Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6674037, 27 pages
https://doi.org/10.1155/2021/6674037

Research Article

Simulation-Based Optimization for the Operation of Toll Plaza at Car Park Exit with Mixed Types of Tollbooths and Waiting-Time-Dependent Service

Shanchuan Yu,[1] **Yuchuan Du**,[2] **Jindong Wang**,[3] **Yishun Li**,[2] **and Yong Zhu**[1]

[1]*Research and Development Center of Transport Industry of Self-Driving Technology, China Merchants Chongqing Communications Research & Design Institute Co. Ltd., Chongqing 400067, China*
[2]*Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, 201804, China*
[3]*Shanghai Jinqiao (Group) Co. Ltd., Shanghai, 201206, China*

Correspondence should be addressed to Shanchuan Yu; yushanchuan@cmhk.com

Received 10 December 2020; Revised 26 January 2021; Accepted 8 February 2021; Published 23 February 2021

Academic Editor: Giulio E. Cantarella

This study presents an approach of simulation-based optimization to the operation of the toll plaza at the car park exit. We first propose a simulation model, as the representation of the queueing system for the toll plaza with mixed-type customers and servers where the service time is dependent on the waiting time of customer. Then, a simulation-based integer programming model is developed to design more traffic-efficient yet cost-effective operation schemes. It is decomposed by a rolling horizon approach into subproblems which are all solved via the Kriging metamodel algorithm. A numerical example is presented to illustrate the model and offer insight on how to achieve traffic efficiency and cost-effectiveness.

1. Introduction

Pay-at-exit (PAE) is a manual parking fee payment system. Cars' plates are identified by the camera when drivers enter a parking area and drivers pay a fee on exiting based on the duration of stay. The manual payment process is relatively long and could result in long queues at the exits when the departing car flows are high. In recent years, some parking operators have installed a new parking fee payment system, which is called pay-on-foot (POF). In the POF system, drivers are required to walk to pay at centralized pay stations before they return to the cars, or pay in their smartphone after scanning the Quick Response (QR) code inside the car park. Upon exiting the car park, the cars' plates become the verified paid tickets as exit passes. The POF system could not only reduce the manpower expense, but also fasten the exit time and reduce the delays of customers at the exit.

For the convenience of customers, parking operators usually set up both the POF and PAE tollbooths in the toll plazas at the car park exits. Drivers could pay cash at the PAE tollbooths if they have not paid at centralized pay stations or in their smartphone. POF tollbooths serve POF cars exclusively. But PAE tollbooths can serve either the POF or PAE cars. Therefore, the long queues on the toll lanes will form if the numbers of POF and PAE tollbooths are not suited to the proportions of POF and PAE cars, as Figure 1(a) illustrates.

To improve the service at the toll plaza, there are more tollbooths than the approaching lanes. Hence, a transition area is needed where the number of approaching lanes has widened to the number of toll lanes in front of the toll plaza. In this area, the conflicts derived from the lane changing behaviors incur delays for drivers, as Figure 1(b) presents. In addition, when the queues in front of some tollbooths stretch to the bottleneck between transition area and approaching lanes (see Figure 1(c)), drivers have to wait before the transition area even though they plan to head to the toll lanes with a few queues.

Therefore, improper allocation of POF and PAE tollbooths in service will incur serious delays for drivers, causing

