

## Transportation Science

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Uncertainty Estimation of Connected Vehicle Penetration Rate

Shaocheng Jia, S. C. Wong, Wai Wong

To cite this article:

Shaocheng Jia, S. C. Wong, Wai Wong (2023) Uncertainty Estimation of Connected Vehicle Penetration Rate. Transportation Science

Published online in Articles in Advance 22 May 2023

. <https://doi.org/10.1287/trsc.2023.1209>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Uncertainty Estimation of Connected Vehicle Penetration Rate

Shaocheng Jia,<sup>a</sup> S. C. Wong,<sup>a,\*</sup> Wai Wong<sup>b</sup>

<sup>a</sup>Department of Civil Engineering, The University of Hong Kong, Hong Kong; <sup>b</sup>Department of Civil and Natural Resource Engineering, University of Canterbury, Christchurch 8041, New Zealand

\*Corresponding author

Contact: [shaocjia@connect.hku.hk](mailto:shaocjia@connect.hku.hk),  <https://orcid.org/0000-0001-9586-5151> (SJ); [hhecwsc@hku.hk](mailto:hhecwsc@hku.hk),  <https://orcid.org/0000-0003-1169-7045> (SCW); [waiwong1012@connect.hku.hk](mailto:waiwong1012@connect.hku.hk),  <https://orcid.org/0000-0001-8150-3280> (WW)

Received: September 25, 2022

Revised: February 11, 2023

Accepted: April 10, 2023

Published Online in Articles in Advance:  
May 22, 2023

<https://doi.org/10.1287/trsc.2023.1209>

Copyright: © 2023 The Author(s)

**Abstract.** Knowledge of the connected vehicle (CV) penetration rate is crucial for realizing numerous beneficial applications during the prolonged transition period to full CV deployment. A recent study described a novel single-source data penetration rate estimator (SSDPRE) for estimating the CV penetration rate solely from CV data. However, despite the unbiasedness of the SSDPRE, it is only a point estimator. Consequently, given the typically nonlinear nature of transportation systems, model estimations or system optimizations conducted with the SSDPRE without considering its variability can generate biased models or suboptimal solutions. Thus, this study proposes a probabilistic penetration rate model for estimating the variability of the results generated by the SSDPRE. An essential input for this model is the constrained queue length distribution, which is the distribution of the number of stopping vehicles in a signal cycle. An exact probabilistic dissipation time model and a simplified constant dissipation time model are developed for estimating this distribution. In addition, to improve the estimation accuracy in real-world situations, the braking and start-up motions of vehicles are considered by constructing a constant time loss model for use in calibrating the dissipation time models. VISSIM simulation demonstrates that the calibrated models accurately describe constrained queue length distributions and estimate the variability of the results generated by the SSDPRE. Furthermore, applications of the calibrated models to the next-generation simulation data set and a simple CV-based adaptive signal control scheme demonstrate the readiness of the models for use in real-world situations and the potential of the models to improve system optimizations.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as “*Transportation Science*. Copyright © 2023 The Author(s). <https://doi.org/10.1287/trsc.2023.1209>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.”

**Funding:** This work was supported by The University of Hong Kong [Francis S Y Bong Professorship in Engineering and Postgraduate Scholarship] and by the Council of the Hong Kong Special Administrative Region, China [Grants 17204919 and 17205822].

**Supplemental Material:** The online appendices are available at <https://doi.org/10.1287/trsc.2023.1209>.

**Keywords:** connected vehicle penetration rate • uncertainty estimation • constrained queue length estimation • stochastic modeling • signal control with uncertainty

## 1. Introduction

The rapid development of communication technologies has expedited the evolution of the Internet of Things (IoT). Transportation systems are a part of the IoT and are therefore following the trend by increasingly adopting connected vehicle (CV) technology, which allows CVs to share travel information (e.g., timestamps, speeds, and locations) with nearby CVs and infrastructures. The huge volumes of CV data that are generated present numerous opportunities for improving the efficiency, safety, and resilience of transportation systems via various applications. However, the range of practical difficulties associated with accessing and processing CV

data, such as budget constraints, privacy concerns, and corporate confidentiality, means that the transition to full CV deployment is likely to be prolonged, and that 100% CV deployment may never be achieved. Due to this current absence of complete CV data, various beneficial applications of available CV data typically employ inferences about traffic data.

The CV penetration rate is the CV-to-total traffic ratio within a spatiotemporal volume and is an essential input for traffic data inference and many other useful applications. For instance, the CV penetration rate is a necessary input for the location and speed algorithm developed by Feng et al. (2015), which infers arrival tables for the

controlled optimization of a phase algorithm (Sen and Head 1997) in CV-based adaptive signal optimization. By assuming that a CV penetration rate and queue length distribution are known, Comert and Cetin (2009, 2011) and Comert (2013) have derived various methods for cycle-by-cycle queue estimation at isolated junctions. Similarly, under the assumption of a known CV penetration rate, Hao et al. (2014) established a Bayesian network-based model for estimating cycle-by-cycle queue length distributions. Based on shockwave theory, Argote et al. (2011) developed a method for queue length estimation and examined the minimum CV penetration rate required to ensure accuracy. CV penetration rates based on linear data projection (Wong and Wong 2015, 2016a, 2019; Wong, Wong, and Liu 2019) have also been applied to infer traffic flow for estimating macroscopic Bureau of Public Roads scenarios (Wong and Wong 2015, 2016a, 2016c) and traffic density or accumulation in the macroscopic fundamental diagram (Geroliminis and Daganzo 2008; Ambühl and Menendez 2016; Du, Rakha, and Gayah 2016; Wong and Wong 2019; Wong, Wong, and Liu et al. 2019, 2021). Other relevant applications include origin–destination estimations (Yang, Lu, and Hao 2017; Wang et al. 2020; Cao et al. 2021), travel time and speed estimations (Jenelius and Koutsopoulos 2013, 2015; Rahmani, Jenelius, and Koutsopoulos 2015; Tian et al. 2015; Mousa and Ishak 2017; Khan, Dey, and Chowdhury 2017; Iqbal, Hadi, and Xiao 2018; Lu et al. 2019), traffic incident impact evaluation (Wong and Wong 2016b), and time exposure estimation in road safety studies (Meng et al. 2017b). The aforementioned studies have demonstrated that knowledge of the CV penetration rate is fundamentally important during the transition period to full CV deployment, and, hence, CV penetration rate estimation is a critical current research topic in the field of transportation.

Current methods for measuring CV penetration rates primarily rely on both CV data and loop detector data. The CV penetration rate for a link outfitted with a loop detector can be easily obtained by dividing the CV flow across the link over a certain period by the total traffic flow passing the detector over the same period. However, the high capital and maintenance cost of loop detectors typically hinders their universal deployment. As there are also circumstances in which a detector may be out of service for a certain period, a simple but useful approach for estimating the CV penetration rates of links without detectors has been devised that models the CV penetration rates of links within a network as a probability distribution (Wong and Wong 2015, 2016a, 2019; Wong, Wong, and Liu 2019). This probability distribution is approximated by the distribution of CV penetration rates sampled from the links in the target network that are outfitted with detectors. Due to the geographical proximity, the mean of this distribution can be taken as the expected CV penetration rate for links without detectors. It follows that the variance of this distribution

represents the spatial distribution. However, despite the simplicity of this approach, its underlying assumption of independent and identically distributed CV penetration rates within a network may not always hold, because of factors such as land-use heterogeneities. Meng et al. (2017a) therefore used data from Hong Kong to establish an empirical model for estimating a CV penetration rate that exploits land-use variables as inputs. However, the use of local data greatly limits its universal application.

To overcome the aforementioned deficiencies, researchers have devoted much effort to estimating CV penetration rates using only CV data. Comert (2016) assumed a Poisson-distributed arrival pattern for deriving several estimators for CV penetration rates, but this assumption limits their generalizability. More recently, Wong et al. (2019) devised the first analytical and nonparametric method—the single-source data penetration rate estimator (SSDPRE)—which unbiasedly estimates the CV penetration rate solely from CV data. The SSDPRE uses the information on stopping locations of CVs at a signalized intersection to deduce the number of non-CVs in front of the last CV. Then, the SSDPRE applies this partial queue information to subtly fuse two estimation mechanisms—(1) the direct estimation of the probability of the first stopping vehicle being a CV, and (2) estimation of the CV penetration rate of the partial queue—to afford an unbiased estimation of the true CV penetration rate. Several novel methods for estimating CV penetration rates have been developed by Zhao et al. (2019a, 2019b, 2022); these adopt a maximum likelihood estimation approach to approximate and exploit the distribution of stopping positions of vehicles in queues. Although the aforementioned methods for CV penetration rate estimation are valuable, they are point estimators, which means that their direct application (i.e., without considering their uncertainties or fluctuations) to transportation systems for the estimation of transport models or the optimization of systems can generate biased models or suboptimal solutions (Yin 2008; Wong and Wong 2015, 2016a, 2019; Wong, Wong, and Liu 2019). As such, the variability of an estimator must be considered to enable unbiased models and optimal solutions to be obtained. Nevertheless, methods for estimating such variability remain unexplored.

This paper fills the aforementioned research gap by deriving a generic probabilistic penetration rate (PPR) model for estimating the variability of the results provided by the SSDPRE. The essential input for the PPR model is the constrained queue length distribution, which is formed by the number of stopping vehicles in a cycle. A probabilistic dissipation time (PDT) model and a constant dissipation time (CDT) model are derived to model the constrained queue length distribution. The PDT model assumes a random arrival, such that the dissipation time is probabilistic, and uses the simple notion of time interval partitioning to subtly determine the exact constrained queue length distribution of a random arrival. In contrast,

for the purpose of deriving the average constrained queue length, the CDT model approximates the random arrivals by an average arrival pattern, such that the dissipation time is constant for a given average arrival rate. While generally keeping a high level of accuracy, this simplification significantly improves computation efficiency. Simulation studies based on the vertical queue assumption demonstrate the excellent accuracy and efficiency of these models.

However, these models' ignorance of real-world braking and start-up motions of vehicles constrains their performance. Therefore, a constant time loss (CTL) model that incorporates real-world braking and start-up motions of vehicles is devised and applied to calibrate the models. A comprehensive VISSIM simulation study demonstrates that the calibrated models accurately represent the constrained queue length distribution and estimate the variability of the results afforded by the SSDPRE. The calibrated models' suitability for real-world use is demonstrated by their application to a next-generation simulation (NGSIM) data set collected on Peachtree Street in Atlanta, Georgia, USA (Federal Highway Administration 2006). A simple illustrative application of CV-based adaptive signal control that also demonstrates the significant potential improvement in system optimization via the incorporation of CV penetration rate variability is presented. Thus, this study contributes to the transport field by providing the missing piece of the puzzle of uncertainty estimation, which must be performed when employing the SSDPRE. Consequently, equipped with information on the variability of results generated by the SSDPRE, optimal solutions or unbiased models can be obtained for system optimizations or model estimations.

