systems

MDPI

Article

Robust Traffic Signal Retiming Based on Queue Service Time Estimation Using Low-Penetration Connected Vehicle Data

Chengchuan An *, Weihua Zhang, Yinpu Wang, Siping Ke and Jingxin Xia

Intelligent Transportation System Research Center, Southeast University, Nanjing 210096, China; 230219399@seu.edu.cn (W.Z.); 230198278@seu.edu.cn (Y.W.); kesiping@seu.edu.cn (S.K.); xiajingxin@seu.edu.cn (J.X.)

* Correspondence: ccan@seu.edu.cn

Abstract: Signal retiming is the most cost-efficient measure to reduce vehicle delay and alleviate congestion on urban roads. Previous studies have explored the potential of using connected vehicle data for signal retiming specifically under the current low-penetration environment, which will significantly reduce the cost and increase the productivity of signal retiming. However, the existing methods are mostly deterministic and do not well consider the uncertainty in both traffic demand and capacity. This compromises their robustness in a real application. In this study, a novel traffic state measure—queue service time (QST)—is introduced and used as the only input to generate a robust signal plan at isolated intersections for a time-of-day period. First, a Bayesian-based model is proposed to estimate the QST distribution by collectively using the lower and upper boundary observations reported by connected vehicles. Then, a goal programming-based signal optimization model is formulated using quantiles of QST as input, which accounts for the combined uncertainty in both traffic demand and capacity. Simulation experiments validate the effectiveness and robustness of the proposed method. It is shown that the proposed QST estimation model is reliable to use under a penetration rate as low as 0.05 and can effectively estimate the actual distribution in both under- and oversaturation conditions. Compared with a demand-based method that only accounts for uncertainty in traffic demand, the proposed QST-based signal timing optimization method shows its superiority in reducing the occurrence of oversaturation or phase failure, as well as enhancing performance against the worst cases.

Keywords: connected vehicle; vehicle trajectory data; queue service time; low penetration rate; robust signal retiming

Academic Editors: Renata Żochowska, Grzegorz Karoń and Marcin Kłos

Received: 5 December 2024
Revised: 18 December 2024
Accepted: 28 December 2024
Published: 30 December 2024

Citation: An, C.; Zhang, W.; Wang, Y.; Ke, S.; Xia, J. Robust Traffic Signal Retiming Based on Queue Service Time Estimation Using Low-Penetration Connected Vehicle Data. *Systems* **2025**, *13*, 15. https://doi.org/10.3390/systems13010015

1. Introduction

Signal retiming is the most cost-efficient measure to reduce vehicle delay and alleviate congestion on urban roads. However, due to the difficulty and high cost of detector installation to collect traffic data, signal timing plans cannot be updated timely to accommodate changes in traffic patterns. A previous study indicated that an outdated signal timing plan would lead to a 3–5% increase in delay per year [1]. Recently, with the prevalence of ride-hailing services and the boosting development of onboard mobile sensors, more vehicles become connected and can consecutively report their location and moving status in a large spatial coverage, which provides new opportunities to develop detector-free and low-cost signal retiming applications.

In the past decade, various approaches have been proposed in the line of detector-free signal control using connected vehicle data. In general, there are two critical procedures

involved. One is the traffic state estimation, and the other is the signal timing optimization. Assuming a relatively high penetration rate of connected vehicles (i.e., 20% or higher), it is possible to achieve real-time adaptive signal control. He et al. (2012) [2] estimated real-time traffic state by recognizing vehicle platoons with different sizes and speeds, and a platoon-based predictive model is formulated in a mixed integer linear program framework to determine future signal status. Feng et al. (2015) [3] divided an urban segment into queuing, slow-down, and free-flow regions to estimate their traffic states, and the adaptive signal control problem was solved by a dynamic programming algorithm. More recently, researchers have extended adaptive signal control to cooperative signal and vehicle control leveraging connected and autonomous vehicles [4,5]. However, the accessibility of adequate and real-time connected vehicle data is crucial to the success of these studies, which will not be fully satisfied in the near future.

In the current stage, only a small fleet of vehicles are connected and can report their trajectories, and most of them are vehicles such as taxi cabs and buses. With limited data, the most practical expectation is to optimize the signal timing for a time-of-day period by aggregating all historical data. Thus, many studies proposed models to estimate a pattern-based traffic state to support signal timing optimization. Traffic volume and queue length are the two most important measures that have gained great attention in the literature. As for traffic volume estimation, Zheng and Liu (2017) [6] described the traffic arrival process over the signal cycle as a time-varying Poisson process and estimated mean arrival rates using historical vehicle trajectory data in a period. Lloret-Batlle et al. (2023) [7] further extended Zheng's method by considering the correlation of vehicle trajectories between two consecutive signal cycles under the oversaturated condition. Tan et al. (2022) [8] utilized the number of probe vehicles as the prior information on the magnitude of mean arrival rates among different movements to jointly estimate multi-phase traffic demand at the intersection. As for queue length estimation, Zhao et al. (2019) [9] proposed various probability models to simultaneously estimate the total queue length and volume for a historical period. Mei et al. (2019) [10] estimated the distribution of the maximum queue length over several neighboring cycles using the lower and upper boundaries observed by the stopped and non-stop vehicle trajectories. Zhao et al. (2022) [11] proposed an expectation-maximization algorithm regarding the actual queue length as the hidden variable and estimated probe vehicle penetration rates and queue length distribution. Most of these studies assumed the queue length can be cleared within the current signal cycle, which may be problematic when overflow occurs.

In contrast to the traffic state estimation, fewer studies utilized the estimated results to develop a signal retiming application with low-penetration connected vehicle data. Feng et al. (2018) [12] predicted vehicle delays using real-time vehicle trajectory data and historical estimated arrival rates and developed a real-time adaptive signal control algorithm for isolated intersections. Tan et al. (2024) [13] proposed a cumulative flow diagram (CFD) model to optimize signal timings for fixed-time signal control. With the estimated mean arrival rate and queue discharge wave speed in a historical period, the CFD model can predict the queue formation and discharge processes given different signal timing schemes. A bi-level signal timing optimization problem was formulated with the primary objective of clearing the queue and the secondary objective of delay minimization. Instead of using arrival rate as input, Ma et al. (2020) [14] directly analyzed the changes in vehicle trajectories under different settings of cycle length and green duration using a same-ratio principle under a uniform arrival assumption. These studies have shown promising results for signal retiming with low-penetration connected vehicle data. However, the signal timing optimization is all based on mean traffic demand (i.e., arrival flow rate) and

capacity estimation (e.g., saturation flow rate). This may compromise their effectiveness in a real-world application due to the stochastic nature of traffic flow.

In terms of time-of-day signal retiming, the robustness of signal plans is of particular importance in a real application to ensure a stable performance with fluctuating traffic flow. A pioneering study by Yin (2008) [15] systemically introduced robust signal timing optimization methodology at isolated intersections. This work was further extended to urban arterial [16] and network [17] in the following research. The goal of robust signal optimization is to enhance the performance in the worst cases. A similar technology is stochastic signal timing optimization, which attempts to achieve the optimal solution in the sense of probability expectation [18,19]. However, these studies mostly consider the uncertainty of traffic demand in a control period. On the capacity side, the saturation flow rate is typically regarded as a static variable. The Highway Capacity Manual (HCM) has indicated that many factors affect the saturation flow rate, such as geometry, traffic composition, and weather. It was found in field experiments that the saturation flow rate is mostly site-dependent and likely to be time-dependent [20]. It is still an open question to jointly consider the uncertainty of both demand and capacity to enhance the robustness of the signal timing optimization.

