Transportation Research Part B 200 (2025) 103292

Contents lists available at ScienceDirect

Transportation Research Part B

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

Real-time vehicle location estimation in signalized networks using partial connected vehicle trajectory data

Shaocheng Jia [a], S.C. Wong [a,*], Wai Wong [b,*]

[a] Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
[b] Department of Civil and Environmental Engineering, University of Canterbury, Christchurch 8041, New Zealand

ARTICLE INFO

Keywords:
Vehicle location estimation
Connected vehicles
Arrival pattern estimation
Travel time estimation
Signal control with connected vehicles

ABSTRACT

Real-time vehicle location estimation is essential for diverse transportation applications, such as travel time estimation, arrival pattern estimation, and adaptive signal control. Existing connected vehicle-based studies rely on either black-box neural networks requiring large training datasets or computationally intensive time-continuous movement simulations grounded in car-following models. However, they often overlook the distinct vehicle location patterns in source lanes, which define network boundaries and experience random arrivals, and intermediate lanes, situated between intersections and receiving traffic discharged from upstream. These patterns are critical for accurate vehicle location estimation. To address these limitations, this study proposes a generic and fully analytical CV-based vehicle location (CVVL) model for estimating vehicle locations within a signalized lane in a network using readily available partial CV trajectory data. The proposed model is applicable to any signal timing, traffic demand, and CV penetration rate and consists of two sub-models: CVVL-S and CVVL-I. The CVVL-S sub-model estimates vehicle locations in source lanes, where vehicle distribution tends to be relatively homogeneous owing to random arrivals. In contrast, the CVVL-I sub-model focuses on estimating vehicle locations in intermediate lanes, where sequential discharges from different upstream lanes can lead to the formation of multiple platoons, adding complexity to vehicle location estimation. The proposed model decomposes the complex task into three sequential sub-problems: identifying candidate platoons (CPs), estimating the number of vehicles in each CP, and determining the spatial distribution of vehicles within each CP. Extensive numerical experiments were conducted under various traffic conditions, CV penetration rates, and times of interest using the VISSIM platform and the real-world Next Generation Simulation dataset. The results demonstrate that the proposed CVVL model achieved improvements of 0–45 %, 0–37 %, and 4–34 % in precision, recall, and F1 score, respectively, compared with the competing method. These results highlight the model's potential to enhance the accuracy and reliability of various downstream applications.

1. Introduction

The integration of Internet of Things technologies has catalyzed the rise of connected vehicles (CVs) within transportation systems. These vehicles are seamlessly connected to cloud servers, facilitating the real-time sharing of trajectory information. The accessibility of CV trajectory data holds immense promise for enhancing intelligent transportation system applications, resulting in improved

* Corresponding authors.
E-mail addresses: shaocjia@connect.hku.hk (S. Jia), hhecwsc@hku.hk (S.C. Wong), wai.wong@canterbury.ac.nz (W. Wong).

https://doi.org/10.1016/j.trb.2025.103292
Received 23 July 2024; Received in revised form 16 July 2025; Accepted 12 August 2025
Available online 20 August 2025

flexibility, cost-efficiency, and operational effectiveness. However, the prolonged transition period to complete CV adoption and the persistence of partial CV connectivity poses significant challenges in transport modeling and system optimization owing to the limited availability of traffic information. Thus, recovering comprehensive traffic information using only partial CV data is crucial for developing effective transportation applications.

The CV penetration rate, defined as the ratio of the number of CVs to the total number of vehicles within a specified road segment and period, serves as a crucial parameter for inferring various traffic datapoints and is a prerequisite for diverse applications. In road segments equipped with on-road detectors, the CV penetration rate can be easily obtained by dividing the number of CVs by the total number of vehicles counted by the detector. However, not all road segments have detectors installed. In segments without detectors, estimating the CV penetration rates requires alternative methods. One approach involves estimating the distribution of CV penetration rates across different road segments with detectors in the network, using the available CV counts and total traffic counts (Wong and Wong, 2015, 2016a, 2019; Wong et al., 2019). The mean of this distribution is then used as the most probable estimate for the CV penetration rate in nearby road segments without detectors. This approach assumes independent and identically distributed CV penetration rates across road segments. Alternatively, Meng et al. (2017) developed an empirical model that explicitly considers land-use properties to identify correlations between CV penetration rates in different segments. Despite these advancements, the existing approaches rely on the presence of on-road detectors, limiting their universal deployment.

Estimating CV penetration rates using partial CV data has gained popularity owing to its flexibility, convenience, and cost-effectiveness. Consider a signalized lane, where vehicles must stop at the stop bar during a red signal. The formation of a traffic queue presents valuable opportunities for estimating the CV penetration rate. For instance, Comert (2016) used the stop location of the last CV in the queue to derive the CV penetration rate, assuming Poisson arrivals. Wong et al. (2019) proposed an unbiased estimator for average CV penetration rate estimation without any vehicle arrival assumption. Additionally, Zhao et al. (2019a), (2019b), (2022) used a maximum likelihood estimation method to estimate the vehicle stopping location and CV penetration rate. Despite significant advancements, these methods provide only first-order point estimators for average CV penetration rates, neglecting the inherent uncertainty. Given the highly dynamic and nonlinear nature of transportation systems, ignoring uncertainty in CV penetration rates can lead to biased models and suboptimal solutions in transportation system optimization (Wong and Wong, 2015, 2016a, 2019; Wong et al., 2019; Yin, 2008). To address this limitation, Jia et al. (2023) derived a probabilistic penetration rate (PPR) model that accurately models the uncertainty in CV penetration rates. However, the constrained queue length distribution, an essential input for the PPR model, was estimated only under undersaturation conditions without holding vehicles (Jia et al., 2023). Holding vehicles at any given instant are defined as vehicles that, based on their projected trajectories using cruise speeds, should have been discharged by that instant but remain held by the system. In real-world scenarios, holding vehicles are common, and their presence significantly affects the model performance. To overcome this challenge, Jia et al. (2024a, 2024b) developed a Markov-constrained queue length (MCQL) model, which accounts for complex holding-vehicle effects in the estimation of uncertainty related to the CV penetration rate. Using the MCQL model, the PPR model can effectively handle undersaturation conditions, regardless of the presence of holding vehicles.

The CV penetration rate serves as a key parameter for numerous transportation applications, including queue length estimation (Comert and Cetin 2009, 2011; Comert 2013; Hao et al., 2014), arrival table determination (Feng et al., 2015; Jenelius et al., 2013, 2015; Rahmani et al., 2015; Tian et al., 2015; Khan et al., 2017; Mousa et al., 2017; Iqbal et al., 2018; Lu et al., 2019), traffic flow prediction (Wong and Wong, 2015, 2016a, 2016c), traffic density estimation (Geroliminis and Daganzo, 2008; Ambühl and Menendez, 2016; Du et al., 2016; Wong and Wong, 2019; Wong et al., 2019, 2021), origin–destination matrix determination (Yang et al., 2017; Wang et al., 2020; Cao et al., 2021), traffic incident impact assessment (Wong and Wong, 2016b), holding vehicle estimation (Jia et al., 2024c), and traffic signal control (Feng et al., 2015; Jia et al., 2023, 2024a, 2024b, 2024d; Wang et al., 2024). Nonetheless, real-time vehicle location estimation is another important yet underexplored traffic state, serving as an essential input for numerous transportation applications. For example, knowing the current location of a vehicle enables the prediction of its travel time to a downstream destination based on cruise speed. Similarly, identifying the locations of all vehicles in a lane facilitates the estimation of their arrival times at the stop bar, which are crucial for determining arrival patterns and estimating delays in traffic signal control (Xia et al., 2023). Feng et al. (2015) demonstrated that incorporating estimated vehicle locations into adaptive signal control could reduce total intersection delays by 0.81 %, 9.30 %, 4.24 %, and 16.33 % at CV penetration rates of 0.25, 0.5, 0.75, and 1, respectively, under demand settings of 667 veh/h/lane on eastbound and westbound approaches and 500 veh/h/lane on northbound and southbound approaches. Moreover, detailed vehicle location data are valuable for estimating traffic density and improving CV applications. For instance, Goodall et al. (2016) found that integrating vehicle location estimation into a ramp metering algorithm enhanced its performance at low CV penetration rates and maintained its effectiveness at higher rates. These examples underscore the importance of accurately estimating vehicle locations for reflecting traffic flow characteristics, deriving other critical traffic parameters, and optimizing traffic signal timings.

In traditional transportation systems, accurately estimating vehicle locations in lanes, whether equipped with on-road point detectors or not, is extremely challenging. Although vehicles visible before CVs can be accurately located using on-board camera data in a connected environment (Jia et al., 2020, 2021, 2022; Jia and Yao, 2023, 2024), limited transmission bandwidth and occlusions render image-based methods difficult to implement in real-time applications. In comparison, vehicle location estimation based on CV trajectory data is more suitable for real-time applications, as trajectory data exhibit low volume, are widely available, and can be easily processed. Recognizing these advantages, several studies have leveraged CV trajectory data to estimate vehicle locations. For instance, Goodall (2013) and Goodall et al. (2016) proposed two algorithms tailored for estimating real-time vehicle locations in urban arterial networks and freeways, respectively. The arterial location estimation algorithm utilized partial CV location data to infer the locations of non-CVs (NCs). It involved inserting NCs based on car-following models in slow-down regions or using vehicle occupancy lengths in queues. The algorithm then simulated the movements of these inserted NCs to predict their future locations. If a simulated NC location

overlapped with a reported CV location, the NC was deemed inaccurately estimated and subsequently removed. While effective in certain scenarios, this method relied heavily on assumptions about specific car-following behaviors and required continuous movement simulation, making it computationally intensive and unsuitable for real-time applications. Similarly, the freeway location estimation algorithm applied car-following models and CV trajectory data to predict NC positions on freeways. However, its accuracy was limited to congested regions, as the method depended on vehicle interactions, which are more pronounced in such conditions. Other approaches, such as the probabilistic method of Cao et al. (2014), derive the joint distribution of vehicle locations and speeds on arterial roads using probe vehicle data. While this method accounts for signal timings, critical flow density, and driving behavior, its complexity hinders real-time implementation. Moreover, it only estimates long-term probability distributions, failing to capture real-time dynamics.

Feng et al. (2015) introduced the novel Estimation of Vehicle Location and Speed (EVLS) algorithm, which estimates vehicle locations using only CV trajectory data. The EVLS algorithm segments a lane into three distinct regions: queue, slow-down, and free-flow. Each region employs a specific method to estimate vehicle locations, assuming a known CV penetration rate. Typically, network approaching lanes can be categorized as source lanes and intermediate lanes (Lo, 1999). Source lanes, located at the network boundaries, experience exogenous random vehicle arrivals, resulting in relatively uniform vehicle distributions. Intermediate lanes, situated between signalized intersections, receive traffic discharged from upstream intersections, producing cyclic vehicle arrival patterns. However, the EVLS algorithm does not adequately capture the complexity of these differing vehicle patterns, often leading to suboptimal vehicle location estimation. Recently, Ye et al. (2024) developed a CV-based vehicle location and speed estimation network that simultaneously infers the locations and speeds of NCs using CV trajectory data as inputs. In this approach, roads are divided into discrete cells, enabling variable traffic states to be represented as fixed-size vectors. A coding-rate transformer serves as the backbone network, learning the relationship between CV and NC states. While comprehensive numerical experiments demonstrate the superiority of this learning framework, its reliance on complex neural networks presents challenges. These models operate as black boxes, complicating the interpretation of their mechanisms. Moreover, machine learning-based methods require extensive training datasets and often exhibit limited generalization capabilities, restricting their robustness in diverse scenarios. With the integration of automation into CVs, connected and automated vehicles (CAVs) emerge, capable of sharing real-time traffic information and perceiving their surroundings. Leveraging CAV data and roadside detectors, Li et al. (2021) proposed a cooperative perception framework using particle filtering to estimate vehicle locations and speeds. Similarly, Xia et al. (2023) addressed the estimation of speeds and locations of human-driven vehicles in detection blind spots by formulating nonlinear optimization problems. Although CAV-based methods achieve high accuracy in estimating vehicle locations and speeds, they demand multi-source data and involve computationally intensive processes, limiting their practicality for real-time applications.

This paper proposes a generic and fully analytical CV-based vehicle location (CVVL) model to accurately estimate vehicle locations in both source and intermediate lanes within a network. Given the distinct vehicle arrival patterns in source and intermediate lanes, the CVVL model includes two sub-models: CVVL-S and CVVL-I. CVVL-S focuses on estimating vehicle locations in source lanes, while CVVL-I is tailored for intermediate lanes. The primary distinction between these lane types lies in their locations in a network and vehicle arrival characteristics (Lo, 1999). Source lanes, located at the network boundaries, experience exogenous random vehicle arrivals, leading to relatively homogeneous vehicle distributions. In contrast, intermediate lanes, positioned between signalized intersections, receive traffic discharged sequentially from upstream lanes, often forming multiple vehicle platoons. The proposed CVVL model takes into account these complex vehicle arrival patterns and systematically decomposes the problem into sequential sub-problems: identifying candidate platoons (CPs), estimating the number of vehicles in each CP, and determining the distribution of vehicles within CPs. Comprehensive simulation experiments conducted using VISSIM demonstrate the effectiveness and superiority of the proposed models. Furthermore, applying the CVVL model to the real-world Next Generation Simulation (NGSIM) dataset (Federal Highway Administration, 2006) validates its practicality and applicability.

The key contributions of this work are summarized as follows:

- The analytical and generic CVVL model, based solely on accessible CV trajectory data, is established to estimate real-time vehicle locations at any time of interest (ToI). Unlike previous methods that rely on continuous movement simulations using car-following models or black-box neural networks, the proposed model is fully analytical, requiring only current traffic information in a lane. This analytical approach facilitates efficient vehicle location estimation for diverse real-time transport applications, eliminating the computational burden associated with time-continuous simulations or neural network inferences.
- The CVVL model systematically estimates vehicle locations in both source and intermediate lanes within a network. By explicitly accounting for the differences in vehicle arrival patterns between these lane types, often overlooked in existing studies, this work achieves significantly improved integration of CV-based vehicle location estimation into network-wide transportation applications.
- The CVVL model introduces a novel decomposition strategy, dividing the complex vehicle location estimation task into three manageable sub-problems: (i) identifying candidate platoons (CPs), (ii) estimating the number of vehicles within each CP, and (iii) determining the spatial distribution of vehicles within each CP. This structured approach not only enhances the understanding of CV-based problems but also provides valuable byproducts that can be leveraged in other transportation applications.

This work advances the fundamental problem of vehicle location estimation using only CV trajectory data, paving the way for broader innovative CV-based transportation applications.

The remainder of the paper is organized as follows. Section 2 defines the research problem. Sections 3 and 4 illustrate the derivation of the CVVL-S and CVVL-I sub-models, respectively. Section 5 presents numerical experiments conducted in VISSIM to evaluate the effectiveness of the proposed models. Section 6 validates the model using the real-world NGSIM dataset. Section 7 presents the

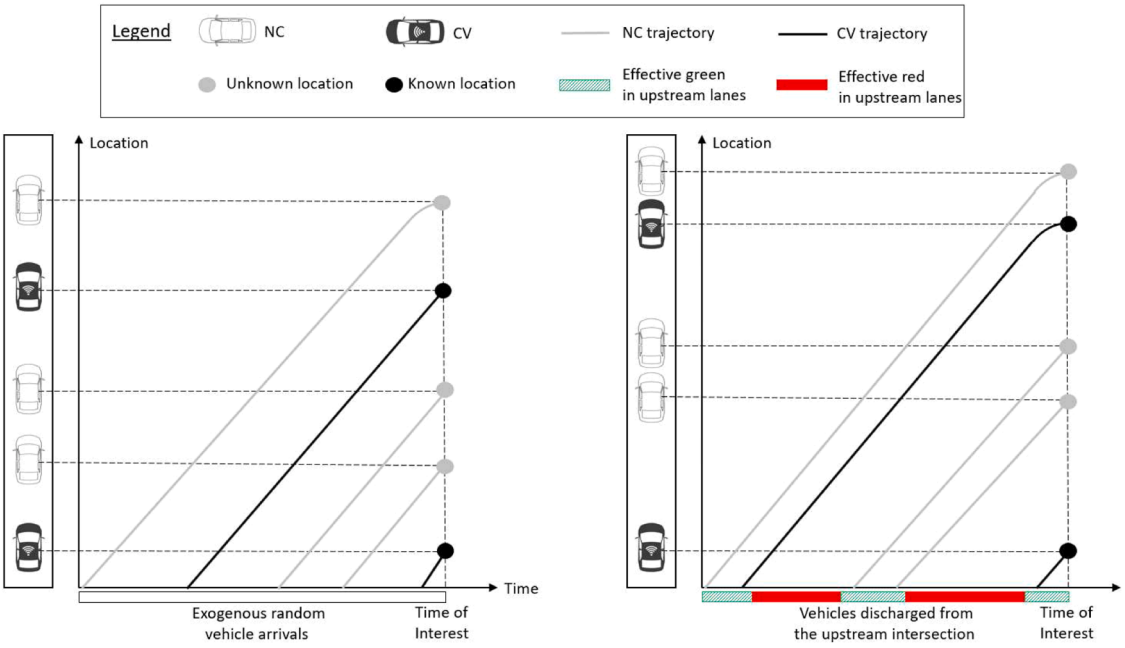

Fig. 1. Vehicle location patterns in (a) a source lane and (b) an intermediate lane.

concluding remarks.

2. Problem statement and preliminaries

2.1. Problem statement

Signalized approaching lanes in a road network can be classified into two categories: source lanes and intermediate lanes. Source lanes define the boundaries of the network and are characterized by exogenous and random vehicle arrivals. In contrast, intermediate lanes, which are located between signalized intersections within the network, receive traffic discharged from upstream intersections, leading to cyclic vehicle arrivals. The vehicle location patterns in a source lane (Fig. 1(a)) typically differ from those in an intermediate lane (Fig. 1(b)). As shown in Fig. 1(a), vehicles in a source lane are expected to appear randomly, exhibiting a homogeneous distribution. In contrast, as illustrated in Fig. 1(b), vehicles in an intermediate lane exhibit a notable platooning effect owing to sequential discharges of vehicles from upstream lanes. Therefore, tailored models are required to accurately estimate vehicle locations in both source and intermediate lanes within a network.

In a partially connected environment, only some of the vehicles are CVs, while NCs remain unobservable, posing significant challenges in obtaining complete vehicle location information. Given the known locations of CVs, the objective of this study is to estimate the real-time locations of unobservable NCs in any given signalized lane of a network at any ToI using only partial CV trajectory data. To facilitate the model derivations, the following assumptions are made:

Assumption 1. CVs and NCs are assumed to be well mixed in lanes, meaning that each vehicle has a probability of p of being a CV and a probability of $1-p$ of being an NC, where p represents the true CV penetration rate.

Assumption 2. Vehicles located between two stopped CVs, or between a stopped CV and the stop bar during a red signal, are assumed to have zero speed and to be uniformly distributed between the two stopped CVs or between the stopped CV and the stop bar, respectively.

Assumption 3. Vehicles located between two CVs with at least one moving, or between a moving CV and the lane entrance, are assumed to have a linear speed profile and are uniformly distributed between the two CVs or between the moving CV and the lane entrance, respectively.

Assumption 4. The dispersion effects of individual vehicles within a platoon are assumed to be consistent, meaning that all individual vehicles share the same spatial dispersion ratio.