The remainder of this paper is organized as follows. Section 2 defines the problem statement. Section 3 derives the PPR model, and Section 4 details the PDT and CDT models. Section 5 introduces the CTL model and presents a comprehensive simulation study. Section 6 provides real-world validation of the model on an NGSIM data set and an illustrative application of CV-based adaptive signal control. Section 7 concludes the paper.

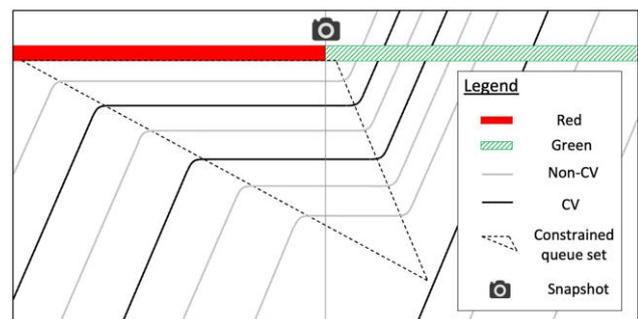
## 2. Problem Statement

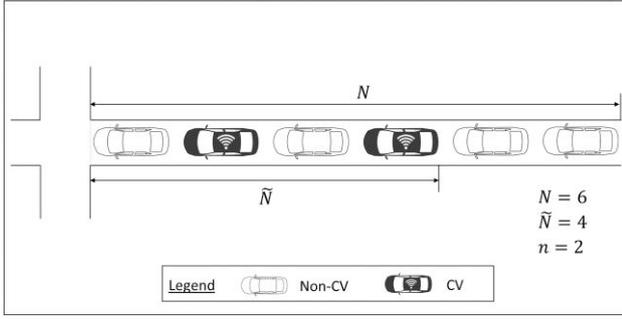
Efficient, safe, and resilient transportation systems rely on system optimizations and models based on complete and accurate traffic information. Nevertheless, such information is typically unavailable. The emergence of CV technology advances further toward complete and accurate traffic information. CVs are probe vehicles circulating with regular vehicles on road networks and provide detailed travel information. However, the prolonged transition period toward full CV deployment means that data projection or scaling is necessary to infer complete information from these data. The CV penetration rate is indispensable in bridging the gap between partial and

complete information. One simple example is traffic flow estimation across a link. Given that the hourly CV flow across a link is 10 CVs/hour, if the CV penetration rate is known to be 10%, then the hourly total traffic flow can be easily estimated by dividing 10 CVs/hour by 10%, which gives 100 vehicles/hour (veh/h). Similarly, other applications using the CV penetration rate as essential inputs include the aforementioned arrival table estimation (Feng et al. 2015), queue length estimations (Comert and Cetin 2009, 2011), and time exposure estimation (Meng et al. 2017b). Despite the central role of CV penetration rates, existing methods only provide their point estimators. The uncertainty of the CV penetration rate, governing the optimality of system optimizations and the biasedness of model estimations, remains unexplored. On the basis of the SSDPRE presented by Wong et al. (2019), the present study aims at deriving analytical models that quantify such uncertainty.

More specifically, consider Figure 1, which illustrates the trajectories of a set of vehicles, comprising CVs and non-CVs, traversing a signalized intersection. A constrained queue set,  $\Psi$ , is defined as the set of vehicles that have been stopped by the red signal, and is enclosed by a triangular spatiotemporal area formed by shockwaves. Figure 2 shows a snapshot of  $\Psi$  taken at the end of the red signal. Vehicles that enter the triangular spatiotemporal area after the end of the red signal are also counted as members of  $\Psi$ . The constrained queue length,  $|\Psi|$ , is defined as the number of stopping vehicles,  $N$ ; here,  $N$  is 6. CVs continuously broadcast their travel information in their basic safety messages, which include information such as their timestamps, locations, speeds, and headings. Thus, the number of CVs,  $n$ , is known; here,  $n$  is 2. Define average effective vehicle length,  $L_e$ , as the average distance between the rear end of a preceding stopping vehicle and the rear end of its following stopping vehicle. If all of the vehicles are identical in physical size, then each vehicle has the same effective vehicle length. For cases with multiple-vehicle classes, the average effective vehicle length can be easily updated using empirical data. Given the location of the stop bar, the number of

Figure 1. (Color online) Vehicle Trajectories Traversing a Signalized Intersection



**Figure 2.** Snapshot of the Constrained Queue

stopped vehicles in front of the last CV (including itself),  $\tilde{N}$ , can be easily determined by dividing the distance from the stop bar to the rear end of the last CV,  $d_s$ , by the effective vehicle length,  $L_e$ . In Figure 2,  $\tilde{N}$  is 4. Non-CVs behind the last CV are unobservable.

In such a scenario, the penetration rate exhibits a spatiotemporal pattern, both CVs and non-CVs are sufficiently well mixed during any time period at any link, and  $N > 0$ . Accordingly, Wong et al. (2019) constructed a novel, simple, analytical, and—most importantly—unbiased SSDPRE to estimate the CV penetration rate solely from CV data. Let  $i$  and  $m$  denote the  $i$ th constrained queue and the total number of constrained queues, respectively. Thus, for  $\forall i \in [1, m]$ , the SSDPRE is defined as follows (Equation (1)):

$$\text{SSDPRE} = \frac{\sum_{i=1}^m \tilde{p}_i}{m}, \quad (1)$$

where

$$\tilde{p}_i = S(n_i, \tilde{N}_i) = \begin{cases} \frac{n_i - 1}{\tilde{N}_i - 1} & \text{if } n_i > 1 \text{ and } \tilde{N}_i > 1 \\ 1 & \text{if } n_i = 1 \text{ and } \tilde{N}_i = 1 \\ 0 & \text{if } n_i = 1 \text{ and } \tilde{N}_i > 1 \\ 0 & \text{if } n_i = 0 \text{ and } \tilde{N}_i = 0. \end{cases} \quad (2)$$

It follows that  $\tilde{p}_i, \forall i \in [1, m]$  form a distribution with a mean that is the result generated by the SSDPRE. However, although the SSDPRE is an unbiased estimator of the underlying true CV penetration rate  $p$ , it is a point estimator. Therefore, given that transportation systems are typically nonlinear, SSDPRE-based system optimizations or model estimations that do not consider this estimator's variability could afford suboptimal solutions and biased models. Accordingly, as the variance of the distribution of  $\tilde{p}_i$ ,  $\text{Var}(\tilde{p}_i)$ , remains unexplored, the objective of this study is to develop an analytical method for estimating the uncertainty of results generated by the SSDPRE.

### 3. Uncertainty Estimation for the SSDPRE

This section introduces the PPR model for estimating the uncertainty of the results generated by the SSDPRE. In reality, both  $N$  and  $n$  are random variables. For the sake

of clarity, the proof of the PPR model is divided into three subsections. Section 3.1 considers the simplest cases, in which  $N$  and  $n$  are constant. Section 3.2 relaxes the constraint of a constant  $n$ , by considering cases with a constant  $N$  and a varying  $n$ . Section 3.3 also relaxes the constraint of a constant  $N$ , by considering cases with a varying  $N$  and  $n$ .

#### 3.1. Constant $N$ and $n$

For any given set of constant  $N$  and  $n$ ,  $\tilde{p}_i$  is solely dependent on the stopping location of the last CV. Thus, its uncertainty depends on the number of permutations of the stopping vehicles for each possible stopping location of the last CV. Proposition 1 states the solutions for  $E(\tilde{p})$  and  $\text{Var}(\tilde{p})$  under the condition of constant  $N$  and  $n$ .

**Proposition 1.** Given that  $\Psi \neq \emptyset$  with a constant  $N$  and  $n$ , where  $N > 0$  and  $N \geq n \geq 0$ ,  $E(\tilde{p})$  and  $\text{Var}(\tilde{p})$  are given as follows:

$$E(\tilde{p}) = \begin{cases} \frac{n}{N} & \text{if } n \geq 1 \\ 0 & \text{if } n = 0, \end{cases} \quad (3a)$$

$$\text{Var}(\tilde{p}) = V_1(n, N)$$

$$= \begin{cases} \frac{\sum_{i=1}^{N-n+1} \frac{n-1}{N-i} \binom{N-i-1}{n-2}}{\binom{N}{n}} - \frac{n^2}{N^2} & \text{if } n > 1 \\ \frac{n^2 - 2n + N}{N^2} & \text{if } n = 1 \\ 0 & \text{if } n = 0. \end{cases} \quad (3b)$$

**Proof.** A detailed proof is presented in Online Appendix A.

To validate and demonstrate the superiority of Proposition 1, a series of comprehensive simulation experiments based on the vertical queue assumption were conducted and are presented in Online Appendix D. Various combinations of fixed  $N$  and  $n$  (i.e., different  $p$ ) were considered. For example, with  $N = 30$  and  $n = 15$  (i.e.,  $p = 0.5$ ), all of the possible permutations were enumerated and the corresponding  $\tilde{p}$  were evaluated using the SSDPRE. The mean and variance of  $\tilde{p}$  were 0.5 and 0.00061, respectively. In a machine with an Intel Core i7-10700 CPU, the computation costs for the full enumeration and evaluation of the mean and variance of  $\tilde{p}$  were 944 seconds (s) and 944 s, respectively. In contrast, when Proposition 1 was used, identical results were obtained with a negligible computation cost.

#### 3.2. Constant $N$ and Varying $n$

Assuming that all of the CVs and non-CVs are sufficiently well mixed within a link, each vehicle shares the

same probability of  $p$  being a CV and of  $1 - p$  being a non-CV. Consequently, a binomial distribution,  $B(N, p)$ , can be used to model  $n$ . For any given  $N$ , the variability of  $\tilde{p}_i$  is dependent on (i) the variation of  $n$ , and (ii) the number of permutations of the stopping vehicles for each possible stopping location of the last CV. Proposition 2 presents  $E(\tilde{p})$  and  $Var(\tilde{p})$  under the condition of a constant  $N$  and a varying  $n$ .

**Proposition 2.** Given that  $\Psi \neq \emptyset$  with a constant  $N$  and a varying  $n \sim B(N, p)$ , where  $N > 0$  and  $N \geq n \geq 0$ ,  $E(\tilde{p})$  and  $Var(\tilde{p})$  are given as follows:

$$E(\tilde{p}) = \binom{N}{0} p^0 (1-p)^N + \sum_{i=1}^N \sum_{j=1}^{N-i+1} \frac{\binom{N-j}{i-1}}{\binom{N}{i}} \binom{N}{i} p^i (1-p)^{N-i} S(i, N-j+1) = p, \quad (4a)$$

$$Var(\tilde{p}) = V_2(N, p) = \begin{cases} \sum_{i=2}^N p^i (1-p)^{N-i} \left[ V_1(i, N) + \left(\frac{i}{N}\right)^2 \right] \binom{N}{i} - p^2 + p(1-p)^{N-1} & \text{if } N > 1 \\ p(1-p) & \text{if } N = 1. \end{cases} \quad (4b)$$

**Proof.** A detailed proof is presented in Online Appendix E.

Similarly, to validate Proposition 2, a series of comprehensive simulation experiments based on the vertical queue assumption were conducted, as shown in Online Appendix H. Various combinations of fixed  $N$  and varying  $n \sim B(N, p)$  were considered. For example, with  $N = 30$  and  $n \sim B(10, 0.5)$ , all of the possible  $n$  were enumerated. Then, for each  $n$ , both  $N$  and  $n$  were fixed, and the problem was reduced to the conditions stated in Proposition 1. Thus, all of the permutations could be enumerated, and the corresponding  $\tilde{p}$  were evaluated using the SSDPRE. The mean and variance of  $\tilde{p}$  were 0.5 and 0.00895, respectively. The computation costs for the enumeration and evaluation of the mean and variance of  $\tilde{p}$  were 11,161 s and 11,617 s, respectively. In contrast, when Proposition 2 was used, identical results were obtained in a fraction of a second: the computation costs for the mean and variance of  $\tilde{p}$  were 0 s and 0.002 s, respectively.