This study aims to develop a robust signal retiming approach for fix-time signal control in time-of-day periods at isolated intersections using historical trajectories of connected vehicles. An alternative traffic state measure, namely, queue service time (QST), is introduced. The QST is defined as the green time needed to discharge the standing queue for a signal cycle. Different from the traffic volume and queue length, which only represent the traffic demand, the QST can be regarded as a combined measure implying both the traffic demand and capacity and directly related to the desired signal timings. To estimate QST and characterize its uncertainty during a control period, a probability-based model is first proposed using the number of stops and departure times of the connected vehicles as inputs. Then, the estimated distribution of QST is used to generate a robust signal timing plan for each control period. A goal programming framework is designed to determine the optimal signal timing given the desired green ratio of each movement. The robustness of the signal timing plan is enhanced in two aspects. First, the optimal signal timing plan can be derived only based on the QST, rather than depending on both traffic demand and saturation flow rate. Second, the traffic fluctuations are considered by leveraging the quantiles of the QST distribution. The proposed method is validated in simulation experiments. The results show the effectiveness and robustness of the proposed method compared with a demand-based method only accounting for the uncertainty in traffic demand. The contributions of this study can be summarized as follows:

- A novel traffic state measure—queue service time (QST)—is introduced, and its distribution is estimated by a Bayesian-based model collectively using the low-penetration connected vehicle data in historical periods.
- A goal programming-based signal timing optimization model is formulated using quantiles of QST distribution as the only input, which accounts for the combined uncertainty in both traffic demand and capacity.
- Compared with a demand-based method, the effectiveness and robustness of the proposed method are validated under a penetration rate as low as 0.05 through comprehensive simulation experiments.

The rest of the paper is organized as follows: Section 2 describes the QST estimation method, including observation extraction, probability-based model, and estimation approach. Section 3 formulates the signal timing optimization model based on QST estimation. Section 4 conducts simulation experiments and discusses the comparative results. Section 5 ends with the conclusion and future work.

2. Queue Service Time Estimation

2.1. QST Observation via Connected Vehicles

In this study, we assume that the vehicle stops are all caused by signal control. The QST is formally defined in both under- and oversaturation conditions, as shown in Figure 1. In the undersaturation condition, vehicles experience no more than one stop and the queue can be dissipated within each signal cycle. The green time needed to clear the last stopped vehicle is regarded as the QST. In the oversaturation condition, vehicles experience more than one stop, and a residual queue is presented which will be dissipated in the next signal cycles. The QST is the summation of the green time of the current signal cycle and the green time needed to clear the residual queue in the following signal cycles. In the case of Figure 1b where the oversaturation lasts two signal cycles, the time to clear the residual queue is indicated by the last twice stopped vehicle in the subsequent signal cycle.

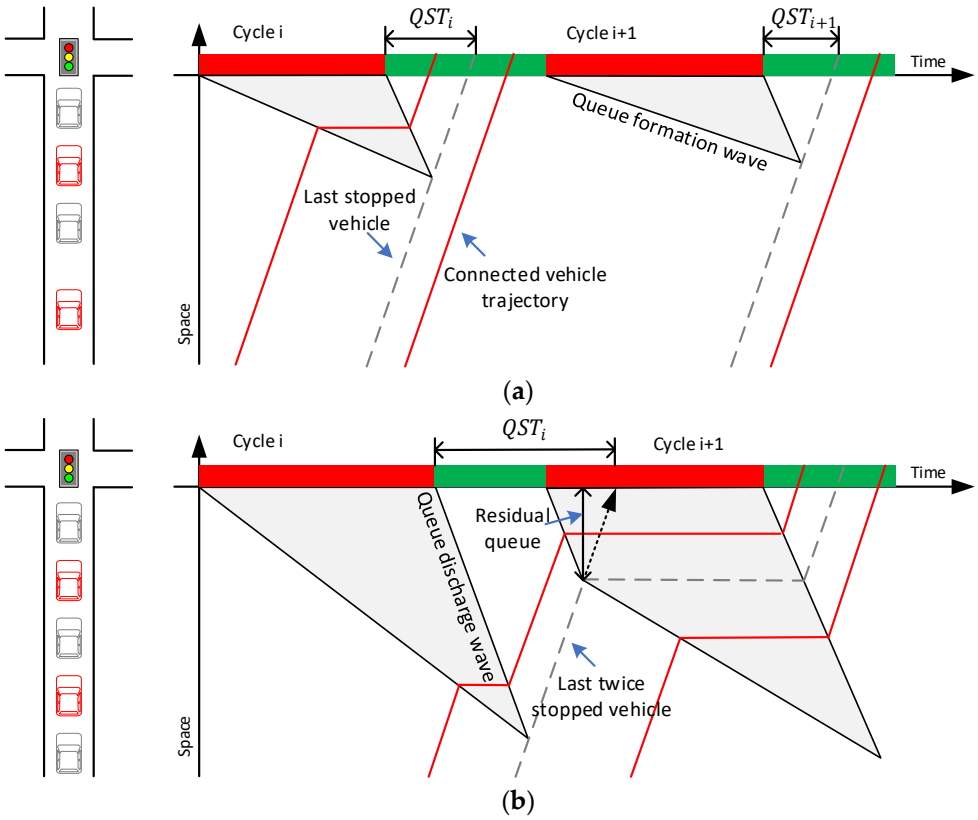

Figure 1. QST definition in both under- and oversaturation conditions: (**a**) undersaturation and (**b**) oversaturation.

With the above definition, the QST of a signal cycle can be determined by investigating the number of stops of vehicles in both under- and oversaturation conditions. In Figure 1, the trajectories of connected vehicles are shown as red curves. A connected vehicle j is assumed to provide its number of stops (i.e., n_j) and departure time since the start of the current green interval (i.e., t_j) when it passes the stop bar of the intersection, notated as a tuple vector $X_j = (n_j, t_j)$. Note X_j can be readily obtained when the full vehicle trajectory is available. Also, such information can be collected as event-based data reported by connected vehicles whenever they drive through the intersection. When considering a low penetration rate, the connected vehicles can only provide the upper and lower boundaries

of the QST. Suppose there are two connected vehicles observed in the same signal cycle and $n_j - n_{j+1} \geq 1$, the following relationship holds for the two conditions:

$$\sum_{p=1}^{n_j} g_{i-p+1} - g_i + t_j \leq QST_{i-n_j+1} < \sum_{p=1}^{n_j} g_{i-p+1} - g_i + t_{j+1} \tag{1}$$

where QST_i and g_i are the QST and the green duration of signal cycle i, respectively, and n_j is an integer number. The index $i - n_j + 1$ represents the signal cycle where the connected vehicle j experiences its first stop. In undersaturation condition, (1) is simply as $t_j \leq QST_i < t_{j+1}$. To distinguish the under- and oversaturation conditions, if any connected vehicle in the signal cycle has experienced two or more stops, it is considered as the oversaturation condition. When there is only one connected vehicle in the signal cycle, the relationship (1) is still valid but limited to providing either the lower or upper boundary for the QST. Accordingly, two types of observation can be collected in signal cycles within the historical periods.

- Lower boundary observations $\boldsymbol{\mathcal{L}}$: the connected vehicles whose n_j equal their maximum number of stops (≥ 1) in signal cycle i.
- Upper boundary observations $\boldsymbol{\mathcal{U}}$: the connected vehicles whose n_j are less than their maximum number of stops (≥ 1) in signal cycle i.

Thus, the final observation can be obtained as $O_j = (q_j, b_j)$. Here, q_j is a binary variable with $q_j = 0$ when $q_j \in \boldsymbol{\mathcal{L}}$ and $q_j = 1$ when $q_j \in \boldsymbol{\mathcal{U}}$, and b_j is the boundary observation determined by (1). The QST estimation problem in this study is to find the distribution of QST in a given period that has the most likelihood of observing all historical data $\{O_j\}$.