Assumption 1 is considered reasonable due to frequent lane-changing, overtaking, and cutting-in behaviors observed in practice. **Assumption 2** describes vehicle speeds and locations in a queue, where vehicles are stationary and uniformly distributed between

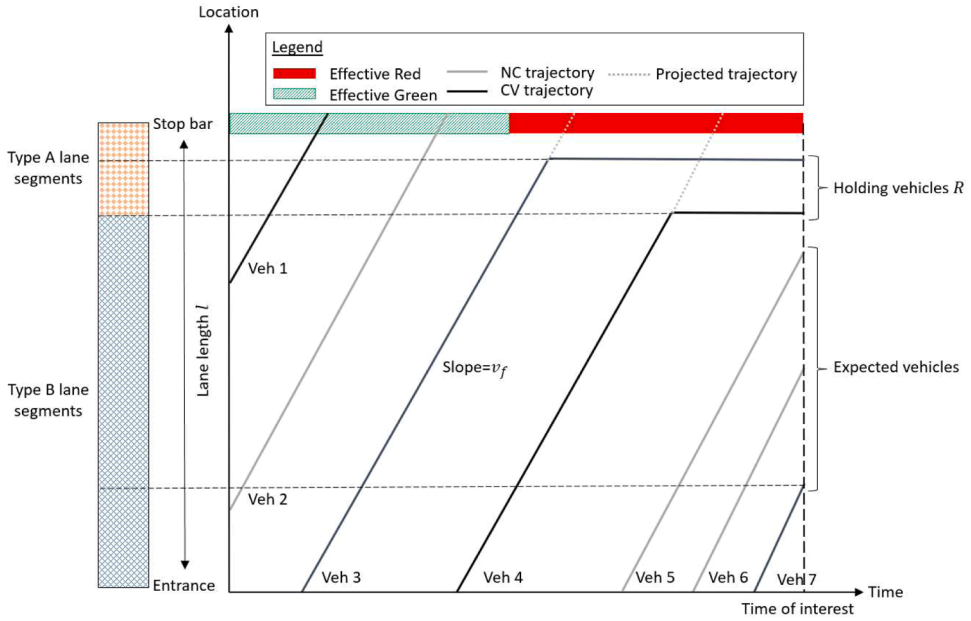

Fig. 2. Holding vehicles, expected vehicles, and different lane segments.

stopped CVs or a stopped CV and the stop bar. **Assumption 3** models vehicle speeds and locations in moving traffic. As vehicles generally adjust their speeds gradually by following leading vehicles, a simple linear speed profile is applied for NCs along a lane segment enclosed by two CVs with at least one moving or by a moving CV and the lane entrance. While the actual speed profile may be complex, this first-order model is deemed reasonable for a limited portion of vehicles along the enclosed lane segment. Additionally, the estimated speeds are only intermediate quantities used to determine the number of NCs in the enclosed lane segment. Therefore, as long as the speed estimation process is consistent, the estimated speeds have minimal impact on NC location estimation. In cases with no additional information, NCs are assumed to be uniformly distributed along the enclosed lane segment. **Assumption 4** describes a homogeneous spatial dispersion process. Specifically, if the observed locations of two CVs in a downstream platoon deviate from their expected locations estimated based on their cruise speeds, the remaining vehicles in the platoon are expected to exhibit the same dispersion (either divergence or shrinkage) ratio. This assumption is plausible given the similar driving behaviors observed within a platoon. Further details on these assumptions are provided in Sections 3 and 4.

2.2. Preliminaries

The proposed CVVL model consists of two sub-models: CVVL-S and CVVL-I. The CVVL-S sub-model, which estimates vehicle locations in source lanes, is derived in Section 3. The CVVL-I sub-model, which estimates vehicle locations in intermediate lanes, is established in Section 4. The key inputs for the CVVL model include the average arrival rate $\overline{q}$, average CV penetration rate $\overline{p}$, and number of holding vehicles R. In a signalized lane, vehicles stop during red signals and form queues. Using the inputs $\overline{q}$, $\overline{p}$, and signal timing plan, the queue distribution can be derived. Based on the estimated queue distribution, the numbers of CVs and vehicles located before the last stopped CV in a queue can be determined through probabilistic analysis. This indicates that $\overline{q}$ and $\overline{p}$ govern the observable numbers of CVs and vehicles preceding the last stopped CV in queues. Consequently, a maximum likelihood estimation approach can be employed to estimate $\overline{q}$ and $\overline{p}$ from the observed numbers of CVs and vehicles preceding the last stopped CV in queues. Once $\overline{q}$ and $\overline{p}$ are obtained, the number of holding vehicles R at any ToI can be estimated by considering components such as stopped holding CVs and NCs and moving holding CVs and NCs. Based on these principles and **Assumption 1**, previous studies (Jia et al., 2023, 2024a, 2024b, 2024c) have shown that these parameters can be estimated using partial CV trajectory data. Therefore, these parameters are assumed to be known inputs for the proposed model. **Appendices A** and **B** briefly summarize the methods to estimate these essential inputs.

3. CVVL-S sub-model

The CVVL-S sub-model is used to estimate the locations of NCs in source lanes using only accessible CV trajectory data. Given the expected random vehicle arrival in a source lane, vehicle distribution tends to be relatively homogeneous. The two sequential sub-problems under the CVVL-S sub-model are the estimation of the total number of NCs and the distribution of NCs, as described in the following text.

3.1. Estimation of the total number of NCs

This subsection describes the estimation of the total number of NCs in a lane at any ToI. In a signal-free lane, vehicles are expected to travel at the cruise speed, and the expected number of NCs in a lane at any time can be estimated by multiplying the average arrival rate of NCs by the travel time through the road segment. However, in a signal-controlled lane, vehicles must stop during red signals, resulting in additional vehicles on the road. These additional vehicles are holding vehicles that can be generically defined as undischarged vehicles at any ToI (Jia et al., 2024c). Thus, the total number of NCs in a lane can be estimated as the sum of the expected number of NCs and the number of holding NCs in the lane. For instance, considering the scenario illustrated in Fig. 2, if the traffic signal were absent, only vehicles 5, 6, and 7 would be present in the lane at the ToI. In contrast, with the traffic signal in place, two additional vehicles, vehicles 3 and 4, are held by the system as holding vehicles, resulting in a total of five vehicles in the lane at the ToI. Based on this analysis, Proposition 1 is presented for estimating the total number of NCs in a lane at any ToI.

Proposition 1. Given the average arrival rate, $\overline{q}$, average CV penetration rate, $\overline{p}$, number of holding vehicles, R, number of CVs in R, R_C, cruise speed, v_f, and lane length, l, the total number of NCs in a lane at any ToI, Q, can be estimated as

$$Q = \overline{q}(1-\overline{p})\frac{l}{v_f} + R - R_C. \tag{1}$$

Proof. A detailed proof is provided in **Appendix C**.

Based on Proposition 1, the total number of NCs in a lane at any ToI is obtained. The next step is to estimate the distribution of these NCs in the lane.

3.2. Estimation of the distribution of NCs

To define the spatial distribution of NCs, the entire lane is divided into multiple lane segments based on the locations of CVs. Subsequently, the number of NCs in each lane segment is estimated, and their exact locations within each lane segment are determined. All lane segments are categorized into two types: Type A lane segments, characterized by two stopped CVs or by a stopped CV and the stop bar during a red signal; and Type B lane segments, characterized by two CVs with at least one moving or by a moving CV and the upstream lane entrance. An illustration of different lane segments is shown in Fig. 2. At the ToI, three CVs are present in the lane, vehicles 3, 4, and 7, with vehicles 3 and 4 being stopped. Based on the definitions above, the lane segments between vehicle 3 and the stop bar, as well as between vehicles 3 and 4, are categorized as Type A lane segments. Meanwhile, the lane segments between vehicles 4 and 7, and between vehicle 7 and the lane entrance, are classified as Type B lane segments. In Type A lane segments, NCs are assumed to have zero speed and are uniformly distributed, as stated in **Assumption 2**. In contrast, in Type B lane segments, the speeds of NCs are assumed to follow a linear relationship, as outlined in **Assumption 3**, which serves as the basis for estimating their locations. Based on these principles, Proposition 2 is proposed to define the spatial distribution of NCs.

Proposition 2. Consider a given set of the locations and speeds of m CVs ($m > 0$) in a lane ordered from the stop bar to the upstream lane entrance, $\{(L_1,\ V_1), \ldots, (L_m,\ V_m)\}$, and boundary conditions (L_0, V_0) and (L_{m+1}, V_{m+1}), where $L_0 = l$ represents the location of the stop bar, and $L_{m+1} = 0$ represents the location of the upstream entrance, such that

$$V_0 = \begin{cases} v_f \ \textit{if green signal} \\ 0 \ \textit{if red signal} \end{cases} \tag{2}$$

and $V_{m+1} = v_f$. If the total number of NCs in the lane is Q, the lane length is l, and the effective vehicle length is l_e, the spatial distributions of NCs for Type A and B lane segments can be derived as follows:

A. If $V_i = 0 \cap V_{i+1} = 0, \forall i \in [0, m-1]$, the location set of NCs between L_i and L_{i+1} of the i^{th} Type A lane segment, denoted as $\left\{L_i^j\right\}^{Q_i^A}$, can be estimated as follows:

$$L_i^j = \begin{cases} L_1 + j\dfrac{l - L_1}{Q_0^A} \ \textit{if } i = 0 \\ L_{i+1} + j\dfrac{L_i - l_e - L_{i+1}}{Q_i^A} \ \textit{if } i \in [0, m-1] \end{cases}, \forall j \in \left[1, Q_i^A\right], \tag{3}$$

where Q_i^A represents the number of NCs between L_i and L_{i+1} in the i^{th} Type A lane segment with $V_i = 0 \cap V_{i+1} = 0$ and is defined as

$$Q_i^A = \begin{cases} \left\lfloor \dfrac{l - L_1}{l_e} \right\rfloor \ \textit{if } i = 0 \\ \left\lfloor \dfrac{L_i - l_e - L_{i+1}}{l_e} \right\rfloor \ \textit{if } i \in (0, m-1] \end{cases}; \tag{4}$$

and $\lfloor \cdot \rceil$ represents the rounding function.

B. If $V_i \neq 0 \cup V_{i+1} \neq 0, \forall i \in [0,m]$, the location set of NCs between L_i and L_{i+1} in the i^{th} Type B lane segment, denoted as $\left\{L_i^j\right\}^{Q_i^B}$, is estimated as follows:

$$L_i^j = \begin{cases} \frac{1}{2}\left(L_i^{(u)} - L_i^{(l)}\right) \ if\ Q_i^B = 1 \\ L_i^{(l)} + (j-1)\frac{L_i^{(u)} - L_i^{(l)}}{Q_i^B - 1} \ otherwise \end{cases}, \forall j \in \left[1, Q_i^B\right], \tag{5}$$

where

$$Q_i^B = \min\left\{\left\lfloor\left(Q - \sum_{\substack{i=0 \\ V_i=0 \ \cap \ V_{i+1}=0}}^{m-1} Q_i^A\right)\frac{\widetilde{Q}_i^B}{\sum_{\substack{i=0 \\ V_i\neq 0 \ \cup \ V_{i+1}\neq 0}}^{m} \widetilde{Q}_i^B}\right\rceil + e_{i-1}, \widetilde{Q}_i^B\right\}; \tag{6}$$

$$\widetilde{Q}_i^B = \begin{cases} \max\left\{\left\lfloor\frac{2(L_i - L_{i+1}) + (V_i - V_{i+1})\Delta t}{\Delta t(V_i + V_{i+1})}\right\rceil, 0\right\} \ if\ i \in \{0, m\} \\ \max\left\{\left\lfloor\frac{2(L_i - L_{i+1}) + (V_i - V_{i+1})\Delta t}{\Delta t(V_i + V_{i+1})}\right\rceil - 1, 0\right\} \ otherwise \end{cases}; \tag{7}$$

$$e_i = \begin{cases} \max\left\{\left\lfloor\left(Q - \sum_{\substack{i=0 \\ V_i=0 \ \cap \ V_{i+1}=0}}^{m-1} Q_i^A\right)\frac{\widetilde{Q}_i^B}{\sum_{\substack{i=0 \\ V_i\neq 0 \ \cup \ V_{i+1}\neq 0}}^{m} \widetilde{Q}_i^B}\right\rceil + e_{i-1} - \widetilde{Q}_i^B, 0\right\} \ if\ i \in [0,m] \\ 0 \ otherwise \end{cases}; \tag{8}$$

$$L_i^{(l)} = \begin{cases} 0 \ if\ i = m \\ L_{i+1} + \max\{V_{i+1}\Delta t, l_e\} \ otherwise \end{cases}; \tag{9}$$

$$L_i^{(u)} = \begin{cases} l \ if\ i = 0 \\ L_i - \max\left\{V_i^{Q_i^B}\Delta t, l_e\right\} \ otherwise \end{cases}; \tag{10}$$

$$V_i^j = \begin{cases} V_1 + j\frac{V_0 - V_1}{Q_0^B} \ if\ i = 0 \\ V_{i+1} + j\frac{V_i - V_{i+1}}{Q_i^B + 1} \ if\ i \in (0,m) \\ V_{m+1} + (j-1)\frac{V_m - V_{m+1}}{Q_m^B} \ if\ i = m \end{cases}, \forall j \in \left[1, Q_i^B\right]; \tag{11}$$

here, Q_i^B represents the number of NCs between L_i and L_{i+1} in the i^{th} Type B lane segment with $V_i \neq 0 \cup V_{i+1} \neq 0$; $\widetilde{Q}_i^B$ represents the maximum number of NCs that can be inserted between L_i and L_{i+1} in the i^{th} Type B lane segment; e_i is the number of NCs exceeding the maximum number of NCs that can be inserted between L_i and L_{i+1} in the i^{th} Type B lane segment; $L_i^{(l)}$ is the lower bound of the feasible space between L_i and L_{i+1} in the i^{th} Type B lane segment where NCs can be inserted; $L_i^{(u)}$ is the upper bound of the feasible space between L_i and L_{i+1} in the i^{th} Type B lane segment where NCs can be inserted; V_i^j represents the speed of the j^{th} inserted NC between L_i and L_{i+1} of the i^{th} Type B lane segment; $max\{\cdot,\cdot\}$ is the maximum function; $min\{\cdot,\cdot\}$ is the minimum function; and Δt represents the minimum safe time headway between two vehicles.

Proof. This proof involves the derivations of the spatial distribution of NCs for Type A and B lane segments, as visualized in Fig. 3. For Type A lane segments, NCs are in queues and are therefore compactly distributed. This necessitates only two sequential steps to determine the NC distribution. The first step is to estimate the number of NCs, followed by the identification of their locations based on the average effective vehicle length. This process is illustrated on the left-hand side of Fig. 3. In contrast, the derivation for Type B lane segments is more complex due to the presence of moving traffic. Four sequential steps are required. The process begins by estimating

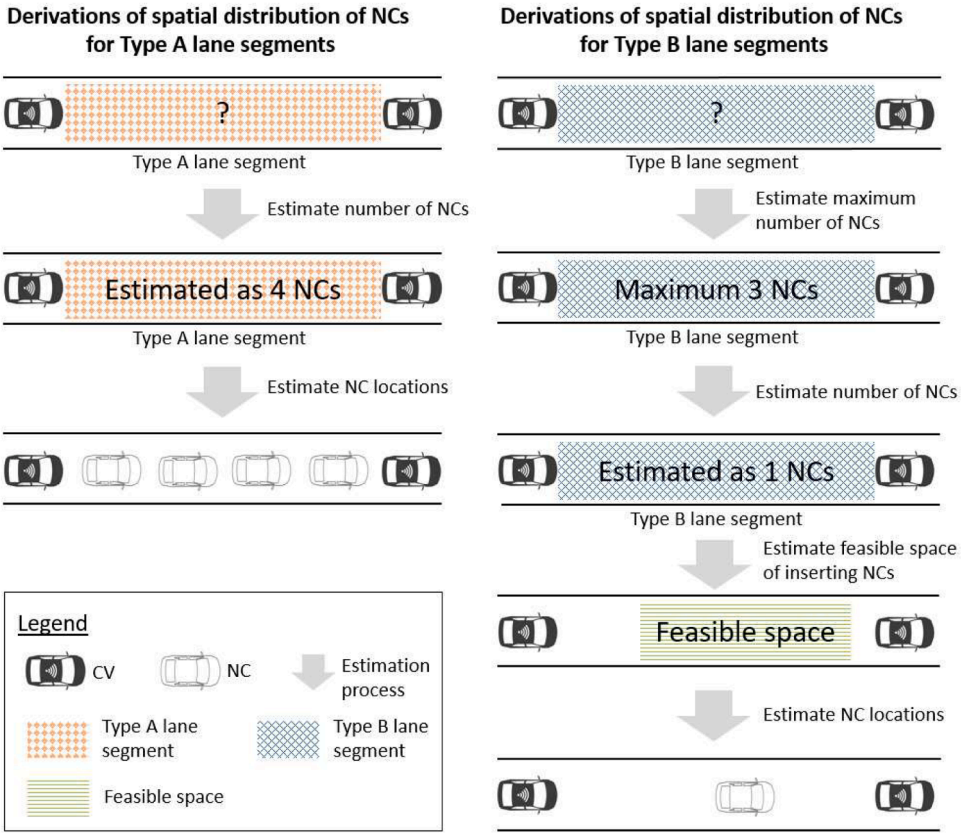

Fig. 3. Derivations of the spatial distribution of NCs for Type A and B lane segments.

the maximum number of NCs that can be inserted into Type B lane segments based on the minimum safe time headway. Next, considering the total number of NCs estimated in Section 3.1 and the allocation of some NCs to Type A lane segments, the remaining NCs are distributed to Type B lane segments. This is achieved by treating the maximum number of NCs as weighting factors and estimating the numbers of NCs in Type B lane segments using a scaling method. As the vehicles in Type B lane segments are moving, NCs are sparsely distributed within these segments. Feasible spaces for NC insertion are then derived in terms of the minimum safe time headway in the third step. Finally, the detailed locations of NCs within these feasible spaces are determined. The complete process is depicted on the right-hand side of Fig. 3. The mathematical derivations for the spatial distribution of NCs for both Type A and Type B lane segments are provided in **Appendix D**.

Proposition 2 provides a systematic and analytical approach to spatially distribute NCs in a lane when some CVs are observed. A special case arises when no CV is observed in the lane. Corollary 1 can be used to spatially distribute NCs in such situations.

Corollary 1. If no CV is observed in the lane, i.e., $m = 0$, the location set of Q NCs, $\{L^j\}^Q$, can be estimated as

$$L^j = \begin{cases} \dfrac{l}{2} \ \textit{if}\ Q = 1 \\ \dfrac{l(j-1)}{Q-1} \ \textit{otherwise} \end{cases}, \forall j \in [1, Q]. \tag{12}$$

Proof. The proof is completed by setting the set of the locations and speeds of CVs to an empty set in Proposition 2**.**

Proposition 2, along with Corollary 1, provides a comprehensive and generic approach for spatially distributing NCs in any given lane segment, ranging from 0 to l, of a source lane. This approach can be further generalized to any given lane segment of an intermediate lane to spatially distribute NCs, as discussed in the next section.

4. CVVL-I sub-model

This section presents the CVVL-I sub-model, which is used to estimate the locations of NCs in intermediate lanes using only CV trajectory data. While vehicle arrivals in source lanes are random, intermediate lanes receive vehicles discharged sequentially from upstream lanes. Consequently, multiple platoons can be formed within intermediate lanes, adding complexity to vehicle location

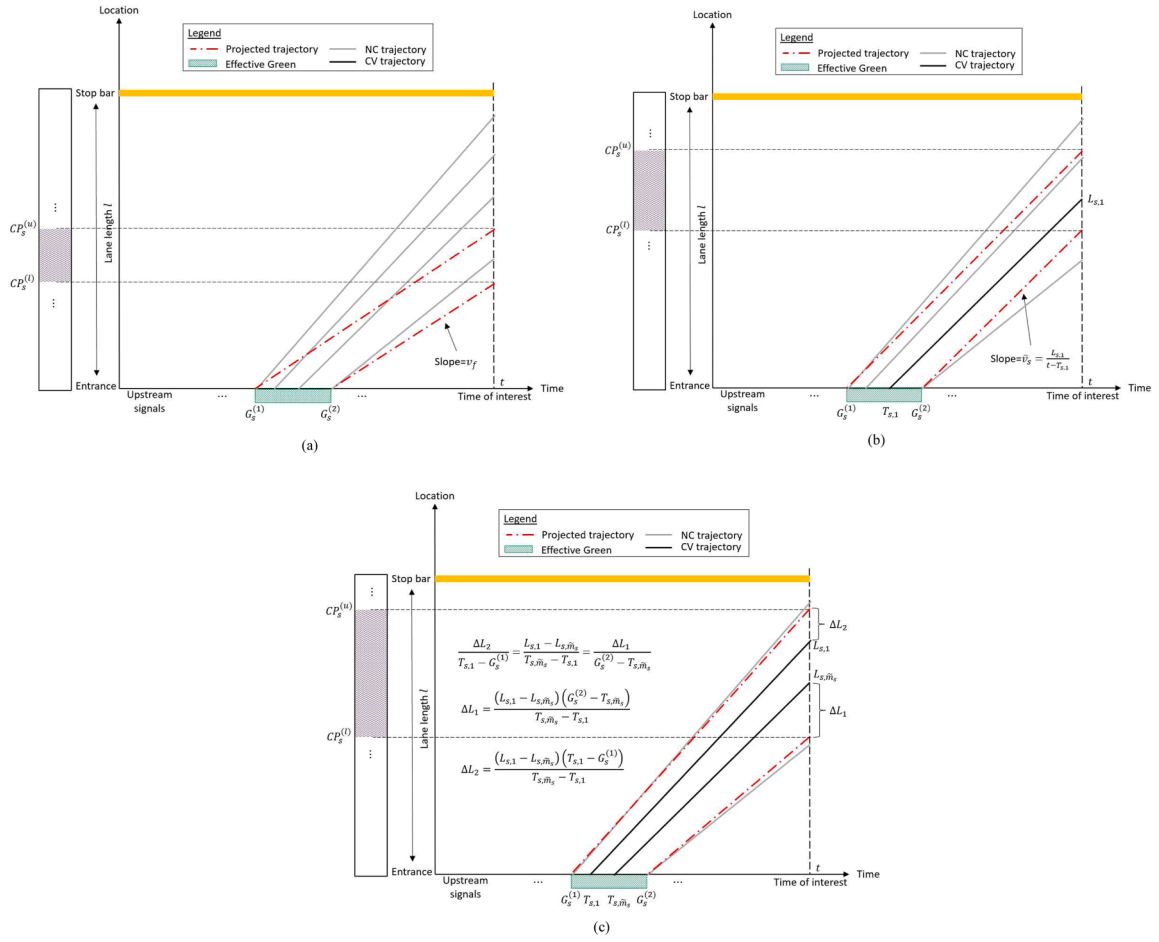

Fig. 4. Identification of CPs under various conditions: (a) without CVs, (b) with only one CV, and (c) with two or more CVs.

estimation. The proposed CVVL-I sub-model addresses this complexity by decomposing the task into three sequential sub-problems: identifying CPs, estimating the number of vehicles in each CP, and determining the spatial distribution of vehicles within each CP.