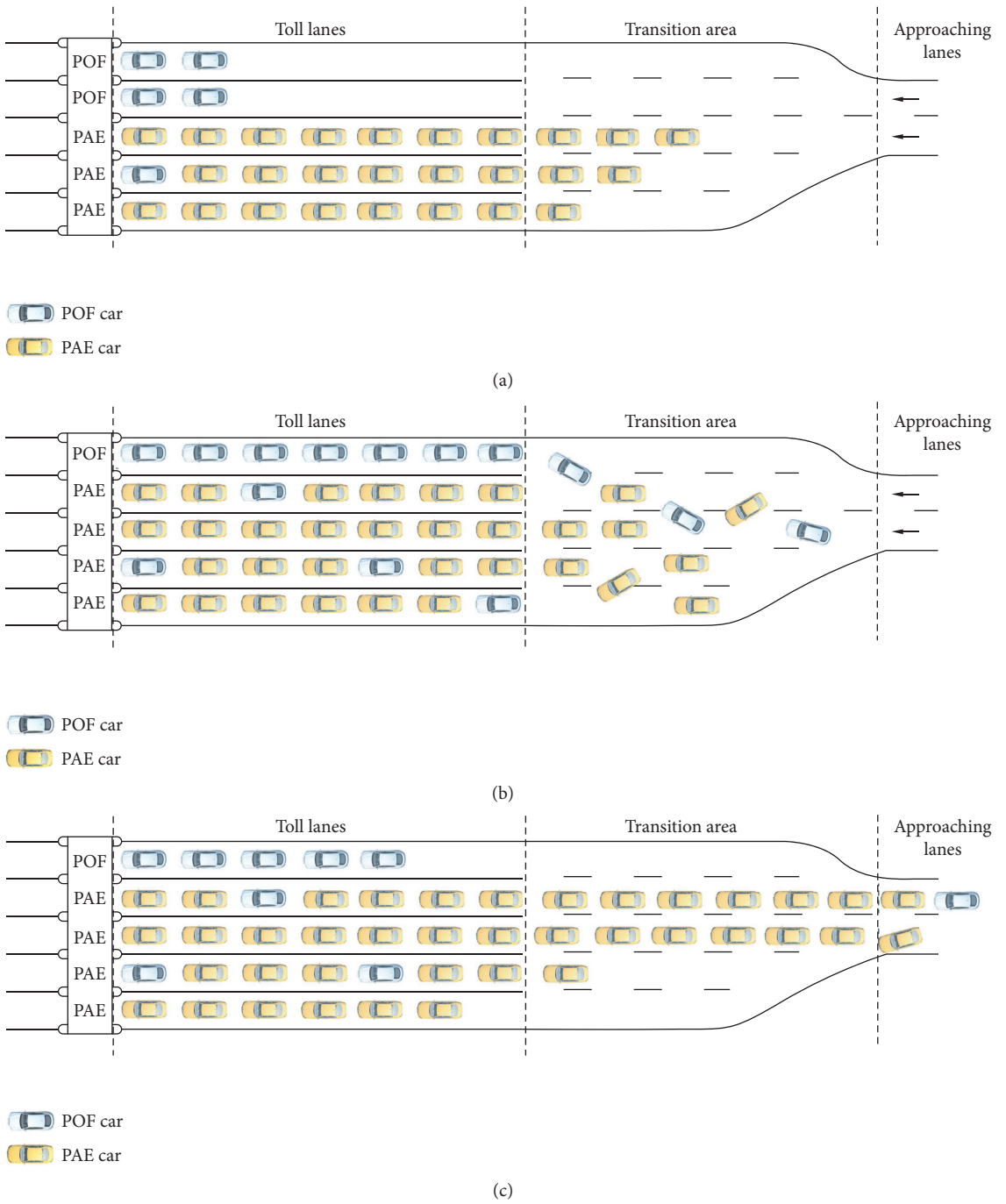


Figure 1: Causes for delays at the car park exit. (a) Queueing on the toll lanes. (b) Lane changing in the transition area. (c) Bottleneck at the approaching lanes.

traffic inefficiency of the exit. However, the efficiency evaluation of POF or PAE tollbooth is not simply multiplying average efficiency by the number of tollbooths. The reasons are threefold. First, the car type is not exclusive to the tollbooth type. PAE tollbooths can serve either the POF or PAE cars. Second, the service time is waiting-time-dependent. In most cases, PAE tollbooths have more service time than POF ones. However, if a POF car unfortunately overstays a specified time due to the long queues after paying inside the car park, the driver has to make an extra payment in the smartphone after scanning the QR code at POF tollbooth. In this case, the service time in POF tollbooth considerably increases. In addition, car parks grant drivers certain time for free parking and a driver needs to pay only if the one overstays the time. If a car stays within a predetermined time, the car does not have to pay and becomes PAE car arriving at the exit, spending only a few seconds in PAE tollbooths. Third, the positions of POF and PAE tollbooths will affect the efficiency if the queues stretch to the bottleneck between transition area and approaching lanes. Therefore, a more sophisticated

mathematical model describing the queueing process is needed for the efficiency evaluation of toll plaza.

More tollbooths in service can improve the traffic efficiency, while it will lead to higher operational costs. Thus, the tradeoff between traffic efficiency and operational cost should be taken into account. In addition, since the arrival intensity of the cars fluctuates throughout the day, the allocation scheme of POF and PAE tollbooths in service should be time-varying in day-to-day operation. Car park operators usually estimate the exiting intensity of cars in a period before the beginning of the period, especially for the car park of airport where the number of cars is highly dependent on the flight schedule. Thus, they adjust the tollbooth operation scheme gradually, dividing the whole tollbooth operation problem into several stages, where each stage is only for exiting cars in the coming period. Only the plan for the near future is updated and executed. Based on this practice, the allocation is changed on a period-by-period basis.

Consequently, in this study, the operation of toll plaza is formulated as a discrete-time dynamic tollbooth allocation problem (DTAP) where the traffic efficiency and operational cost for the toll plaza are balanced in the objective function. The traffic efficiency evaluations of POF and PAE tollbooths are derived from a mathematical model that describes the queueing process at the toll plaza.

1.1. Literature Review

1.1.1. Operation Analysis of Toll Plaza. Few researchers have studied the operation of the toll plaza at a car park exit, especially the one with both POF and PAE tollbooths. Nevertheless, various studies have evaluated and proposed strategies to improve the operation of the highway toll collection system, which is similar to the parking fee collection system. The operation is described with the dynamic traffic states [1–3] or queue evolution [4–7]. After the introduction of Electronic Toll Collection (ETC) system, research efforts have been made to appropriately allocate the ETC and manual toll collection (MTC) tollbooths [8–11]. The operation of toll plaza on highway with ETC and MTC tollbooths is similar to that at car park exit with POF and PAE tollbooths. ETC tollbooths serve ETC cars exclusively while MTC tollbooths can serve either the ETC or MTC cars. However, the billing rule is the key distinction between the highway fare collection system and the parking fee collection system. For highway application, the fare is distance-based and is independent of the amount of time when a car waits at the toll plaza while for parking application, the fare is waiting-time-based. Some car parks grant drivers certain time for free parking. A driver needs to pay only if the one overstays the time. In addition, if a POF car unfortunately overstays a specified time after paying inside the car park, the driver has to make an extra payment in the smartphone after scanning the QR code at POF tollbooth.