### 3.3. Varying $N$ and $n$

Here,  $n$  follows  $B(N, p)$ , as described in Proposition 2, whereas  $N$  follows any counting distribution. As such, the variability of  $\tilde{p}_i$  is dependent on (i) the variation of  $N$ , (ii) the variation of  $n$ , and (iii) the number of permutations of the stopping vehicles for each possible stopping location of the last CV. Proposition 3 completes the PPR model and presents  $E(\tilde{p})$  and  $Var(\tilde{p})$  under the condition of varying  $N$  and  $n$ .

**Proposition 3.** Considering any  $\Psi$  with  $N$  following any counting distribution, such that  $P(N = i) = \pi_i, \forall i = 0, 1, 2, \dots, k$ , and a varying  $n \sim B(N, p)$ , where  $N \geq n \geq 0$ ,  $E(\tilde{p})$  and  $Var(\tilde{p})$  are given as follows:

$$E(\tilde{p}) = \lim_{k \rightarrow +\infty} \left[ \left( \pi_0 p + \sum_{i=1}^k \sum_{j=1}^i \sum_{m=1}^{i-j+1} \pi_i \frac{\binom{i-m}{j-1}}{\binom{i}{j}} \binom{i}{j} p^j (1-p)^{i-j} S(i, N-j+1) \right) \right] = p, \quad (5a)$$

$$Var(\tilde{p}) = \lim_{k \rightarrow +\infty} \left[ \sum_{i=1}^k \pi_i V_2(i, p) \right]. \quad (5b)$$

**3.3.1. Proof of Proposition 3.** In a constrained queue set,  $\Psi$ , with a varying  $N$  following any counting distribution and a varying  $n$  following  $B(N, p)$ , both  $N$  and  $n$  can vary from sample to sample. As such, the permutations of the constrained queues can be arranged into three levels. The first level contains all of the permutations grouped in terms of  $N$ . The possible outcomes  $N$  and the corresponding probabilities,  $\pi_i, \forall i = 0, 1, 2, \dots, k$ , are determined by an observed, assumed, or derived counting distribution. However, the  $S(\cdot)$  function in the SSDPRE is not defined for empty queues, where  $N = 0$ . In such a case, the absence of information means that  $\tilde{p}$  can be directly replaced by  $p$ . Aside from this  $N = 0$  case, under each  $N$ , all of the permutations are arranged into subgroups in terms of  $n$  in the second level, and Proposition 2 can be applied to evaluate  $E(\tilde{p})$  and  $Var(\tilde{p})$ . For instance, when  $N = 1$ , the corresponding probability,  $\pi_1$ , can be obtained based on the counting distribution. In the second level,  $n$  can either be 0 or 1 with the probability of  $\binom{1}{0} p^0 (1-p)^1$  or  $\binom{1}{1} p^1 (1-p)^0$ , respectively. In the third level,  $\tilde{p}$  for each permutation group of any fixed pair of  $N$  and  $n$  and the probability of observing a permutation from each group are obtained. When  $N = 1$  and  $n = 0$ ,  $\tilde{p}$  is given by  $S(0, 1)$ , and the corresponding probability is 1. Similarly, when  $N = 1$  and  $n = 1$ ,  $\tilde{p}$  is given by  $S(1, 1)$  and the corresponding probability is also 1. Table 1 enumerates all of the possible permutations in the three levels.

(1) Proof of mean:

$$E(\tilde{p}) = \lim_{k \rightarrow +\infty} \left\{ \pi_0 p + \sum_{i=1}^k \left[ \pi_i \binom{i}{0} p^0 (1-p)^i \cdot 1 \cdot S(0, i) + \sum_{j=1}^i \sum_{m=1}^{i-j+1} \pi_i \binom{i-m}{j-1} p^j (1-p)^{i-j} S(i, j-m+1) \right] \right\}. \quad (6)$$

As  $\sum_{j=1}^i \sum_{m=1}^{i-j+1} \binom{i-m}{j-1} p^j (1-p)^{i-j} S(i, j-m+1)$  is exactly

**Table 1.** Enumeration of All of the Possible Permutations Under Varying  $N$  and  $n$ 

First level: $N$		Second level: $n$		Third level: $l_{CV}$	
$N$	Probability	$n$	Probability	$\tilde{p}$	Probability
0	$\pi_0$	0	1	$p$	1
1	$\pi_1$	0	$\binom{1}{0} p^0 (1-p)^1$	$S(0,1)$	1
		1	$\binom{1}{1} p^1 (1-p)^0$	$S(1,1)$	1
...	...	...	...	...	...
$k$ ( $k \rightarrow +\infty$ )	$\pi_k$	0	$\binom{k}{0} p^0 (1-p)^k$	$S(0,k)$	1
	$\pi_k$	1	$\binom{k}{1} p^1 (1-p)^{k-1}$	$S(1, k-m+1),$ $\forall m=1, 2, \dots, k$	$\binom{k-m}{1-1} / \binom{k}{1},$ $\forall m=1, 2, \dots, k$
		2	$\binom{k}{2} p^2 (1-p)^{k-2}$	$S(2, k-m+1),$ $\forall m=1, 2, \dots, k-1$	$\binom{k-m}{2-1} / \binom{k}{2},$ $\forall m=1, 2, \dots, k-1$
	$\pi_k$	$k$	$\binom{k}{k} p^k (1-p)^0$	$S(k, k-m+1),$ $\forall m=1$	$\binom{k-m}{k-1} / \binom{k}{k},$ $\forall m=1$

equivalent to Equation (E1) in Online Appendix E, which is proven to be equal to  $p$ , Equation (6) can be rewritten as follows:

$$E(\tilde{p}) = \lim_{k \rightarrow +\infty} \left( \pi_0 p + \sum_{i=1}^k \pi_i p \right) = p \lim_{k \rightarrow +\infty} \sum_{i=0}^k \pi_i = p. \quad (7)$$

(2) Proof of variance:

$$\begin{aligned} \text{Var}(\tilde{p}) &= [p - E(\tilde{p})]^2 \pi_0 + \lim_{k \rightarrow +\infty} \sum_{i=1}^k \left\{ \pi_i \binom{i}{0} p^0 (1-p)^i \right. \\ &\quad \cdot [0 - E(\tilde{p})]^2 + \sum_{j=1}^i \sum_{m=1}^{i-j+1} \pi_i \binom{i-m}{j-1} p^j (1-p)^{i-j} \\ &\quad \left. [S(j, i-m+1) - E(\tilde{p})]^2 \right\}. \end{aligned} \quad (8)$$

As

$$\begin{aligned} &\binom{i}{0} p^0 (1-p)^i \cdot [0 - E(\tilde{p})]^2 + \sum_{j=1}^i \sum_{m=1}^{i-j+1} \binom{i-m}{j-1} p^j \\ &\quad (1-p)^{i-j} [S(j, i-m+1) - E(\tilde{p})]^2 = V_2(i, p) \end{aligned}$$

according to Equations (E3)–(E7) in Online Appendix E, we have

$$V(\tilde{p}) = \lim_{k \rightarrow +\infty} \left[ \sum_{i=1}^k \pi_i V_2(i, p) \right]. \quad (9)$$

Q.E.D.

**3.3.2. Vertical Queue Experiments.** The application of Proposition 3 requires an observed or assumed constrained

queue length distribution. To enable validation of Proposition 3,  $N$  is assumed to follow a Poisson distribution,  $Pois(\lambda)$ , where  $\lambda$  is the average constrained queue length. The following corollaries are obtained.

**Corollary 1.** Given that  $N \sim Pois(\lambda)$  and  $n \sim B(N, p)$ ,  $E(\tilde{p})$  and  $\text{Var}(\tilde{p})$  are given as follows:

$$\begin{aligned} E(\tilde{p}) &= \lim_{k \rightarrow +\infty} \left[ e^{-\lambda} p + \sum_{i=1}^k \sum_{j=1}^i \sum_{m=1}^{i-j+1} \frac{\lambda^i e^{-\lambda}}{i!} \binom{i-m}{j-1} \right. \\ &\quad \left. p^j (1-p)^{i-j} S(i, N-j+1) \right] = p \end{aligned} \quad (10a)$$

$$\text{Var}(\tilde{p}) = \lim_{k \rightarrow +\infty} \left[ \sum_{i=1}^k \frac{\lambda^i e^{-\lambda}}{i!} V_2(i, p) \right]. \quad (10b)$$

**Proof.** By replacing the counting distribution in Proposition 3 with a  $Pois(\lambda)$ , the proof is completed.

**Corollary 2.** Given that  $N \sim Pois(\lambda)$  and  $n \sim B(N, p)$ , the joint probability distribution of  $n$  and  $\tilde{N}$  is given as follows:

$$P(n=i, \tilde{N}=j) = \begin{cases} \pi_0 + \sum_{z=1}^k \pi_z (1-p)^z, & i=0, j=0 \\ \sum_{z=j}^k \pi_z \binom{j-1}{i-1} p^i (1-p)^{z-i}, & \forall i, j=1, 2, \dots, k, j \geq i \end{cases} \quad (11)$$

where  $\pi_i = P(N=i)$ ,  $\forall i=0, 1, 2, \dots, k$ .

