical basis of a flow-travel time relationship for use in transportation planning.

Davidson, K.B., 1966. A flow travel time relationship for use in transportation planning. Presented at the Australian Road Research Board (ARRB) Conference, 3rd, Sydney.

Deshpande, N., Park, H.J., 2025. Physics-informed deep learning with Kalman filter mixture for traffic state prediction. Int. J. Transp. Sci. Technol. 17, 161–174.

Dion, F., Rakha, H., Kang, Y.-S., 2004. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transp. Res. Part B Methodol. 38, 99–122.

Doll, A., Abbasi, M., Zhao, M., Zhou, X.S., 2024. Oversaturated intersections: A real-world assessment of polynomial fluid queue models. Physica A Stat. Mech. Appl. 651, 129864.

Dong, J., Mahmassani, H.S., 2009. Flow breakdown and travel time reliability. Transp. Res. Rec. 2124, 203–212.

Dowling, R., Skabardonis, A., 1993. Improving average travel speeds estimated by planning models. Transp. Res. Rec. 68–74.

Dowling, R.G., Singh, R., Cheng, W.W.K., 1998. The accuracy and performance of improved speed-flow curves. Transp. Res. Rec. 1646(1), 9–17.

Dowling, R., Margiotta, R., Cohen, H., Skabardonis, A., Elias, A., 2011. Methodology to evaluate active transportation and demand management strategies. Procedia Soc. Behav. Sci. 16, 751–761.

Duan, Y., Lv, Y., Wang, F.Y., 2016. Travel time prediction with LSTM neural network. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 1053–1058. IEEE.

Duffin, R.J., 1947. Nonlinear networks. IIa. Bull. Am. Math. Soc. 53(10), 963–971.

Engelson, L., van Amelsfort, D., 2015. The role of volume–delay functions in forecasting and evaluating congestion charging schemes: The Stockholm case. Transp. Plan. Technol. 38, 684–707.

Fambro, D.B., Rouphail, N.M., 1997. Generalized delay model for signalized intersections and arterial streets. Transp. Res. Rec. J. Transp. Res. Board 1572, 112–121.

Fellendorf, M., Vortisch, P., 2010. Microscopic traffic flow simulator VISSIM, in: Barceló, J. (Ed.), Fundamentals of Traffic Simulation, International Series in Operations Research & Management Science. Springer New York, New York, NY, pp. 63–93.

Fisk, C.S., 1991. Link travel time functions for traffic assignment. Transp. Res. Part B Methodol. 25, 103–113.

Fosgerau, M., Small, K.A., 2012. Marginal congestion cost on a dynamic expressway network. J. Transp. Econ. Policy 46, 431–450.

Gayah, V.V., Daganzo, C.F., 2011. Clockwise hysteresis loops in the macroscopic fundamental diagram: An effect of network instability. Transp. Res. Part B Methodol. 45, 643–655.

Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transp. Res. Part B Methodol. 45, 605–617.

Genser, A., Kouvelas, A., 2022. Dynamic optimal congestion pricing in multi-region urban networks by application of a multi-layer neural network. Transp. Res. Part C Emerg. Technol. 134, 103485.

Guerrieri, M., 2024. A theoretical model for evaluating the impact of connected and autonomous vehicles on the operational performance of turbo roundabouts. Int. J. Transp. Sci. Technol. 14, 202–218.

Guo, L., Huang, S., Sadek, A.W., 2013. An evaluation of environmental benefits of time-dependent green routing in the greater Buffalo–Niagara region. J. Intell. Transp. Syst. 17(1), 18–30.

Hadi, M., Zhou, X., Hale, D., 2022. Multiresolution modeling for traffic analysis: Guidebook (No. FHWA-HRT-22-055). Federal Highway Administration.

Hall, F.L., Agyemang-Duah, K., 1991. Freeway capacity drop and the definition of capacity. Transp. Res. Rec. 1320.