Hence, it is important to consider the dependence of the service time in tollbooths on the waiting time in front of the toll plaza when we model the operation of the parking fee collection system.

As the aforementioned studies indicate, queueing analysis is the mainstream of evaluating the operation of toll plaza. Approaches to queueing analysis usually fall into two categories: analytical modeling and simulation. Analytical modeling is based on the queueing theory, which uses known probability distributions to describe the cars' interarrival and tollbooths' service patterns. A large body of literature studies exists on the queueing system with dependent services, which are related to arrival rate [12], waiting time [13], queue length [14, 15], or workload [16, 17]. Nevertheless, the queueing theory fails when the probability distribution of either cars' interarrival or tollbooths' service is not all mathematically explicit or when the car's arrival rate is greater than the processing capacity in which the queue tends to be infinite. Simulation tools [4, 10, 18–20] could capture the behaviors of individual cars, including deceleration, waiting, acceleration, lane choice, lane changing, and paying. Although emulating the entire queueing process could be complex, simulation provides an access to evaluate the system performance when the interarrival of car platoons and service patterns of servers do not yield to the distributions with explicit mathematical forms, and the servers are of mixed types and are dependent on the waiting times of customers.

1.1.2. Simulation-Based Optimization Approach. The optimization for the operation of toll plaza tends to be developed based on the analytical model for the queueing system [8, 21]. Although simulation has been traditionally used as a tool to understand and experiment with a system, connecting the simulation model to the optimization engine also gives an effective solution to the optimization problem [22, 23]. Simulation-based optimization is an approach whereby an optimization engine provides the decision variables for the simulation model. The simulation model provides the results of the optimization objective function. This process will continue iteratively between the simulation model and the optimization engine until it results in a satisfactory solution or a termination due to prescribed conditions [24, 25].

Hence, the discrete-time DTAP in this paper developed by an integer programming model is formulated based on the simulation results from the queueing system. The simulation-based optimization approach has been widely used to solve the problems of congestion pricing, traffic signal control, transit scheduling, vehicle sharing, supply chain management, liner shipping, etc. [23]. However, the above simulation-based optimization problem is computationally expensive and cannot be solved by derivative-based solvers because the objective function and/or constraint set have to be treated as black boxes for the algebraic description of the simulation is not directly available. To overcome the challenges of simulation-based optimization, significant research efforts have been done to generate surrogate models of the black-box functions [26]. The surrogate method includes the response surface method, multivariate adaptive regression

splines, the regression polynomials method, the Kriging method, the radial basis function (RBF) method, and the neural network method [27, 28]. The differences of these models are mainly in the approximating functions. The Kriging model is a widely used surrogate model. We use the Kriging metamodel to solve the discrete-time DTAP in this study because it is more flexible than polynomial regression models in fitting arbitrary smooth response functions and less sensitive than other metamodels (such as RBF) to a small change in the design of the experiment [23].

The complexity of discrete-time DTAP is also given by the number of integer variables, which increases with the number of periods [29]. Furthermore, it is difficult to obtain an optimal or even a near-optimal solution, since a long time horizon needs to be considered [30]. One way to handle this problem is to use a rolling horizon heuristic. Such a procedure only considers a portion of the entire time horizon, solves the reduced problem, and fixes parts of the solution. Rolling horizon schemes have been applied to several problems within operational problems under uncertainty (see Chand et al. [31] for an extensive review). Rolling horizon approaches decompose the time horizon into several stages. The offset between the starting times of two consecutive stages is defined as roll period. Roll periods also can be fixed [32–34] or event-based [30]. Usually, car park operators change the tollbooth plan at fixed intervals for the convenience of tollmen's scheduling, even though such operation cannot give an immediate response to the intensity of exiting cars. Hence, in this study, we utilize a rolling horizon approach to decompose the discrete-time DTAP based on fixed roll period.

1.2. Objectives and Contributions. The objective of this study is to determine the optimal operation scheme of POF and PAE tollbooths for a toll plaza at the car park exit. The problem is formulated as a discrete-time DTAP that a simulation-based integer programming (SIMIP) model is formulated where the objective function balances the traffic efficiency and operational cost. Then, the discrete-time DTAP is decomposed into period-based subproblems via a rolling horizon approach. Each subproblem is iteratively solved by the Kriging metamodel-based algorithm (KMA). The contributions of this study are twofold. First, the car park exit is simulated as a queueing system where the customers and servers are of mixed types, and the service time is dependent on the waiting time of customer (hereinafter we use "customer," "car," and "driver" interchangeably, and "server" and "tollbooth" interchangeably). The simulation also takes into account the time-varying allocations of POF and PAE tollbooths with respect to arrival intensities of car platoons and the blockage from the queueing spillover on the adjacent toll lanes in the transition area. Second, the operation of a toll plaza is modeled as a discrete-time DTAP where a SIMIP is formulated. The problem is decomposed by the rolling horizon approach into subproblems that are solved via KMA.

The remainder of this paper is organized as follows. Section 2 develops a SIMIP model to describe the discrete-time DTAP for the toll plaza where a simulation model is formulated as the representation of the queueing system. The validation for the simulation model is also presented. A rolling horizon approach is proposed in Section 3 where the problem is decomposed into subproblems which are solved via KMA. Section 4 demonstrates the simulation results and recommended operation scheme for the toll plaza in the numerical example. Finally, the conclusions and future work are discussed in Section 5.