**Proof.** The detailed proof is presented in Online Appendix I.

Based on Corollary 1, a series of comprehensive simulation experiments under the vertical queue assumption were conducted and are presented in Online Appendix J. These demonstrate the superiority and efficiency of Proposition 3 for various combinations of  $\lambda$  and  $p$ . As  $N \sim \text{Pois}(\lambda)$ ,  $N$  has infinitely many possible outcomes, and the computation cost increases exponentially as the number of possible  $N$  considered increases. Therefore, it is impossible to enumerate all of the possible permutations for the validation. Instead, random sampling of the possible permutations was conducted, based on the selected  $\text{Pois}(\lambda)$  and  $B(N, p)$  with various sample sizes (i.e., 100,000, 1,000,000, and 10,000,000), to estimate the mean and variance of  $\tilde{p}$ . Only when the sample size approaches infinity do such estimates asymptotically approach their true values. Thus, the estimates given here cannot be regarded as the ground truths for this case. Nevertheless, a gradual convergence of the mean and variance of  $\tilde{p}$  can be expected and their orders serve as a useful reference for the validation of Corollary 1 and thus Proposition 3. For example, with  $\lambda = 10$  and  $p = 0.1$ , the use of random sampling resulted in the computation time increasing linearly, from approximately 3.5s to 35s to 350s, as the sample sizes increased from 100,000 to 1,000,000 to 10,000,000, respectively. The means of  $\tilde{p}$  were 0.1 for all three selected sample sizes, indicating good convergence when the sample size reached 100,000. Conversely, as the sample size of random sampling increased as above, the variance of  $\tilde{p}$  gradually increased from 0.04870 to 0.04875 to 0.04892. Although clear convergence was not observed for this parameter, the order of the estimates served as a useful reference for the validation.

To obtain estimates using Corollary 1, the number of terms must be determined and summed. The greater the value of  $k$ , the higher the accuracy of the estimates; thus, as  $k$  tends to infinity, the estimates asymptotically approach the true values. In the case of  $\lambda = 10$  and  $p = 0.1$ ,  $k$  was chosen to be 20, 30, and 40; subsequently, the means of  $\tilde{p}$  were instantly determined to be 0.1 for all of these three values of  $k$ . In addition, as  $k$  increased from 20 to 30 to 40, the variance of  $\tilde{p}$  gradually increased from 0.05068 to 0.05071 and converged at 0.05071, with a minimal increase in computation time (from 0.010s to 0.025s to 0.054s, respectively). These results clearly demonstrate that estimations based on Corollary 1 were obtained much more efficiently than those based on random sampling. In addition, identical means for  $\tilde{p}$  were estimated by the two approaches, and the variances of  $\tilde{p}$  estimated by the two approaches were of the same order of magnitude and were only slightly different. To ascertain the accuracy of Corollary 1, all of the possible permutations for the case with  $\lambda = 10$ ,  $p = 0.1$ , and  $k = 20$  were fully enumerated, as this was computationally tractable. The

ground truth of the variance of  $\tilde{p}$  was 0.05068, which was the same as the result obtained based on Corollary 1.

## 4. Estimation of Constrained Queue Length Distribution

Proposition 3 states the PPR model for uncertainty estimation for the SSDPRE. The method is generic and flexible, as the constrained queue length distribution can be any counting distribution. In this section, two models are introduced for estimating the essential constrained queue length distribution: the PDT model and the CDT model.

### 4.1. PDT Model

Section 4.1.1 derives the base PDT model, and Section 4.1.2 presents the modified PDT model, which is more efficient. Section 4.1.3 compares the performance of the PDT model with that of the modified PDT model.

**4.1.1. Base PDT Model.** By considering random arrivals, the arrival pattern can be assumed to follow a Poisson distribution,  $\text{Pois}(qt)$ , where  $q$  is the average arrival rate;  $t$  is a chosen observation period; and  $qt$  is the average number of arrivals in  $t$ . The probability mass function,  $f(\eta; qt)$ , is thus given as follows:

$$f(\eta; qt) = \frac{(qt)^\eta e^{-qt}}{\eta!}, \quad (12)$$

where  $\eta$  is the number of arrivals.

For any constrained queue set,  $\Psi$ , with a constrained queue length  $N = k$ ,  $k$  can be decomposed into a set of sequences, where the sum of each sequence equals  $k$ . In each sequence, the elements represent the numbers of arrivals in different partitioned time intervals. Let the first partitioned interval be the red period  $r$ , the second partitioned interval be the discharge period for the vehicles that have arrived during the first partitioned interval, the third partitioned interval be the discharge period for the vehicles that have arrived during the second partitioned interval, and so forth. Similarly, the first element in a sequence represents the number of arrivals during the first partitioned interval, the second element in a sequence represents the number of arrivals during the second partitioned interval, and so forth. A sequence terminates with an element of zero value. Formally,  $\Omega_k = \{\Phi_{k,1}, \Phi_{k,2}, \dots, \Phi_{k,\omega_k}\}$ , which is the set of possible sequences where the combinational sum of each sequence equals  $k$ , where  $\omega_k = |\Omega_k|$ , the cardinality of the set  $\Omega_k$ , representing the number of these possible sequences;  $\Phi_{k,i} = \{\eta_{k,i,1}, \eta_{k,i,2}, \dots, \eta_{k,i,\varphi_{k,i}}\}$ , which is the  $i$ th sequence,  $\forall i \in [1, \omega_k]$ ;  $\varphi_{k,i} = |\Phi_{k,i}|$ , the cardinality of the set  $\Phi_{k,i}$ , representing the number of elements in  $\Phi_{k,i}$ ;  $\eta_{k,i,j} \in \mathbb{N}^+$ ,  $\forall j \in [1, \varphi_{k,i} - 1]$ ;  $\eta_{k,i,\varphi_{k,i}} = 0$ , indicating the termination of a sequence; and  $\sum_{j=1}^{\varphi_{k,i}} \eta_{k,i,j} = k$ ,  $\forall i \in [1, \omega_k]$ . For example, when  $k = 5$ , the  $i$ th sequence,  $\Phi_{5,i}$ , could be  $\{\eta_{5,i,1}, \eta_{5,i,2}, \eta_{5,i,3}, \eta_{5,i,4}$ ,

$\eta_{5,i,5} = \{1, 2, 1, 1, 0\}$  subject to (s.t.).  $\sum_{j=1}^5 \eta_{5,i,j} = 5$ . Based on the notion of time interval partitioning, Proposition 4 describes the constrained queue length distribution under the condition of random arrival.

**Proposition 4.** Given an arrival pattern following a Poisson distribution with an average arrival rate of  $q$ , a red period of  $r$ , and a saturation headway of  $\tau$ , the constrained queue length distribution is given as follows:

$$P(N = k) = \begin{cases} \sum_{i=1}^{\omega_k} f(\eta_{k,i,1}; qr) \prod_{j=2}^{\varphi_{k,i}} f(\eta_{k,i,j}; q\eta_{k,i,j-1}\tau) & \text{if } k \in \mathbb{N}^+ \\ f(0; qr) & \text{if } k = 0. \end{cases} \quad (13)$$

**Proof.** Considering the commonly observed random arrivals, the vehicle arrival pattern can be assumed to follow a Poisson distribution with an average arrival rate of  $q$ . Let  $r$  and  $\tau$  be the red period and saturated discharge headway, respectively.

It follows that  $\forall \Psi$  with  $|\Psi| = k$ ,  $\exists \omega_k$  sequences s.t.  $\sum_{j=1}^{\varphi_{k,i}} \eta_{k,i,j} = k$ ,  $\forall i \in [1, \omega_k]$ . Considering the  $i$ th sequence  $\Phi_{k,i} = \{\eta_{k,i,1}, \eta_{k,i,2}, \dots, \eta_{k,i,\varphi_{k,i}}\}$ , the  $j$ th element  $\eta_{k,i,j}$ ,  $\forall j \in [1, \varphi_{k,i}]$ , represents the number of arrivals during the  $j$ th partitioned interval. For  $j = 1$ , the probability of observing  $\eta_{k,i,1}$  arrivals during the first partitioned interval is given by  $f(\eta_{k,i,1}; qr)$ . For  $1 < j \leq \varphi_{k,i}$ , the probability of observing  $\eta_{k,i,j}$  arrivals during the  $j$ th partitioned interval is given by  $f(\eta_{k,i,j}; q\eta_{k,i,j-1}\tau)$ . Thus, when  $k > 0$ , considering all of the possible sequences, the probability of  $N = k$  is given as follows:

$$P(N = k) = \sum_{i=1}^{\omega_k} f(\eta_{k,i,1}; qr) \prod_{j=2}^{\varphi_{k,i}} f(\eta_{k,i,j}; q\eta_{k,i,j-1}\tau). \quad (14)$$

In particular, when  $k = 0$ ,  $\omega_k = 1$  and  $\Phi_0 = \{0\}$ . In such cases,  $P(N = 0) = f(0; qr)$ . Q.E.D.

For illustrative purposes, Table 2 presents all of the possible sequences and probabilities for the case when  $k = 4$ . Thus, the probability of  $N = 4$  is the sum of all of the probabilities in the last column.

**Table 2.** Enumeration of All of the Possible Sequences When  $k = 4$

$\Phi_{4,i}$	$\eta_{4,i,1}$	$\eta_{4,i,2}$	$\eta_{4,i,3}$	$\eta_{4,i,4}$	$\eta_{4,i,5}$	Probability
$\Phi_{4,1}$	4	0				$f(4; qr) \times f(0; q4\tau)$
$\Phi_{4,2}$	3	1	0			$f(3; qr) \times f(1; q3\tau) \times f(0; q1\tau)$
$\Phi_{4,3}$	2	2	0			$f(2; qr) \times f(2; q2\tau) \times f(0; q2\tau)$
$\Phi_{4,4}$	2	1	1	0		$f(2; qr) \times f(1; q2\tau) \times f(1; q1\tau) \times f(0; q1\tau)$
$\Phi_{4,5}$	1	3	0			$f(1; qr) \times f(3; q1\tau) \times f(0; q3\tau)$
$\Phi_{4,6}$	1	2	1	0		$f(1; qr) \times f(2; q1\tau) \times f(1; q2\tau) \times f(0; q1\tau)$
$\Phi_{4,7}$	1	1	2	0		$f(1; qr) \times f(1; q1\tau) \times f(2; q1\tau) \times f(0; q2\tau)$
$\Phi_{4,8}$	1	1	1	1	0	$f(1; qr) \times f(1; q1\tau) \times f(1; q1\tau) \times f(1; q1\tau) \times f(0; q1\tau)$

**4.1.2. Modified Model.** Proposition 4 models the constrained queue length distribution based on the enumeration of all of the possible sequences of arrivals in various partitioned time intervals. As  $k$  increases, the number of possibilities and the computation time increase exponentially. By adopting a recursive computation structure, Proposition 5 omits the repeated computation steps and thus efficiently models the constrained queue length distribution.

**Proposition 5.** Given an arrival pattern following a Poisson distribution with an average arrival rate of  $q$ , a red period of  $r$ , and a saturation headway of  $\tau$ , the constrained queue length distribution is given as follows:

$$P(N = k) = \begin{cases} f(k; qr)f(0; qk\tau) + \sum_{i=1}^{k-1} \sum_{j=1}^{J_i} f(i; qr)\tilde{P}_j(N = k, M = i)W_j(N = k, M = i) & \text{if } k \in \mathbb{N}^+, \\ f(0; qr) & \text{if } k = 0 \end{cases} \quad (15)$$

where  $\tilde{P}_j(N = k, M = i)$  is the  $j$ th unique value of the product of the probabilities of observing the remaining  $k - i$  vehicles in the subsequent partitioned time intervals,  $\forall k \in \mathbb{N}^+$ ,  $i \in [1, k]$ ,  $j \in [1, J_i]$ ; and  $W_j(N = k, M = i)$  is the weight of  $\tilde{P}_j(N = k, M = i)$ .  $\tilde{P}_j(N = k, M = i)$  and  $W_j(N = k, M = i)$ ,  $\forall j \in [1, J_i]$ , can be obtained using Algorithm 1.