4.1. Identification of CPs

Vehicle arrivals in intermediate lanes are governed by the signal timing plan of their upstream lanes. Vehicles can only proceed from an upstream lane to a downstream intermediate lane during the green period for that upstream lane. To prevent conflicts, vehicles are discharged sequentially from upstream lanes to the downstream intermediate lane according to the signal timing plan at the upstream intersection. This sequential discharge can result in the formation of multiple platoons in the downstream intermediate lane, leading to an uneven spatial distribution of vehicles. This subsection describes the identification of the CPs formed owing to this process.

At any given time t of interest, CPs in an intermediate lane can be formed by holding vehicles or newly arrived vehicles from different upstream lanes. Each CP can be spatially defined by a lower bound and an upper bound. Here, $CP_s^{(l)}$ and $CP_s^{(u)}$ are defined as the lower and upper bounds of the s^{th} CP, respectively, $\forall s \in [0, S]$. If $s = 0$, the corresponding CP is formed by holding vehicles. The upper bound of the CP formed by holding vehicles can be determined assuming that this CP starts at the stop bar, while the lower bound can be derived based on the number of holding vehicles and average speeds of holding CVs. If $s \neq 0$, the corresponding CP is formed by newly arrived vehicles from the s^{th} upstream lane. Define $\left[G_s^{(1)}, G_s^{(2)}\right], \forall s \in [1, S]$, as the interval of the effective green period of the s^{th} upstream lane, which can be either part or all of its effective green period, during which vehicles discharged from the s^{th} upstream lane proceed to the downstream intermediate lane at the assumed cruise speed at time t of interest. The upper and lower bounds of the s^{th} CP can be estimated by the locations of these newly arrived vehicles to the downstream intermediate lane from the s^{th} upstream lane at time t of interest. However, as the assumption of cruise speed does not always hold, the bounds of the CP can be adjusted based on the available CV trajectory data.

Fig. 4 illustrates how CV trajectories are used to adjust the bounds of the CPs. In the absence of a CV observed within a CP, the

bounds are determined by projecting vehicles that enter the lane at the start and end of the green period, using the cruise speed. However, due to the lack of observation, the estimated CP bounds may be inaccurate, as shown in Fig. 4(a). Fig. 4(b) depicts a scenario where only one CV is observed within a CP. This observed CV provides valuable information regarding the actual speed of the platoon. By using the current location and entrance time of the CV, its average speed can be calculated, which then serves as the projection speed for determining the CP bounds. As a result, the bounds in Fig. 4(b) are more accurate compared with those in Figure 4(a). When two or more CVs are observed in a CP, the region between the first and last CVs becomes exact. The remaining regions before the first CV and after the last CV have to be estimated. With **Assumption 4**, simultaneous equations can be established to estimate the lengths of these regions, and the upper and lower bounds can then be obtained, as shown in Fig. 4(c), yielding more precise results than those in Fig. 4(a) and (b). Based on these principles, Proposition 3 is derived to determine the lower and upper bounds of the CPs.

Proposition 3. Given the interval of the effective green period of the s^{th} upstream lane, $\left[G_s^{(1)}, G_s^{(2)}\right], \forall s \in [1,S]$, during which vehicles discharged from the s^{th} upstream lane proceed to the downstream intermediate lane at the assumed cruise speed at time t of interest; the number of CVs traveling to the downstream intermediate lane in the s^{th} CP, $\widetilde{m}_s$, $\forall s \in [0,S]$; the entrance time of the i^{th} CV in the s^{th} CP, $T_{s,i}$, $\forall s \in [0,S]$; and the location of the i^{th} CV in the s^{th} CP at any time t of interest, $L_{s,i}$, $\forall s \in [0,S]$, the lower and upper bounds of the s^{th} CP, denoted by $CP_s^{(l)}$ and $CP_s^{(u)}$, respectively, can be derived as follows:

A. If $s = 0$, the lower and upper bounds of the 0^{th} CP, denoted by $CP_0^{(l)}$ and $CP_0^{(u)}$, respectively, can be derived as follows:

$$CP_0^{(l)} = \begin{cases} l - \left(\frac{g}{\tau} + R\right) l_e \ \text{if} \ \widetilde{m}_0 = 0 \\ \frac{l}{v_f \widetilde{m}_0} \sum_{i=1}^{\widetilde{m}_0} \frac{L_{0,i}}{t - T_{0,i}} \ \text{if} \ \widetilde{m}_0 > 0 \end{cases}, \tag{13}$$

$$CP_0^{(u)} = l, \tag{14}$$

where g represents the effective green after the effective red for the signal group controlling the downstream intermediate lane, τ represents the saturation headway, and R represents the number of holding vehicles in the downstream intermediate lane.

A. If $s > 0$, the lower and upper bounds of the s^{th} CP, denoted by $CP_s^{(l)}$ and $CP_s^{(u)}$, respectively, can be derived as follows:

$$CP_s^{(l)} = \begin{cases} \left(t - G_s^{(2)}\right) v_f \ \text{if} \ \widetilde{m}_s = 0 \\ \max\left\{\left(t - G_s^{(2)}\right) \frac{L_{s,1}}{t - T_{s,1}}, 0\right\} \ \text{if} \ \widetilde{m}_s = 1 \\ \max\left\{L_{s,\widetilde{m}_s} - \frac{\left(L_{s,1} - L_{s,\widetilde{m}_s}\right)\left(G_s^{(2)} - T_{s,\widetilde{m}_s}\right)}{T_{s,\widetilde{m}_s} - T_{s,1}}, 0\right\} \ \text{if} \ \widetilde{m}_s > 1 \end{cases}, \forall s \in [1,S], \tag{15}$$

$$CP_s^{(u)} = \begin{cases} \left(t - G_s^{(1)}\right) v_f \ \text{if} \ \widetilde{m}_s = 0 \\ \min\left\{\left(t - G_s^{(j)}\right) \frac{L_{s,1}}{t - T_{s,1}}, l\right\} \ \text{if} \ \widetilde{m}_s = 1 \\ \min\left\{L_{s,1} + \frac{\left(L_{s,1} - L_{s,\widetilde{m}_s}\right)\left(T_{s,1} - G_s^{(1)}\right)}{T_{s,\widetilde{m}_s} - T_{s,1}}, l\right\} \ \text{if} \ \widetilde{m}_s > 1 \end{cases}, \forall s \in [1,S]. \tag{16}$$

Proof. A detailed proof is provided in **Appendix E**.

Based on Proposition 3, the lower and upper bounds of CPs in an intermediate lane can be estimated. The identified CPs determine different lane segments to spatially distribute vehicles.

4.2. Estimation of numbers of NCs in CPs

This subsection describes the estimation of the number of NCs in each CP. The number of holding NCs in the 0^{th} CP can be derived by subtracting the number of holding CVs from the total number of holding vehicles. The number of newly arrived NCs in the s^{th} CP is the product of the number of NCs discharged from the s^{th}upstream lane during a specific interval of the effective green period, $\left[G_s^{(1)}, G_s^{(2)}\right], \forall s \in [1,S]$, and the turning proportion from the s^{th}upstream lane to the downstream intermediate lane, β_s, as illustrated in Fig. 5. $\widetilde{Q}_s$ can be estimated using queue information along with the average arrival rate and CV penetration rate in the upstream lane, while β_s can be estimated by dividing the number of CVs traveling to the target intermediate lane from the upstream lane by the total number of CVs in the upstream lane over the past few cycles. Based on these principles, Proposition 4 is derived to estimate the number of NCs in

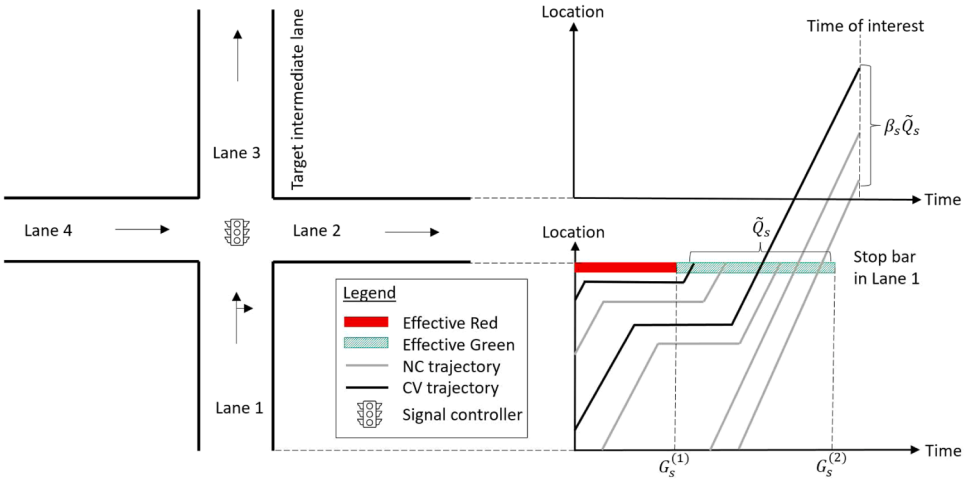

Fig. 5. Estimation of the number of NCs in a CP.

CPs.

Proposition 4. Given the interval of the effective green period of the s^{th} upstream lane, $\left[G_s^{(1)}, G_s^{(2)}\right], \forall s \in [1, S]$, during which vehicles discharged from the s^{th} upstream lane proceed to the downstream intermediate lane at the assumed cruise speed at time t of interest; the average CV penetration rate of the s^{th} upstream lane, $\overline{p}_s$; the average arrival rate of the s^{th} upstream lane, $\overline{q}_s$; the number of CVs traveling to the downstream intermediate lane from the s^{th} upstream lane in cycle k, $\widetilde{m}_s^k$; the total number of CVs from the s^{th} upstream lane in cycle k, m_s^k; the number of past cycles considered, ω; the turning proportion from the s^{th} upstream lane to the downstream intermediate lane, β_s; the number of holding vehicles in the downstream intermediate lane at any time t of interest, R; and the number of holding CVs in the downstream intermediate lane at any time t of interest, R_C, the number of NCs in the s^{th} CP of the downstream intermediate lane in the range $\left[CP_s^{(l)}, CP_s^{(u)}\right]$, Q_s, $\forall s \in [0, S]$, is estimated as follows:

A. If $s = 0$,

$$Q_0 = R - R_C. \tag{17}$$

B. If $s > 0$,

$$Q_s = \left\lfloor \beta_s \left(\widetilde{Q}_s^{(2)} - \widetilde{Q}_s^{(1)} \right) \right\rceil, \tag{18}$$

where

$$\beta_s = \frac{\sum_{k=1}^{\omega} \widetilde{m}_s^k}{\sum_{k=1}^{\omega} m_s^k}, \tag{19}$$

$\forall w \in \{1, 2\}$, $\widetilde{Q}_s^{(w)}$ representing the number of NCs discharged from the s^{th} upstream lane between time period $\left[g_s - r_s, G_s^{(w)}\right]$ can be estimated as

$$\widetilde{Q}_s^{(w)} = \begin{cases} \max\left\{ \overline{q}_{Ns} \left(G_s^{(w)} - g_s + r_s \right) - R_{Ns}^{(w)}, 0 \right\} \text{ if } \widetilde{n}_s^{(w)} = 0 \\ \max\left\{ \dfrac{l_s - L_s'^{(w)}}{l_e} + 1 - \widetilde{n}_s^{(w)} + \overline{q}_{Ns} \left(G_s^{(w)} - \dfrac{l_s}{v_f} - T_s'^{(w)} \right) - R_{Ns}^{(w)}, 0 \right\} \text{ if } \widetilde{n}_s^{(w)} > 0 \end{cases}, \tag{20}$$

$$R_{Ns}^{(w)} = R_s^{(w)} - R_{Cs}^{(w)}, \tag{21}$$

$$\overline{q}_{Ns} = \overline{q}_s (1 - \overline{p}_s), \tag{22}$$

$\widetilde{n}_s^{(w)}$ represents the number of CVs that stop between $g_s - r_s$ and $G_s^{(w)}$ in the s^{th} upstream lane, g_s represents the start of the effective green of the s^{th} upstream lane, r_s represents the duration of the effective red of the s^{th} upstream lane, $R_s^{(w)}$ represents the number of holding vehicles at $G_s^{(w)}$ in the s^{th} upstream lane, $R_{Cs}^{(w)}$ represents the number of holding CVs at $G_s^{(w)}$ in the s^{th} upstream lane, $R_{Ns}^{(w)}$ represents the number of holding NCs at $G_s^{(w)}$ in the s^{th} upstream lane, l_s represents the length of the s^{th} upstream lane, $L_s'^{(w)}$ represents the stopping location of the last stopped CV between $g_s - r_s$ and $G_s^{(w)}$ in the s^{th} upstream lane, and $T_s'^{(w)}$ represents the entrance time of the last stopped CV between $g_s - r_s$ and $G_s^{(w)}$ in the s^{th} upstream lane.

Algorithm 1
Merging adjacent CPs with overlaps.

Input: $\left[CP_s^{(l)}, CP_s^{(u)}\right]$ and Q_s, $\forall s \in [0, S]$
Initialization: $CP_m = \{\}$, $Q_m = \{\}$
For s in $\{0, \ldots, S\}$ **do**
  If $s = 0$ **then**
    $CP_{m,temp} = \left[CP_s^{(l)}, CP_s^{(u)}\right]$
    $Q_{m,temp} = [Q_s]$
  Else
    If $CP_s^{(u)} \geq \min\left(CP_{m,temp}\right)$ **then**
      $CP_{m,temp} \leftarrow CP_{m,temp} + \left[CP_s^{(l)}, CP_s^{(u)}\right]$
      $Q_{m,temp} \leftarrow Q_{m,temp} + [Q_s]$
    Else
      $CP_m \leftarrow CP_m + [\min(CP_{m,temp}), \max(CP_{m,temp})]$
      $Q_m \leftarrow Q_m + \text{Sum}(Q_{m,temp})$
      $CP_{m,temp} = \left[CP_s^{(l)}, CP_s^{(u)}\right]$
      $Q_{m,temp} = [Q_s]$
    End if
    If $s = S$ **then**
      $CP_m \leftarrow CP_m + [\min(CP_{m,temp}), \max(CP_{m,temp})]$
      $Q_m \leftarrow Q_m + \text{Sum}(Q_{m,temp})$
    End if
  End if
End for
Output: CP_m and Q_m

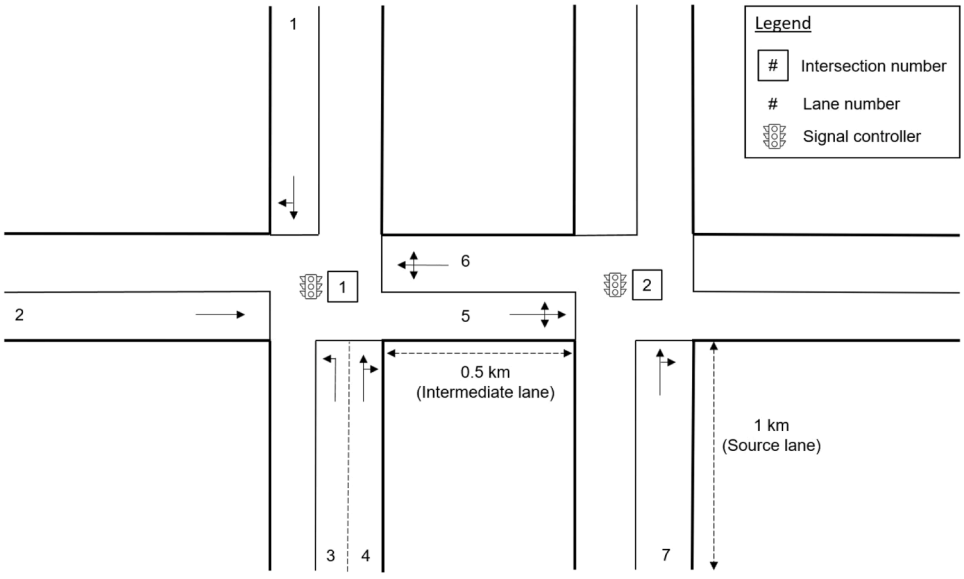

Fig. 6. Illustration of the two-signalized controlled intersection network.

Proof. A detailed proof is provided in **Appendix F**.

Using the information from the upstream lanes, Proposition 4 estimates the numbers of NCs in CPs within the downstream intermediate lane. Given the CPs and number of NCs in each CP, the final sub-problem is to spatially distribute these NCs in the CPs, as detailed in the next subsection.

4.3. Distributions of NCs in CPs

Given all the lower and upper bounds of CPs, $\left[CP_s^{(l)}, CP_s^{(u)}\right]$, and the corresponding numbers of NCs in CPs, $Q_s, \forall s \in [0, S]$, the final sub-problem is to distribute these NCs within each CP. Before estimating the NC locations, the potential overlap issue between adjacent CPs needs to be addressed. To handle this issue, a merging algorithm, as outlined in Algorithm 1, is adopted. The principle is to merge adjacent CPs with overlaps into a single CP. The number of NCs in the merged CP is the sum of all NCs in the CPs prior to merging. After merging, Proposition 2 and Corollary 1 are used to estimate the locations of NCs in each CP by adjusting the lane segment range to that

Table 1
Results of vehicle location estimation in the source lane using the CVVL-S sub-model.

No.	r	V/C	ρ	ToI	Precision (%)			Recall (%)			F_1 (%)		
					EVLS	CVVL-S	↑	EVLS	CVVL-S	↑	EVLS	CVVL-S	↑
Baseline	30	0.5	0.4	EoR	47	58	+11	71	76	+5	57	66	+9
A-1	15	0.5	0.4	EoR	45	55	+10	66	72	+6	54	63	+9
A-2	45	0.5	0.4	EoR	49	63	+14	71	74	+3	58	68	+10
B-1	30	0.3	0.4	EoR	45	56	+11	66	72	+6	54	63	+9
B-2	30	0.7	0.4	EoR	52	62	+10	74	76	+2	61	68	+7
C-1	30	0.5	0.1	EoR	23	32	+9	32	39	+7	26	35	+9
C-2	30	0.5	0.7	EoR	63	81	+18	85	85	+0	72	83	+11
D-1	30	0.5	0.4	MoR	43	51	+8	65	74	+9	52	60	+8
D-2	30	0.5	0.4	EoG	43	46	+3	63	70	+7	51	56	+5
D-3	30	0.5	0.4	MoG	43	47	+4	63	71	+8	52	56	+4

of the corresponding CP. This completes the CVVL-I sub-model.

5. VISSIM simulations

Comprehensive and realistic simulation experiments were conducted using the VISSIM platform to demonstrate the effectiveness and applicability of the proposed models in real-time vehicle location estimation.

5.1. General settings

A two-signalized controlled intersection network (Fig. 6) was established for the simulation experiments. Lane 7, spanning 1 km, was designated as a source lane with random vehicle arrival. Lane 5, covering 0.5 km, served as an intermediate lane receiving vehicles discharged sequentially from upstream lanes 2 and 4. Both intersections operated with a common cycle length of 60 s and employed a signal structure of red–green–amber, featuring a 3-s amber period and 5-s clearance time. At Intersection 1, a four-group signal plan was implemented, where signal groups 1, 2, 3, and 4 controlled lanes 2, 3, 1 and 4, and 6, respectively. The green periods for signal groups 1, 2, 3, and 4 were set to be 20, 5, 10, and 5 s, respectively. This signal plan at Intersection 1 remained unchanged throughout the experiment. However, Intersection 2 employed different signal timing plans to explore different testing scenarios for lanes 7 and 5. All generated vehicles were cars. Each vehicle was randomly assigned to be a CV with a probability of ρ or an NC with a probability of $1-\rho$, where ρ is the average CV penetration rate and is unknown during estimation. The cruise speed was set as 50 km/h for all lanes. The effective vehicle length and saturation headway were found to be 6.44 m and 1.59 s, respectively.