2. Problem Description and Model Formulation

In this study, the operation scheme of the toll plaza is determined in the time horizon $[0, T]$, where T is the ending time of the cars' arrival at the car park exit. We discretize the time horizon $[0, T]$ into N number of δT time periods, $\{1, 2, \ldots, n, \ldots, N\}$ where $T = N\delta T$. Period n can also be denoted as $[T_{n-1}, T_n]$, where $T_0 = 0$, $T_n = n\delta T$, and $T_N = T$. δT can be one hour in day-to-day operations. Before we formulate the optimization problem, the following assumption is proposed initially.

Assumption 1. The allocation of PAE and POF tollbooths can only be changed at the beginning of each period.

It will heavily take up the memory of the computer if we simulate the queueing system in the entire time horizon $[0, T]$, as shown in Figure 2(a). In addition, the allocation of PAE and POF tollbooths changes at the beginning of each period according to the arrival intensity of car platoons. Hence, we discretize the simulation into a number of subsimulations on a period-by-period basis, as Figure 2(b) presents. In simulation for each period, inputs are the car platoons arriving in this period and the queueing evolution vectors from the last period, while outputs are queueing evolution vectors and the performance measures in this period.

In Section 2.1 the SIMIP model for discrete-time DTAP is presented. Then, we describe the queueing system of the toll plaza with mixed-type customers and servers where the service time is dependent on the waiting time of customer in Section 2.2.

2.1. SIMIP Model for Discrete-Time DTAP. According to the brief analysis in the Introduction, the number of tollbooths in service, the proportion, and the location of POF and PAE tollbooths in service affect the waiting in queues in front of tollbooths and also the blockage from the queueing spillover on the adjacent tollbooths. Hence, we should find the appropriate tollbooth operation scheme which delivers the highest traffic efficiency. More tollbooths in service can improve the traffic efficiency of the toll plaza. However, it will also lead to higher operational cost. Therefore, we need to balance the traffic efficiency and the cost for the toll plaza at the car park exit.

Let B denote the label set of tollbooths and $B = \{1, \ldots, j, \ldots, N_T\}$ where N_T denotes the number of tollbooths (toll lanes). The number of tollbooths is described as $|B|$ where $|\cdot|$ is the cardinality of the set. B_n^{IN} denotes the

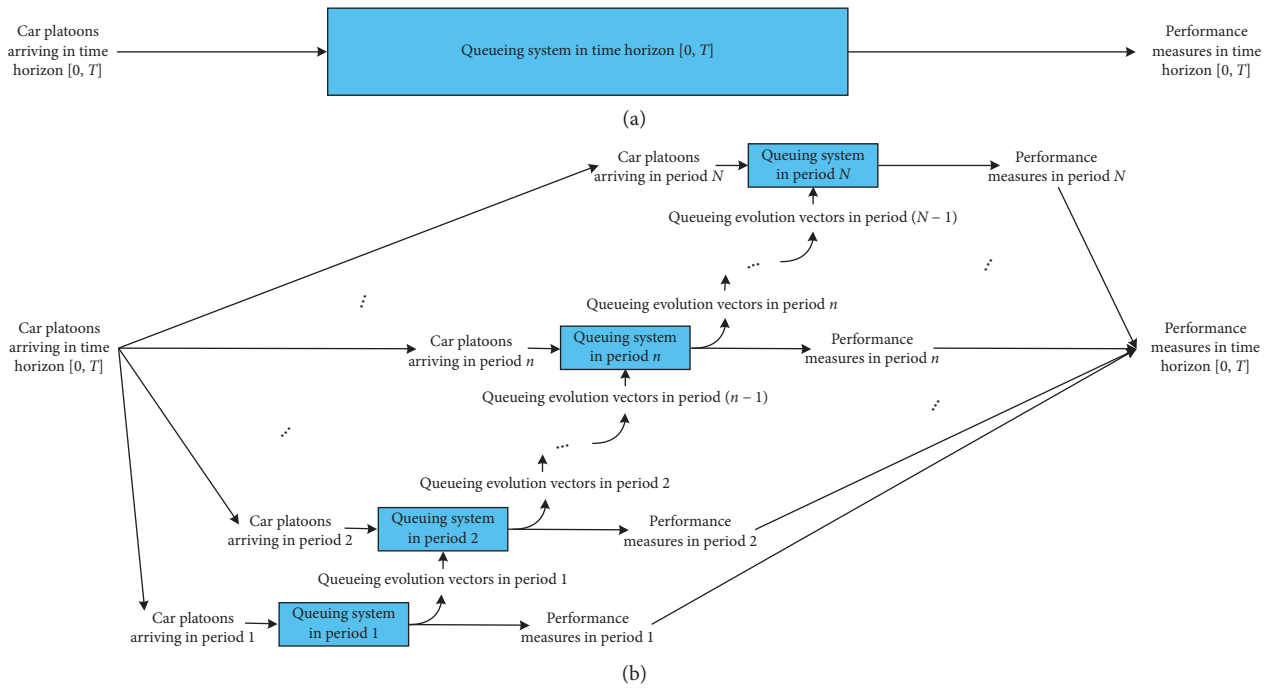


FIGURE 2: Simulation for the entire time horizon and subsimulations based on periods. (a) Simulation for the entire time horizon. (b) Simulations on a period-by-period basis.

label set of tollbooths in service in the period n. The set B_n^{IN} is divided into two disjoint subsets B_n^{POF} and B_n^{PAE}, where B_n^{POF} and B_n^{PAE} denote the label sets of POF and PAE tollbooths in service in the period n, respectively.