2.2. Probability-Based Model Formulation

Due to the stochastic nature of QST in different signal cycles, it is motivated to develop a probability-based model. The key question is how to define the likelihood of observing the data given a distribution of QST. According to Section 2.1, we can define two random events, which are to observe a lower boundary (i.e., $\boldsymbol{\mathcal{L}}$) and observe an upper boundary (i.e., $\boldsymbol{\mathcal{U}}$) of the QST at a particular time. As the two random events are mutually exclusive and have a one-way transition over time (i.e., the transition only starts from observing the lower boundary to the upper boundary), it is reasonable to adopt the logistic function (also known as the sigmoid function) to describe the probability of observing the two mutually exclusive events as a function of time. At any time, this is equivalent to making a binary classification/decision about whether it is under or over the QST. Another important reason to choose the logistic function is that the logistic function is the cumulative distribution function of the logistic distribution, which provides convenience in capturing the QST distribution via logistic distribution. To this end, the probability-based model can be mathematically expressed as follows:

$$QST \sim Logistic(\mu, s) = \frac{e^{-\frac{t-\mu}{s}}}{s\left(1 + e^{-\frac{t-\mu}{s}}\right)^2} \tag{2}$$

$$p(\boldsymbol{\mathcal{U}}|t, \mu, s) = \frac{1}{1 + e^{-(t-\mu)/s}} \forall t \in [0, \infty) \tag{3}$$

$$p(q_j|b_j, \mu, s) = p(\boldsymbol{\mathcal{U}}|b_j, \mu, s)^{q_j} \left(1 - p(\boldsymbol{\mathcal{U}}|b_j, \mu, s)\right)^{1-q_j} \tag{4}$$

where μ and s are the mean and scale parameters of the logistic distribution, Equation (3) is the cumulative distribution function of the logistic distribution (i.e., logistic function), and $1 - p(\boldsymbol{\mathcal{U}}|t, \mu, s)$ and $p(\boldsymbol{\mathcal{U}}|t, \mu, s)$ represent the probabilities of observing the lower and upper

boundaries at time t, respectively, given the distribution of QST. Therefore, the likelihood of observing q_j at time b_j (i.e., $p(q_j|b_j, \mu, s)$) can be determined by a Bernoulli distribution shown in Equation (4).

As the QST is a positive-scale value and potentially has a tailed distribution in real observations, a logarithmic transformation of the QST is further considered, which is equivalent to say the QST following the log-logistic distribution. This would enhance the model to capture occasionally longer QST. In addition, the QST is highly correlated to the signal timing. For an isolated intersection, a larger red duration will lead to longer QTS. This may not be a critical issue if the signal timing is fixed in the analysis period but could have misleading results if the signal timing varies. The varying signal timing should be considered in a real application when signal control is actuated or during the transition of two signal plans. To address this issue, we use the red duration to normalize the QST. The observed lower and upper boundaries are normalized as $b'_j = b_j / r_{i-n_j^{max}+1}$, where $n_j^{max} = \max(n_j, 1)$ is the maximum number of stops of connected vehicles departed in the *i-th* signal cycle and $r_{i-n_j^{max}+1}$ is the red duration of the signal cycle indexed as $i - n_j^{max} + 1$.

2.3. Bayesian Estimation of Model Parameters

In the proposed probability-based model, the unknown parameters are μ and s. Here, we proposed to adopt the Bayesian estimation approach with the principle:

$$p(\mu, s|\boldsymbol{\mathcal{O}}) \propto p(\boldsymbol{\mathcal{O}}|\mu, s)p(\mu)p(s) \tag{5}$$

where $\boldsymbol{\mathcal{O}}$ is the observation set in historical periods $O_j \in \boldsymbol{\mathcal{O}}$, $p(\boldsymbol{\mathcal{O}}|\mu, s)$ is the likelihood function defined as Equation (4), $p(\mu)$ and $p(s)$ are prior distributions of model parameters, respectively, and $p(\mu, s|\boldsymbol{\mathcal{O}})$ is the posterior distribution. The Bayesian estimation is particularly suitable to estimate the QST when the penetration rate of the connected vehicles is low because (i) the Bayesian estimation can accommodate small sample cases when introducing a reasonable prior distribution and (ii) the Bayesian estimation is recursive when new observations are available. For a well-defined Bayesian estimation problem, the proper definition of the prior is necessary, especially with limited observations. In this study, we define the priors as follows:

$$\mu \sim Normal\left(\mu_{piror}, \sigma^2_{piror,\mu}\right) \tag{6}$$

$$s \sim HalfNormal\left(\sigma_{piror,s}\right) \tag{7}$$

where μ_{piror}, $\sigma_{piror,\mu}$, and $\sigma_{piror,s}$ are hyperparameters of the prior distributions. The values of the hyperparameters should be determined based on historical observations to reflect our faith in the solution space of the model parameters. In this study, the hyperparameters are determined as follows:

- μ_{piror} and $\sigma_{piror,\mu}$ are determined by the mean and standard deviation of all the boundary observations in the historical periods.
- $\sigma_{piror,s}$ is determined by the standard deviation of the lower boundary observations in signal cycles, which reflects our faith that the dispersion of the QST distribution should be similar to the lower boundary observations.

With the continuous and differential nature of the proposed model, there are many techniques qualified to solve the Bayesian estimation problem, such as the maximum a posteriori (MAP), variational inference (VI), and Markov chain Monte Carlo (MCMC). In this study, we resort to a Hamiltonian Monte Carlo Method, namely, No-U-Turn sampler (NUTS), to achieve a generalized and reliable estimation performance. Compared to

the traditional samplers such as the Metropolis–Hastings (M-H) algorithm, the NUTS avoids the random walk process when proposing a new sample and largely improves the effectiveness and efficiency by evaluating the gradients of the solution space. More technique details of the NUTS are referred to [21]. The NUTS is readily available in advanced probability programming toolkits, such as PyMC [22].

2.4. Estimation of QST Posterior Distribution

With the posterior samples of model parameters generated by the NUTS, the posterior distribution of QST can be estimated. For computational convenience, we approximate the posterior distribution by discretizing the solution space of QST into small bins. The solution space of the QST is denoted as $[0, T]$ and the width of the small discrete bin is denoted as δ. For the *t-th* bin $[(t-1)\delta, t\delta]$, the probability of the posterior distribution can be calculated as follows:

$$p(t) = \frac{\sum_n^N (F(t\delta; \mu_n, s_n) - F((t-1)\delta; \mu_n, s_n))}{N} t = \left[1, 2, \cdots, \frac{T}{\delta}\right] \tag{8}$$

where N is the total number of the posterior samples, μ_n and s_n are the *n-th* posterior samples of model parameters, and $F(t\delta;\ \mu_n, s_n)$ is the cumulative distribution function of the logistic distribution, which is defined as Equation (3). Accordingly, the quantile of the posterior distribution can be further calculated as follows:

$$Q(p) = \frac{\sum_n^N \left(\mu_n + s_n \ln\left(\frac{p}{1-p}\right)\right)}{N}, 0 < p < 1 \tag{9}$$

where $Q(p)$ is the p-quantile of the QST posterior distribution. In the next section, the quantile of QST is used to determine the optimal signal timing.

3. Signal Timing Optimization

In this section, the estimated distribution of QST will be utilized to generate a robust signal timing plan. Here, the isolated signalized intersection under or near saturation is considered. Though the QST estimation is also applicable in oversaturation, the signal timing optimization for persistent oversaturation may need additional considerations. Typically, the maximum throughput should be considered over the entire congestion period, and the coordination of upstream intersections is also necessary to prevent the queue from spillback [23]. Nevertheless, due to the stochastic nature of traffic flow, oversaturation still occurs in signal cycles, and its occurrence frequency will increase when the intersection approaches saturation.

3.1. Calculation of Desired Green Ratio Based on QST

For an isolated intersection under or near saturation, the capacity provided by the signal timing should be sufficient to serve the traffic demand. The demand-to-capacity ratio is defined as the degree of saturation (DoS) for a signal group:

$$x = \frac{vC}{gs} = \frac{vh_s}{\theta} \tag{10}$$

where x is the DoS, v represents traffic arrival rate (i.e., traffic demand), C and g are the cycle length and effective green, respectively, $\theta = g/C$ is the green ratio, s is the saturation flow rate, and $h_s = 1/s$ is the saturation headway.