**Proof.** When  $k = 0$ ,  $P(N = 0) = f(0; qr)$ .

For  $k > 0$ , consider a generic sequence of arrivals in various partitioned time intervals,  $\{\eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi\}$ , s.t.

$$k = \eta_1 + \eta_2 + \eta_3 + \eta_4 + \dots + \eta_\varphi, \quad (16)$$

where  $\eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi$  represent the numbers of arrivals during the 1st, 2nd, 3rd, 4th, ...,  $\varphi$ th partitioned intervals, respectively,  $\forall \eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_{\varphi-1} \in \mathbb{N}^+$ , and  $\eta_\varphi = 0$ . The probability of observing this sequence is given by

$$P(N = k | \eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi) = f(\eta_1; qr)f(\eta_2; q\eta_1\tau)f(\eta_3; q\eta_2\tau) \dots f(\eta_\varphi; q\eta_{\varphi-1}\tau). \quad (17)$$

Rearranging the terms in Equation (14) affords

$$k - \eta_1 = \eta_2 + \eta_3 + \eta_4 + \dots + \eta_\varphi. \quad (18)$$

Similarly, the probability of observing such a sequence is given by

$$P(N = k - \eta_1 | \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi) = f(\eta_2; qr) f(\eta_3; q\eta_2\tau) f(\eta_4; q\eta_3\tau) \dots f(\eta_\varphi; q\eta_{\varphi-1}\tau). \quad (19)$$

By substituting Equation (17) into Equation (19),

$$\frac{P(N = k | \eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi)}{f(\eta_1; qr)} = f(\eta_2; q\eta_1\tau) \frac{P(N = k - \eta_1 | \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi)}{f(\eta_2; qr)}. \quad (20)$$

By defining

$$\tilde{P}(N = k, M = \eta_1) = \frac{P(N = k | \eta_1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi)}{f(\eta_1; qr)} \quad (21)$$

and

$$\tilde{P}(N = k - \eta_1, M = \eta_2) = \frac{P(N = k - \eta_1 | \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi)}{f(\eta_2; qr)}, \quad (22)$$

and substituting Equations (21) and (22) into Equation (20), a generic recursive formula (Equation (23)) is obtained for the product of the probabilities of observing the remaining vehicles in the subsequent partitioned time intervals for any  $k$  and  $\eta_1$ :

$$\tilde{P}(N = k, M = \eta_1) = f(\eta_2; q\eta_1\tau) \tilde{P}(N = k - \eta_1, M = \eta_2). \quad (23)$$

For  $\eta_1 = 1$ , the probability of  $N$  being  $k$  with the sequence  $\{1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi\}$  is given by

$$P(N = k | \eta_1 = 1, \eta_2, \eta_3, \eta_4, \dots, \eta_\varphi) = f(1; qr) \tilde{P}(N = k, M = 1), \quad (24)$$

where  $\tilde{P}(N = k, M = 1)$  can be obtained recursively using Equation (19) for various possible sequences formed by  $\eta_2, \eta_3, \eta_4, \dots, \eta_\varphi$ . Some of the possible sequences are identical in terms of their values of  $\tilde{P}(N = k, M = 1)$ . Let  $\tilde{P}_j(N = k, M = 1)$ ,  $W_j(N = k, M = 1)$ , and  $J_1$  be the  $j$ th unique value of the product of the probabilities of observing the remaining vehicles in the subsequent partitioned time intervals, where the weight of the  $j$ th unique value, and the number of unique values,  $\forall j \in [1, J_1]$ . Then,  $\tilde{P}_j(N = k, M = 1)$  and  $W_j(N = k, M = 1)$  can be obtained using Algorithm 1. It follows that by considering all of the possible sequences, the probability of  $N$  being  $k$  given that  $\eta_1 = 1$  is as follows:

$$P(N = k | \eta_1 = 1) = \sum_{j=1}^{J_1} f(1; qr) \tilde{P}_j(N = k, M = 1) W_j(N = k, M = 1). \quad (25)$$

Similarly, given  $\eta_1 = 2$ , the probability of  $N$  being  $k$  is as follows:

$$P(N = k | \eta_1 = 2) = \sum_{j=1}^{J_2} f(2; qr) \tilde{P}_j(N = k, M = 2) W_j(N = k, M = 2). \quad (26)$$

The probability of  $N$  being  $k$  for  $\eta_1 = 3, 4, \dots$ , or  $k$ , can be obtained in a similar manner. In particular, when  $\eta_1 = k$ ,

$$P(N = k | \eta_1 = k) = f(k; qr) f(0; qk\tau). \quad (27)$$

Summing all of the possibilities from  $\eta_1 = 1$  to  $\eta_1 = k$  reveals that the probability of  $N$  being  $k$  is given by

$$P(N = k) = f(k; qr) f(0; qk\tau) + \sum_{i=1}^{k-1} \sum_{j=1}^{J_i} f(i; qr) \tilde{P}_j(N = k, M = i) W_j(N = k, M = i). \quad (28)$$

Q.E.D.

Let  $\{\tilde{P}(N = k, M = i)\}$  and  $\{W(N = k, M = i)\}$  be the arrays storing  $\tilde{P}_j(N = k, M = i)$  and  $W_j(N = k, M = i)$ , respectively, where  $\forall k \in \mathbb{N}^+$ ,  $i \in [1, k]$ , and  $j \in [1, J_i]$ .  $\{\tilde{P}(N = k, M = i)\}$  and  $\{W(N = k, M = i)\}$  can then be obtained using Algorithm 1.

**Algorithm 1** (Computing  $\{\tilde{P}(N = k, M = i)\}$  and  $\{W(N = k, M = i)\}$ ,  $\forall k \in \mathbb{N}^+$ ,  $i \in [1, k]$ )

- 1: **Initialization:**  $\{\tilde{P}(N = 1, M = 1)\} \leftarrow \{f(0; q\tau)\}$ ,  
 $\{W(N = 1, M = 1)\} \leftarrow \{1\}$ .
- 2: **For**  $k$  in  $\mathbb{N}^+ \setminus \{1\}$  **do**
- 3:     **For**  $i$  in  $\{1, 2, 3, \dots, k\}$  **do**
- 4:         **If**  $i < k$  **then**
- 5:             **For**  $l$  in  $\{1, 2, 3, \dots, k - i\}$  **do**
- 6:                 **If**  $l = 1$  **then**
- 7:                      $\{\tilde{P}(N = k, M = i)\} \leftarrow f(l; qi\tau) \cdot$   
 $\{\tilde{P}(N = k - i, M = l)\}$
- 8:                      $\{W(N = k, M = i)\} \leftarrow \{W(N = k - i,$   
 $M = l)\}$
- 9:                 **Else**
- 10:                     **For each value** in  $f(l; qi\tau) \cdot \{\tilde{P}(N = k - i,$   
 $M = l)\}$  **do**
- 11:                         **If value**  $\in \{\tilde{P}(N = k, M = i)\}$  **then**
- 12:                             **Find**  $j$  such that  $\tilde{P}_j(N = k, M = i)$   
 $= \text{value}$
- 13:                              $W_j(N = k, M = i) \leftarrow W_j(N = k,$   
 $M = i) + 1$
- 14:                         **Else**
- 15:                              $\{\tilde{P}(N = k, M = i)\} \leftarrow \{\tilde{P}(N = k,$   
 $M = i)\} + \{\text{value}\}$
- 16:                              $\{W(N = k, M = i)\} \leftarrow \{W(N = k,$   
 $M = i)\} + \{1\}$
- 17:                         **End if**
- 18:                 **End for**
- 19:             **End if**
- 20:     **End for**

```

21:   Else
22:     { $\tilde{P}(N = k, M = i)$ }  $\leftarrow$  { $f(0; qi\tau)$ }
23:     { $W(N = k, M = i)$ }  $\leftarrow$  {1}
24:   End if
25: End for
26: End for
27: Output: { $\tilde{P}(N = k, M = i)$ } and { $W(N = k, M = i)$ },
         $\forall k \in \mathbb{N}^+, i \in [1, k]$ .

```

**4.1.3. Numerical Experiments Using Propositions 4 and 5.** To compare the computation efficiency of Propositions 4 and 5, a series of comprehensive simulation experiments based on the vertical queue assumption were conducted, as shown in Online Appendix K. Experiments with various combinations of red periods and volume-to-capacity (V/C) ratios were performed. For example, with a red period of 30 s and a V/C ratio of approximately 0.5, the cycle length, saturation flow, saturation headway, and traffic demand were 60 s, 2,268 veh/h, 1.59 s, and 567 veh/h, respectively. For various ranges of constrained queue length  $k$ , namely,  $k \in [0, 10]$ ,  $k \in [0, 13]$ , and  $k \in [0, 15]$ , the constrained queue length distributions obtained based on Propositions 4 and 5 were identical. However, the computation times of the three cases using Proposition 4 increased drastically, from 0.353 s to 50.682 s to 2,045.356, whereas the computation times of the three cases based on Proposition 5 increased only slightly, from 0.028 s to 0.064 s to 0.112 s, respectively. Similar results were obtained for other cases, demonstrating the superiority of Proposition 5 over Proposition 4 in terms of computation efficiency.

#### 4.2. CDT Model

Under the assumption of random arrivals, Proposition 4 can elegantly model the constrained queue length distribution based on the notion of time interval partitioning. As it adopts a recursive computation structure, Proposition 5 is more efficient than Proposition 4. Nevertheless, both models are rather complex. Therefore, the CDT model is introduced here as an approximation model that greatly simplifies the procedures of constrained queue length estimation.

Despite the random nature of the arrival pattern, there exists an average arrival rate,  $q$ . For the purpose of deriving the average constrained queue length, the random arrival can be approximated by its average arrival pattern, and there exists a constant dissipation time for that given average arrival rate. Given the red period  $r$ , average arrival rate  $q$ , and saturation flow  $s$ , the dissipation time  $t_0$  and constrained queue length  $N_0$  can be modeled by Equations (29) and (30), respectively, as follows:

$$q(r + t_0) = st_0, \quad (29)$$

$$N_0 = (r + t_0) \times q. \quad (30)$$

Equation (29) models the flow conservation at the stop bar, based on which  $t_0$  can be deduced. As shown in

Equation (30),  $N_0$  is directly proportional to  $r + t_0$ , during which the constrained vehicles are captured. Therefore, integrating Equations (29) and (30) affords  $N_0$ , as follows:

$$N_0 = \frac{sqr}{s - q}. \quad (31)$$

Here,  $N_0$  can be regarded as the average constrained queue length. Thus, Equation (31) can be used in conjunction with Corollary 1. Under the assumption that the constrained queue length follows a Poisson distribution,  $Pois(\lambda)$ , the parameter  $\lambda$  in Corollary 1 can be set to  $N_0$  to enable estimation of the variability of the estimated penetration rate.