Hajbabaie, A., Tajalli, M., Bardaka, E., 2024. Effects of connectivity and automation on saturation headway and capacity at signalized intersections. Transp. Res. Rec. 2678(5), 31–46.

Han, Y., Wang, M., Li, L., Roncoli, C., Gao, J., Liu, P., 2022. A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering. Transp. Res. Part C Emerg. Technol. 137, 103584.

Harks, T., Klimm, M., 2012. On the existence of pure Nash equilibria in weighted congestion games. Math. Oper. Res. 37(3), 419–436.

Highway Capacity Manual, 2000. Highway capacity manual. Washington, DC.

Hörcher, D., Graham, D.J., Anderson, R.J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. Transp. Res. Part B Methodol. 95, 105–125.

Horowitz, A., Creasey, T., Pendyala, R., Chen, M., National Cooperative Highway Research Program, Transportation Research Board, National Academies of Sciences, Engineering, and Medicine, 2014. Analytical travel forecasting approaches for project-level planning and design. Transportation Research Board, Washington, D.C.

Horowitz, A.J., 1991. Delay/Volume relations for travel forecasting based upon the 1985 Highway Capacity Manual (FHWA Report).

Huang, Y., Ye, Y., Sun, J., Tian, Y., 2023. Characterizing the impact of autonomous vehicles on macroscopic fundamental diagrams. IEEE Trans. Intell. Transp. Syst.

Huang, A.J., Agarwal, S., 2023. On the limitations of physics-informed deep learning: Illustrations using first-order hyperbolic conservation law-based traffic flow models. IEEE Open J. Intell. Transp. Syst. 4, 279–293.

Huntsinger, L.F., Rouphail, N.M., 2011. Bottleneck and queuing analysis: Calibrating volume-delay functions of travel demand models. Transp. Res. Rec. J. Transp. Res. Board 2255, 117–124.

Huo, J., Wu, X., Lyu, C., Zhang, W., Liu, Z., 2022. Quantify the road link performance and capacity using deep learning models. IEEE Trans. Intell. Transp. Syst. 23(10), 18581–18591.

Hurdle, V.F., 1984. Signalized intersection delay models—A primer for the uninitiated. Transp. Res. Rec. 971(112), 96–105.

Jia, A., Zhou, X., Li, M., Rouphail, N.M., Williams, B.M., 2011. Incorporating stochastic road capacity into day-to-day traffic simulation and traveler learning framework: Model development and case study. Transp. Res. Rec. J. Transp. Res. Board 2254, 112–121.

Ka, E., Xue, J., Leclercq, L., Ukkusuri, S.V., 2024. A physics-informed machine learning for generalized bathtub model in large-scale urban networks. Transp. Res. Part C Emerg. Technol. 164, 104661.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. 3(6), 422–440.

Kerner, B.S., 2000. Theory of breakdown phenomenon at highway bottlenecks. Transp. Res. Rec. J. Transp. Res. Board 1710, 136–144.

Ke, J., Zheng, H., Yang, H., Chen, X.M., 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. Transp. Res. Part C Emerg. Technol. 85, 591–608.

Kim, J., Mahmassani, H.S., Dong, J., 2010. Likelihood and duration of flow breakdown: Modeling the effect of weather. Transp. Res. Rec. J. Transp. Res. Board 2188, 19–28.

Kim, Y.G., Mahmassani, H.S., 1987. Link performance functions for urban freeways with asymmetric car-truck interactions. Transp. Res. Rec.

Kimber, R.M., Hollis, E.M., 1979. Traffic queues and delays at road junctions.

Kosun, C., Tayfur, G., Celik, H.M., 2016. Soft computing and regression modelling approaches for link-capacity functions. Neural Netw. World 26, 129–140.

Kucharski, R., Drabicki, A., 2017. Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows. J. Adv. Transp. 2017, 1–10.

Lai, Y.-C. (Rex), Barkan, C.P.L., 2009. Enhanced parametric railway capacity evaluation tool. Transp. Res. Rec. J. Transp. Res. Board 2117, 33–40.