For isolated intersections, it is reasonable to assume a uniform traffic arrival; hence, the traffic arrival rate can be estimated by traffic arriving during the red light and queue

clearance stage. The number of arrivals should equal the number of vehicles discharged during the queue clearance stage. Therefore, the traffic arrival rate can be estimated by the QST for an isolated intersection:

$$v = \frac{QST/h_s}{QST + r} = \frac{1}{h_s}\frac{\beta}{1+\beta} \tag{11}$$

where QST/h_s represents the number of vehicles discharged during the queue clearance stage, and $\beta = QST/r$ is the normalized QST by red duration r. Equation (11) also implies that the traffic fundamental diagram is triangle-shaped, meaning that a vehicle is either traveling at free flow speed or fully stopped. This ensures that there are QST/h_s vehicles arrive within the time window $QST + r$. One can readily verify this in both under- and oversaturation conditions with a uniform arrival in a signal cycle, see Figure 1a,b. Substituting (11) into (10), we can calculate the desired green ratio for a signal group:

$$\theta = \frac{1}{x}\frac{\beta}{1+\beta} \tag{12}$$

As h_s is canceled out in Equation (12), we can see the desired green ratio is only determined by the target DoS and normalized QST of a signal group. To enhance the robustness, the desired green ratio should reserve spare capacity to accommodate traffic fluctuations in a control period. The most intuitive and simple way is to set a lower target DoS (e.g., 0.85) or a higher QST. With the estimated QST distribution, we would use the quantile of QST (larger than 0.5) to calculate the desired green ratio. The target DoS is initially set as one unless oversaturation occurs. Though there are more sophisticated models tackling traffic fluctuation [15], we intend to highlight the robustness of using QST to optimize signal timings. As shown in Equation (12), we can derive the desired green ratio without the exact knowledge of traffic demand and capacity. In the experiment, it is evidenced that even a small bias introduced in demand or capacity will incur an unstable signal performance.

3.2. Goal Programming Formulation

With the desired green ratios for all signal groups, the next question is how to find a signal plan to achieve these goals. In this study, we formulate the signal timing optimization problem as a linear goal program. The basic idea is to achieve the desired green for each signal group as much as possible. It is equivalent to the objective of minimizing the occurrence of oversaturation or phase failure. For K signal groups, their green time is denoted as $\{g_1, g_{2,} \cdots g_K\}$ and desired green ratio as $\{\theta_1, \theta_{2,} \cdots \theta_K\}$. The signal plan is defined in J stages, including the green and inter-green (i.e., yellow and all red) stages. The signal timing optimization problem can be formulated as follows.

Objective Function:

$$\min Z = \sum_{k=1}^{K} z_k = \sum_{k=1}^{K} \left(w_k^+ d_k^+ + w_k^- d_k^-\right) \tag{13}$$

Goal Constraints:

$$\sum_{j=1}^{J} \varphi_j^k D_j - d_k^+ + d_k^- - l_k = \theta_k C \tag{14}$$

Stage-Cycle Constraint:

$$\sum_{j=1}^{J} D_j = C \tag{15}$$

Min and Max Green Constraints:

$$g_k^{min} \leq g_k = \sum_{j=1}^{J} \varphi_j^k D_j \leq g_k^{max} \tag{16}$$

Min and Max Cycle Constraints:

$$C_{min} \leq C \leq C_{max} \tag{17}$$

Equation (13) is the objective or cost function for minimization. z_k is the cost of deviation from the desired green for the *k-th* signal group. The cost is determined by two deviation variables and their weights. d_k^+ and d_k^- are the two deviation variables that indicate the amount of time that is over and under the desired green, respectively. w_k^+ and w_k^- are the weights to punish the two deviations. Equation (14) imposes a goal constraint for each signal group. φ_j^k is a binary indicator showing if the signal group k is green in signal stage j. Equations (15)–(17) guarantee that the cycle length and green time of each signal group are bounded. The decision variables are D_j, d_k^+, d_k^-, and C. As the objective function and constraints are all linearly formulated, it is easy to solve the signal timing optimization problem. In this study, the optimization of phase sequence is not considered and assumed to be given.

By solving the optimization problem, if all signal groups result in $d_k^+ \geq 0$ and $d_k^- = 0$, it is regarded as a "fully achieved case" because all signal groups have green time no less than the desired green. The desired greens are exactly realized only for the signal group with critical movements. When any $d_k^- > 0$, it suggests certain signal groups have insufficient green and oversaturation occurs. This case is regarded as a "partially achieved case". In this study, we will dissolve the pressure of oversaturation over all the signal groups. This can be realized by increasing the targe DoS in a small incremental step (e.g., 0.01) and re-conducting the desired green calculation and signal timing optimization processes iteratively. Such iteration will terminate and output the optimal results once all signal groups reach $d_k^+ \geq 0$ and $d_k^- = 0$. Thus, the signal plan would have the maximum cycle length, and the green durations would be allocated proportionally according to the magnitude of the QST. Note that this strategy may not guarantee the maximum throughput for an oversaturated intersection, but it is a reasonable choice for under and near-saturated conditions. One can also set different lengths of the incremental steps to prioritize certain signal groups.

4. Experiment

4.1. Simulation Scenario

A typical four-leg intersection with different traffic scenarios is simulated by the microscopic simulator SUMO to validate the proposed method. Without loss of generality, we only consider the eastbound and southbound left and through movements (i.e., EL, ET, SL, and ST), assuming other movements are non-critical. The lane layout of the intersection is shown in Figure 2. The signal phase sequence is fixed as $ET \rightarrow EL \rightarrow ST \rightarrow SL$. Five 15 min intervals are configured in the simulation with different traffic demands as listed in Table 1.

To account for uncertainty in both traffic demand and capacity, the traffic demand is generated randomly following the Poisson distribution, and stochastic saturation headway is created by adopting the stochastic car-following model in SUMO. The field tests and analysis in previous studies have indicated that the Poisson distribution is a rational choice to describe traffic arrivals at isolated signalized intersections under both unsaturated and oversaturated conditions [6–8]. A baseline signal timing plan with a cycle length of 100 s is

used to represent an inferior and outdated signal plan. There are eight stages including four green (i.e., Stage 1, 2, 5, and 7) and four inter-green stages (i.e., Stages 2, 4, 6, and 8). The green stages are 30, 15, 30, and 13 s, respectively, and all the inter-green stages are 3 s. The baseline signal plan is applied in all five intervals. By observing the simulation, Intervals 1, 2, and 5 are typical undersaturation scenarios, where green time is often wasted with no service on traffic. Interval 3 is the congested scenario and frequently experiences phase failure leading to residual queues. Interval 4 is the after-congestion scenario, which may be affected by the residual queue left from Interval 3.

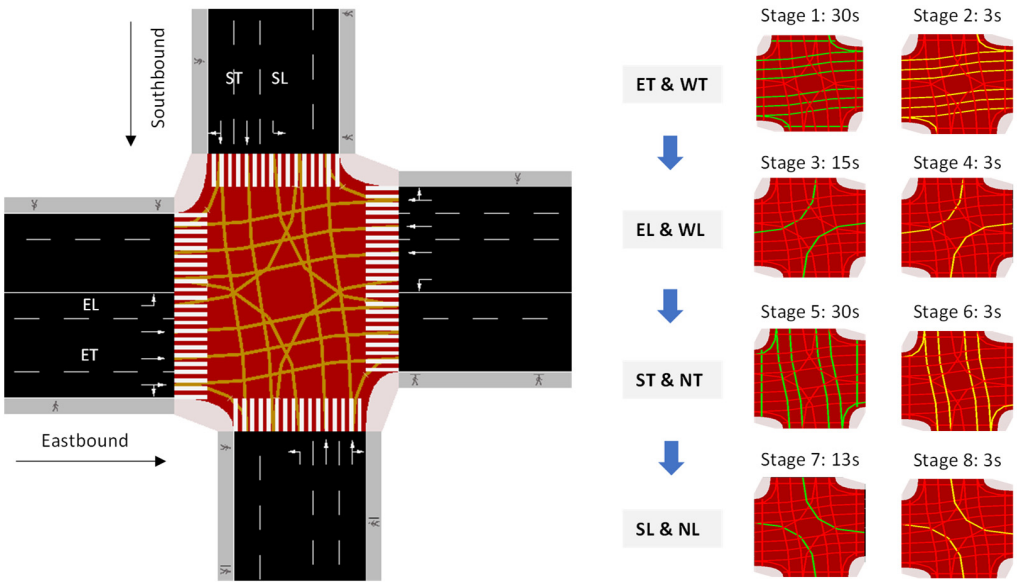

Figure 2. Simulation intersection and the baseline signal timing plan.