## 5. VISSIM Simulation

The numerical experiments presented in previous sections were based on a vertical queue assumption. Thus, to more realistically mimic vehicle movement and queuing processes, comprehensive simulation studies were performed on a VISSIM platform in a Windows 10 environment on a machine equipped with an Intel Core i7-10700 CPU. Either the CDT model or the PDT model can be used for constrained queue length estimation. If the CDT model is used, then the estimated  $N_0$  is substituted into Corollary 1 (denoted “CDT model + Corollary 1”); if the PDT model is adopted, then Proposition 3 is subsequently applied for the estimation (denoted “PDT model + Proposition 3”). However, despite the elegance and accuracy of these models, the results show that their ignorance of the braking and start-up motions of vehicles prevents them from exhibiting their full potential. Thus, a CTL model is devised that incorporates the braking and start-up motions of vehicles into its estimation and is used to calibrate the CDT and PDT models. The simulation results demonstrate that the calibrated CDT and PDT models accurately model the constrained queue length distribution and hence guarantee the accuracy of variability estimation.

### 5.1. Uncertainty Estimation Directly Based on the PDT and CDT Models

“CDT model + Corollary 1” and “PDT model + Proposition 3” were directly applied to examine their performance. A single-lane road with a length of one kilometer (km) connecting to an isolated junction was considered. A fixed time-signal plan was set: 30 seconds red, 27 seconds green, and 3 seconds amber. Vehicles were generated according to a Poisson distribution with an average arrival rate of  $q = 700$  veh/h and an initial average speed of  $u = 50$  km/h. Default values were chosen for other settings, such as the car-following model, driver behavior, and vehicle characteristics. Based on several simulation runs, the saturation flow  $s$  and saturation headway  $\tau$  were determined to be 2,268 veh/h and 1.59 s, respectively. One thousand cycles of trajectory data were collected. The CV penetration rate  $p$  was chosen to be 0.4;

**Table 3.** Results of Simulation with Direct Application of CDT and PDT Models

	Ground truth	“CDT model + Corollary 1”	“PDT model + Proposition 3”
Variance	0.09897	0.06381	0.07531
APE (%)	—	35.53%	23.92%

thus, each of the vehicles was randomly assigned as a CV or a non-CV based on a probability of 0.4 and 0.6, respectively. The SSDPRE was applied to each of these cycles, which generated a total of 1,000 results that were used to determine the CV penetration rate distribution. The variance of the CV penetration rate distribution served as the ground truth.

Table 3 presents the simulation result. Despite the difficulties of obtaining second-moment estimations, both approaches accurately estimated the variances in the correct order, with absolute percentage errors (APE) of only approximately 20%–30%. Moreover, as expected, “PDT model + Proposition 3” performed better than “CDT model + Corollary 1” in terms of the computational expense of closely modeling the constrained queue length distribution. Nevertheless, the potential of this method remains to be fully realized.

### 5.2. Losses in Red Time and Dissipation Time

Given that the main focus is modeling the constrained queue length distribution, the derivations of PDT and CDT models in Section 4 are simplified by assuming that vehicles are able to instantly decelerate and accelerate to their desired speeds (see the ideal trajectories in Figure 3). However, in reality, it takes time for a driver to react and for a vehicle to accelerate or decelerate (see the actual trajectories in Figure 3). Such delays lead to a shorter effective red time and a shorter dissipation time. In this context, ignorance of the braking and start-up motions of vehicles results in slight overestimates of the constrained

queue length. For instance, in the simulation case presented in Section 5.1, the estimated average constrained queue lengths based on the PDT and CDT were approximately 8.438 vehicles, whereas the ground truth was 6.547 vehicles. This illustrates that ignorance of the losses in the red time and dissipation time was the major contributor to the remaining errors in the estimates.

### 5.3. Constant Time-Loss Model

To further improve estimation accuracy, the time losses due to the reaction times of drivers and the deceleration and acceleration times of vehicles must be incorporated into the models. As shown in Figure 3, the delay in the start of the effective red time can be represented by  $\Delta_1$ , whereas the delay in the start of the dissipation time can be represented by  $\Delta_2$ . Thus, the net loss of red time,  $\Delta$ , is given by

$$\Delta = \Delta_1 - \Delta_2. \quad (32)$$

Note that  $\Delta$  should only be dependent on the drivers’ reaction time and the vehicles’ start-up time and braking time. Thus, in general,  $\Delta$  only varies within a small range and can be taken as a constant.

For a generic constrained queue set,  $\Psi$ , with a constrained queue length  $N = k$ ,  $k$  can be decomposed into a sequence of arrivals in various partitioned time intervals,  $\{\eta_1, \eta_2, \eta_3, \dots, \eta_\phi\}$ , s.t.

$$k = \eta_1 + \eta_2 + \eta_3 + \dots + \eta_\phi. \quad (33)$$

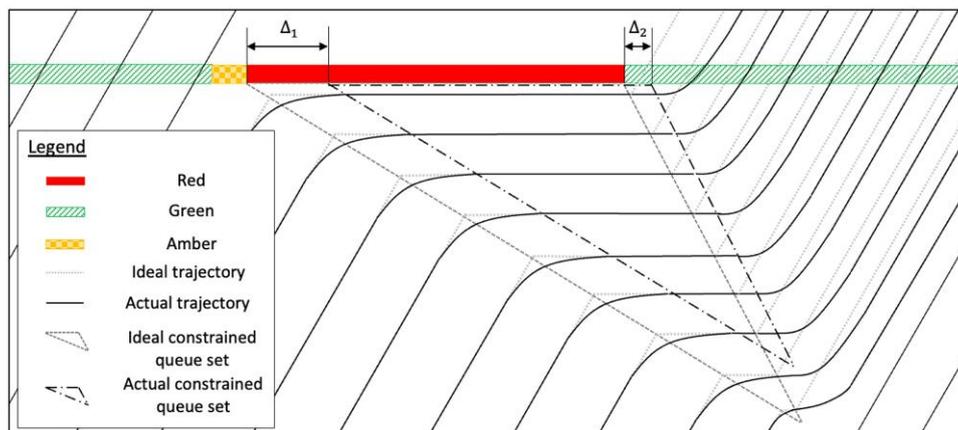
By considering random arrivals with an average arrival rate of  $q$ , based on flow conservation, on average,

$$\begin{aligned} q(r + \eta_1\tau + \eta_2\tau + \dots + \eta_\phi\tau) \\ = q(r + \tau k) = s(\eta_1\tau + \eta_2\tau + \dots + \eta_\phi\tau) = s\tau k. \end{aligned} \quad (34)$$

Let  $t_d$  be the dissipation time of the ideal constrained queue set, which is given by

$$t_d = \tau k = \frac{qr}{s - q}. \quad (35)$$

**Figure 3.** (Color online) Illustration of Ideal Trajectories and Actual Trajectories



Let  $r'$  be the effective red, and let  $t'_d$  be the dissipation time of the actual constrained queue set, where  $r'$  is given by

$$r' = r - \Delta. \quad (36)$$

By substituting  $r'$  and  $t'_d$  into Equation (33), the following equation for the dissipation time of the actual constrained queue set is obtained:

$$t'_d = \frac{q r'}{s - q} = \frac{q (r - \Delta)}{s - q}. \quad (37)$$

The dimensionless ratio  $\mu$  of the dissipation time of the actual constrained queue set to the dissipation time of the ideal constrained queue set or the effective red to the red is obtained by dividing Equation (37) by Equation (35), as follows:

$$\mu = \frac{t'_d}{t_d} = \frac{r'}{r} = \frac{r - \Delta}{r} = 1 - \frac{\Delta}{r}. \quad (38)$$

Thus,  $t'_d$  can be alternatively expressed as  $\mu t_d$  or  $\mu \tau k$ . Based on Equation (36), detailed calibration procedures for  $\Delta$  of the PDT and CDT models, based on the CTL model, are provided in Online Appendices L and M, respectively. Once  $\Delta$  is calibrated, the effective red time  $r'$  can be obtained and incorporated into the PDT and CDT models, thereby enabling a more accurate estimation of constrained queue lengths.

#### 5.4. Uncertainty Estimation Based on the Calibrated PDT and CDT Models

Nine simulations with various signal plans and traffic demands were conducted, with all of the other settings identical to those presented in Section 5.1. In all of the simulations, the cycle time was set to 60 s and the amber period was set to 3 s; however, various  $r$  (i.e., 15, 30, or 45 s) and V/C ratio (i.e., 0.3, 0.5, or 0.7) combinations were considered. The resulting nine sets of data were used to calibrate the PDT and CDT models using the methods introduced in Online Appendices L and M. The

$\Delta$  were determined to be 5.048 s and 7.270 s for the calibrated PDT and CDT models, respectively.

Based on the calibrated models, the constrained queue length distributions and the variances of the CV penetration rates of the nine cases were estimated. Table 4 reports the  $p$ -values of the Kolmogorov–Smirnov (KS) tests, the root-mean-square errors (RMSEs), and the relative root-mean-square errors (RRMSEs) of the estimated constrained queue length distributions, relative to the observed distributions obtained from the VISSIM simulations. The null hypothesis of the KS tests is that the estimated distribution is consistent with the observed distribution. As the  $p$ -value of each case was greater than the level of significance (0.05), there was insufficient statistical evidence to reject the null hypothesis, indicating that the calibrated PDT and CDT models accurately replicated the observed constrained queue length distributions. Moreover, the mean and variance of the RMSEs and RRMSEs demonstrate that the calibrated PDT model generally performed better than the calibrated CDT model. Figure 4 illustrates the goodness-of-fit of the estimated distributions for the calibrated PDT and CDT models, with reference to the observed distributions.