In this paper, we use the total delay for the cars arriving in the entire time horizon as the measure of traffic efficiency, where the delay for a car is defined as the waiting time before receiving a service plus the service time in the toll plaza. The waiting for a car includes the waiting due to the busyness of tollbooth the car has chosen and the one due to the blockage of queues in front of other tollbooths. Let $d_t^{[n]}$ and d_t denote the total delays for the cars arriving in period n and entire time horizon $[0, T]$, respectively and we have

$$d_t\left(B_1^{\mathrm{POF}}, B_1^{\mathrm{PAE}}, \ldots, B_N^{\mathrm{POF}}, B_N^{\mathrm{PAE}}\right) = \sum_{n=1}^{N} d_t^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}\right), \tag{1}$$

where $d_t^{[n]}(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}})$ represents the total delay under the scheme $\{B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}\}$ in period n and can be obtained from the nth subsimulation. Hereinafter, the superscript $[n]$ on the notations of variables indicates that the variables are in period n (i.e., in the nth subsimulation).

We use the manpower and electricity expense to represent the operational cost. There is a tollman for each PAE tollbooth in service. The operational cost of the toll plaza under the scheme $\{B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}\}$ in period n, $c_p^{[n]}$, is given by

$$c_p^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}\right) = \begin{cases} c_m\left|B_n^{\mathrm{PAE}}\right|\delta T + c_e\left(\left|B_n^{\mathrm{PAE}}\right| + \left|B_n^{\mathrm{POF}}\right|\right)\delta T, & \text{if } 1 \le n \le N-1, \\ c_m\left|B_n^{\mathrm{PAE}}\right|\max\left(\delta T, (T_l)^{[n]}\right) + c_e\left(\left|B_n^{\mathrm{PAE}}\right| + \left|B_n^{\mathrm{POF}}\right|\right)\max\left(\delta T, (T_l)^{[n]}\right), & \text{if } n = N, \end{cases} \tag{2}$$

where c_m and c_e denote the manpower and electricity expense for each tollbooth in the unit time of the studied time horizon, respectively. $(T_l)^{[n]}$ denotes the moment when the last car leaves the toll plaza in the nth subsimulation. The operational cost of the toll plaza for the entire time horizon, c_p, is given by

$$c_p\left(B_1^{\mathrm{POF}}, B_1^{\mathrm{PAE}}, \ldots, B_N^{\mathrm{POF}}, B_N^{\mathrm{PAE}}\right) = \sum_{n=1}^{N} c_p^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}\right). \tag{3}$$

Let α and α_n denote the set of the indicators for tollbooth types in the entire time horizon and period n, respectively, where $\alpha = (\alpha_1, \ldots, \alpha_N)$ and $\alpha_n = (\ldots, \alpha_j^{[n]}, \ldots)$, $j \in B$. $\alpha_j^{[n]} = -1$ if tollbooth j is out of service for period n. $\alpha_j^{[n]} = 0$ if tollbooth j is a PAE one in service for period n. $\alpha_j^{[n]} = 1$ if tollbooth j is a POF one in service for period n. Then, the sets B_n^{IN}, B_n^{POF}, and B_n^{PAE} can be derived from the set B and the variable vector α_n. Hence, the tollbooth operation problem, i.e., discrete-time DTAP can be described as a simulation-based integer programming (SIMIP) model, presented as follows:

[SIMIP-DTAP]

$$\boldsymbol{\alpha}^* = \left(\boldsymbol{\alpha}_1^*, \ldots, \boldsymbol{\alpha}_N^*\right) \in \underset{(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)}{\arg\min}\ p_{\mathrm{cf}}(\boldsymbol{\alpha}) = \mathbb{E}\left[(1-w)\beta d_t\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\right) + w c_p\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\right)\right], \tag{4}$$

$$\text{subject to } \alpha_j^{[n]} = -1, 0, 1, \quad j \in B, n = 1, \ldots, N, \tag{5}$$

$$\prod_{j \in B} \alpha_j^{[n]} = 0, \quad n = 1, \ldots, N, \tag{6}$$

$$\mathbb{E}\left[q_j^{[n-1]}\left(\delta T; \boldsymbol{\alpha}_{n-1}, \mathbf{q}^{[n-2]}\right)\right]\left(\alpha_j^{[n]} - \alpha_j^{[n-1]}\right) = 0, \quad j \in B, n = 2, \ldots, N, \tag{7}$$

where α^* and α_n^* denote the optimal tollbooth allocation schemes for the entire time horizon and the period n, respectively. In the objective function, $d_t(\alpha_1, \ldots, \alpha_N)$ and $c_p(\alpha_1, \ldots, \alpha_N)$ are derived from equations (1) and (3), respectively. w and β are the weight for operational cost and the value of time (VOT), respectively. Note that the outputs are stochastic from simulation model. Therefore, $\mathbb{E}[\cdot]$ indicates the mean or median values after multiple parallel simulations. Constraint (6) is proposed to guarantee that at least one PAE tollbooth is on service, since the POF tollbooth cannot serve PAE cars while PAE tollbooth can serve both PAE and POF cars. In addition, a server cannot be closed or switch to another type in a period if it is still busy at the end of the last period, as presented in constraint (7) where $q_j^{[n-1]}(\delta T; \alpha_{n-1}, \mathbf{q}^{[n-2]})$ denotes the queue length in front of tollbooth j at the time δT in the $(n-1)$th subsimulation, i.e., $q_j^{[n-1]}(\delta T)$, which is derived from the inputs α_{n-1} and $\mathbf{q}^{[n-2]}$.