Lam, W.H.K., Cheung, C., 2000. Pedestrian speed/flow relationships for walking facilities in Hong Kong. J. Transp. Eng. 126, 343–349.

Lam, W.H.K., Lee, J.Y.S., Cheung, C.Y., 2002. A study of the bi-directional pedestrian flow characteristics at Hong Kong signalized crosswalk facilities. Transportation 29, 169–192.

Larsson, T., Patriksson, M., 1995. An augmented lagrangean dual algorithm for link capacity side constrained traffic assignment problems. Transp. Res. Part B Methodol. 29, 433–455.

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

Lazar, D.A., Coogan, S., Pedarsani, R., 2020. Routing for traffic networks with mixed autonomy. IEEE Trans. Autom. Control 66(6), 2664–2676.

Levin, M.W., Boyles, S.D., 2015. Effects of autonomous vehicle ownership on trip, mode, and route choice. Transp. Res. Rec. J. Transp. Res. Board 2493, 29–38.

Li, X., 2022. Trade-off between safety, mobility and stability in automated vehicle following control: An analytical method. Transp. Res. Part B Methodol. 166, 1–18.

Li, Y., Zheng, J., Qin, L., Li, H., 2025. Highway capacity of mixed traffic flow with autonomous vehicles: A review. Physica A Stat. Mech. Appl. 671, 130653.

Li, Z.C., Huang, H.J., Yang, H., 2020. Fifty years of the bottleneck model: A bibliometric review and future research directions. Transp. Res. Part B Methodol. 139, 311–342.

Lianeas, T., Nikolova, E., Stier-Moses, N.E., 2018. Risk-averse selfish routing. Math. Oper. Res. 43, 146–167.

Liu, J., Jiang, R., Zhao, J., Shen, W., 2023. A quantile-regression physics-informed deep learning for car-following model. Transp. Res. Part C Emerg. Technol. 154, 104275.

Liu, T., Meidani, H., 2024. End-to-end heterogeneous graph neural networks for traffic assignment. Transp. Res. Part C Emerg. Technol. 165, 104695.

Long, J., Gao, Z., Szeto, W.Y., 2011. Discretised link travel time models based on cumulative flows: Formulations and properties. Transp. Res. Part B Methodol. 45, 232–254.

Long, K., Sheng, Z., Shi, H., Li, X., Chen, S., Ahn, S., 2025. Physical enhanced residual learning (PERL) framework for vehicle trajectory prediction. Commun. Transp. Res. 5, 100166.

Long, K., Shi, X., Li, X., 2024. Physics-informed neural network for cross-dynamics vehicle trajectory stitching. Transp. Res. Part E Logist. Transp. Rev. 192, 103799.

Lorenz, M., Elefteriadou, L., 2000. A probabilistic approach to defining freeway capacity and breakdown (Doctoral dissertation). Pennsylvania State University.

Lu, J., Li, C., Wu, X.B., Zhou, X.S., 2023. Physics-informed neural networks for integrated traffic state and queue profile estimation: A differentiable programming approach on layered computational graphs. Transp. Res. Part C Emerg. Technol. 153, 104224.

Ma, X., Lo, H.K., 2012. Modeling transport management and land use over time. Transp. Res. Part B Methodol. 46, 687–709.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C Emerg. Technol. 54, 187–197.

Mahmassani, H.S., 2016. Autonomous vehicles and connected vehicle systems: Flow and operations considerations. Transp. Sci. 50, 1140–1162.

Mahmassani, H.S., Hou, T., Saberi, M., 2013. Connecting networkwide travel time reliability and the network fundamental diagram of traffic flow. Transp. Res. Rec. J. Transp. Res. Board 2391, 80–91.

Mesbah, M., Sarvi, M., Ouveysi, I., Currie, G., 2011. Optimization of transit priority in the transportation network using a decomposition methodology. Transp. Res. Part C Emerg. Technol. 19, 363–373.