Table 1. Traffic demand configuration in simulation.

Interval	Simulation Time (s)	Average Traffic Demand (veh/h/lane)			
		ET	EL	ST	SL
1	[0, 900]	324	108	230	81
2	[900, 1800]	486	162	230	81
3	[1800, 2700]	486	162	459	162
4	[2700, 3600]	324	108	459	162
5	[3600, 4200]	324	108	230	81

4.2. Model Implementation

The connected vehicle and traffic signal data are collected from SUMO through the Traci interface. The connected vehicles are randomly created in the traffic flow according to a given penetration rate. Each connected vehicle will report its number of stops and departure time when it crosses the stop bar of the intersection. The QST estimation model is implemented in PyMC. The NUTS is configured to collect 2000 posterior samples of which the first 1000 samples would be discarded as the burn-in samples. The signal optimization is solved by linprog in SciPy, which implements the simplex algorithm. The maximum signal cycle length, minimum green duration, and inter-green duration are set as 180, 6, and 3 s, respectively. The lost time for each signal group is 3 s.

4.3. Estimation Results of QST

Five simulations are conducted with the baseline signal plan and different random seeds to provide historical data for QST estimation. Five penetration rates of connected

vehicles (i.e., 0.02, 0.05, 0.1, 0.15, and 0.2) are considered to investigate their impacts. At each penetration rate, the estimation is repetitively conducted 10 times to produce sufficient results. The ground truth of QST is recorded according to the criterion described in Section 2.1. Since the output of the proposed model is the posterior distribution of QST, we adopt the Hellinger distance to evaluate the estimation results. In statistics, the Helligner distance is designed to measure the similarity between two distributions, which is regarded as a probabilistic analog of the Euclidean distance. According to Equation (8), the Hellinger distance can be calculated as follows:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T/\delta}\left(\sqrt{p(t)} - \sqrt{q(t)}\right)} \tag{18}$$

where $p(t)$ and $q(t)$ are the discrete probabilities of the estimated and ground truth distributions at the *t-th* bin, respectively. The Hellinger distance ranges from 0 to 1, and smaller values indicate higher similarity between the two distributions.

Figure 3 illustrates the Hellinger distance of QST estimations at different penetration rates of connected vehicles. The error bars indicate the 95% confidence interval of all 10 estimations. For ST and ET, it is observed that the Hellinger distance decreases with the increase in penetration rate, and further decrease becomes marginal when the penetration rate is over 0.1. In contrast, SL and EL do not show such a trend. The impact of a higher penetration rate is minor. The evidence observed in the cases of SL and EL indicates that the posterior distribution is primarily affected by the prior distribution, and the additional observations are not sufficient to be predominated. The results also reflect the fact the prior distributions do show their effectiveness in enhancing the reliability of estimations with limited observations.

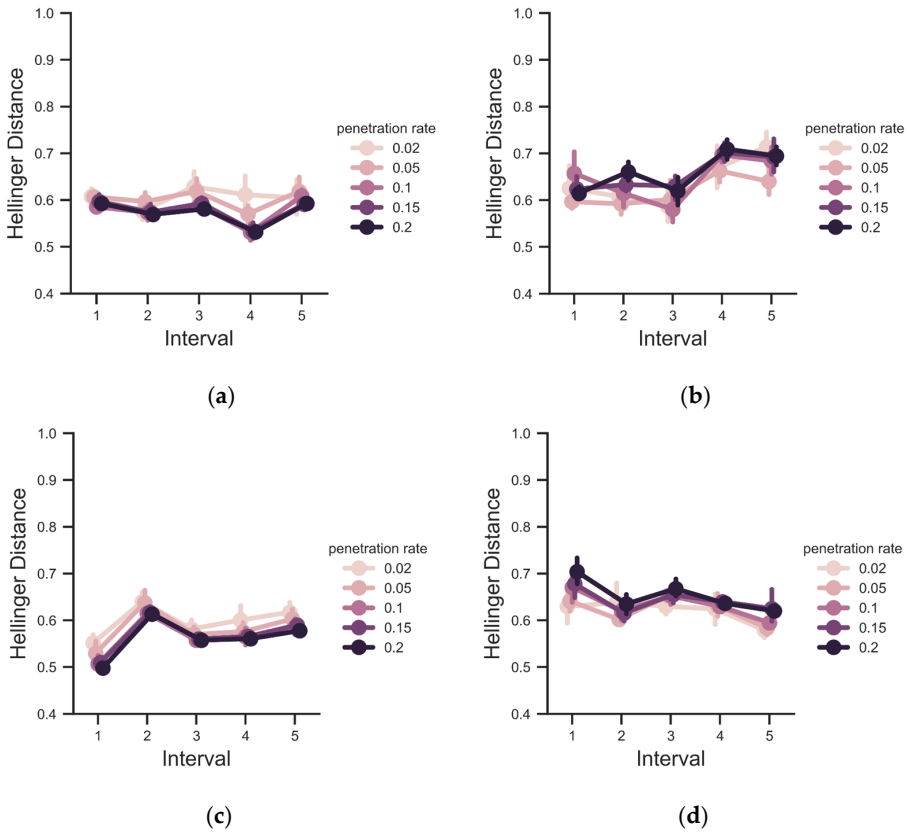

Figure 3. Hellinger distance of QST estimations at different penetration rates: (**a**) ST, (**b**) SL, (**c**) ET, and (**d**) EL.

To gain more insight into the proposed Bayesian estimation, the relation between the number of observations and Hellinger distance is analyzed, shown in Figure 4. Twelve groups are divided according to the number of observations. For each group, the Hellinger distance is summarized with its mean and standard deviation, shown as the dots and error bars, respectively. In Figure 4, when the number of observations exceeds 50, it is shown that the Hellinger distance will reach a convergence. Further increase in the number of observations will not result in a significant decrease in the Hellinger distance. The results suggest that not only the penetration rate matters but also the total number of observations in the historical periods. In a real application, at least 50 connected vehicles are recommended to achieve a desirable performance of the proposed model.

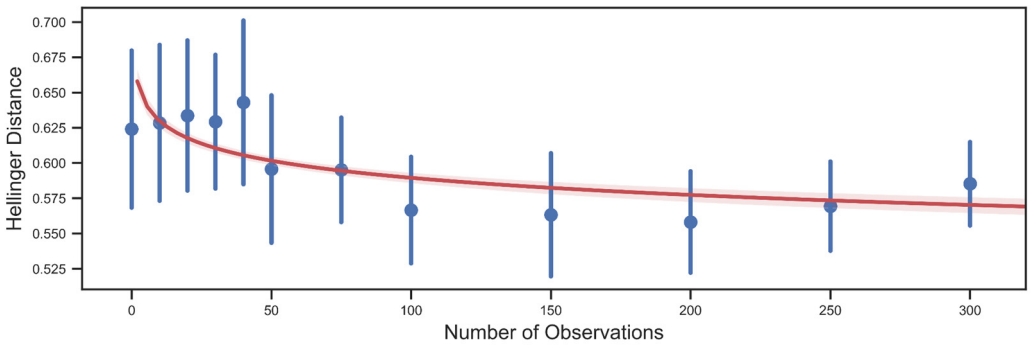

Figure 4. Relation between the number of observations and Hellinger distance.