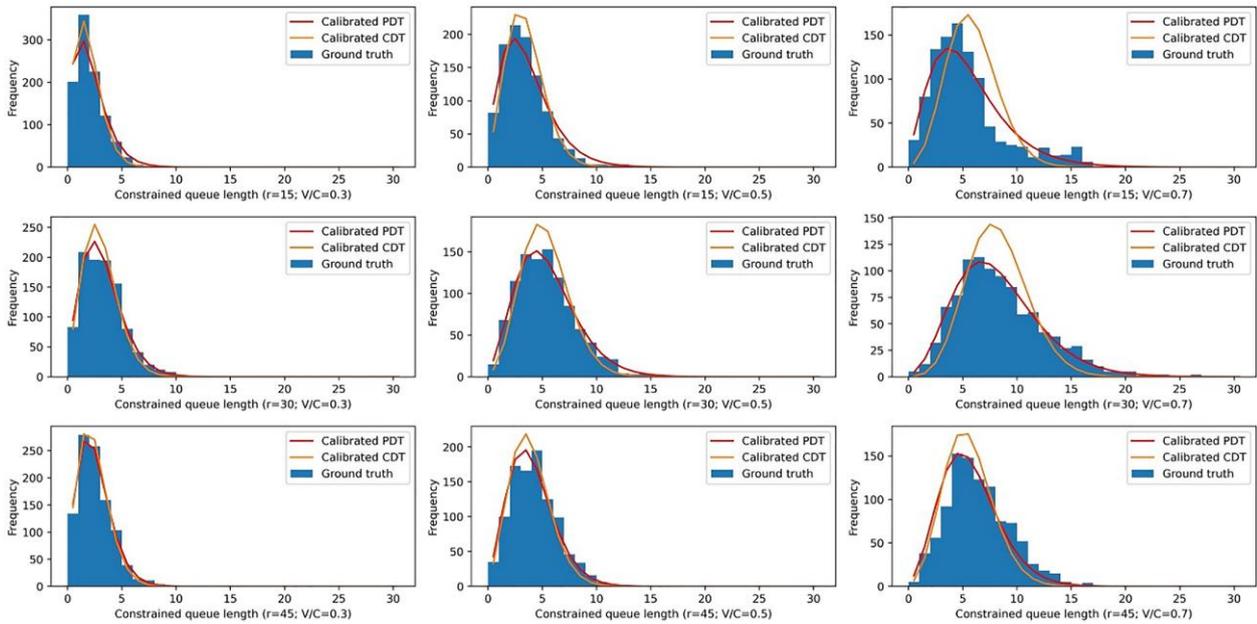
Table 5 presents the estimated variances of the nine cases based on the calibrated PDT and CDT models. Overall, the estimated variances of both models were extremely close to the ground truths, with significantly reduced APEs. All of the APEs of the estimates based on “Calibrated CDT model + Corollary 1” were much less than 15%, while the APEs of the estimates based on “Calibrated PDT model + Proposition 3” were much less than 10%. In general, the calibrated PDT model more closely modeled the constrained queue length distribution than the calibrated CDT model.

Subsequently, the simulation case presented in Section 5.1 was reperformed using the PDT and CDT models calibrated based on the above nine sets of data. Table 6 presents a comparison of the results of the uncalibrated models and the calibrated models. The results of the uncalibrated models are extracted from Table 3. By using

**Table 4.**  $p$ -Values of KS Tests, RMSEs, and RRMSEs of the Estimated Distributions

$r$	V/C	Calibrated CDT model			Calibrated PDT model		
		$p$ -value	RMSE	RRMSE (%)	$p$ -value	RMSE	RRMSE (%)
15	0.3	0.833	16.312	17.94	0.997	23.287	25.62
15	0.5	0.365	12.997	27.29	0.603	9.824	20.63
15	0.7	0.503	36.154	65.08	0.781	15.565	28.02
30	0.3	0.997	21.013	23.11	1.000	12.982	14.28
30	0.5	0.422	13.764	31.66	0.422	4.768	10.97
30	0.7	0.351	17.997	50.39	0.944	4.676	13.09
45	0.3	0.994	10.034	10.03	1.000	8.792	8.79
45	0.5	0.751	16.752	28.48	1.000	12.731	21.64
45	0.7	0.172	17.346	45.10	0.172	13.705	35.63
Mean		—	18.041	33.23	—	11.814	19.85
Variance		—	49.988	0.02649	—	29.464	0.00701

**Figure 4.** (Color online) Illustrations of the Goodness-of-Fit of the Estimated Distributions Based on the Calibrated PDT and CDT Models



the calibrated models, the APEs of the estimates based on the CDT model and PDT model were significantly reduced (from 35.53% and 23.92%–3.06% and 2.48%, respectively). Similarly, the calibrated PDT model outperformed the calibrated CDT model.

## 6. Applications

This section first validates the proposed models on the real-world NGSIM data set. A simple illustrative application of CV-based adaptive signal control is then presented to demonstrate the potential improvement in system optimization by further incorporation of CV penetration rate variability.

### 6.1. Real-World Validation

To demonstrate the readiness of applications in real-world scenarios, the calibrated PDT and CDT models

defined in Section 5.4 were applied to the real-world NGSIM data set to obtain variance estimations for CV penetration rates. Two periods of 15-minute trajectory data (i.e., 12:45–13:00 and 16:00–16:15) for November 8, 2006, were extracted from the arterial road data for Peachtree Street in Atlanta, Georgia (USA). The southbound through-lane between intersections 1 and 2 was chosen for the validation. According to the signal plans, the cycle times were 95 s and 100 s, and the red periods were 62 s and 64 s, respectively, for the two 15-minute periods. After observing several cycles, the saturation flow  $s$  was determined to be 1,761 veh/h, and the saturation headway  $\tau$  was found to be 2.044 s. Counting the number of vehicles traveling through the southbound through-lane revealed that the traffic demands of the two periods were 7.0 veh/cycle and 8.8 veh/cycle, respectively. In reality, when only CV data are available, the

**Table 5.** Variance Estimations Based on Calibrated CDT and PDT Models

$r$	V/C	Ground truth Variance	"Calibrated CDT model + Corollary 1"		"Calibrated PDT model + Proposition 3"	
			Variance	APE (%)	Variance	APE (%)
15	0.3	0.17408	0.17034	2.15	0.16213	6.86
15	0.5	0.16807	0.17649	5.01	0.16000	4.80
15	0.7	0.13811	0.11758	14.86	0.12922	6.44
30	0.3	0.17076	0.18209	6.64	0.17246	0.10
30	0.5	0.12770	0.13244	3.71	0.13013	1.90
30	0.7	0.08266	0.07350	11.08	0.08345	0.96
45	0.3	0.18275	0.18260	0.08	0.17757	2.83
45	0.5	0.15336	0.16671	8.71	0.16055	4.69
45	0.7	0.11308	0.12575	11.20	0.12411	9.75
Mean	—	—	—	7.05	—	4.36
Variance	—	—	—	0.00206	—	0.00079

**Table 6.** Comparison of the Results of the Uncalibrated Models and the Calibrated Models

	Ground truth	“CDT model + Corollary 1”		“PDT model + Proposition 3”	
		Uncalibrated	Calibrated	Uncalibrated	Calibrated
Variance	0.09897	0.06381	0.09594	0.07531	0.10142
APE (%)	—	35.53%	3.06%	23.92%	2.48%

traffic demand can be accurately estimated by dividing the CV flow by the result generated by the SSDPRE (Wong et al. 2019).

The CV penetration rate  $p$  was set to 0.1, 0.4, or 0.7. In each case, each vehicle was randomly assigned to be a CV or a non-CV, as per the predefined CV penetration rate. Thus, by applying the SSDPRE to each constrained queue, a CV penetration rate distribution and its variance were obtained. Because of the small sample size (only nine complete cycles in a 15-minute period), the variance estimates fluctuated due to the random seeds during CV assignments. To minimize this sampling error, the nine constrained queues were replicated 10,000 times. Accordingly, by repeating the above steps, 10,000 CV penetration rate distributions and variances were obtained. The mean of the 10,000 variances served as the ground truth.

The PDT and CDT models calibrated in Section 5.4, with  $\Delta$  being 5.048 s and 7.270 s, respectively, were applied for the variance estimations. Table 7 summarizes the results. Both the calibrated PDT and CDT models accurately estimated the variances in all six real-world scenarios, with APEs much less than 20%. “Calibrated PDT model + Proposition 3” afforded average APEs of 10.91% and 5.22% for the two 15-minute periods, respectively. “Calibrated CDT model + Corollary 1” resulted in comparable performance, with average APEs of 11.32% and 6.57% for the two 15-minute periods, respectively. Although the simplified CDT model led to slightly larger errors, the average computation time was only 0.001 s, which was much shorter than that of the “Calibrated PDT model + Proposition 3” approach (2.355 s). Thus, in certain real-time applications requiring instant responses, the “Calibrated CDT model + Corollary 1” approach

could be advantageous. These results demonstrate the robustness of the models developed in this study and illustrate their readiness for real-world applications.

### 6.2. Illustrative Application of CV-Based Adaptive Signal Control

To demonstrate the importance of the potential improvement in system optimization by further incorporating the CV uncertainty in system optimization, a simple CV-based adaptive signal control application is constructed using VISSIM in this subsection. Two adaptive signal control schemes are compared. The first scheme (Scheme A—without uncertainty) optimizes the signal plan by minimizing the total delay according to the expected traffic demand estimated using the CV penetration rate. The second scheme (Scheme B—with uncertainty) searches for the most robust signal plan by further considering the traffic demand variability estimated using the CV penetration rate variability.

Consider a crossroad with two approaches to an isolated intersection. Traffic demands for the two approaches were generated using Poisson distributions, one with an average arrival rate of 800 veh/hour and the other with an average arrival rate of 400 veh/hour. Setting the ground-truth CV penetration rate as 0.4, each of the generated vehicles had a 40% or 60% probability of being assigned as a CV or non-CV, respectively. A simple red-green-amber signal structure was adopted for each approach. The cycle length, amber time, and clearance time were fixed at 60, 3, and 5 seconds, respectively. The signal plan was optimized at the end of each cycle according to the estimated traffic demand.

**Table 7.** Real-World Application of Variance Estimations Based on Calibrated CDT and PDT Models

Period	$p$	Ground truth Variance	“Calibrated CDT model + Corollary 1”		“Calibrated PDT model + Proposition 3”	
			Variance	APE (%)	Variance	APE (%)
12:45–13:00	0.1	0.07318	0.07463	1.98	0.07336	0.25
	0.4	0.15138	0.13321	12.00	0.13161	13.06
	0.7	0.11441	0.09156	19.97	0.09217	19.44
Mean	—	—	—	11.32	—	10.91
16:00–16:15	0.1	0.06249	0.06852	9.65	0.06774	8.40
	0.4	0.10184	0.10248	0.63	0.10435	2.46
	0.7	0.07313	0.06624	9.42	0.06962	4.80
Mean	—	—	—	6.57	—	5.22

To estimate the traffic demand, the total number of vehicle arrivals in cycle  $i$  on approach  $j$ ,  $M_{i,j}$ , is given by

$$M_{i,j} = m_{i,j} + q_{i,j}(1 - p_{i,j})C, \quad \forall j \in \{1, 2\}, \quad (39)$$

where  $m_{i,j}$ ,  $q_{i,j}$ ,  $p_{i,j}$ , and  $C$  represent the number of CV arrivals in cycle  $i$  on approach  $j$ , the real-time average arrival rate in cycle  $i$  on approach  $j$ , the real-time CV penetration rate in cycle  $i$  on approach  $j$ , and the cycle length, respectively. Thus,  $M_{i,j}$  and its variability depend on  $p_{i,j}$  and its variability. Besides  $m_{i,j}$ ,  $n_{i,j}$ , and  $\tilde{N}_{i,j}$  are also observable in the CV environment. A likelihood function, as shown below, can be established by adopting the CDT model and Corollary 2. The likelihood function can be maximized by estimating the parameters—the average arrival rate,  $q_{i,j}$ , and the CV penetration rate,  $p_{i,j}$ —such that the observables  $n_{i,j}$  and  $\tilde{N}_{i,j}$  are the most probable:

$$\max_{q_{i,j}, p_{i,j}} \prod_{k=0}^T P(n_{i-k,j}, \tilde{N}_{i-k,j}), \quad (40)$$

where  $T = 0, 1, 2, \dots, i - 1$  is the number of past cycles considered in the likelihood function formulation ( $T$  was set to 2 in this illustration). The maximum likelihood estimators,  $q_{i,j}^*$ ,  $p_{i,j}^*$ , can be taken as the real-time average arrival rate in cycle  $i$  on approach  $j$ ,  $q_{i,j}$ , and the real-time CV penetration rate in cycle  $i$  on approach  $j$ ,  $E(p_{i,j})$ , which can then be used as inputs of the CDT model and Corollary 1 to estimate the real-time CV penetration rate variance,  $Var(p_{i,j})$ .