Miandoabchi, E., Farahani, R.Z., Dullaert, W., Szeto, W.Y., 2012. Hybrid evolutionary metaheuristics for concurrent multi-objective design of urban road and public transit networks. Netw. Spat. Econ. 12(3), 441–480.

Montanino, M., Monteil, J., Punzo, V., 2021. From homogeneous to heterogeneous traffic flows: Lp string stability under uncertain model parameters. Transp. Res. Part B Methodol. 146, 136–154.

Moses, R., Mtoi, E., Ruegg, S., McBean, H., 2013. Development of speed models for improving travel forecasting and highway performance evaluation. Florida Dept. of Transportation.

Mosher, W.W., 1963. A capacity-restraint algorithm for assigning flow to a transport network. Highw. Res. Rec. 6.

Müller, S., Schiller, C., 2015. Improvement of the volume-delay function by incorporating the impact of trucks on traffic flow. Transp. Plan. Technol. 38, 878–888.

Nagurney, A., Qiang, Q., 2012. Fragile networks: Identifying vulnerabilities and synergies in an uncertain age. Int. Trans. Oper. Res. 19, 123–160.

Newell, G.F., 1982. Applications of queueing theory. Springer, Dordrecht.

Newell, G.F., 1968a. Queues with time-dependent arrival rates I: The transition through saturation. J. Appl. Probab. 5(2), 436–451.

Newell, G.F., 1968b. Queues with time-dependent arrival rates II: The maximum queue and the return to equilibrium. J. Appl. Probab. 5(3), 579–590.

Ni, D., Leonard, J.D., Jia, C., Wang, J., 2016. Vehicle longitudinal control and traffic stream modeling. Transp. Sci. 50(3), 1016–1031.

Nie, X., Zhang, H.M., 2005a. A comparative study of some macroscopic link models used in dynamic traffic assignment. Netw. Spat. Econ. 5, 89–115.

Ordónez, F., Stier-Moses, N.E., 2010. Wardrop equilibria with risk-averse users. Transp. Sci. 44(1), 63–86.

Pan, Y., Guo, J., Chen, Y., 2022. Calibration of dynamic volume-delay functions: A rolling horizon-based parsimonious modeling perspective. Transp. Res. Rec. J. Transp. Res. Board 2676, 606–620.

Pan, Y.A., Guo, J., Chen, Y., Cheng, Q., Li, W., Liu, Y., 2024. A fundamental diagram-based hybrid framework for traffic flow estimation and prediction by combining a Markovian model with deep learning. Expert Syst. Appl. 238, 122219.

Pan, Y.A., Li, F., Li, A., Niu, Z., Liu, Z., 2025. Urban intersection traffic flow prediction: A physics-guided stepwise framework utilizing spatio-temporal graph neural network algorithms. Multimodal Transp. 4(2), 100207.

Pan, Y., Cheng, Q., Li, A., Zhang, J., Guo, J., Chen, Y., 2025. Analysis of congestion key parameters, dynamic discharge process, and capacity estimation at urban freeway bottlenecks: A case study in Beijing, China. Transp. Lett. 17(6), 984–1003.

Patriksson, M., 2015. The traffic assignment problem: Models and methods. Dover Publications, Mineola, NY.

Peeta, S., Ziliaskopoulos, A.K., 2001. Foundations of dynamic traffic assignment: The past, the present and the future. Netw. Spat. Econ. 1, 233–265.

Petersen, E.R., 1974. Over-the-road transit time for a single track railway. Transp. Sci. 8, 65–74.

Powell, W.B., Sheffi, Y., 1982. The convergence of equilibrium algorithms with predetermined step sizes. Transp. Sci. 16(1), 45–55.

Prokopy, J.C., Richard, B.R., 1975. Parametric analysis of railway line capacity (No. FRA-OPPD-75-1).

Qian, Z. (Sean), Li, J., Li, X., Zhang, M., Wang, H., 2017. Modeling heterogeneous traffic flow: A pragmatic approach. Transp. Res. Part B Methodol. 99, 183–204.