Figure 5 illustrates an example of the estimated QST distributions at a penetration rate of 0.05. It is shown that the proposed model successfully estimates the actual QST distributions under different traffic scenarios. The logarithm transformation is also proven to be effective in capturing the tailed distributions, which are mostly presented in more congested intervals (i.e., Intervals 3 and 4). Note that the simulation provides a noisy-free data collection, but the accuracy of the observation could be a critical issue in a real-world application. In field observation, not all vehicle stops are related to signal control when there is a bus stop or pedestrian crossing near the intersection. The exact time when a vehicle crosses the stop bar is also affected by the localization accuracy, updating frequency, and time reference consistency of the vehicle trajectory. The potential approach to enhance the accuracy of the observation data is to incorporate the shockwave theory for data preprocessing. For instance, the vehicle with signal-related stops should depart near the discharge shockwave path. The time drift between the signal timing and vehicle trajectory can also be corrected by analyzing the vehicle start-up process and the discharge shockwave.

4.4. Performance of Signal Timing Optimization

With the estimated QST distributions, the signal timing plans are optimized for the five control intervals. Here, a demand-based optimization method is compared assuming the demand distribution and mean saturation headway are known. The demand distributions are Poisson distributions identical to the simulation settings to generate traffic. The saturation headways are estimated using simulation data, with 2.0 and 2.3 s for the through and left-turn lanes, respectively. Note that the actual demand distribution and mean saturation headway are not easy to obtain with sampled connected vehicle trajectory data. Nevertheless, the significance of the comparison is to show the effectiveness and robustness of the proposed method using only connected vehicle data.

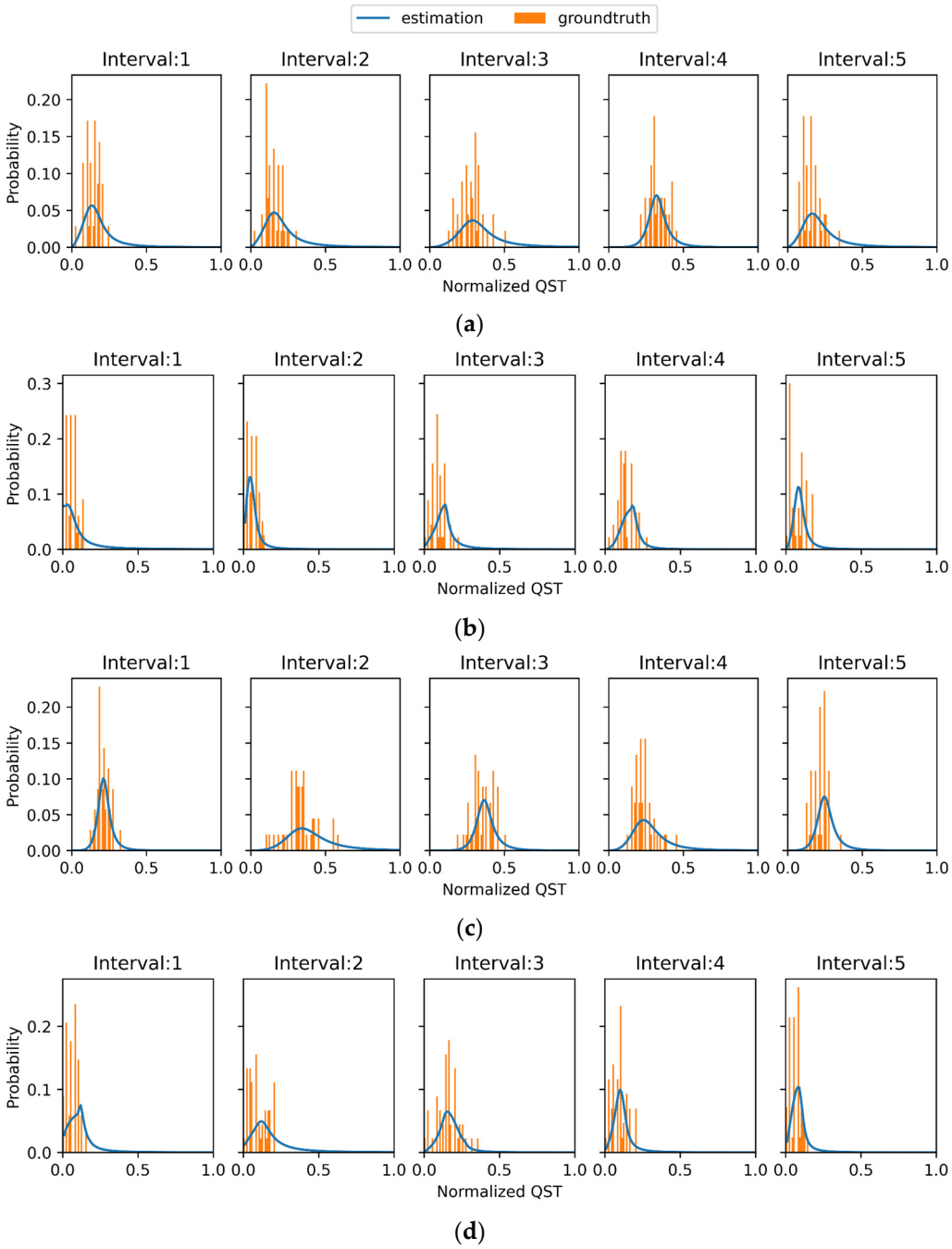

Figure 5. Estimated QST distributions at a penetration rate of 0.05: (**a**) ST, (**b**) SL, (**c**) ET, and (**d**) EL.

The QST-based and demand-based methods are both solved by the optimization model described in Section 3. Different quantiles of QST and demand are used to generate robust signal timing plans. For a particular quantile, 10 simulations are repeated given the signal timing plans generated by each method. The vehicle delay and number of stops of all movements are recorded, as illustrated in Figure 6. The dots and error bars represent the mean and 95% confidence interval of the two performance measures, respectively. In Figure 6, it is observed that the number of vehicle stops in the QST-based method gradually decreases with the increase in QST quantile and incurs fewer stops than in the demand-based method. This suggests that the goal of minimizing the occurrence of oversaturation or phase failure is better achieved by the QST-based method. By investigating the optimization results, it is found the QST-based method tends to reserve more capacity by applying a longer signal cycle length than the demand-based method at the same quantile. When the

quantile exceeds 0.7, the cycle length will significantly increase in the QST-based method than in the demand-based method, which results in a larger vehicle delay. In the higher quantile cases, the QST-based method may be too conservative because the QST quantiles are less likely to occur simultaneously for all the movements. When the QST quantile exceeds 0.9, the target DoS (i.e., $x = 1$) cannot be realized in the QST-based method and the signal plans become ill-posed by overemphasizing the less likely cases. Nevertheless, the QST-based method under the quantile of 0.7 shows superior performance over the demand-based method.