Lane 7, serving as a source lane, was used to test the CVVL-S sub-model. Various signal timing plans controlling lane 7, volume-to-capacity (V/C) ratios, CV penetration rates, and ToIs of a cycle were considered to assess the robustness of the proposed model. Specifically, signal timing plans with red periods r of 15 s, 30 s, and 45 s were evaluated. The selected V/C ratios were 0.3, 0.5, and 0.7. The CV penetration rates were 0.1, 0.4, and 0.7. The ToI was chosen to be the end of red (EoR), middle of red (MoR), end of green (EoG), or middle of green (MoG) of a cycle. For each testing scenario, 1000 cycles were simulated and recorded using VISSIM after a 30-cycle warm up period. At each ToI, the CVVL-S sub-model was applied to estimate the locations of NCs in lane 7. For comparison, the EVLS algorithm (Feng et al., 2015) was also used for the estimation. The estimated locations of NCs from both methods were compared with the ground truths for evaluation.

Lane 5, serving as an intermediate lane and fed by upstream lanes 2 and 4 at Intersection 1, was used to evaluate the CVVL-I sub-model. The traffic demand ratio from upstream lane 2 to lane 4 was set as 2 to simulate an asymmetric traffic demand for lane 5. As traffic from lanes 1, 3, and 6 was irrelevant to lane 5, the associated traffic demands were randomly configured. Similar to the testing scenarios considered for the CVVL-S sub-model, various signal timing plans controlling lane 5, V/C ratios, CV penetration rates, and ToIs of a cycle were tested to evaluate the robustness of the proposed CVVL-I sub-model. For each scenario, 1000 cycles were simulated and recorded using VISSIM after a 30-cycle warm up period. At each ToI, both the CVVL-I sub-model and EVLS algorithm were used to estimate the locations of NCs. The estimated locations were then compared with the ground truths for evaluation.

5.2. Evaluation metrics

To evaluate the performance of the proposed CVVL model and EVLS algorithm, three evaluation metrics were selected: precision, recall rate, and F_1 score. These metrics are based on true positive (TP), false positive (FP), and false negative (FN), which can be defined as follows:

- TP: The number of instances where the model correctly identifies the presence of vehicles, with estimated locations falling within or equal to a specified threshold (set as 10 m in this experiment) of the ground truth locations.
- FP: The number of instances where the model incorrectly identifies the presence of vehicles, or where estimated locations exceed the specified threshold compared with the ground truth locations.

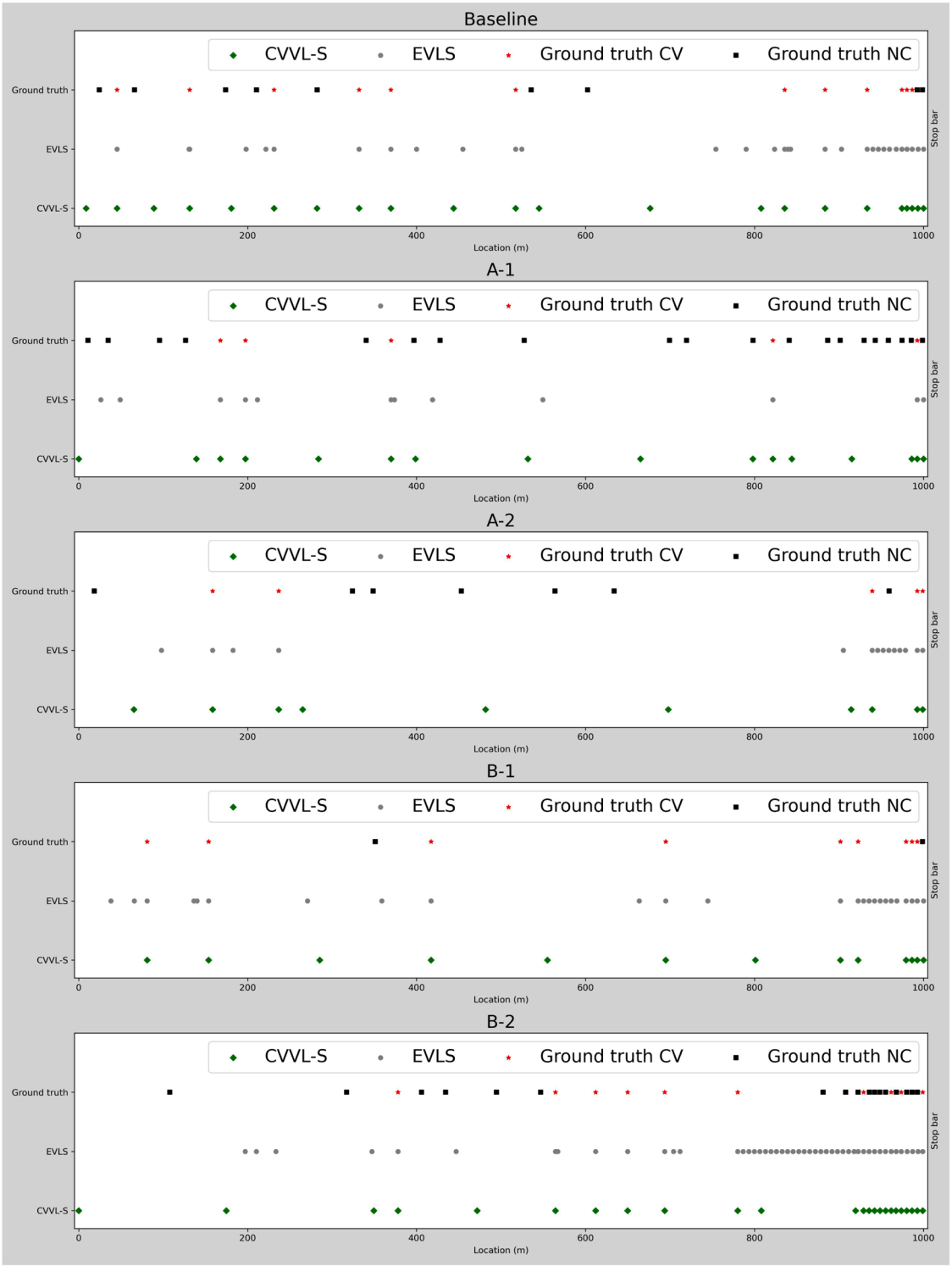

Fig. 7. Results of vehicle location estimations in the source lane using the proposed CVVL-S sub-model and EVLS algorithm for randomly selected cycles in Baseline, A-1, A-2, B-1, and B-2 cases.

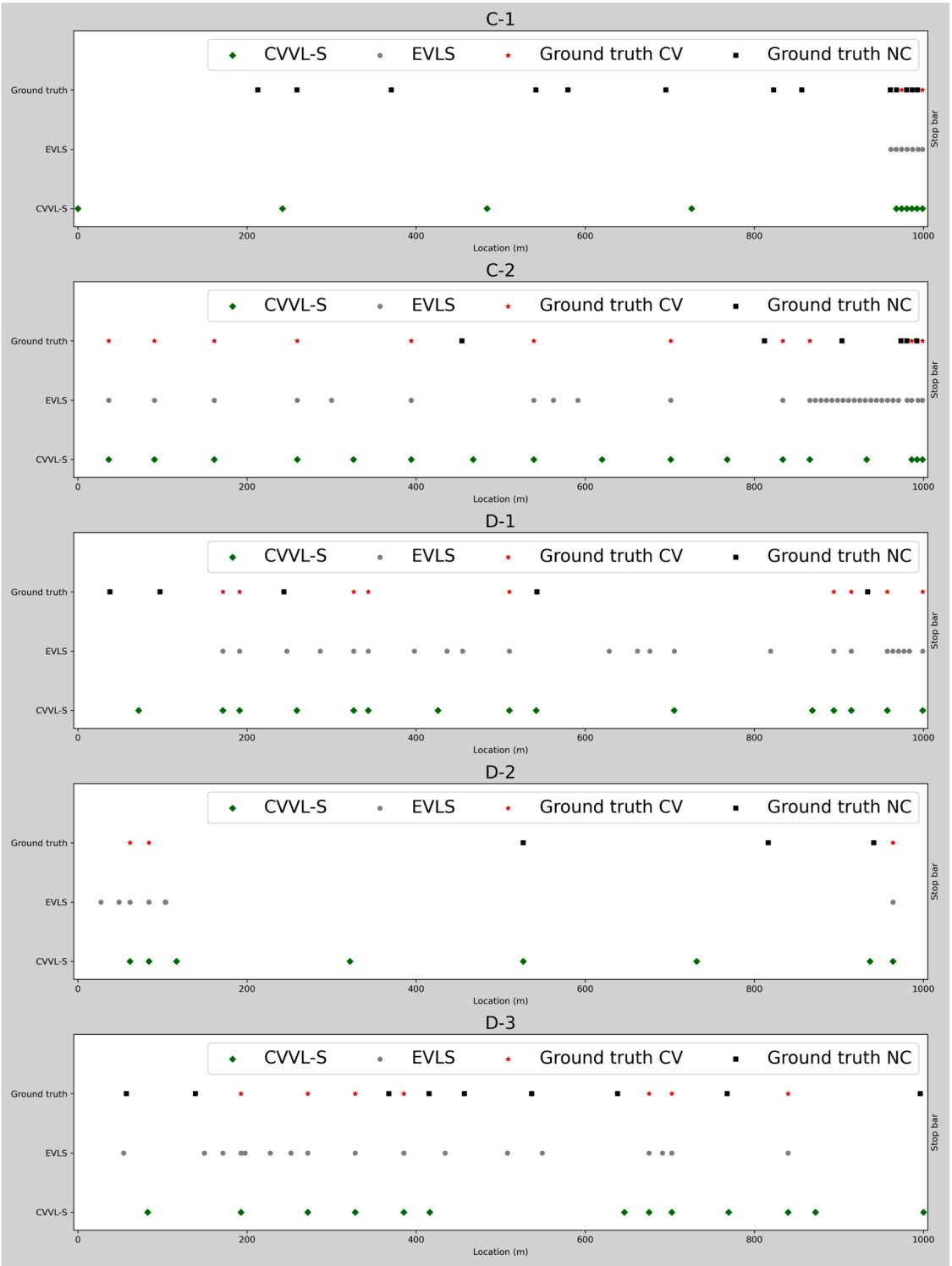

Fig. 8. Results of vehicle location estimations in the source lane using the proposed CVVL-S sub-model and EVLS algorithm for randomly selected cycles in C-1, C-2, D-1, D-2, and D-3 cases.

Table 2
Results of location estimation in the intermediate lane using the CVVL-I sub-model.

No.	r	V/C	ρ	ToI	Precision (%)			Recall (%)			F_1 (%)		
					EVLS	CVVL-I	↑	EVLS	CVVL-I	↑	EVLS	CVVL-I	↑
Baseline	30	0.5	0.4	EoR	33	63	+30	70	78	+8	45	70	+25
A-1	15	0.5	0.4	EoR	46	61	+15	62	77	+15	53	68	+15
A-2	45	0.5	0.4	EoR	55	61	+6	69	80	+11	61	70	+9
B-1	30	0.3	0.4	EoR	30	57	+27	63	80	+17	41	67	+26
B-2	30	0.7	0.4	EoR	37	66	+29	74	79	+5	49	72	+23
C-1	30	0.5	0.1	EoR	25	46	+21	28	52	+24	26	49	+23
C-2	30	0.5	0.7	EoR	34	79	+45	87	87	+0	49	83	+34
D-1	30	0.5	0.4	MoR	45	55	+10	61	77	+16	52	65	+13
D-2	30	0.5	0.4	EoG	42	42	+0	55	75	+20	47	54	+7
D-3	30	0.5	0.4	MoG	45	51	+6	60	70	+10	51	59	+8

§ FN: The number of instances where the model fails to identify the presence of vehicles that are actually present in the ground truth, indicating missed detections.

The three chosen metrics, derived from TP, FP, and FN, formed the basis for evaluating the effectiveness of the models in vehicle location estimation. These metrics can be computed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{23}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{24}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{25}$$

5.3. Results

Table 1 presents the results of vehicle location estimation in the source lane (lane 7) using the CVVL-S sub-model and EVLS algorithm. ↑ represents performance improvement of the CVVL model compared with the EVLS algorithm (in percentage). In the baseline case, with a red period r of 30 s, V/C ratio of 0.5, CV penetration rate of 0.4, and ToI at EoR, the CVVL-S sub-model outperformed the EVLS algorithm by 11 %, 5 %, and 9 % in terms of the precision, recall rate, and F_1, respectively. These significant improvements demonstrate the effectiveness and superiority of the CVVL-S sub-model. Furthermore, comprehensive sensitivity analysis was conducted. The results of various cases of groups A, B, C, and D (Table 1) demonstrated the effects of various signal timing plans, V/C ratios, CV penetration rates, and ToIs on the estimation performance of the proposed model and EVLS algorithm. The consistent and remarkable improvements achieved by the CVVL-S sub-model further confirm the applicability and generalizability of the proposed model.

Figs. 7 and 8 illustrate the results of vehicle location estimations in the source lane using the proposed CVVL-S sub-model and EVLS algorithm for randomly selected cycles in each case. The results indicate that the CVVL-S sub-model could accurately estimate vehicle locations, significantly outperforming the EVLS algorithm. Notably, the EVLS algorithm tended to erroneously insert more vehicles in slow-down regions, whereas the CVVL-S sub-model provided more reasonable estimations. These findings further affirm the superiority of the CVVL-S sub-model.

Table 2 presents the results of vehicle location estimation in the intermediate lane (lane 5) using the CVVL-I sub-model and EVLS algorithm. The EVLS algorithm does not account for differences in vehicle arrival patterns between source and intermediate lanes, failing to capture the discrete formation of CPs in the intermediate lane due to sequential discharges from upstream lanes. In contrast, the CVVL-I sub-model effectively models these patterns. Consequently, the CVVL-I sub-model significantly outperformed the EVLS algorithm in cases with an intermediate lane. Compared with the results in Table 1, the pronounced improvements across all cases presented in Table 2 further demonstrate the effectiveness and superiority of the CVVL-I sub-model.

Figs. 9 and 10 show the results of vehicle location estimations in the intermediate lane using the proposed CVVL-I sub-model and EVLS algorithm for randomly selected cycles in each case. As expected, the platooning effect of vehicles was evident in the ground truth vehicle distributions. The CVVL-I sub-model accurately estimated various vehicle platoons, while the EVLS algorithm could not model this platooning effect, as in the case of source lanes. Notably, this platooning effect is crucial for determining the vehicle arrival pattern at the stop bar and then optimizing traffic signal control. These observations further confirm the effectiveness of the CVVL-I sub-model. The impact of the accuracy of average arrival rate and CV penetration rate estimation on the CVVL model was also investigated. Due to the inherent trade-off between precision and recall, these metrics exhibited opposite trends as expected. However, their variations remained within acceptable limits. In contrast, the F_1 score, which balances precision and recall, serves as the most representative metric of the robustness of the proposed model. The consistently stable F_1 score across varying input parameters indicates the model's robustness or insensitivity to inaccuracies in the input parameters. Additional details are provided in **Appendix G**.

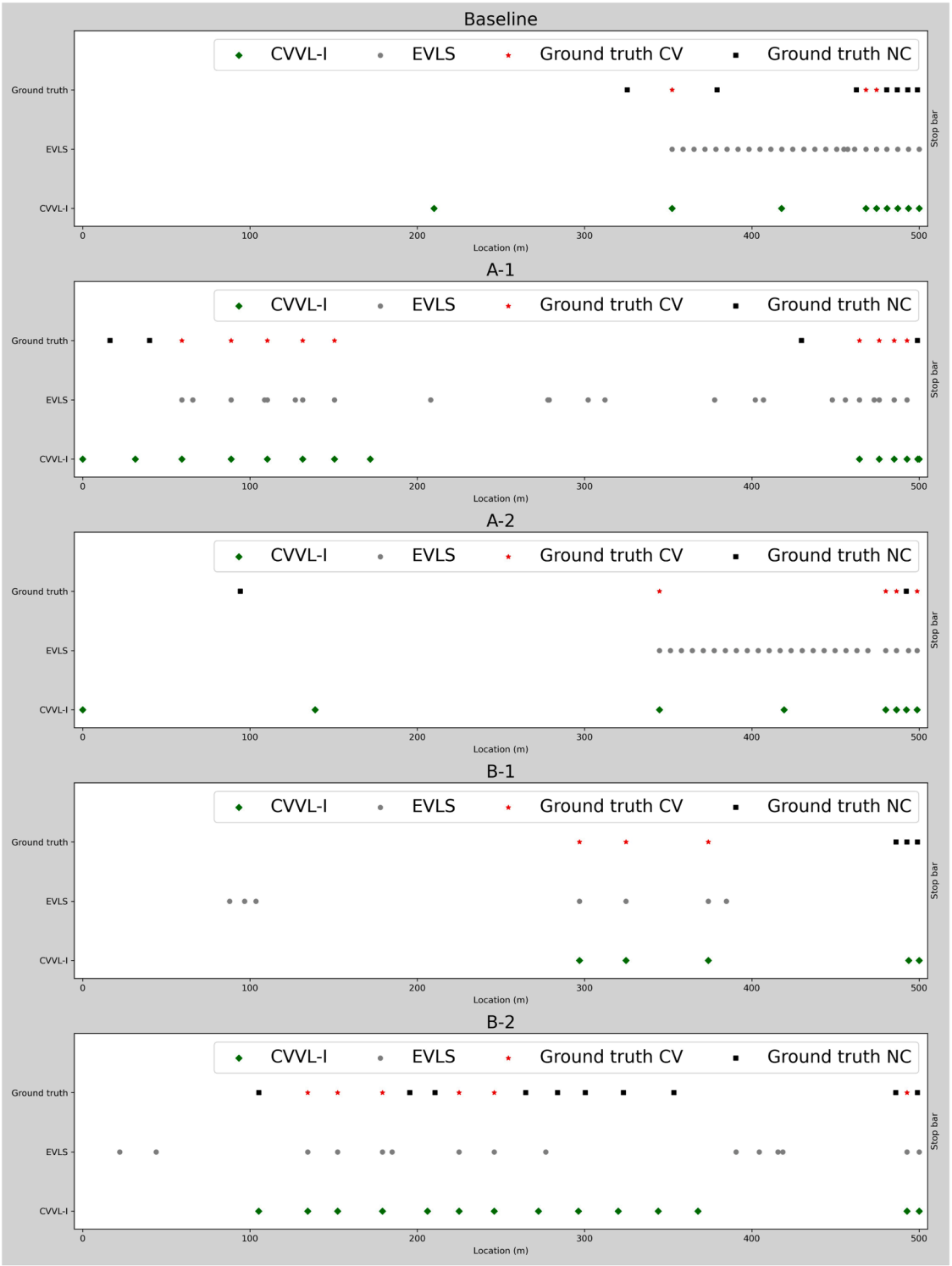

Fig. 9. Results of vehicle location estimations in the intermediate lane using the proposed CVVL-I sub-model and EVLS algorithm for randomly selected cycles in Baseline, A-1, A-2, B-1, and B-2 cases.