Raadsen, M.P.H., Bliemer, M.C.J., 2019. Steady-state link travel time methods: Formulation, derivation, classification, and unification. Transp. Res. Part B Methodol. 122, 167–191.

Raadsen, M.P.H., Bliemer, M.C.J., Bell, M.G.H., 2020. Aggregation, disaggregation and decomposition methods in traffic assignment: Historical perspectives and new trends. Transp. Res. Part B Methodol. 139, 199–223.

Rakha, H., Zhang, W., 2005. Consistency of shock-wave and queuing theory. Presented at the 84th Annual Meeting of the Transportation Research Board, Washington, DC, 219–226.

Ran, B., Boyce, D.E., 1996. A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. Transp. Res. Part C Emerg. Technol. 4, 1–12.

Ran, B., Hall, R.W., Boyce, D.E., 1996. A link-based variational inequality model for dynamic departure time/route choice. Transp. Res. Part B Methodol. 30, 31–46.

Roughgarden, T., Tardos, E.V., 2002. How bad is selfish routing? J. ACM 49(2), 236–259.

Rouphail, N., Tarko, A., Li, J., 1992. Traffic flow at signalized intersections.

Rowan, D., He, H., Hui, F., Yasir, A., Mohammed, Q., 2025. A systematic review of machine learning-based microscopic traffic flow models and simulations. Commun. Transp. Res. 5, 100164.

Samaranayake, S., Chand, S., Sinha, A., Dixit, V., 2024. Impact of connected and automated vehicles on the travel time reliability of an urban network. Int. J. Transp. Sci. Technol. 13, 171–185.

Saric, A., Albinovic, S., Dzebo, S., Pozder, M., 2019. Volume-delay functions: A review. In: Avdaković, S. (Ed.), Advanced Technologies, Systems, and Applications III, Lecture Notes in Networks and Systems. Springer International Publishing, Cham, 3–12.

Schwarz, J.A., Selinka, G., Stolletz, R., 2016. Performance analysis of time-dependent queueing systems: Survey and classification. Omega 63, 170–189.

Sengupta, A., Guler, S.I., 2025. Deep learning–based spatial translation of traffic prediction using Newell's theory. J. Transp. Eng. Part A Syst. 151(7), 04025041.

Sheffi, Y., 1984. Urban transportation networks: Equilibrium analysis with mathematical programming methods. Prentice-Hall, Englewood Cliffs, NJ.

Shi, X., Li, X., 2021. Constructing a fundamental diagram for traffic flow with automated vehicles: Methodology and demonstration. Transp. Res. Part B Methodol. 150, 279–292.

Mo, Z., Shi, R., Di, X., 2021. A physics-informed deep learning paradigm for car-following models. Transp. Res. Part C Emerg. Technol. 130, 103240.

Shi, R., Mo, Z., Huang, K., Di, X., Du, Q., 2021. A physics-informed deep learning paradigm for traffic state and fundamental diagram estimation. IEEE Trans. Intell. Transp. Syst. 23(8), 11688–11698.

Skabardonis, A., Dowling, R., 1997. Improved speed-flow relationships for planning applications. Transp. Res. Rec. J. Transp. Res. Board 1572, 18–23.

Small, K.A., 1983. The incidence of congestion tolls on urban highways. J. Urban Econ. 13, 90–111.

Small, K.A., Verhoef, E.T., Lindsey, R., 2007. The economics of urban transportation. Routledge.

Smock, R., 1962. An iterative assignment approach to capacity restraint on arterial networks. Highw. Res. Board Bull. 347.

Sogin, S.L., Lai, Y.-C. (Rex), Dick, C.T., Barkan, C.P., 2016. Analyzing the transition from single- to double-track railway lines with nonlinear regression analysis. Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit 230, 1877–1889.

Soltman, T.J., 1965. Selected bibliography for transportation planning (No. 97).

Song, L., Fan, W., 2023. Intersection capacity adjustments considering different market penetration rates of connected and automated vehicles. Transp. Plan. Technol. 46(3), 286–303.