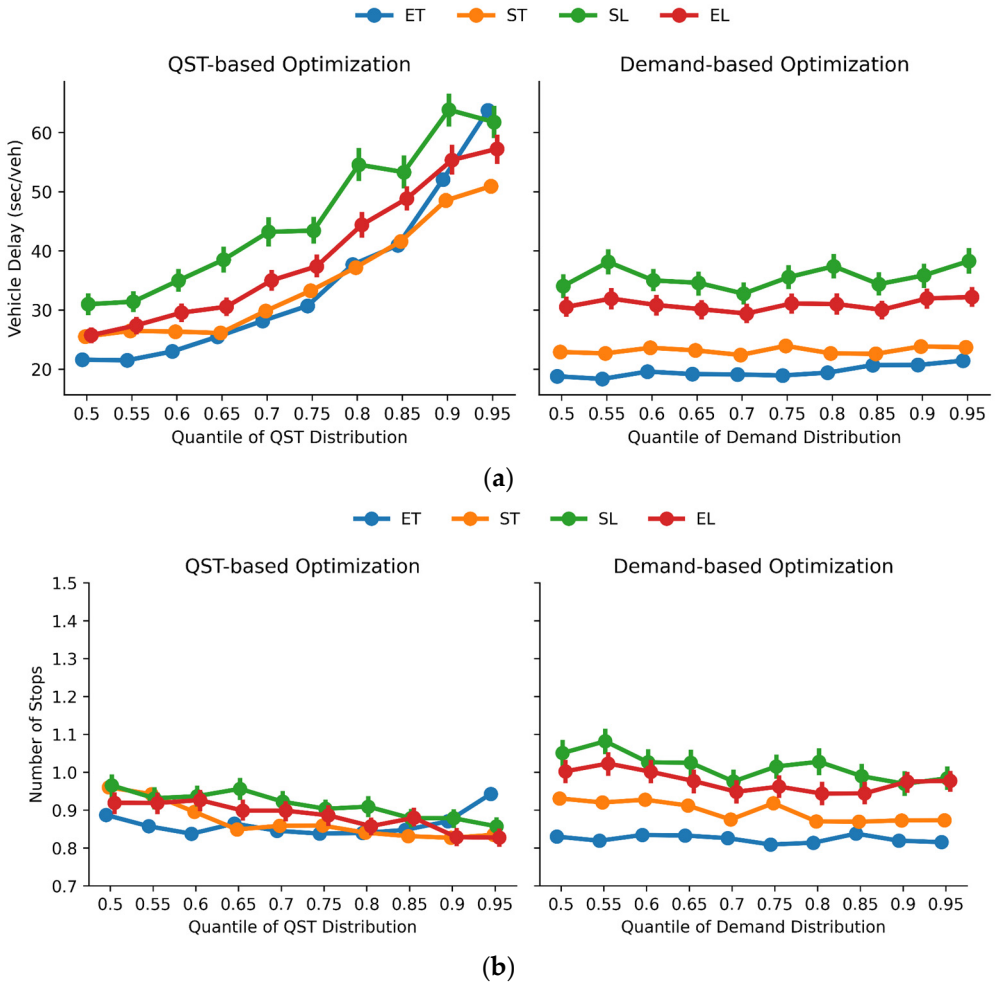

Figure 6. Evaluation of vehicle delay and number of stops at different quantiles: (**a**) vehicle delay and (**b**) number of stops.

Compared with the QST-based method, the demand-based method is less sensitive to the changes in demand quantiles, and all the target DoS can be realized even at the quantile of 0.95. However, this may indicate a robustness issue that the demand-based method is too optimistic by ignoring the stochastic saturation headway. To verify this, we investigate the worst cases of the two methods, as summarized in Table 2. The demand-0.5 and QST-0.5 represent the demand-based and QST-based methods using the quantile of 0.5 for signal timing optimization. The quantiles of 0.5, 0.55, 0.6, and 0.65 are investigated as larger quantiles in the QST-based method have significantly longer cycle lengths, making it less meaningful to evaluate vehicle delay. Here, the worst cases in each method are represented as the vehicle delay distribution larger than its 95th percentiles. The mean and standard deviation (SD) of vehicle delays are calculated for the worst cases.

Table 2. Performance in the worst cases.

Method	Mean Delay	95th Delay	Worst Mean	Worst SD	Max. Delay	Changes (%)			
						Mean	95th	Worst Mean	Worst SD
Baseline	30.6	66.0	84.6	33.1	327	-	-	-	-
Demand-0.5	21.7	58.0	73.2	21.6	186	−29.1	−12.1	−13.5	−34.7
QST-0.5	23.6	59.0	72.2	15.8	162	−22.9	−10.6	−14.7	−52.3
Demand-0.55	21.7	58.0	75.3	23.4	185	−29.1	−12.1	−11.0	−29.3
QST-0.55	24.0	63.0	74.5	14.8	154	−21.6	−4.5	−11.9	−55.3
Demand-0.6	22.4	61.0	76.2	19.4	184	−26.8	−7.6	−9.9	−41.4
QST-0.6	25.1	69.0	79.9	13.1	174	−18.0	4.5	−5.6	−60.4
Demand-0.65	22.0	60.0	74.6	20.1	163	−28.1	−9.1	−11.8	−39.3
QST-0.65	26.7	73.0	85.3	14.5	193	−12.7	10.6	0.8	−56.2

Compared with the baseline signal timing plan, the mean vehicle delays are all reduced by the demand-based and QST-based methods. At the quantile of 0.5, the signal timing is optimized with the mean QST and traffic demand. The QST-based method shows its robustness in reducing the worse mean and standard deviation. This verifies that the QST can better describe the combined stochastic effects of traffic demand and capacity. When the quantile increases, it is also observed that the standard deviations of the worst cases in the QST-based method are further decreased and lower than the demand-based method. Except for the quantile of 0.65, the delays of the worst cases are all bounded by a smaller maximum delay in the QST-based method.

To further demonstrate the robustness of the proposed method, we fix the cycle length as 100 s and optimize the signal timing using the two methods. With a fixed cycle length, the optimization problem is reduced to allocate green time to each signal group. Figure 7 shows the vehicle delay in the fixed cycle signal timing optimization cases. Here, we also introduce a small bias (i.e., 0.1 s) into the mean saturation headway in the demand-based method, see Figure 7c,d. With a fixed cycle length constraint, the demand-based method will allocate the green time to each signal group proportional to the flow ratio (i.e., the ratio of traffic demand and saturation flow rate). With the identical demand distribution and static saturation headway, the flow ratio will remain constant for all the quantiles in the demand-based method. In contrast, the QST quantiles will vary as each movement has a different QST distribution with different levels of uncertainty, which explains the varying delay when increasing the quantile in the QST-based method. In general, the QST-based method will prioritize the movement with larger uncertainty when increasing the quantile but may become over-conservative with high quantiles. By comparing Figure 7b,c, it is shown that the demand-based method is very sensitive to the saturation headway. The performance may be significantly diverse even with a small bias in the saturation headway.

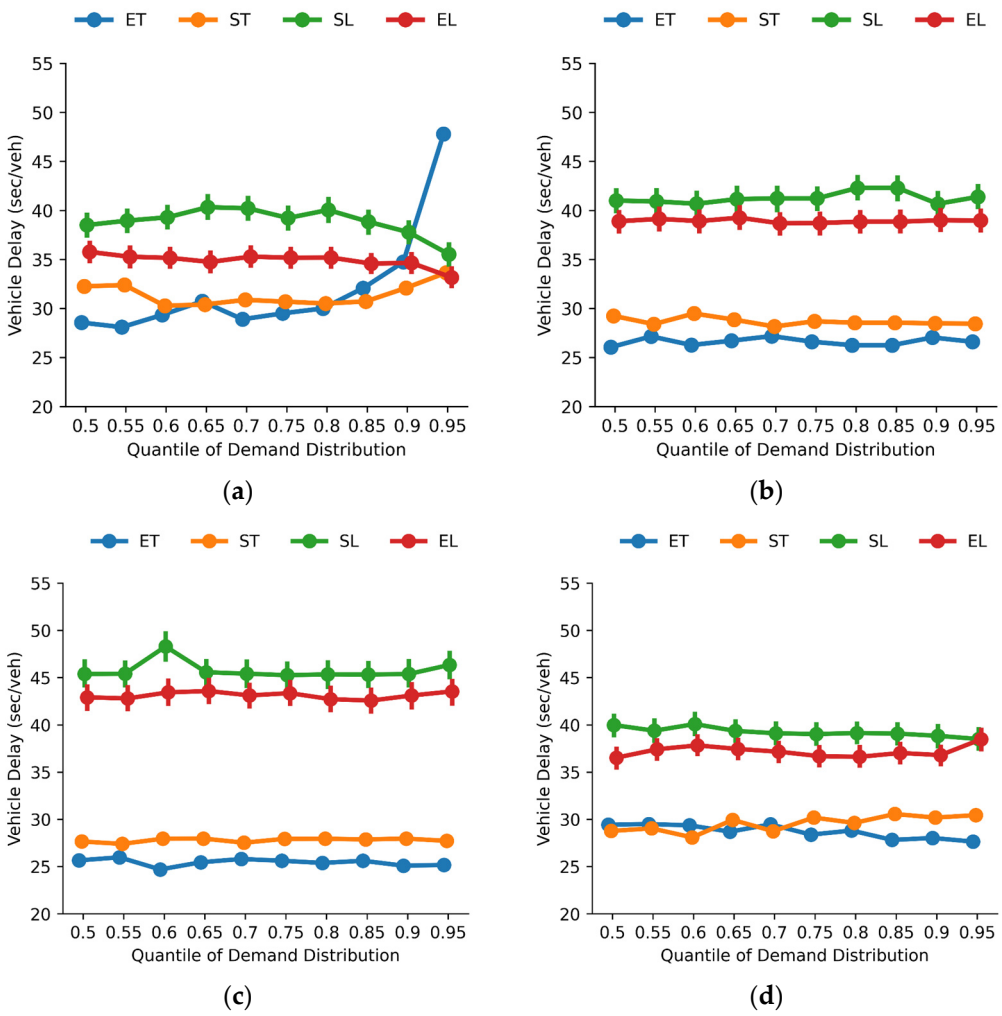

Figure 7. Vehicle delay by the fixed cycle signal timing optimization: (**a**) QST-based, (**b**) Demand-based ($h_s^T = 2.0, h_s^L = 2.3$), (**c**) Demand-based ($h_s^T = 2.1, h_s^L = 2.2$), and (**d**) Demand-based ($h_s^T = 1.9, h_s^L = 2.4$).