Here, $q_j^{[n]}(\cdot)$ denotes the queue length in front of tollbooth j with respect to the time in the nth subsimulation. $\mathbf{q}^{[n]}$ represents the matrix of the queue length evolution for all tollbooths in period n. The queue length in our study is defined as the total number of cars staying in an area at a moment. The queue length for the exit is the total number of cars in the exit at a moment, including the cars moving, queueing, and served at tollbooth and blocked by queues at other tollbooths. The queue length in front of a tollbooth is the total number of cars which have selected this tollbooth at a moment.

The derivation of all outputs from the simulation model such as $d_t^{[n]}$, $(T_l)^{[n]}$, and $q_j^{[n]}(\cdot)$ is elaborated in the next section.

2.2. Simulation Model for Queueing System of Toll Plaza. In this section, we develop the simulation model to represent the queueing system on a period-by-period basis and obtain the performance of the toll plaza. In each period of the queueing system (see Figure 3), the locations of PAE and POF servers in service are fixed. The customers arrive randomly with a probability distribution and can be presented as a random arrival profile. The service times for customers in a server are also random variables that are dependent with the types of customers and the server, and the customers in queue. Customers are served based on First-In-First-Out (FIFO) rule in a server. In addition, the customers' behaviors before receiving a service are also taken into account such as server selection and queueing due to the blockage from the queueing spillover on the adjacent servers.

In addition, car parks grant drivers certain time for free parking. If a car stays within a predetermined time, the car does not have to pay and becomes PAE car arriving at the exit. In addition, POF cars to exit the car park within a required time after they pay at centralized pay stations or in their smartphone. Otherwise, the POF cars have to pay by cash at PAE tollbooths or in smartphone at POF tollbooths for the extra times. The above regulations are also captured in the simulation.

2.2.1. Assumptions and Simulation Inputs. To simplify the representation of the queueing system, the following behavioral assumptions are proposed in the simulation.

Assumption 2. Travel speeds of cars are homogeneous. Decelerations and accelerations of cars are not considered.

Assumption 3. Before a driver arrives at the toll lanes, he or she selects the tollbooth in service with the shortest expected waiting time to wait. Once choosing a toll lane to queue on, he or she will not switch to another tollbooth.

Assumption 4. Blockage from the queueing spillover on the adjacent toll lanes is evaluated, while the delay due to the conflicts from the lane changing is not taken into account.

The exit is divided by three cross sections, denoted as CS_1, CS_2, and CS_3, respectively, in Figure 3. CS_1 is the place where drivers decide which tollbooth they will wait for exit. It is located upstream of the exit and is assumed to be rarely reached by queues. CS_2 is the place where cars start to be served at the toll plaza and CS_3 is the place where cars leave the toll plaza. The studied area is determined as the one between CS_1 and CS_3.

N_A and N_T denote the numbers of approaching lanes and toll lanes, respectively. There are N_L more toll lanes at the left-hand side of approaching lanes and N_R more toll lanes at the right-hand side of approaching lanes. Hence, $N_T = N_A + N_R + N_L$, as presented in Figure 3.