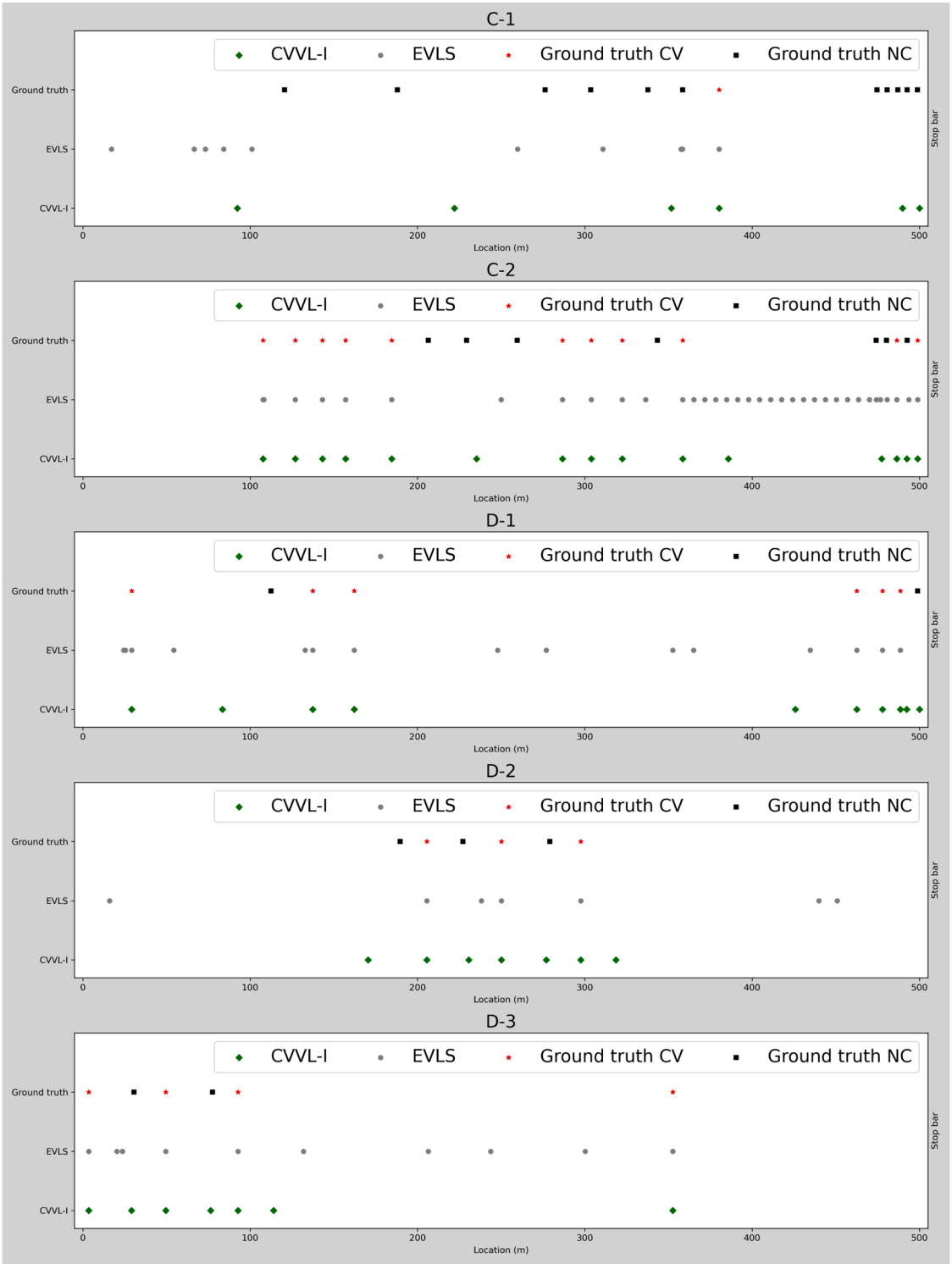

Fig. 10. Results of vehicle location estimations in the intermediate lane using the proposed CVVL-I sub-model and EVLS algorithm for randomly selected cycles in C-1, C-2, D-1, D-2, and D-3 cases.

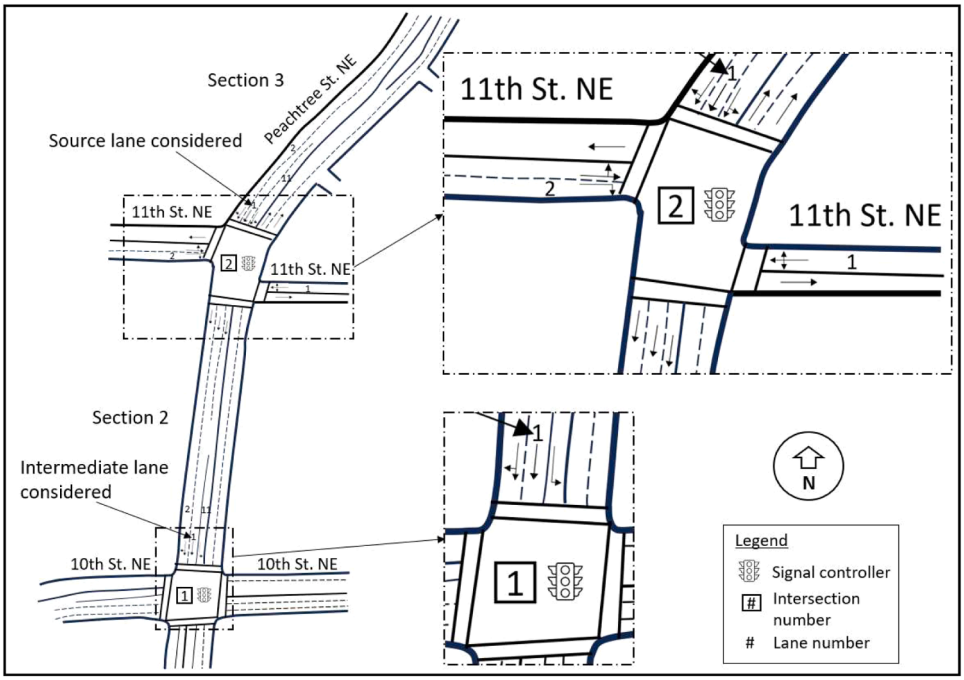

Fig. 11. Schematic representation of the selected network from the NGSIM dataset.

Table 3
Results of vehicle location estimation using the CVVL-S sub-model on the NGSIM dataset.

Period	No.	ρ	ToI	Precision (%)			Recall (%)			F_1 (%)		
				EVLS	CVVL-S	↑	EVLS	CVVL-S	↑	EVLS	CVVL-S	↑
12:45	Baseline	0.4	EoR	38	51	+13	57	64	+7	45	56	+11
-	C-1	0.1	EoR	13	23	+10	14	51	+37	14	31	+17
13:00	C-2	0.7	EoR	50	71	+21	72	76	+4	59	74	+15
	D-1	0.4	MoR	42	51	+9	45	61	+16	44	55	+11
	D-2	0.4	EoG	37	39	+2	47	67	+20	42	50	+8
	D-3	0.4	MoG	40	44	+4	56	60	+4	47	51	+4
16:00	Baseline	0.4	EoR	59	78	+19	70	74	+4	64	76	+12
-	C-1	0.1	EoR	36	48	+12	44	64	+20	40	54	+14
16:15	C-2	0.7	EoR	67	88	+21	85	87	+2	75	88	+13
	D-1	0.4	MoR	48	60	+12	68	79	+11	56	68	+12
	D-2	0.4	EoG	NaN	0	-	NaN	NaN	-	NaN	NaN	-
	D-3	0.4	MoG	41	49	+8	61	61	+0	49	54	+5

Overall, the results demonstrate significant advancements in vehicle location estimation using only partial CV trajectory data, highlighting the potential of the proposed CVVL model for transport modeling and system optimization applications.

6. Real-world validations

This section applies the proposed CVVL model to the real-world NGSIM dataset to validate its practical applicability.

6.1. General settings

A two-junction signalized network from the NGSIM dataset, located on Peachtree Street in Atlanta, Georgia, USA, was selected for this study, as shown in Fig. 11. In this network, the southbound through lanes (i.e., lane 1) in Road Sections 2 and 3 were designated as the target intermediate and source lanes, respectively. The upstream lanes for lane 1 in Road Section 2 included signalized lane 1 in Road Section 3 and control-free lanes 1 and 2 on 11th Street. Lane-changing behavior was not considered. The lengths of lane 1 in Road Sections 2 and 3 were approximately 127 m and 122 m, respectively. The cruise speed for both lanes was approximately 10.5 m/s. Trajectory data for the designated network were extracted for two periods: from 12:45 to 13:00 and from 16:00 to 16:15 on November 8, 2006. The common cycle lengths for Intersections 1 and 2 during these periods were 95 s and 100 s, respectively. The red durations

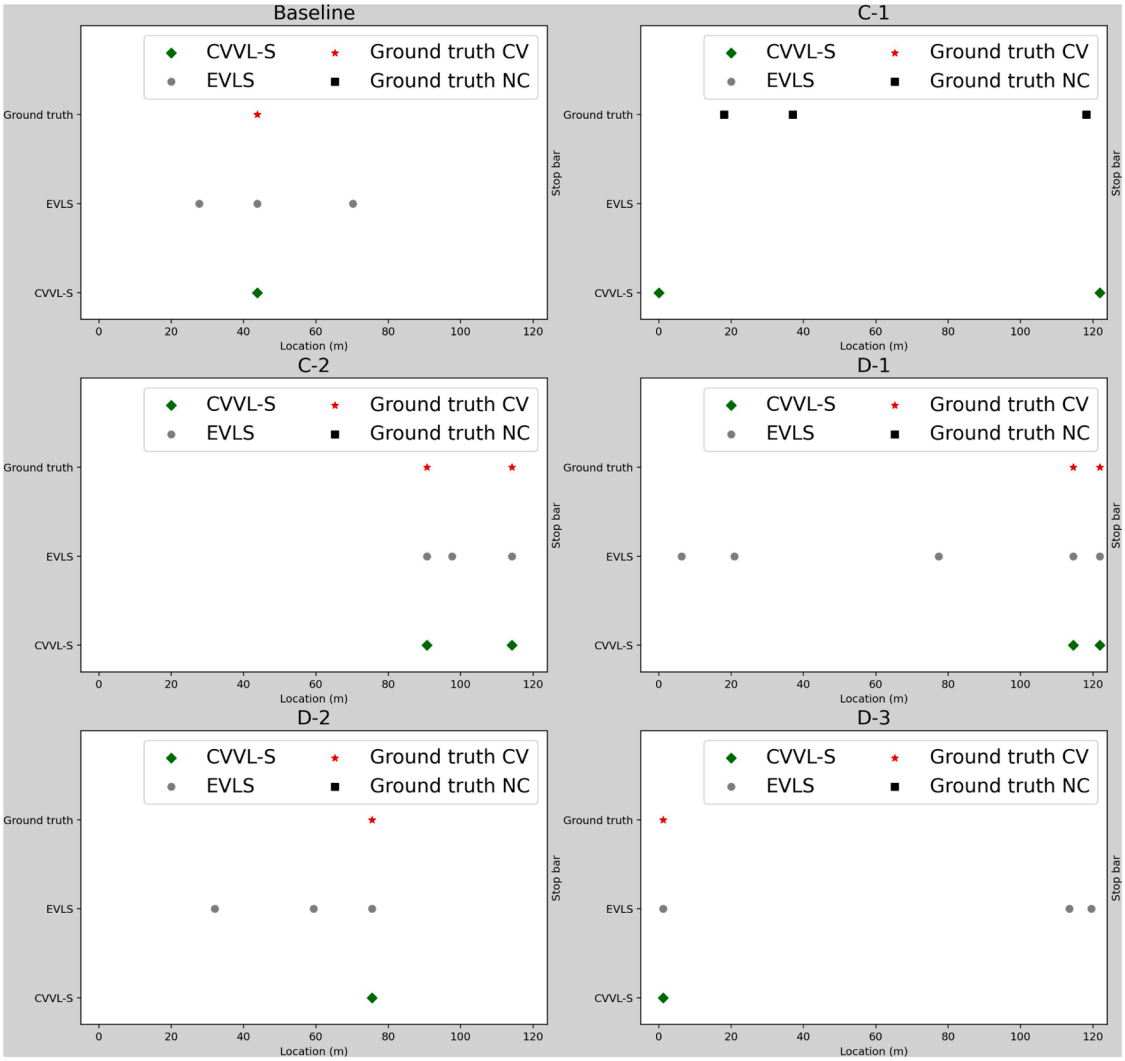

Fig. 12. Vehicle location estimation using the CVVL-S sub-model for the period from 12:45 to 13:00.

for lane 1 in Road Section 2 during these periods were 62 s and 64 s, respectively, while the red durations for lane 1 in Road Section 3 were 40 s. The saturation headway and average effective vehicle length for both lanes were determined to be 2.04 s and 8.47 m, respectively.

To simulate the CV environment, each vehicle in the dataset was randomly assigned as either a CV or an NC, with probabilities p_c or $1 - p_c$, respectively. Here, p_c represents the underlying true CV penetration rate, which was set to 0.1, 0.4, and 0.7. Similar to the VISSIM simulation experiments, various ToIs, including EoR, MoR, EoG, and MoG, were considered. As the data for each period include only nine complete cycles, and the most recent three cycles were used to estimate the average CV penetration rate and average arrival rate, only seven cycles remained available for validation. This resulted in only seven estimates for each case. The random assignment of vehicle identities could significantly affect the performance assessment. To address this issue and ensure a more accurate and representative evaluation of the estimation performance, the identity assignment process was conducted 100 times using different random seeds, resulting in data from a total of 700 cycles for evaluation.

The CVVL-S and CVVL-I sub-models were applied to estimate the real-time vehicle locations for source lane 1 in Road Section 3 and intermediate lane 1 in Road Section 2, using partial CV data, respectively. As there was no traffic from upstream lanes 1 and 2 on 11th Street to intermediate lane 1 in Road Section 2 during the two periods under consideration, and because signalized upstream lanes are required for the CVVL-I sub-model, only signalized lane 1 in Road Section 3 was considered the effective upstream lane for intermediate lane 1 in Road Section 2 in these experiments. Additionally, the EVLS algorithm was utilized as a benchmark for vehicle location estimation. The estimated vehicle locations were then compared with the ground truth data to compute the relevant evaluation metrics.

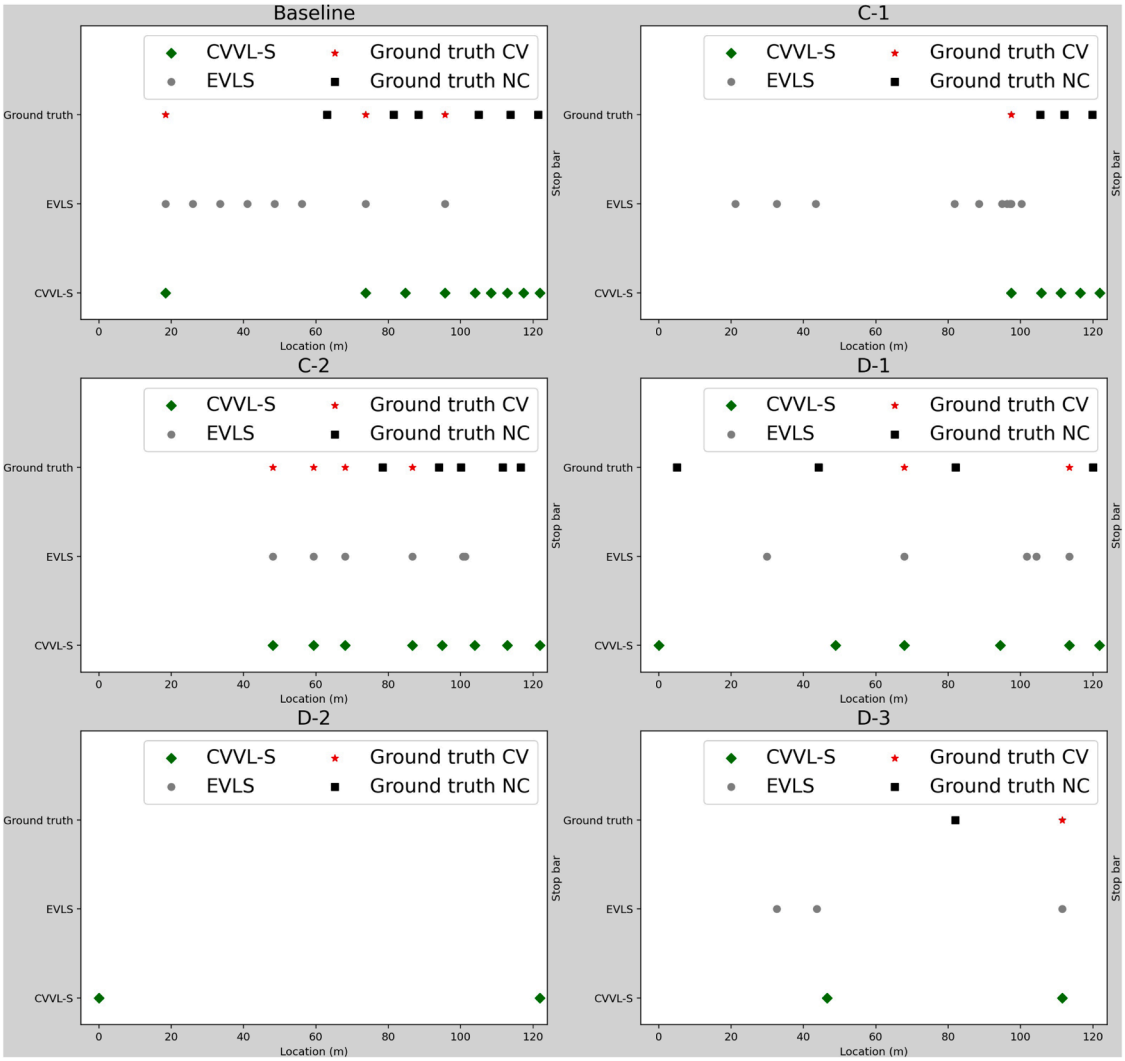

Fig. 13. Vehicle location estimation using the CVVL-S sub-model for the period from 16:00 to 16:15.

Table 4
Results of vehicle location estimation using the CVVL-I sub-model on the NGSIM dataset.

Period	No.	ρ	ToI	Precision (%)			Recall (%)			F_1 (%)		
				EVLS	CVVL-I	↑	EVLS	CVVL-I	↑	EVLS	CVVL-I	↑
12:45	Baseline	0.4	EoR	59	67	+8	67	89	+22	63	77	+14
-	C-1	0.1	EoR	40	45	+5	31	66	+35	35	54	+19
13:00	C-2	0.7	EoR	67	79	+12	85	93	+8	75	85	+10
	D-1	0.4	MoR	51	73	+22	67	86	+19	58	79	+21
	D-2	0.4	EoG	39	63	+24	52	67	+15	44	65	+21
	D-3	0.4	MoG	47	70	+23	63	65	+2	54	67	+13
16:00	Baseline	0.4	EoR	66	81	+15	72	83	+11	69	82	+13
-	C-1	0.1	EoR	43	76	+33	43	49	+6	43	60	+17
16:15	C-2	0.7	EoR	73	87	+14	86	89	+3	79	88	+9
	D-1	0.4	MoR	65	73	+8	72	80	+8	68	77	+9
	D-2	0.4	EoG	50	58	+8	67	74	+7	57	65	+8
	D-3	0.4	MoG	45	45	+0	56	87	+31	50	60	+10

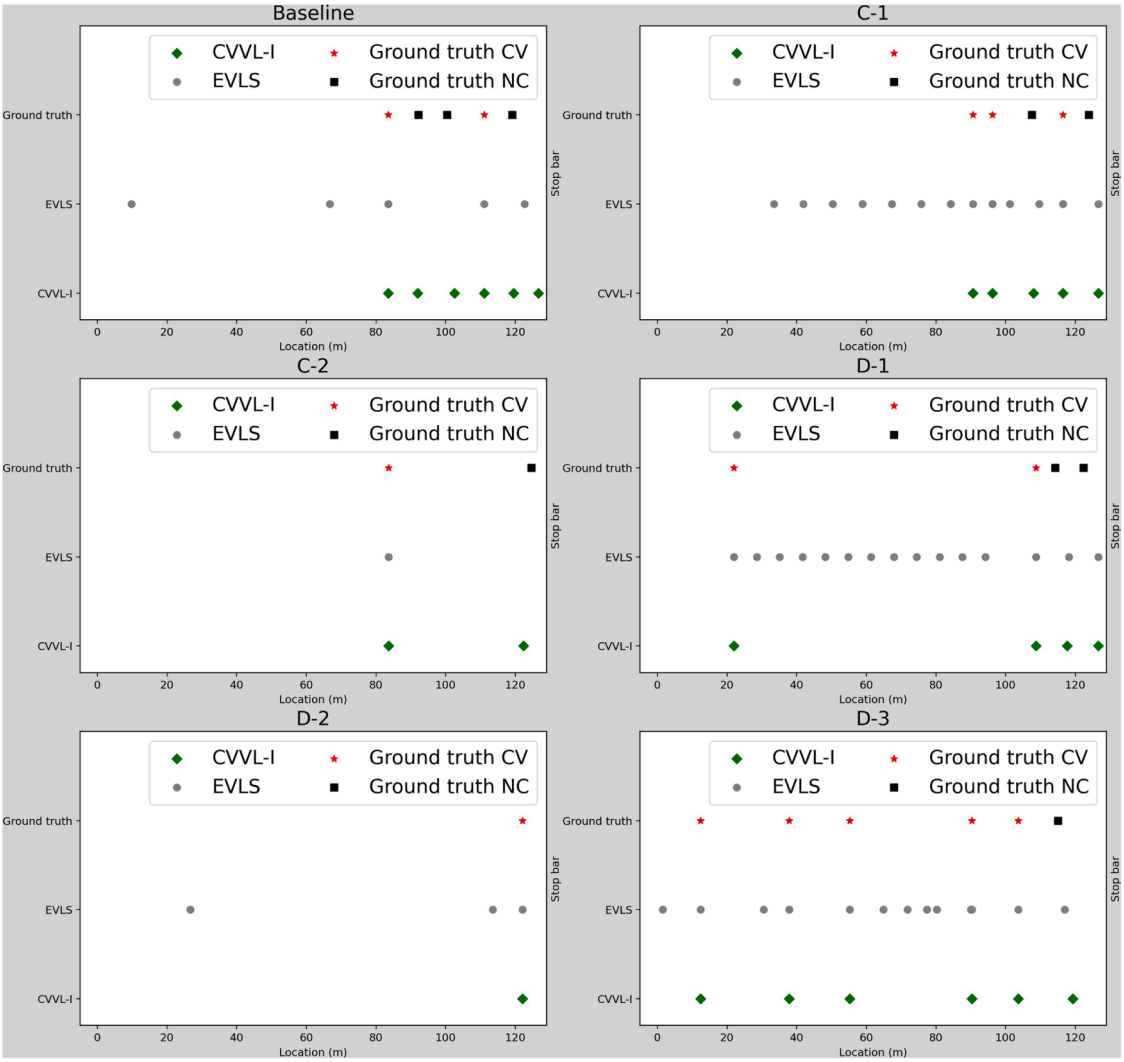

Fig. 14. Vehicle location estimation using the CVVL-I sub-model for the period from 12:45 to 13:00.