5. Conclusions and Future Work

This study proposes a robust signal retiming approach for isolated intersections using low-penetration connected vehicle data. A novel traffic state measure—queue service time (QST)—is introduced and used as the only input to optimize signal timing. A probability-based model and Bayesian estimation approach are developed to effectively estimate the QST distribution by collectively using the lower and upper boundary observations reported by connected vehicles. A goal programming-based signal optimization model is formulated using quantiles of QST as input, which accounts for the combined uncertainty in both traffic demand and capacity. The proposed method has been validated through simulation experiments. It is shown that the proposed QST estimation model is reliable to use under a penetration rate as low as 0.05 and effectively estimates the actual distribution in both under- and oversaturation conditions. Compared with a demand-based method that only accounts for uncertainty in traffic demand, the proposed QST-based signal timing optimization method shows its superiority in reducing the occurrence of oversaturation or phase failure, as well as enhancing performance against the worst cases. The results also suggest that the demand-based method is sensitive to a small bias in the settings of saturation headway.

There are still several aspects that need further consideration in future work. First, the simple utilization of QST quantiles to generate a robust signal timing plan may lead

to over-conservative results. It would be more effective to consider the joint distribution of QST among different signal groups. Second, the proposed method now is restricted to isolated intersections and implies that traffic arrival is not significantly different before and after queue clearance. This restriction can be further addressed by introducing the penetration rate as an additional model parameter in the estimation model. Finally, more field experiments are needed to further validate the proposed model.

Author Contributions: The authors confirm their contributions to the paper as follows: C.A.: conceptualization, methodology, validation, and writing—original draft and editing; W.Z.: methodology, validation, and writing—original draft and editing; Y.W.: validation and writing—review and editing; S.K.: validation and writing—review and editing; J.X.: conceptualization, writing—review and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 52202398 and Grant 52272309 and partially supported by the International Science and Technology Cooperation Project of Jiangsu Province under Grant BZ2023015.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Luyanda, F.; Gettman, D.; Head, L.; Shelby, S.; Bullock, D.; Mirchandani, P. ACS-Lite algorithmic architecture: Applying adaptive control system technology to closed-loop traffic signal control systems. *Transp. Res. Rec.* **2003**, *1856*, 175–184. [CrossRef]
2. He, Q.; Head, K.L.; Ding, J. PAMSCOD: Platoon-based arterial multi-modal signal control with online data. *Transp. Res. Part C Emerg. Technol.* **2012**, *20*, 164–184. [CrossRef]
3. Feng, Y.; Head, K.L.; Khoshmagham, S.; Zamanipour, M. A real-time adaptive signal control in a connected vehicle environment. *Transp. Res. Part C Emerg. Technol.* **2015**, *55*, 460–473. [CrossRef]
4. Tajalli, M.; Hajbabaie, A. Traffic signal timing and trajectory optimization in a mixed autonomy traffic stream. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6525–6538. [CrossRef]
5. Cheng, Y.; Hu, X.; Chen, K.; Yu, X.; Luo, Y. Online longitudinal trajectory planning for connected and autonomous vehicles in mixed traffic flow with deep reinforcement learning approach. *J. Intell. Transp. Syst.* **2023**, *27*, 396–410. [CrossRef]
6. Zheng, J.; Liu, H.X. Estimating traffic volumes for signalized intersections using connected vehicle data. *Transp. Res. Part C Emerg. Technol.* **2017**, *79*, 347–362. [CrossRef]
7. Lloret-Batlle, R.; Wang, Z.H.; Zheng, J. Traffic volume estimation for both undersaturated and oversaturated signalized intersections with stopbar location estimation using trajectory data. *Transp. Res. Rec.* **2023**, *2677*, 343–354. [CrossRef]
8. Tan, C.; Yao, J.; Tang, K. Joint estimation of multi-phase traffic demands at signalized intersections based on connected vehicle trajectories. *arXiv* **2022**, arXiv:2210.10516. [CrossRef]
9. Zhao, Y.; Zheng, J.; Wong, W.; Wang, X.; Meng, Y.; Liu, H.X. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 70–91. [CrossRef]
10. Mei, Y.; Gu, W.; Chung, E.C.; Li, F.; Tang, K. A Bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data. *Transp. Res. Part C Emerg. Technol.* **2019**, *109*, 233–249. [CrossRef]
11. Zhao, Y.; Wong, W.; Zheng, J.; Liu, H.X. Maximum likelihood estimation of probe vehicle penetration rates and queue length distributions from probe vehicle data. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 7628–7636. [CrossRef]
12. Feng, Y.; Zheng, J.; Liu, H.X. Real-time detector-free adaptive signal control with low penetration of connected vehicles. *Transp. Res. Rec.* **2018**, *2672*, 35–44. [CrossRef]
13. Tan, C.; Cao, Y.; Ban, X.; Tang, K. Connected vehicle data-driven fixed-time traffic signal control considering cyclic time-dependent vehicle arrivals based on cumulative flow diagram. *IEEE Trans. Intell. Transp. Syst.* **2024**. [CrossRef]
14. Ma, W.; Wan, L.; Yu, C.; Zou, L.; Zheng, J. Multi-objective optimization of traffic signals based on vehicle trajectory data at isolated intersections. *Transp. Res. Part C Emerg. Technol.* **2020**, *120*, 102821. [CrossRef]
15. Yin, Y. Robust optimal traffic signal timing. *Transp. Res. Part B Methodol.* **2008**, *42*, 911–924. [CrossRef]
16. Zhang, L.; Yin, Y.; Lou, Y. Robust signal timing for arterials under day-to-day demand variations. *Transp. Res. Rec.* **2010**, *2192*, 156–166. [CrossRef]
17. Tettamanti, T.; Luspay, T.; Kulcsar, B.; Péni, T.; Varga, I. Robust control for urban road traffic networks. *IEEE Trans. Intell. Transp. Syst.* **2013**, *15*, 385–398. [CrossRef]

18. Ma, W.; An, K.; Lo, H.K. Multi-stage stochastic program to optimize signal timings under coordinated adaptive control. *Transp. Res. Part C Emerg. Technol.* **2016**, *72*, 342–359. [CrossRef]
19. Li, L.; Huang, W.; Chow, A.H.; Lo, H.K. Two-stage stochastic program for dynamic coordinated traffic control under demand uncertainty. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 12966–12976. [CrossRef]
20. An, C.; Shen, H.; Xu, Y.; Lu, Z.; Xia, J. Hidden mixture vehicle discharge state inference at signalized intersection using vehicle travel time and discharge headway data. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21700–21711. [CrossRef]
21. Hoffman, M.D.; Gelman, A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
22. Patil, A.; Huard, D.; Fonnesbeck, C.J. PyMC: Bayesian stochastic modelling in Python. *J. Stat. Softw.* **2010**, *35*, 1. [CrossRef]
23. Sun, W.; Wang, Y.; Yu, G.; Liu, H.X. Quasi-optimal feedback control for an isolated intersection under oversaturation. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 109–130. [CrossRef]