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

Srinivasan, D., Choy, M.C., Cheu, R.L., 2006. Neural networks for real-time traffic signal control. IEEE Trans. Intell. Transp. Syst. 7(3), 261–272.

Szeto, W.Y., Jiang, Y., Wang, D.Z.W., Sumalee, A., 2015. A sustainable road network design problem with land use transportation interaction over time. Netw. Spat. Econ. 15, 791–822.

Talebpour, A., Mahmassani, H.S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transp. Res. Part C Emerg. Technol. 71, 143–163.

Tarko, A.P., Perez-Cartagena, R.I., 2005. Variability of peak hour factor at intersections. Transp. Res. Rec. 1920(1), 125–130.

Tak, S., Kim, S., Jang, K., Yeo, H., 2014. Real-time travel time prediction using multi-level k-nearest neighbor algorithm and data fusion method. In: Computing in Civil and Building Engineering (2014), 1861–1868.

Tang, Y., Jin, L., Ozbay, K., 2024. Physics-informed machine learning for calibrating macroscopic traffic flow models. Transp. Sci. 58(6), 1389–1402.

Tisato, P., 1991. Suggestions for an improved Davidson travel time function. Aust. Road Res.

Van Ommeren, J., Fosgerau, M., 2009. Workers' marginal costs of commuting. J. Urban Econ. 65, 38–47.

Vickrey, W.S., 1969. Congestion theory and transport investment. Am. Econ. Rev. 59(2), 251–260.

Wang, D., Zhang, J., Cao, W., Li, J., Zheng, Y., 2018. When will you arrive? Estimating travel time based on deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Wang, H., Li, J., Chen, Q.Y., Ni, D., 2011. Logistic modeling of the equilibrium speed-density relationship. Transp. Res. Part A Policy Pract. 45(6), 554–566.

Webster, F.V., 1958. Traffic signal settings.

Wei, Y., Avcı, C., Liu, J., Belezamo, B., Aydın, N., Li, P., Zhou, X., 2017. Dynamic programming-based multi-vehicle longitudinal trajectory optimization with simplified car-following models. Transp. Res. Part B Methodol. 106, 102–129.

Wei, Z., Cheng, Y., Zhu, J., Huang, Q., 2019. Genipin-crosslinked ovotransferrin particle-stabilized Pickering emulsions as delivery vehicles for hesperidin. Food Hydrocoll. 94, 561–573.

Wong, W., Wong, S.C., 2016. Network topological effects on the macroscopic Bureau of Public Roads function. Transp. Transp. Sci. 12, 272–296.

Wu, C.H., Ho, J.M., Lee, D.T., 2004. Travel-time prediction with support vector regression. IEEE Trans. Intell. Transp. Syst. 5(4), 276–281.

Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Transp. Res. Part C Emerg. Technol. 96, 321–346.

Wu, X., Dutta, A., Zhang, W., Zhu, H., Livshits, V., Zhou, X., 2021. Characterization and calibration of volume-to-capacity ratio in volume delay functions on freeways based on a queue analysis approach.

Xiong, H., Davis, G.A., 2009. Field evaluation of model-based estimation of arterial link travel times. Transp. Res. Rec. J. Transp. Res. Board 2130, 149–157.

Yang, H., Wang, X., 2011. Managing network mobility with tradable credits. Transp. Res. Part B Methodol. 45, 580–594.

Yang, H., Yagar, S., 1995. Traffic assignment and signal control in saturated road networks. Transp. Res. Part A Policy Pract. 29, 125–139.

Yeon, J., Hernandez, S., Elefteriadou, L., 2009. Differences in freeway capacity by day of the week, time of day, and segment type. J. Transp. Eng. 135, 416–426.

Yperman, I., Logghe, S., Immers, B., 2005. The link transmission model: An efficient implementation of the kinematic wave theory in traffic networks. In: Proceedings of the 10th EWGT Meeting, Poznan, Poland.