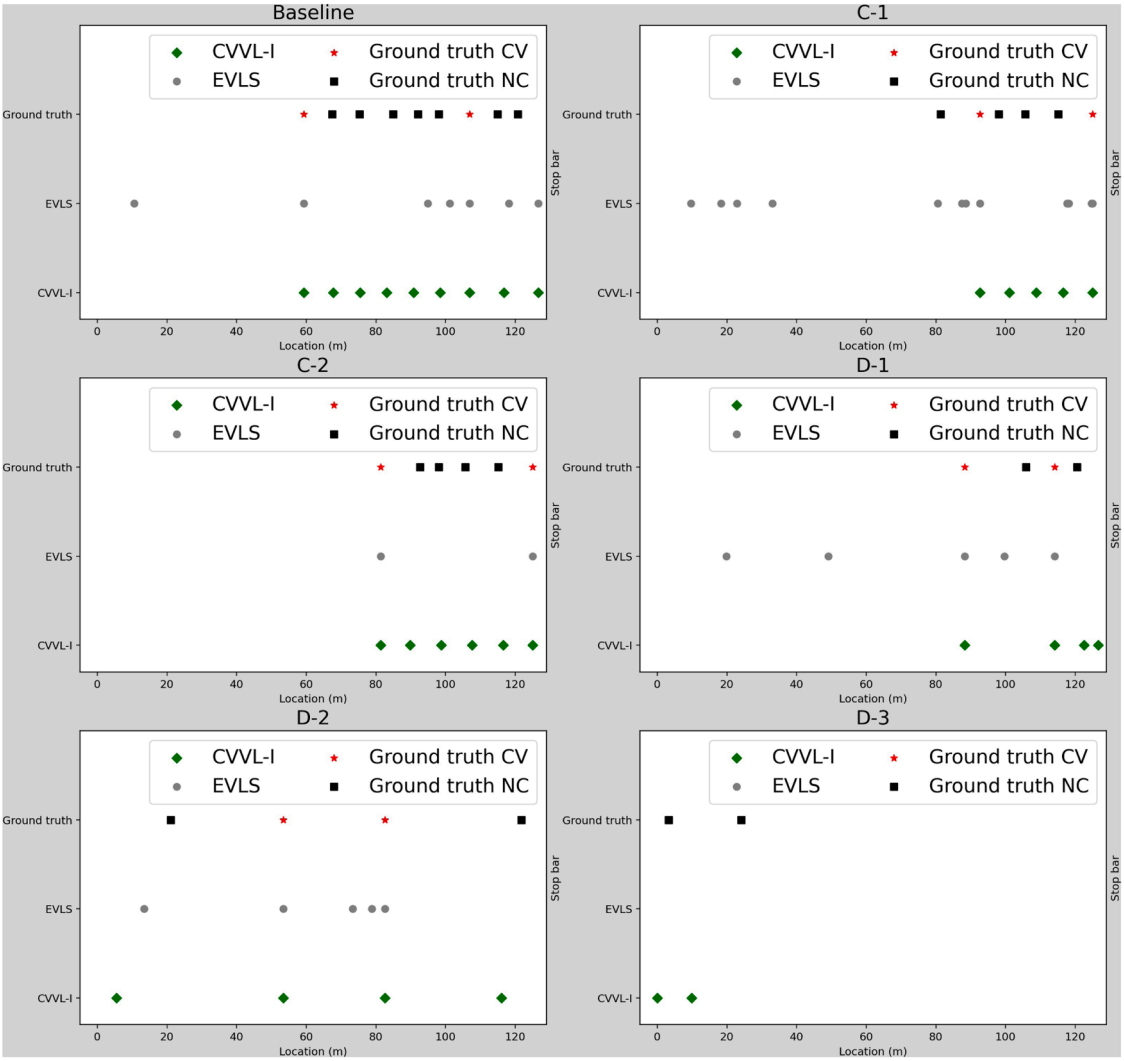

Fig. 15. Vehicle location estimation using the CVVL-I sub-model for the period from 16:00 to 16:15.

6.2. Results

Table 3 presents the results of vehicle location estimation for source lane 1 in Road Section 3 using the CVVL-S sub-model. For data collected during the period from 12:45 to 13:00, the baseline case was set with a CV penetration rate of 0.4 and the ToI at the EoR. The CVVL-S sub-model outperformed the EVLS algorithm, achieving improvements of 13 %, 7 %, and 11 % in terms of precision, recall, and F_1 score, respectively. These results demonstrate the effectiveness and superiority of the CVVL-S sub-model in a real-world context.

The cases in Groups C and D further explored the effects of varying CV penetration rates and ToIs. The consistent improvements observed across these cases reaffirm the practicality and applicability of the CVVL-S sub-model. Applying the CVVL-S sub-model to data during the period from 16:00 to 16:15 yielded similar results. However, as no vehicles were present in source lane 1 of Road Section 3 at the EoG, certain metrics were undefined (NaN) in case D-2 when the denominators were zero.

Figs. 12 and 13 illustrate the vehicle location estimation results using the CVVL-S sub-model and the EVLS algorithm for the periods from 12:45 to 13:00 and from 16:00 to 16:15. Overall, the proposed CVVL-S sub-model provided accurate vehicle location estimates, whereas the EVLS algorithm frequently exhibited inaccuracies.

Table 4 presents the results of vehicle location estimation for intermediate lane 1 in Road Section 2 using the CVVL-I sub-model. The proposed CVVL-I sub-model consistently outperformed the EVLS algorithm across all cases, demonstrating significant improvements. Figs. 14 and 15 illustrate sample estimations using both the CVVL-I sub-model and the EVLS algorithm. These quantitative and qualitative results collectively confirm that the proposed CVVL model accurately estimates vehicle locations in both source and intermediate lanes within a network using real-world data, thereby validating its practicality and applicability.

7. Conclusions

Real-time vehicle location estimation plays a pivotal role in numerous transport applications. However, accurately estimating the locations of all vehicles in a lane is extremely challenging owing to complex vehicle arrival patterns and partial connectivity. Thus, this study introduces a novel CVVL model to analytically estimate the vehicle locations by decomposing this challenging problem into a series of sequential sub-problems. The proposed model is applicable to any lane within a network and can account for any signal timing, traffic demand, and CV penetration rate. The effectiveness of the proposed models was confirmed through extensive VISSIM simulations and real-world validations using the NGSIM dataset. Nonetheless, this study has certain limitations: (1) reliance on average arrival rate and CV penetration rate as essential inputs for the CVVL model, and (2) need for the average effective vehicle length for estimation. To address the reliance on average arrival rate and CV penetration rate, future work can incorporate the uncertainty in the CV penetration rate and arrival rate to yield more robust estimations. To mitigate the need for the average effective vehicle length, its value can be recalibrated when applying the method to a different traffic scenario. Future research directions include addressing these limitations and exploring diverse applications based on the estimated vehicle locations.

CRediT authorship contribution statement

Shaocheng Jia: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **S.C. Wong:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Wai Wong:** Writing – review & editing, Validation, Conceptualization.

Acknowledgments

The first author was supported by a Postgraduate Scholarship from The University of Hong Kong. The second author was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.: 17205822), and the Francis S Y Bong Professorship in Engineering.

Appendix A. Estimation of average arrival rate and CV penetration rate

The average arrival rate, $\overline{q}$, and CV penetration rate, $\overline{p}$, are essential inputs for the proposed CVVL model and can be estimated using the following method. In a signalized lane, vehicles stopping in front of the stop bar form constrained queues (Wong et al., 2019; Jia et al., 2023, 2024a, 2024b). The constrained queue length, N, follows a certain distribution, denoted as $P(N=i)=\pi_i, \forall i=0,\ 1,\ 2,\ \ldots,\ k$. The numbers of CVs and observable vehicles before the last stopped CV in a constrained queue are denoted as n and $\widetilde{N}$, respectively, where n follows a binomial distribution, $B(N,\overline{p})$. The joint probability distribution between n and $\widetilde{N}$ is derived as follows (Jia et al., 2023):

$$P(n=i,\widetilde{N}=j)=\begin{cases}\pi_0+\sum_{z=1}^{k}\pi_z(1-\overline{p})^z, & i=0,\ j=0\\ \sum_{z=j}^{k}\pi_z\binom{j-1}{i-1}\overline{p}^i(1-\overline{p})^{z-i}, & \forall i,j=1,\ 2,\ \ldots,\ k,\ j\geq i\end{cases}. \tag{A1}$$

The distribution of N can be estimated using the following probabilistic dissipation time (PDT) model:

$$P(N=k)=\begin{cases}f(k;\overline{q}r)f(0;\overline{q}k\tau)+\\ \sum_{i=1}^{k-1}\sum_{j=1}^{J_i}f(i;\overline{q}r)\widetilde{P}_j(N=k,\ M=i)W_j(N=k,\ M=i) & \text{if } k\in\mathbb{N}^+,\\ f(0;\overline{q}r) & \text{if } k=0\end{cases} \tag{A2}$$

where $f(k;\overline{q}t)$ represents the probability of k vehicles arriving with the average arrival rate $\overline{q}$ during time interval t and is determined by the given vehicle arrival pattern. $\widetilde{P}_j(N=k,\ M=i)$ and $W_j(N=k,M=i)$, $\forall j\in[1,J_i]$ can be obtained using Algorithm 1 presented by Jia et al. (2023). Explicitly considering the arrival pattern in the PDT model enables accurate estimation of the constrained queue length distribution but reduces the computational efficiency. The use of a simplified model based on the constant dissipation time (CDT) can facilitate the efficient estimation of the distribution of N, by assuming that N follows a Poisson distribution. The governing parameter of the Poisson distribution, N_0, is defined as (Jia et al., 2023)

$$N_0=\frac{s\overline{q}r}{s-\overline{q}}, \tag{A3}$$

where s and r represent the saturation flow rate and red period, respectively. Users can choose either the PDT or CDT model based on their requirements. Furthermore, an MCQL model can be seamlessly integrated into both the PDT and CDT models to account for complex holding-vehicle effects (Jia et al., 2024a, 2024b). Eqs. (A1)–(A3) indicate that the joint probability distribution between n and

$\widetilde{N}$ depends on $\overline{q}$ and $\overline{p}$. This dependence enables the establishment of the following maximum likelihood estimation function:

$$\max_{\overline{q},\,\overline{p}} \prod_{j=0}^{\omega} P\left(n_{i-j}, \widetilde{N}_{i-j}\right), \tag{A4}$$

where $\omega = 0,\ 1,\ 2,\ \ldots, i-1$ represents the number of past cycles considered in the likelihood function. In this study, ω is set as 2, following previous studies (Jia et al., 2023, 2024a, 2024b, 2024c, 2024d). A simple grid search method is used to determine the optimal solutions $\overline{q}^*$ and $\overline{p}^*$ in Eq. (A4). For simplicity, $\overline{q}$ and $\overline{p}$ are used to represent $\overline{q}^*$ and $\overline{p}^*$ in this work.

Appendix B. Estimation of holding vehicles

The number of holding vehicles at any ToI, denoted as R, is defined as the number of vehicles that, based on their projected trajectories using cruise speeds, should have been discharged by that instant but remain held by the system. This quantity serves as an essential input for the CVVL model. Using the average arrival rate, $\overline{q}$, and CV penetration rate, $\overline{p}$, as inputs, the CV-based holding vehicle (CVHV) model, comprising the CVHV-I and CVHV-II sub-models, is employed to estimate the number of holding vehicles at any given ToI (Jia et al., 2024c). Specifically, the CVHV-I sub-model is applied for ToIs where the signal groups fall within their effective red period. In contrast, the CVHV-II sub-model is used for other ToIs. Within CVHV-I, various cases, determined by different combinations of stopped and moving holding CVs, are considered. CVHV-II addresses more complex scenarios, incorporating cases defined by various combinations of moving holding CVs preceding the stopped holding CVs, the stopped holding CVs themselves, and moving holding CVs following the stopped holding CVs. For each case, a fully analytical model has been derived to estimate R. Further details are provided in Jia et al. (2024c).

Appendix C. Proof of Proposition 1

Consider a signal-free lane. Vehicles on the lane are expected to move at the cruise speed. The average arrival rate of NCs, $\overline{q}_N$, can be estimated as

$$\overline{q}_N = \overline{q}(1-\overline{p}). \tag{C1}$$

The expected travel time for each vehicle, t_f, is defined as

$$t_f = \frac{l}{v_f}. \tag{C2}$$

Thus, the expected number of NCs in a lane at any time, Q_1, can be estimated as

$$Q_1 = \overline{q}_N t_f. \tag{C3}$$

However, in a signalized lane, some vehicles will stop at a red signal, resulting in the formation of holding vehicles at the end of an arbitrarily defined cycle that ends at the ToI. This red-signal effect makes the actual number of NCs on the lane larger than the expected number of NCs, Q_1. Given the number of holding vehicles, R, the number of NCs in R, Q_2, is given by

$$Q_2 = R - R_C. \tag{C4}$$

Therefore, the total number of NCs in a lane is estimated as

$$Q = Q_1 + Q_2. \tag{C5}$$

Substituting Eqs. (C1)–(C4) into Eq. (C5) yields Eq. (1).

Appendix D. Proof of Proposition 2

This proof involves derivations for the spatial distribution of NCs for both Type A and Type B lane segments.

A. *Derivation of the spatial distribution of NCs for Type A lane segments*

This part of the proof includes the estimations of (i) Q_i^A and (ii) $\left\{L_i^j\right\}^{Q_i^A}$

(i) If $V_i = 0 \cap V_{i+1} = 0, \forall i \in [0, m-1]$, the speeds of NCs between L_i and L_{i+1} of the i^{th} Type A lane segment can be considered zero. The distance headway between any two consecutive stopped vehicles in this case can be assumed to be the average effective vehicle length l_e. If $i = 0$, meaning that an NC can be inserted immediately after the stop bar, the number of NCs between L_0 and L_1 in the 0^{th} Type A lane segment, where $L_0 = l$, can be estimated as

$$Q_i^A = \left\lfloor \frac{l - L_1}{l_e} \right\rfloor. \tag{D1}$$

In contrast, if $i \neq 0$, meaning that the Type A lane segment is enclosed by two stopped CVs, the feasible space for the insertion of NCs is defined by the range $[L_i - l_e, L_{i+1}]$. In this case, the number of NCs between L_i and L_{i+1} in the i^{th} Type A lane segment can be estimated as

$$Q_i^A = \left\lfloor \frac{L_i - l_e - L_{i+1}}{l_e} \right\rfloor. \tag{D2}$$

Combining Eqs. (D1) and (D2) yields Eq. (4).

(ii) Without any additional information, Q_i^A NCs are assumed to be uniformly distributed within the range of $[L_i, L_{i+1}]$. The distance headway between any two consecutive stopped vehicles, Δl, can be updated as follows:

$$\Delta l = \begin{cases} \dfrac{l - L_1}{Q_0^A} \ \ if\ i = 0 \\ \dfrac{L_i - l_e - L_{i+1}}{Q_i^A} \ \ if\ \forall i \in (0, m-1] \end{cases}. \tag{D3}$$

Therefore, the locations of Q_i^A NCs, $\left\{L_i^j\right\}^{Q_i^A}$, can be estimated as

$$L_i^j = L_{i+1} + j\Delta l, \forall j \in \left[1, Q_i^A\right]. \tag{D4}$$

Combining Eqs. (D3) and (D4) yields Eq. (3).

B. Derivation of the spatial distribution of NCs for Type B lane segments

This part of the proof includes the estimations of (i) $\widetilde{Q}_i^B$, (ii) Q_i^B, (iii) V_i^j, and (iv) L_i^j.

(i) If $V_i \neq 0 \cup V_{i+1} \neq 0, \forall i \in [0, m]$, some NCs may be traveling between L_i and L_{i+1} in the i^{th} Type B lane segment. The maximum number of vehicles between L_i and L_{i+1}, including vehicles located at L_i and L_{i+1}, is defined as $k+2$. Then, the minimum safe distance headway between the j^{th} and $(j+1)^{th}$ vehicles within the range $[L_i, L_{i+1}]$, ΔL_j, can be estimated as

$$\Delta L_j = V_i^j \Delta t, \forall j \in [0, k], \tag{D5}$$

where V_i^j represents the speed of the j^{th} vehicle between L_i and L_{i+1}. Therefore, the following expression can be obtained:

$$\sum_{j=0}^{k} \Delta L_j = \Delta t \sum_{j=0}^{k} V_i^j = L_i - L_{i+1}. \tag{D6}$$

Assuming that the speeds of NCs between L_i and L_{i+1} follow a linear relationship,

$$V_i^j = V_i^0 + j\Delta V, \forall j \in [1, k] \tag{D7}$$

where

$$V_i^0 = V_{i+1}, \tag{D8}$$

and

$$\Delta V = \frac{V_i - V_{i+1}}{k+1}. \tag{D9}$$

Substituting Eqs. (D7)–(D8) into Eq. (D6) leads to the following expression:

$$\Delta t\{V_{i+1} + (V_{i+1} + \Delta V) + \ldots + [V_{i+1} + k\Delta V]\} = L_i - L_{i+1}. \tag{D10}$$

Dividing both sides by Δt, the following expression is obtained,

$$\underbrace{V_{i+1}+\ldots+V_{i+1}}_{k+1}+\underbrace{\Delta V+\ldots+k\Delta V}_{k}=\frac{L_i-L_{i+1}.}{\Delta t} \tag{D11}$$

Applying the formula for the sum of arithmetic sequences, it can be shown that

$$(k+1)V_{i+1}+\frac{(k+1)k}{2}\Delta V=\frac{L_i-L_{i+1}.}{\Delta t} \tag{D12}$$

Substituting Eq. (D9) into Eq. (D12) and rearranging the terms yields the following expression:

$$k=\frac{2(L_i-L_{i+1})-2\Delta t V_{i+1}}{\Delta t(V_i+V_{i+1})}. \tag{D13}$$

If $i=0\cup i=m$, meaning that an NC can be inserted at the stop bar or at the upstream lane entrance, respectively, the maximum number of NCs that can be inserted between L_i and L_{i+1} in the i^{th} Type B lane segment is

$$\widetilde{Q}_i^B=\lfloor k+1\rfloor, \forall i\in\{0,m\}. \tag{D14}$$

If $i\neq 0\cap i\neq m$, meaning that the lane segment is enclosed by two CVs, the maximum number of NCs that can be inserted between L_i and L_{i+1} in the i^{th} Type B lane segment is

$$\widetilde{Q}_i^B=\lfloor k\rfloor,\ \forall i\in(0,m). \tag{D15}$$

Combining Eqs. (D13)–(D15) and considering the nonnegative constraint yields Eq. (7).