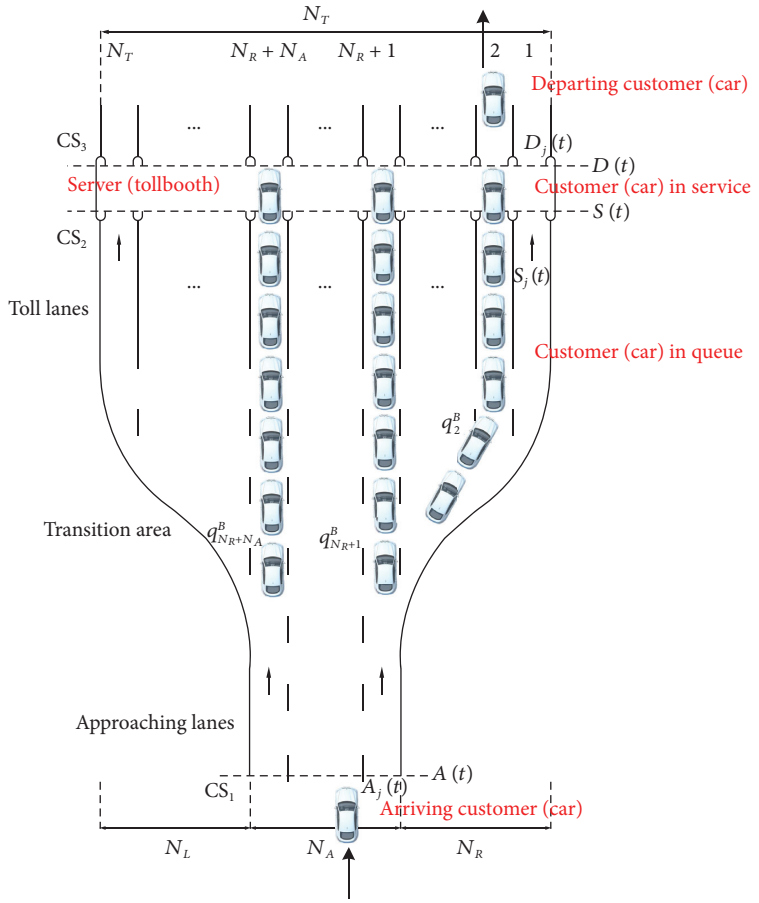


FIGURE 3: Queueing system for a toll plaza.

We denote $K_l^{[n]}$ and λ_n as the number and arrival rate of car platoons in period n, respectively, and we have $?_?? = ?_?^{[?]}$. The arrival pattern is generated via a given stochastic process with the arrival rate, as the representation of the cumulative profile at CS_1 for toll plaza, denoted by $A^{[n]}(t^{[n]}; \lambda_n \delta T)$, $t^{[n]} \in [T_{n-1}, T_n]$ as the cumulative profile at CS_1 for toll plaza, where $t^{[n]}$ denotes the time in the nth subsimulation. Let $(K_l)_j^{[n]}$ denote the number of cars choosing tollbooth j in the period n and we have $K_l^{[n]} = \sum_{j \in B_n^{IN}} (K_l)_j^{[n]}$. Let $(T_l)_j^{[n]}$ denote the moment in the nth subsimulation when the last car leaves the tollbooth j. In addition, we define that $(T_l)_j^{[0]} = 0$ and $(K_l)_j^{[0]} = 0$.

In conclusion, the main inputs in the simulation are the number of cars in period n, $K_l^{[n]}$, and the label sets of POF and PAE tollbooths in service in the period n, B_n^{POF} and B_n^{PAE}.

2.2.2. Driver Behaviors and Output Evaluation. Now we focus on driver behaviors in the simulation and the output evaluation. Consider the Kth car in the period n (denoted by $(K)^{[n]}$) which arrives at CS_1 at time $t_K^{[n]}$ with the parking duration $c_K^{[n]}$. The parking duration is randomly generated via empirical distribution.

Firstly, the type of the car is determined. If $c_K^{[n]} < t_f$, the driver has the chance to be free of charge if one exits the toll plaza immediately, where t_f denotes the free-parking time of the car park. He or she does not need to pay at centralized pay stations or pay in the smartphone after scanning the QR code inside the car park and the car $(K)^{[n]}$ is PAE car. Otherwise, the driver must pay. Whether the car $(K)^{[n]}$ is PAE or POF car is determined using Bernoulli distribution in terms of his or her preference to the POF system. Let P_r denote the proportion of drivers who prefer to the POF system and the car type is generated by P_r.

Then, the driver compares the queues of tollbooths in service. Since the service time of a POF server is usually shorter than that of a PAE server, the driver chooses the tollbooth with the shortest waiting time that he or she estimates (see Assumption 3). If more than one server has the shortest waiting time, the driver can select any one of those servers with identical probabilities. Let $\omega_j^{e[n]}(K)$ denote the estimated waiting time for tollbooth j by the car $(K)^{[n]}$. Here, we assume that drivers only know the average not the variance of a tollbooth's service times. Hence, $\omega_j^{e[n]}(K)$ is given by

$$\omega_j^{e[n]}(K) = q_j^{[n]}\left(t_K^{[n]}\right) \cdot \bar{s}_j, \tag{8}$$

where $\bar{s}_j$ denotes the average service time for tollbooth j. $q_j^{[n]}(\cdot)$ denotes the queue length in front of tollbooth j with respect to the time in the nth subsimulation. $q_j^{[n]}(t_K^{[n]})$ is the summation of the newly formed queue length in the period n and the queue length in the last subsimulation $(n-1)$ at the

same time, $q_j^{[n-1]}(t_K^{[n]} + \delta T)$. The detail for the derivation of $q_j^{[n]}(\cdot)$ is presented in Appendix A.

Once the Kth car makes a decision at CS_1, one is labelled as the kth car choosing tollbooth j and denoted by $(k)_j^{[n]}$. Then, the car may be blocked and wait for the discharge of the queues from other tollbooths in the transition area, and afterwards, it queues on the toll lane until tollbooth j start to serve it. Let τ_j, $\omega_j^{B[n]}(k)$, and $\omega_j^{Q[n]}(k)$ denote the travel time from CS_1 to tollbooth j when there is no queue in front of tollbooth j, the waiting time for the car $(k)_j^{[n]}$ due to the blockage of the queues in front of other tollbooths, and the waiting time for the car $(k)_j^{[n]}$ due to the busyness of tollbooth j, respectively. The derivation of $\omega_j^{B[n]}(k)$ and $\omega_j^{Q[n]}(k)$ is presented in Appendix A. We denote $\omega_j^{[n]}(k)$ as the waiting time for the car $(k)_j^{[n]}$ and we have