Zhang, L., Levinson, D., 2004. Some properties of flows at freeway bottlenecks. Transp. Res. Rec. J. Transp. Res. Board 1883, 122–131.

Zhang, X., Waller, S.T., 2018. Link performance functions for high occupancy vehicle lanes of freeways. Transport 33, 657–668.

Zhang, X., Sun, J., Sun, J., 2025. On the stochastic fundamental diagram: A general micro-macroscopic traffic flow modeling framework. Commun. Transp. Res. 5, 100163.

Zhao, Z., Liang, Y., 2023. A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards. Transp. Res. Part C Emerg. Technol. 149, 104079.

Zhou, X., Cheng, Q., Wu, X., Li, P., Belezamo, B., Lu, J., Abbasi, M., 2022. A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio. Multimodal Transp. 1, 100017.

Zhou, X., Hadi, M., Hale, D., 2021. Multiresolution modeling for traffic analysis: State-of-practice and gap analysis report (No. FHWA-HRT-21-082). Federal Highway Administration, United States.

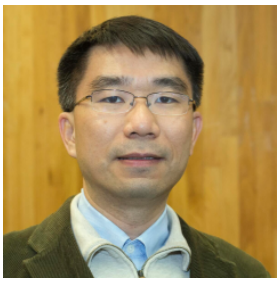
Zhou, D., Gayah, V.V., 2023. Improving deep reinforcement learning-based perimeter metering control methods with domain control knowledge. Transp. Res. Rec. 2677(7), 384–405.

Author Biography

Yuyan (Annie) Pan received her Ph.D. in Transportation Engineering from Beijing University of Technology in 2023. She is currently a Postdoctoral Scholar in the Department of Civil and Environmental Engineering at Pennsylvania State University. Her research focuses on connected and automated vehicles, electric vehicle charging, traffic flow theory, and data-driven modeling. She has published more than 20 peer-reviewed papers and serves as a reviewer for leading journals such as *Transportation Research Part B/C/E*. Her work integrates fundamental traffic theory with AI-based methods to enhance traffic system efficiency and resilience.

Xianbiao Hu received the Ph.D. degree from The University of Arizona, Tucson, AZ, USA, in 2013. He is currently an Associate Professor in the Civil and Environmental Engineering Department, The Pennsylvania State University. His current research interests include smart mobility systems, connected and automated vehicles, electric vehicles, mobility behavior management, transportation big data analytics, and traffic flow and system modeling. He serves as an Associate Editor for *IEEE Transactions on Intelligent Transportation Systems*, an Assistant Editor of the *Journal of Intelligent Transportation Systems*, and a Handling Editor of *Transportation Research Record*.

George F. List received his B.S.E.E. degree from Carnegie Mellon University, Pittsburgh, PA, USA, in 1971; his M.E.E. degree from the University of Delaware, Newark, DE, USA, in 1976; and his Ph.D. degree in Civil Engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1984. He is currently a Professor in the Department of Civil, Construction and Environmental Engineering at North Carolina State University, Raleigh, NC, USA. He is best known for his work on the modeling, simulation, and optimization of transport systems and networks. His professional affiliations include ASCE (Fellow), TRB, IEEE, ITE, and INFORMS.



Xuesong (Simon) Zhou received the Ph.D. degree in Civil Engineering from the University of Maryland, College Park, MD, USA, in 2004. He is currently a Professor in the School of Sustainable Engineering and the Built Environment at Arizona State University, Tempe, AZ, USA. His research focuses on dynamic traffic assignment, traffic estimation and prediction, large-scale routing, and rail scheduling. He previously served as an Associate Editor for *Transportation Research Part C* and is currently the Executive Editor-in-Chief of *Urban Rail Transit*, and an Associate Editor for *Transportation Research Part B*. He has also chaired the INFORMS Rail Application Section (2016 and 2025) and currently serves as Chair of the Transportation Research Board Committee on Transportation Network Analysis (AEP13).