(ii) Given the total number of NCs to be inserted in a lane Q from Proposition 1 and the number of NCs to be inserted between L_i and L_{i+1} in the i^{th} Type A lane segment, i.e., Q_i^A, with $V_i=0\cap V_{i+1}=0, \forall i\in[0,m-1]$, the remaining number of NCs that needs to be inserted is given by

$$Q'=Q-\sum_{\substack{i=0\\ V_i=0\ \cap\ V_{i+1}=0}}^{m-1}Q_i^A. \tag{D16}$$

Using a simple scaling method, the actual number of NCs to be inserted within the range $\left[L_i^{(l)},\ L_i^{(u)}\right]$ of the i^{th} Type B lane segment, i.e., Q_i^B, with $V_i\neq 0\cup V_{i+1}\neq 0, \forall i\in[0,m]$, is estimated as

$$Q_i^B=\lfloor\rho\widetilde{Q}_i^B\rceil, \tag{D17}$$

where ρ represents a scaling factor:

$$\rho=\frac{Q'}{\sum_{\substack{i=0\\ V_i\neq 0\ \cup\ V_{i+1}\neq 0}}^{m}\widetilde{Q}_i^B}. \tag{D18}$$

Owing to rounding errors or inaccuracies in estimating Q and Q', Q_i^B may occasionally exceed $\widetilde{Q}_i^B$, resulting in excess e_i that cannot be inserted within $\left[L_i^{(l)},\ L_i^{(u)}\right]$. Considering the $(i-1)^{th}$ excess e_{i-1} from $\left[L_{i-1}^{(l)},\ L_{i-1}^{(u)}\right]$ and the maximum number of NCs between L_i and L_{i+1}, $\widetilde{Q}_i^B$, Q_i^B can be modified as

$$Q_i^B=\min\left\{\lfloor\rho\widetilde{Q}_i^B\rceil+e_{i-1},\widetilde{Q}_i^B\right\}, \tag{D19}$$

where e_{i-1} can be recursively determined as

$$e_i=\begin{cases}\max\left\{\lfloor\rho\widetilde{Q}_i^B\rceil+e_{i-1}-\widetilde{Q}_i^B,0\right\}\ \textit{if}\ i\in[0,m]\\ 0\ \textit{otherwise}\end{cases}. \tag{D20}$$

Substituting Eqs. (D16)–(D18) into Eqs. (D19) and (D20) yields Eqs. (6) and (8), respectively.

(iii) The speeds of the inserted NCs are the essential inputs for determining the feasible space in any Type B lane segment for the insertion of NCs. If $i = 0 \cup i = m$, meaning that an NC can be inserted at the stop bar or at the upstream lane entrance, the speed difference between any two consecutive vehicles,ΔV, is given by

$$\Delta V = \frac{V_i - V_{i+1}}{Q_i^B}, \forall i \in \{0, m\}. \tag{D21}$$

If $i = 0$, the speeds of the inserted NCs can be estimated as

$$V_0^j = V_1 + j\Delta V, \forall j \in \left[1, Q_i^B\right]. \tag{D22}$$

If $i = m$, the speeds of the inserted NCs can be estimated as

$$V_m^j = V_{m+1} + (j-1)\Delta V, \forall j \in \left[1, Q_i^B\right]. \tag{D23}$$

If $i \neq 0 \cap i \neq m$, meaning that the lane segment is enclosed by two CVs, the speed difference between any two consecutive vehicles, $\Delta V'$, is defined as

$$\Delta V' = \frac{V_i - V_{i+1}}{Q_i^B + 1}, \forall i \in (0, m). \tag{D24}$$

Similarly, the speeds of the inserted NCs can be estimated as

$$V_i^j = V_{i+1} + j\Delta V', \forall j \in \left[1, Q_i^B\right]. \tag{D25}$$

Combining Eqs. (D21)–(D25) yields Eq. (11).

(iv) The feasible space for the insertion of NCs is defined by a lower bound and an upper bound. If $i = m$, the lower bound of the feasible space, $L_m^{(l)}$, is 0, meaning that an NC can be inserted at the upstream lane entrance. If $i \neq m$, $L_i^{(l)}$ should be kept at a minimum safe distance headway from L_{i+1}, as determined by the minimum safe time headway Δt. Assuming that the $(i+1)^{th}$ CV travels at a speed of V_{i+1}, the minimum safe distance headway from its preceding vehicle, ΔL, can be estimated as

$$\Delta L = V_{i+1}\Delta t. \tag{D26}$$

Considering the physical size of a vehicle, the following relation can be obtained:

$$\Delta L = \max\{V_{i+1}\Delta t, l_e\}. \tag{D27}$$

Thus,

$$L_i^{(l)} = L_{i+1} + \Delta L. \tag{D28}$$

Next, combining the cases for $i = m$ and $i \neq m$ yields Eq. (9). Similarly, if $i = 0$, the upper bound of the feasible space, $L_0^{(u)}$, is l, meaning that an NC can be inserted at the stop bar. If $i \neq 0$, $L_i^{(u)}$ should be kept at a minimum safe distance headway from the i^{th} CV. Based on the estimated speed $V_i^{Q_i^B}$, $L_i^{(u)}$ can be estimated as

$$L_i^{(u)} = L_i - V_i^{Q_i^B}\Delta t. \tag{D29}$$

Considering the physical size of a vehicle, $L_i^{(u)}$ can be modified as

$$L_i^{(u)} = L_i - \max\left\{V_i^{Q_i^B}\Delta t, l_e\right\}. \tag{D30}$$

Combining cases when $i = 0$ and $i \neq 0$ yields Eq. (10).

Based on the identified feasible space, Q_i^B NCs can be uniformly distributed in $\left[L_i^{(l)}, L_i^{(u)}\right]$ of the i^{th} Type B lane segment. If $Q_i^B = 1$, the NC is inserted at the center of $\left[L_i^{(l)}, L_i^{(u)}\right]$, i.e.,

$$L_i^1 = \frac{1}{2}\left(L_i^{(u)} - L_i^{(l)}\right). \tag{D31}$$

If $Q_i^B > 1$, all NCs can be uniformly distributed across the range $\left[L_i^{(l)}, L_i^{(u)}\right]$, yielding the following relation:

$$L_i^j = L_i^{(l)} + (j-1)\frac{L_i^{(u)} - L_i^{(l)}}{Q_i^B - 1}, \forall j \in [1, Q_i^B]. \tag{D32}$$

Combining Eqs. (D31) and (D32) gives Eq. (5).

Appendix E. Proof of Proposition 3

This proof involves the derivations of the lower and upper bounds of the 0^{th} CP including holding vehicles and the lower and upper bounds of the s^{th} CP formed by newly arrived vehicles from the s^{th} upstream lane, $\forall s \in [1, S]$.

A. Derivation of the lower and upper bounds of the 0^{th} CP

This part of the proof includes the derivations of (i) $CP_0^{(u)}$ and (ii) $CP_0^{(l)}$

(i) Given that holding vehicles must be in front of newly arrived vehicles, the upper bound of the 0^{th} CP, $CP_0^{(u)}$, is set to be the lane length, l, resulting in Eq. (14).

(ii) To derive the lower bound of the 0^{th} CP, $CP_0^{(u)}$, two scenarios are considered: cases without and with holding CVs. If no holding CVs exist, i.e., $\widetilde{m}_0 = 0$, two sub-cases can be identified, i.e., cases in which the downstream intermediate lane is controlled by a signal group ending with an effective red or with an effective green. When the downstream intermediate lane is controlled by a signal group ending with an effective red signal, holding vehicles are expected to be in a queue. Given the number of holding vehicles, R, the lower bound of the0^{th} CP is estimated as

$$CP_0^{(l)} = l - Rl_e. \tag{E1}$$

In cases where the downstream intermediate lane is controlled by a signal group ending with an effective green, the maximum number of vehicles that can be discharged is estimated using g/τ. Vehicles remaining in the queue behind these discharged vehicles are the holding vehicles. Thus, the lower bound of the0^{th} CP is estimated as the stopping location of the last vehicle in the queue:

$$CP_0^{(l)} = l - \left(\frac{g}{\tau} + R\right) l_e. \tag{E2}$$

Combining Eqs. (E1) and (E2) yields the first case in Eq. (13).

If holding CVs are observed, i.e., $\widetilde{m}_0 > 0$, the available CV information can be used to estimate $CP_0^{(l)}$. As vehicles that enter the downstream intermediate lane before $t - l/v_f$ and remain in the lane at time t are considered holding vehicles, $CP_0^{(l)}$ is estimated by the projected location of a vehicle entering the downstream intermediate lane at $t - l/v_f$ with a speed of the mean of the average speeds of the holding CVs, $\bar{v}_0$:

$$CP_0^{(l)} = \frac{l}{v_f}\bar{v}_0, \tag{E3}$$

where

$$\bar{v}_0 = \frac{1}{\widetilde{m}_0}\sum_{i=1}^{\widetilde{m}_0}\frac{L_{0,i}}{t - T_{0,i}}. \tag{E4}$$

Substituting Eq. (E4) into Eq. (E3) yields the second case in Eq. (13).

B. Derivation of the lower and upper bounds of the s^{th} CP, $\forall s \in [1, S]$

This part of the proof includes the derivations of $CP_s^{(l)}$ and $CP_s^{(u)}$for cases with (i) $\widetilde{m}_s = 0$, (ii) $\widetilde{m}_s = 1$, and (iii) $\widetilde{m}_s > 1$.

(i) If $\widetilde{m}_s = 0$, meaning no CV is present in the s^{th} CP, $\forall s \in [1, S]$, all vehicles from the s^{th} upstream lane are assumed to travel at the cruise speed, v_f. The lower bound of the s^{th} CP, $CP_s^{(l)}$, is estimated by the projected location of a vehicle discharged at $G_s^{(2)}$ from the upstream lane s, defined as

$$CP_s^{(l)} = \left(t - G_s^{(2)}\right) v_f. \tag{E5}$$

The upper bound of the s^{th} CP, $CP_s^{(u)}$, is estimated by the projected location of a vehicle discharged at $G_s^{(1)}$ from the upstream lane s, given by

$$CP_s^{(u)} = \left(t - G_s^{(1)}\right) v_f. \tag{E6}$$

Eqs. (E5) and (E6) provide the first cases in Eqs. (15) and (16), respectively.

(ii) If $\tilde{m}_s = 1$, indicating only one CV is present in the s^{th} CP, $\forall s \in [1,S]$, the actual average speed of this CV is used for the location projection instead of the cruise speed:

$$\bar{v}_s = \frac{L_{s,1}}{t - T_{s,1}}. \tag{E7}$$

Replacing v_f in the first cases of Eqs. (15) and (16) with $\bar{v}_s$ and considering the range of the downstream intermediate lane, [0, l], yield the second cases in Eqs. (15) and (16), respectively.

(iii) If $\tilde{m}_s > 1$, meaning that two or more CVs are present in the s^{th} CP, $\forall s \in [1,S]$, the s^{th} CP must at least encompass the range between the locations of the first and last CVs within it, $\left[L_{s,\tilde{m}_s}, L_{s,1}\right]$. To estimate the lower and upper bounds of the s^{th} CP, the additional distance beyond $L_{s,\tilde{m}_s}$, ΔL_1, and before $L_{s,1}$, ΔL_2, covered by the s^{th} CP must be determined. Given ΔL_1 and ΔL_2, it follows that

$$CP_s^{(l)} = L_{s,\tilde{m}_s} - \Delta L_1, \tag{E8}$$

$$CP_s^{(u)} = L_{s,1} + \Delta L_2. \tag{E9}$$

Assuming that ΔL_1 and ΔL_2 share the same dispersion pattern as that within the range $\left[L_{s,\tilde{m}_s}, L_{s,1}\right]$,

$$\frac{\Delta L_2}{T_{s,1} - G_s^{(1)}} = \frac{L_{s,1} - L_{s,\tilde{m}_s}}{T_{s,\tilde{m}_s} - T_{s,1}} = \frac{\Delta L_1}{G_s^{(2)} - T_{s,\tilde{m}_s}}. \tag{E10}$$

Solving Eq. (E10) yields

$$\Delta L_1 = \frac{\left(L_{s,1} - L_{s,\tilde{m}_s}\right)\left(G_s^{(2)} - T_{s,\tilde{m}_s}\right)}{T_{s,\tilde{m}_s} - T_{s,1}}, \tag{E11}$$

$$\Delta L_2 = \frac{\left(L_{s,1} - L_{s,\tilde{m}_s}\right)\left(T_{s,1} - G_s^{(1)}\right)}{T_{s,\tilde{m}_s} - T_{s,1}}. \tag{E12}$$

Substituting Eqs. (E11) and (E12) into Eqs. (E8) and (E9) and considering the range of the downstream intermediate lane, [0, l], yields the third cases in Eqs. (15) and (16), respectively.

Appendix F. Proof of Proposition 4

This proof involves the derivations of the number of NCs in the 0^{th} CP constituted by holding vehicles and number of NCs in the s^{th} CP formed by newly arrived vehicles from the s^{th} upstream lane, $\forall s \in [1,S]$.

A. *Derivation of the number of NCs in the 0^{th} CP*

The 0^{th} CP is formed by holding vehicles in the downstream intermediate lane at any time t of interest. Given the total number of holding vehicles, R, and the number of holding CVs, R_C, the number of NCs in the 0^{th} CP is intuitively obtained by $R - R_C$, as shown in Eq. (17).

B. *Derivation of the number of NCs in the s^{th} CP, $\forall s \in [1,S]$*

Given the number of NCs discharged between $g_s - r_s$ and $G_s^{(w)}$, $\widetilde{Q}_s^{(w)}$, $\forall w \in \{1,2\}$, the total number of NCs discharged between $G_s^{(1)}$ and $G_s^{(2)}$ is

$$\widetilde{Q}_s = \widetilde{Q}_s^{(2)} - \widetilde{Q}_s^{(1)}. \tag{F1}$$

The number of NCs in the range of the s^{th} CP, $\left[CP_s^{(l)}, CP_s^{(u)}\right]$, Q_s, is obtained by

$$Q_s = \lfloor \beta_s \widetilde{Q}_s \rceil, \forall s \in [1,S]. \tag{F2}$$

where β_s represents the turning proportion from the s^{th} upstream lane to the downstream intermediate lane. This value can be

estimated by dividing the number of CVs traveling from the s^{th} upstream lane to the downstream intermediate lane over the past ω cycles by the total number of CVs observed in the s^{th} upstream lane during those cycles, as shown in Eq. (19). Substituting Eq. (F1) into Eq. (F2) yields Eq. (18).

To estimate $\widetilde{Q}_s^{(w)}, w \in \{1,2\}$, the cases with (i) $\widetilde{n}_s^{(w)} = 0$ and (ii) $\widetilde{n}_s^{(w)} > 0$ are considered.

(i) If no CV stops between $g_s - r_s$ and $G_s^{(w)}$ in the s^{th} upstream lane, i.e., $\widetilde{n}_s^{(w)} = 0$, the expected total number of NCs between $g_s - r_s$ and $G_s^{(w)}$, $A_s^{(w)}$, can be estimated as

$$A_s^{(w)} = \overline{q}_{Ns}\left(G_s^{(w)} - g_s + r_s\right), \tag{F3}$$

where $\overline{q}_{Ns}$ represents the average arrival rate of NCs in the s^{th} upstream lane and is defined using Eq. (22). Moreover, given the number of holding NCs at $G_s^{(w)}$ in the s^{th} upstream lane, $R_{Ns}^{(w)}$, which is estimated by subtracting the number of holding CVs at $G_s^{(w)}$ in the s^{th} upstream lane, $R_{Cs}^{(w)}$, from the total number of holding vehicles at $G_s^{(w)}$ in the s^{th} upstream lane, $R_s^{(w)}$, as shown in Eq. (21), the total number of NCs discharged between $g_s - r_s$ and $G_s^{(w)}$ can be estimated as

$$\widetilde{Q}_s^{(w)} = A_s^{(w)} - R_{Ns}^{(w)}, \forall w \in \{1,2\}. \tag{F4}$$

Substituting Eq. (F3) into Eq. (F4) and considering the nonnegativity constraint yields the first case in Eq. (20).

(ii) If some CVs stop between $g_s - r_s$ and $G_s^{(w)}$ in the s^{th} upstream lane, i.e., $\widetilde{n}_s^{(w)} > 0$, the number of NCs before the last stopped CV in the s^{th} upstream lane, $\widetilde{Q}_{s1}^{(w)}$, can be estimated as

$$\widetilde{Q}_{s1}^{(w)} = \frac{l_s - L_s'^{(w)}}{l_e} + 1 - \widetilde{n}_s^{(w)}, \forall w \in \{1,2\}. \tag{F5}$$

The number of NCs after the last stopped CV in the s^{th} upstream lane, $\widetilde{Q}_{s2}^{(w)}$, can be estimated as

$$\widetilde{Q}_{s2}^{(w)} = \overline{q}_{Ns}\left(G_s^{(w)} - T_s'^{(w)}\right), \forall w \in \{1,2\}. \tag{F6}$$

Thus,

$$\widetilde{Q}_s^{(w)} = \widetilde{Q}_{s1}^{(w)} + \widetilde{Q}_{s2}^{(w)} - R_{Ns}^{(w)}. \tag{F7}$$

Substituting Eqs. (F5) and (F6) into Eq. (F7) and considering the nonnegativity constraint yields the second case in Eq. (20).

Appendix G. Impacts of average arrival rate and CV penetration rate estimation accuracy on the CVVL model

The CVVL model utilizes the average arrival rate $\overline{q}$, average CV penetration rate $\overline{p}$, and the number of holding vehicles as inputs to estimate vehicle locations. Because the number of holding vehicles is estimated based on $\overline{q}$ and $\overline{p}$ as described in **Appendix B**, $\overline{q}$ and $\overline{p}$ are the only independent input parameters for the CVVL model. This appendix investigates the robustness of the proposed model to parameter inaccuracy.

The analysis employed VISSIM simulation data and the baseline settings outlined in Tables 1 and 2, including a red period of 30 s, a V/C ratio of 0.5, a CV penetration rate of 0.4, and the ToI at the EoR. Additionally, a scenario with a low CV penetration rate of 0.1 was analyzed. A control variable approach was employed, meaning that when assessing the impact of one parameter, the other was fixed at its true value. For example, when assessing the impact of the accuracy of $\overline{q}$, the true value of $\overline{p}$ was used in the analysis. This ensured an isolated and fair evaluation of the influence of parameter accuracy on the model's performance.

To comprehensively evaluate the robustness of the proposed model to parameter inaccuracy, the analysis was conducted across a wide range of parameter values. For average CV penetration rate, the full range from 0 to 1 with a step size of 0.1 was tested. For average arrival rate, a wide range from 0 to 0.3 veh/s with a step size of 0.02 veh/s was considered. In practice, the real-world variations in $\overline{q}$ and $\overline{p}$ due to inaccuracy are expected to be much smaller. For each parameter combination, the proposed CVVL model was applied to 1000-cycle VISSIM simulation data to estimate vehicle locations at ToIs.

The estimated vehicle locations were evaluated using three metrics: precision, recall, and F_1 score. Precision measures the proportion of correctly predicted positives among all predicted positives, indicating the model's ability to avoid false positives. Recall measures the proportion of correctly predicted positives among all actual positives, reflecting the model's ability to avoid false negatives. Because precision and recall inherently trade off against each other, they are expected to move in opposite directions as an input parameter varies. In contrast, the F_1 score provides a balanced metric, combining both precision and recall, and is the most representative measure of the robustness of the proposed model. A stable F_1 score across varying input parameters would indicate the model's robustness or insensitivity to input inaccuracies.