In Scheme A, the variability of the CV penetration rate was not considered. Assuming that traffic demands in cycles  $i$  and  $i + 1$  were identical,  $M_{i+1,j}$  was estimated by directly substituting  $q_{i+1,j}$  and  $p_{i+1,j}$  in Equation (39) by  $q_{i,j}^*$  and  $p_{i,j}^*$ , respectively. The real-time delays in cycle  $i + 1$  for the two approaches,  $D_{i+1,1}$  and  $D_{i+1,2}$ , can be readily estimated using the method presented in Online Appendix N. The optimal signal plan can be obtained by solving the following optimization problem with the objective of minimizing the total delay using a simple line search method:

$$\begin{aligned} \min_{g_{i+1,1}, g_{i+1,2}} \{ & D_{i+1,1} + D_{i+1,2} \} \\ \text{s.t. } & g_{i+1,1} + g_{i+1,2} = 52 \\ & g_{i+1,1} \geq 5 \\ & g_{i+1,2} \geq 5. \end{aligned} \quad (41)$$

After the first 30 warm-up cycles with a fixed signal plan, the signal plan was optimized 1,000 times at the end of

each cycle as per the above signal control scheme. The results of the actual delays are given in Table 8.

In Scheme B, the uncertainty in the CV penetration rates, and thus the uncertainty in the traffic demands were considered. As  $p_{i,j}$ , by definition, is confined between 0 and 1, it was assumed to follow a beta distribution. Through Monte Carlo sampling, 1,000 sets of the possible CV penetration rates for the two approaches were sampled from the assumed beta distributions. Using Equation (39), 1,000 sets of possible traffic demands were estimated from the sampled CV penetration rates. For each set of traffic demands, the delays predicted for the two approaches were evaluated using Equations (N1) and (N2). Given a signal plan, the average total delay over the 1,000 sets of traffic demands,  $E(D_{i+1,1} + D_{i+1,2})$ , can thus be estimated. A robust signal plan for cycle  $(i + 1)$  can be formulated in Equation (42) below to minimize the average total delay, which can also be solved by a simple line search method on  $g_{i+1,1}$  (with  $g_{i+1,2} = 52 - g_{i+1,1}$ ),

$$\begin{aligned} \min_{g_{i+1,1}, g_{i+1,2}} & E(D_{i+1,1} + D_{i+1,2}) \\ \text{s.t. } & g_{i+1,1} + g_{i+1,2} = 52 \\ & g_{i+1,1} \geq 5 \\ & g_{i+1,2} \geq 5. \end{aligned} \quad (42)$$

Similarly, after the first 30 warm-up cycles, the robust signal plan was optimized 1,000 times at the end of each cycle as per the described control scheme. The results are given in Table 8.

The results show that the incorporation of the CV penetration rate uncertainty reduced the average actual delay and maximum actual delay by approximately 15% and reduced the variance in the actual delay by approximately 45.5%. Thus, this simple illustrative application of CV-based adaptive signal control clearly demonstrates the potential improvement in system optimization via the incorporation of the CV penetration rate uncertainty.

## 7. Conclusion

This study proposed the PPR model (i.e., Proposition 3) for estimating the variability of the SSDPRE. Constrained queue length distribution is the essential input for the PPR model. Thus, the PDT model and CDT model were derived. The PDT model closely models the constrained queue length distribution under the assumption of random arrival, whereas the CDT model is a simplified model based on the assumption of constant dissipation time. However, due to the PDT and CDT models'

**Table 8.** Comparison of Results Obtained Using Scheme A and Scheme B as Control Schemes

Metric	Scheme A—without uncertainty	Scheme B—with uncertainty	Improvement (%)
Average actual delay (s)	27.2	23.1	15.1
Maximum actual delay (s)	202.5	171.5	15.3
Variance in actual delay (s <sup>2</sup> )	843.0	459.6	45.5

ignorance of the braking and start-up motions of vehicles, their potential cannot be fully realized. The CTL models and calibration procedures for the PDT and CDT models were therefore established. Simulation studies showed that the calibrated PDT and CDT models accurately modeled the constrained queue length distribution and estimated the variances of CV penetration rates. Applications of these models to NGSIM data demonstrated their robustness and readiness for real-world applications. Although the calibrated PDT model usually had better estimation accuracy, the calibrated CDT model had the shortest computation time. A simple illustrative application of CV-based adaptive signal control based on the proposed models clearly demonstrated the potential improvement in system optimization via the incorporation of the CV penetration rate uncertainty. Future work will extend the framework to manage cases of near-capacity conditions.

## References

- Ambühl L, Menendez M (2016) Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Res. Part C Emerging Tech.* 71:184–197.
- Argote J, Christofa E, Xuan Y, Skabardonis A (2011) Estimation of measures of effectiveness based on connected vehicle data. *Proc. 14th Internat. IEEE Conf. Intelligent Transportation Systems (IEEE, Piscataway, NJ)*, 1767–1772.
- 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. *Transportation Res. Part C Emerging Tech.* 129:103241.
- Comert G (2013) Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals. *Transportation Res. Part B Methodological* 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 probe vehicle data. *IEEE Trans. Intelligent Transportation Systems* 12(2):563–573.
- Du J, Rakha H, Gayah VV (2016) Deriving macroscopic fundamental diagrams from probe data: Issues and proposed solutions. *Transportation Res. Part C Emerging Tech.* 66:136–149.
- Federal Highway Administration (2006) Next generation simulation: Peachtree Street data set. Retrieved June 25, 2022, <https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Program-Peachtree/mupt-aksf>.
- Feng Y, Head KL, Khoshmashgham S, Zamanipour M (2015) A real-time adaptive signal control in a connected vehicle environment. *Transportation Res. Part C Emerging Tech.* 55:460–473.
- Geroliminis N, Daganzo CF (2008) Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Res. Part B Methodological* 42(9):759–770.
- Hao P, Ban XJ, Guo D, Ji Q (2014) Cycle-by-cycle intersection queue length distribution Estimation using sample travel times. *Transportation Res. Part B Methodological* 68:185–204.
- Iqbal MS, 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. Intelligent Transportation Systems* 20(1):87–96.
- Jenelius E, Koutsopoulos HN (2013) Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Res. Part B Methodological* 53:64–81.
- Jenelius E, Koutsopoulos HN (2015) Probe vehicle data sampled by time or space: Consistent travel time allocation and estimation. *Transportation Res. Part B Methodological* 71:120–137.
- Khan SM, Dey KC, Chowdhury M (2017) Real-time traffic state estimation with connected vehicles. *IEEE Trans. Intelligent Transportation Systems* 18(7):1687–1699.
- Lu Y, Xu X, Ding C, Lu G (2019) A speed control method at successive signalized intersections under connected vehicles environment. *IEEE Intelligent Transportation Systems Magazine* 11(3):117–128.
- Meng F, Wong SC, Wong W, Li YC (2017a) Estimation of scaling factors for traffic counts based on stationary and mobile sources of data. *Internat. J. Intelligent Transportation Systems Res.* 15(3):180–191.
- Meng F, Wong W, Wong SC, Pei X, Li YC, Huang H (2017b) Gas dynamic analogous exposure approach to interaction intensity in multiple-vehicle crash: Case study of crashes involving taxis. *Anal. Methods Accident Res.* 16:90–103.
- Mousa SR, Ishak S (2017) An extreme gradient boosting algorithm for freeway short-term travel time prediction using basic safety messages of connected vehicles. *Transportation Res. Board 96th Annual Meeting (Transportation Research Board, Washington, DC)*.
- Rahmani M, Jenelius E, Koutsopoulos HN (2015) Non-parametric estimation of route travel time distributions from low-frequency floating car data. *Transportation Res. Part C Emerging Tech.* 58:343–362.
- Sen S, Head KL (1997) Controlled optimization of phases at an intersection. *Transportation Sci.* 31(1):5–17.
- 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. Problems Engrg.* 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. Transportation Land Use* 13(1):517–538.
- Wong W, Wong SC (2015) Systematic bias in transport model calibration arising from the variability of linear data projection. *Transportation Res. Part B Methodological* 75:1–18.
- Wong W, Wong SC (2016a) Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection. *Transportation Res. Part B Methodological* 88:72–92.
- Wong W, Wong SC (2016b) Evaluation of the impact of traffic incidents using GPS data. *Transport* 169(3):148–162.
- Wong W, Wong SC (2016c) Network topological effects on the macroscopic Bureau of Public Roads function. *Transportmetrica A Transportation Sci.* 12(3):272–296.
- Wong W, Wong SC (2019) Unbiased estimation methods of nonlinear transport models based on linearly projected data. *Transportation Sci.* 53(3):665–682.
- Wong W, Wong SC, Liu X (2019) Bootstrap standard error estimations of nonlinear transport models based on linearly projected data. *Transportmetrica A Transportation Sci.* 15(2):602–630.
- Wong W, Wong SC, Liu X (2021) Network topological effects on the macroscopic fundamental diagram. *Transportmetrica B Transport Dynamics* 9(1):376–398.
- Wong W, Shen S, Zhao Y, Liu X (2019) On the estimation of connected vehicle penetration rate based on single-source connected vehicle data. *Transportation Res. Part B Methodological* 126:169–191.
- Yang X, Lu Y, Hao W (2017) Origin-destination estimation using probe vehicle trajectory and link counts. *J. Advanced Transportation* 2017:4341532.
- Yin Y (2008) Robust optimal traffic signal timing. *Transportation Res. Part B Methodological* 42(10):911–924.

Zhao Y, Wong W, Zheng J, Liu HX (2022) Maximum likelihood estimation of probe vehicle penetration rates and queue length distributions from probe vehicle data. *IEEE Trans. Intelligent Transportation Systems* 23(7):7628–7636.

Zhao Y, Zheng J, Wong W, Wang X, Meng Y, Liu HX (2019a) Estimation of queue lengths, probe vehicle penetration rates, and

traffic volumes at signalized intersections using probe vehicle trajectories. *Transportation Res. Record* 2673(11):660–670.

Zhao Y, Zheng J, Wong W, Wang X, Meng Y, Liu HX (2019b) Various methods for queue length and traffic volume estimation using probe vehicle trajectories. *Transportation Res. Part C Emerging Tech.* 107:70–91.