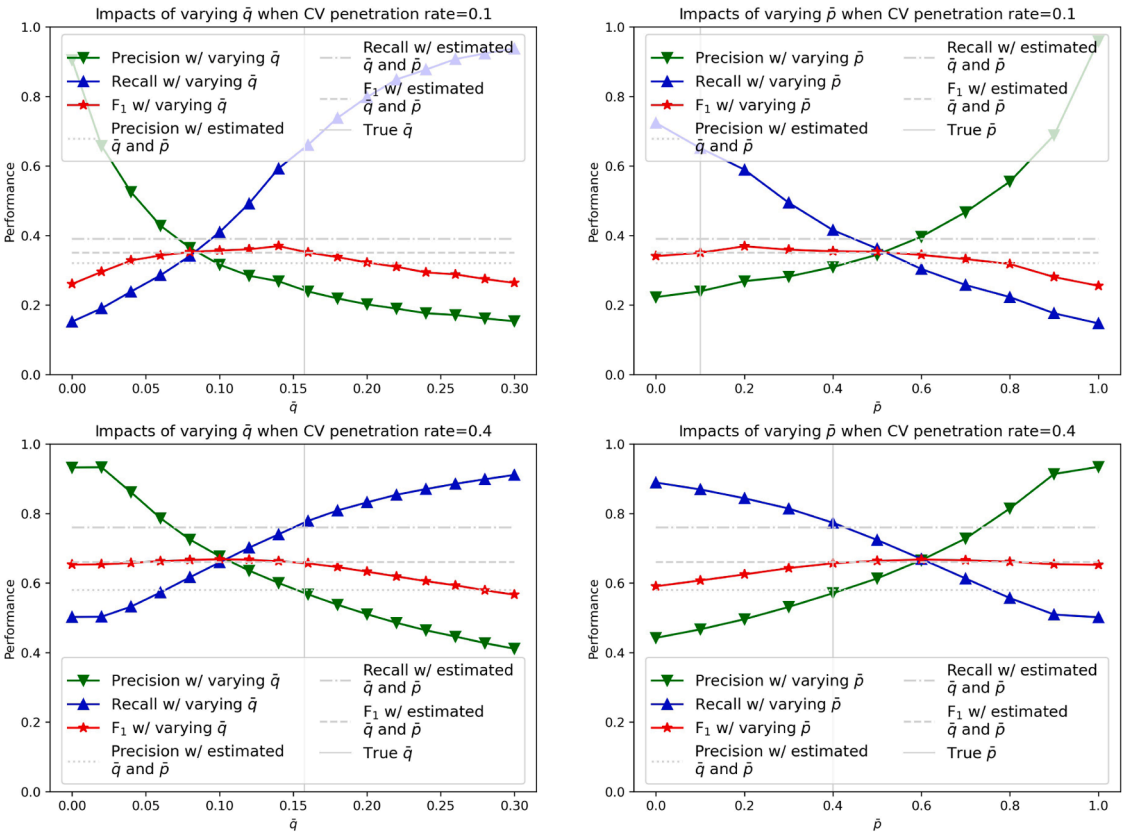

Fig. G1. Impacts of $\bar{q}$ and $\bar{p}$ estimation accuracy in the source lane (lane 7 in Fig. 6) on the CVVL-S sub-model.

Fig. G1 illustrates the impacts of average arrival rate and CV penetration rate estimation accuracy in the source lane (lane 7 in Fig. 6) on the CVVL-S sub-model. The experimental results indicate that precision and recall exhibit the expected trade-off, moving in opposite directions. For instance, in the top-left sub-figure of Fig. G1, which examines the impact of average arrival rate on the CVVL-S sub-model under a low CV penetration rate of 0.1, precision decreases while recall increases with higher average arrival rates. This behavior arises because a higher average arrival rate results in more vehicles predicted in the lane, leading to an increase in FP and a reduction in FN. Similarly, the top-right sub-figure of Fig. G1, which depicts the effect of average CV penetration rate on the CVVL-S sub-model under the same low CV penetration rate, shows an opposite trend for precision and recall. Precision increases as the average CV penetration rate rises, while recall decreases. This occurs because a higher CV penetration rate implies fewer NCs, resulting in fewer FPs but more FNs. When the CV penetration rate is increased to 0.4, similar patterns are observed in the bottom sub-figures of Fig. G1. However, the variations in precision and recall are further reduced. Notably, the variations in precision and recall for all cases remain within reasonable bounds, even across the extensive parameter ranges tested. Most importantly, the F_1 score, which is the most representative measure of the robustness of the proposed model, shows minimal variation across changes in both average arrival rate and CV penetration rate for all the cases considered, confirming the robustness of the CVVL-S sub-model. Additionally, the F_1 scores based on the estimated $\bar{q}$ and $\bar{p}$ using existing models closely align with the F_1 scores obtained using true parameter values, further validating the proposed model's performance.

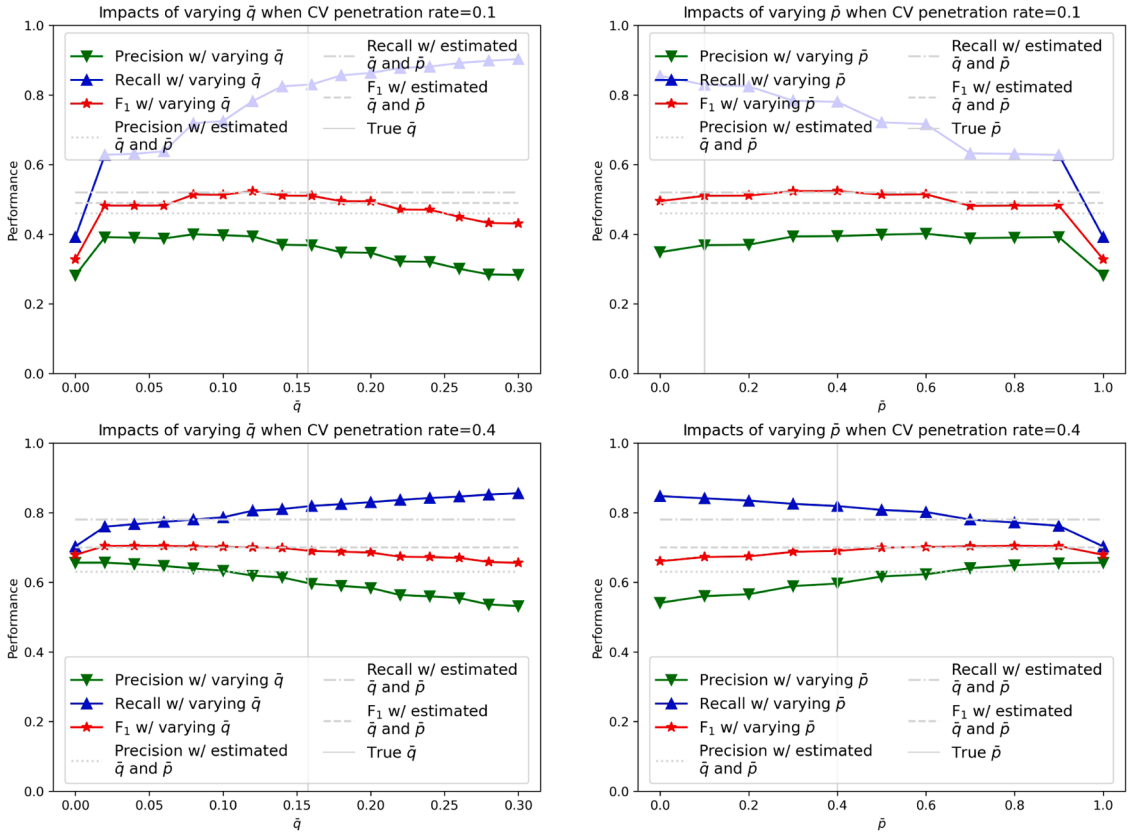

Fig. G2. Impacts of $\bar{q}$ and $\bar{p}$ estimation accuracy in the intermediate lane (lane 5 in Fig. 6) on the CVVL-I sub-model.

Fig. G2 presents the impacts of average arrival rate and CV penetration rate estimation accuracy in the intermediate lane (i.e., lane 5 in Fig. 6) on the CVVL-I sub-model. As with the CVVL-S sub-model, precision and recall exhibit the expected trade-off. The F_1 scores based on the estimated $\bar{q}$ and $\bar{p}$ using existing models align closely with the F_1 scores obtained using the true parameter values. Moreover, the flat F_1 scores across the full test ranges demonstrate that the CVVL-I sub-model is not sensitive to input inaccuracies, thereby confirming its robustness.

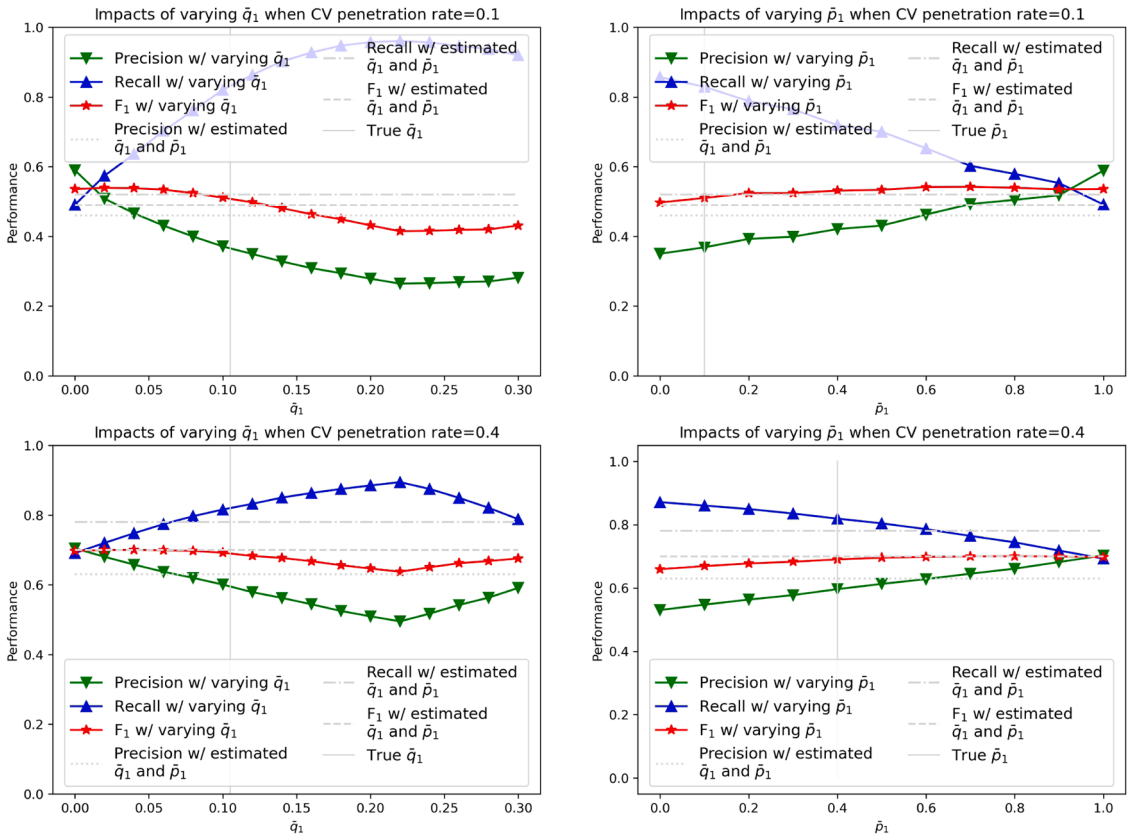

Fig. G3. Impacts of $\bar{q}_1$ and $\bar{p}_1$ estimation accuracy in the upstream lane of the target intermediate lane (lane 2 in Figure 6) on the CVVL-I sub-model.

Finally, the impacts of average arrival rate and CV penetration rate estimation accuracy in the upstream lane of the target intermediate lane (lane 2 in Fig. 6) on the CVVL-I sub-model (lane 2 in Fig. 6, which serves as the only upstream lane of the target intermediate lane for the selected ToI) are shown in Fig. G3. Consistent with the above findings for the source and intermediate lanes, the flat F_1 scores across the full range of tested parameters highlight the CVVL-I sub-model's robustness to input inaccuracies.

Data availability

The authors do not have permission to share data.

References

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for macroscopic fundamental diagram estimation. Transp. Res. C: Emerg. Technol. 71, 184–197.
Cao, P., Miwa, T., Morikawa, T., 2014. Use of probe vehicle data to determine joint probability distributions of vehicle location and speed on an arterial road. Transp. Res. Rec. 2421 (1), 103–114.
Cao, Y., Tang, K., Sun, J., Ji, Y., 2021. Day-to-day dynamic origin–destination flow estimation using connected vehicle trajectories and automatic vehicle identification data. Transp. Res. C: Emerg. Technol. 129, 103241.
Comert, G., 2013. Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals. Transp. Res. B: Methodol. 55, 59–74.
Comert, G., 2016. Queue length estimation from probe vehicles at isolated intersections: estimators for primary parameters. Eur. J. Oper. Res. 252, 502–521.
Comert, G., Cetin, M., 2009. Queue length estimation from probe vehicle location and the impacts of sample size. Eur. J. Oper. Res. 197, 196–202.
Comert, G., Cetin, M., 2011. Analytical evaluation of the error in queue length estimation at traffic signals from prove vehicle data. IEEE Trans. Intell. Transp. Syst. 12, 563–573.
Du, J., Rakha, H., Gayah, V.V., 2016. Deriving macroscopic fundamental diagrams from probe data: issues and proposed solutions. Transp. Res. C: Emerg. Technol. 66, 136–149.
Federal Highway Administration, 2006. Next generation simulation: Peachtree Street dataset. Accessed June 25, 2022, https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Program-Peachtree/mupt-aksf.
Feng, Y., Head, K.L., Khoshmagham, S., Zamanipour, M., 2015. A real-time adaptive signal control in a connected vehicle environment. Transp. Res. C: Emerg. Technol. 55, 460–473.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp. Res. B: Methodol. 42 (9), 759–770.
Goodall, N.J., 2013. Real-time prediction of vehicle locations in a connected vehicle environment (No. FHWA/VCTIR 14-R4).
Goodall, N.J., Smith, B.L., Park, B.B., 2016. Microscopic estimation of freeway vehicle positions from the behavior of connected vehicles. J. Intell. Transp. Syst. 20 (1), 45–54.

Hao, P., Ban, X.J., Guo, D., Ji, Q., 2014. Cycle-by-cycle intersection queue length distribution estimation using sample travel times. Transp. Res. B: Methodol. 68, 185–204.
Iqbal, M.S., Hadi, M., Xiao, Y., 2018. Effect of link-level variations of connected vehicles (CV) proportions on the accuracy and reliability of travel time estimation. IEEE Trans. Intell. Transp. Syst. 20 (1), 87–96.
Jenelius, E., Koutsopoulos, H.N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. Transp. Res. B: Methodol. 53, 64–81.
Jenelius, E., Koutsopoulos, H.N., 2015. Probe vehicle data sampled by time or space: consistent travel time allocation and estimation. Transp. Res. B: Methodol. 71, 120–137.
Jia, S., Pei, X., Yang, Z., Tian, S., Yue, Y., 2020. Novel hybrid neural network for dense depth estimation using on-board monocular images. Transp. Res. Rec. 2674 (12), 312–323.
Jia, S., Pei, X., Jing, X., Yao, D., 2021. Self-supervised 3D reconstruction and ego-motion estimation via on-board monocular video. IEEE Trans. Intell. Transp. Syst. 23 (7), 7557–7569.
Jia, S., Pei, X., Yao, W., Wong, S.C., 2022. Self-supervised depth estimation leveraging global perception and geometric smoothness. IEEE Trans. Intell. Transp. Syst. 24 (2), 1502–1517.
Jia, S., Yao, W., 2023. Joint learning of frequency and spatial domains for dense image prediction. ISPRS J. Photogramm. Remote Sens. 195, 14–28.
Jia, S., Yao, W., 2024. Self-supervised multi-task learning framework for safety and health-oriented road environment surveillance based on connected vehicle visual perception. Int. J. Appl. Earth Obs. Geoinf. 128, 103753.
Jia, S., Wong, S.C., Wong, W., 2023. Uncertainty estimation of connected vehicle penetration rate. Transp. Sci. 57 (5), 1160–1176.
Jia, S., Wong, S.C., Wong, W., 2024a. Modeling residual-vehicle effects on uncertainty estimation of the connected vehicle penetration rate. Transp. Res. C: Emerg. Technol. 168, 104825.
Jia, S., Wong, S.C., Wong, W., 2024b. Modeling residual-vehicle effects in undersaturation conditions on uncertainty estimation of the connected vehicle penetration rate. In: the 25th International Symposium on Transportation and Traffic Theory. Michigan, USA, 15–17 July.
Jia, S., Wong, S.C., Wong, W., 2024c. Estimating real-time traffic state of holding vehicles at signalized intersections using partial connected vehicle trajectory data. Transp. Res. C: Emerg. Technol. 178, 105251.
Jia, S., Wong, S.C., Wong, W., 2024d. Adaptive signal control at partially connected intersections: a stochastic optimization model for uncertain vehicle arrival rates. Transp. Res. B: Methodol. 193, 103161.
Khan, S.M., Dey, K.C., Chowdhury, M., 2017. Real-time traffic state estimation with connected vehicles. IEEE Trans. Intell. Transp. Syst. 18 (7), 1687–1699.
Li, T., Han, X., Ma, J., 2021. Cooperative perception for estimating and predicting microscopic traffic states to manage connected and automated traffic. IEEE Trans. Intell. Transp. Syst. 23 (8), 13694–13707.
Lo, H.K., 1999. A novel traffic signal control formulation. Transp. Res. A: Policy Pract. 33 (6), 433–448.
Lu, Y., Xu, X., Ding, C., Lu, G., 2019. A speed control method at successive signalized intersections under connected vehicles environment. IEEE Intell. Transp. Syst. Mag. 11 (3), 117–128.
Meng, F., Wong, S.C., Wong, W., Li, Y.C., 2017. Estimation of scaling factors for traffic counts based on stationary and mobile sources of data. Int. J. Intell. Transp. Syst. Res. 15 (3), 180–191.
Mousa, S.R., Ishak, S., 2017. An extreme gradient boosting algorithm for freeway short-term travel time prediction using basic safety messages of connected vehicles. In: *Transportation Research Board 96th Annual* Meeting, 2017. Washington, DC, United States.
Rahmani, M., Jenelius, E., Koutsopoulos, H.N., 2015. Non-parametric estimation of route travel time distributions from low-frequency floating car data. Transp. Res. C: Emerg. Technol. 58, 343–362.
Tian, D., Yuan, Y., Qi, H., Lu, Y., Wang, Y., Xia, H., He, A., 2015. A dynamic travel time estimation model based on connected vehicles. Math. Probl. Eng. 2015, 903962.
Wang, P., Zhang, J., Deng, H., Zhang, M., 2020. Real-time urban regional route planning model for connected vehicles based on V2X communication. J. Transp. Land Use 13 (1), 517–538.
Wang, X., Jerome, Z., Wang, Z., Zhang, C., Shen, S., Kumar, V.V., Bai, F., Krajewski, P., Deneau, D., Jawad, A., Jones, R., 2024. Traffic light optimization with low penetration rate vehicle trajectory data. Nat. Commun. 15 (1), 1306.
Wong, W., Wong, S.C., 2015. Systematic bias in transport model calibration arising from the variability of linear data projection. Transp. Res. B: Methodol. 75, 1–18.
Wong, W., Wong, S.C., 2016a. Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection. Transp. Res. B: Methodol. 88, 72–92.
Wong, W., Wong, S.C., 2016b. Evaluation of the impact of traffic incidents using GPS data. Proc. Inst. Civ. Eng. – Transp. 169 (3), 148–162.
Wong, W., Wong, S.C., 2016c. Network topological effects on the macroscopic Bureau of public roads function. Transp. A: Transp. Sci. 12 (3), 272–296.
Wong, W., Wong, S.C., 2019. Unbiased estimation methods of nonlinear transport models based on linearly projected data. Transp. Sci. 53 (3), 665–682.
Wong, W., Wong, S.C., Liu, X., 2019. Bootstrap standard error estimations of nonlinear transport models based on linearly projected data. Transp. A: Transp. Sci. 15 (2), 602–630.
Wong, W., Shen, S., Zhao, Y., Liu, X., 2019. On the estimation of connected vehicle penetration rate based on single-source connected vehicle data. Transp. Res. B: Methodol. 126, 169–191.
Wong, W., Wong, S.C., Liu, X., 2021. Network topological effects on the macroscopic fundamental diagram. Transp. B: Transp. Dyn. 9 (1), 376–398.
Xia, H., Liu, X., Ma, Z., Zhu, F., Zhang, L., Zhao, Y., Wang, Y., 2023. Vehicle speed and position estimation considering microscopic heterogeneous car-following characteristics in connected vehicle environments. J. Adv. Transp. 2023 (1), 6627042.
Yang, X., Lu, Y., Hao, W., 2017. Origin-destination estimation using probe vehicle trajectory and link counts. J. Adv. Transp. 2017, 4341532.
Ye, J., Wang, D., Jia, S., Pei, X., Yang, Z., Zhang, Y., Wong, S.C., 2024. In: CVVLSNet: vehicle location and speed estimation using partial connected vehicle trajectory data. IEEE, pp. 3739–3744.
Yin, Y., 2008. Robust optimal traffic signal timing. Transp. Res. B: Methodol. 42 (10), 911–924.
Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., Liu, H.X., 2019a. Estimation of queue lengths, probe vehicle penetration rates, and traffic volumes at signalized intersections using probe vehicle trajectories. Transp. Res. Rec. 2673 (11), 660–670.
Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., Liu, H.X., 2019b. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transp. Res. C: Emerg. Technol. 107, 70–91.
Zhao, Y., Wong, W., Zheng, J., Liu, H.X., 2022. Maximum likelihood estimation of probe vehicle penetration rates and queue length distributions from probe vehicle data. IEEE Trans. Intell. Transp. Syst. 23 (7), 7628–7636.