$$\omega_j^{[n]}(k) = \omega_j^{B[n]}(k) + \omega_j^{Q[n]}(k). \tag{9}$$

The service time for kth car is determined by the type of tollbooth j, and $c_K^{[n]}$, τ_j, and $\omega_j^{[n]}(k)$ of the car. Let $s_j^{[n]}(k)$ and $d_j^{[n]}(k)$ denote the service time and delay for the car $(k)_j^{[n]}$, respectively. We have

$$d_j^{[n]}(k) = \omega_j^{[n]}(k) + s_j^{[n]}(k). \tag{10}$$

After capturing the movement of each individual car, the performance measures of the toll plaza can be described. We use the total delay of car platoons in order to reflect the overall time costs of drivers' waiting. Total delay for the cars arriving in period n, $d_t^{[n]}$, is given by

$$d_t^{[n]} = \sum_{j \in B} \sum_{k=1}^{(K_l)_j^{[n]}} d_j^{[n]}(k), \quad n = 1, \ldots, N. \tag{11}$$

Total delay for the cars arriving in the entire time horizon $[0, T]$, d_t, is given by

$$d_t = \sum_{n=1}^{N} d_t^{[n]}. \tag{12}$$

In addition, other outputs from the nth subsimulation are obtained that the next subsimulation $(n+1)$ needs. They are the vector of the moments when the last car leaves each server, and the matrix of the queue length evolution for each server. Vector of the moments when the last car leaves corresponding servers in period n, $\mathbf{T}_l^{[n]}$, is given by

$$\mathbf{T}_l^{[n]} = \left((T_l)_j^{[n]}\right)_{N_T \times 1} = \left((T_l)_1^{[n]}, (T_l)_2^{[n]}, \ldots, (T_l)_{N_T}^{[n]}\right)^T, \quad n = 1, \ldots, N, \tag{13}$$

where for tollbooth $i \notin B_n^{\mathrm{IN}}$, we have $(T_l)_i^{[n]} = 0$. Matrix of the queue length evolution for all tollbooths in period n, $\mathbf{q}^{[n]}$, is given by

$$\mathbf{q}^{[n]} = \left(q_j^{[n]}\left(t^{[n]}\right)\right)_{N_T \times (T_l)^{[n]}} = \begin{bmatrix} q_1^{[n]}(1) & q_1^{[n]}(2) & \cdots & q_1^{[n]}\left((T_l)^{[n]}\right) \\ q_2^{[n]}(1) & q_2^{[n]}(2) & \cdots & q_2^{[n]}\left((T_l)^{[n]}\right) \\ \vdots & \vdots & \ddots & \vdots \\ q_{N_T}^{[n]}(1) & q_{N_T}^{[n]}(2) & \ldots & q_{N_T}^{[n]}\left((T_l)^{[n]}\right) \end{bmatrix}, \quad n = 1, \ldots, N, \tag{14}$$

where $(T_l)^{[n]}$ denotes the moment when the last car leaves the toll plaza in the nth subsimulation and $(T_l)^{[n]} = \max_{j \in B}(T_l)_j^{[n]}$. If the tollbooth $i \in B_n^{\mathrm{IN}}$ has $(T_l)_i^{[n]} < (T_l)^{[n]}$, then we have $q_i^{[n]}(t^{[n]}) = 0$, $t^{[n]} \in ((T_l)_i^{[n]}, (T_l)^{[n]}]$. In addition, for tollbooth $i \notin B_n^{\mathrm{IN}}$, we have $q_i^{[n]}(t^{[n]}) = 0$, $t^{[n]} \in [0, (T_l)^{[n]}]$.

Matrix of the queue length evolution for all tollbooths in the entire horizon, $\mathbf{q}$, is given by

$$\mathbf{q} = \left[\widehat{q}^{[1]}, \ldots, \widehat{q}^{[n]}, \ldots, \widehat{q}^{[N-1]}, \mathbf{q}^{[N]}\right], \tag{15}$$

where

$$\widehat{q}^{[n]} = \left(q_j^{[n]}\left(t^{[n]}\right)\right)_{N_T \times \delta T} = \begin{bmatrix} q_1^{[n]}(1) & q_1^{[n]}(2) & \cdots & q_1^{[n]}(\delta T) \\ q_2^{[n]}(1) & q_2^{[n]}(2) & \cdots & q_2^{[n]}(\delta T) \\ \vdots & \vdots & \ddots & \vdots \\ q_{N_T}^{[n]}(1) & q_{N_T}^{[n]}(2) & \ldots & q_{N_T}^{[n]}(\delta T) \end{bmatrix}, \quad n = 1, \ldots, N-1. \tag{16}$$

In conclusion, the main outputs from the subsimulation can be described as black-box functions, presented as follows:

$$\begin{aligned} d_t^{[n]} &= d_t^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}, \mathbf{q}^{[n-1]}; A^{[n]}\left(t^{[n]}; \lambda_n \delta T\right), P_r\right), \quad n = 1, \ldots, N, \\ \mathbf{q}^{[n]} &= \mathbf{q}^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}, \mathbf{q}^{[n-1]}; A^{[n]}\left(t^{[n]}; \lambda_n \delta T\right), P_r\right), \quad n = 1, \ldots, N, \\ \mathbf{T}_l^{[n]} &= \mathbf{T}_l^{[n]}\left(B_n^{\mathrm{POF}}, B_n^{\mathrm{PAE}}, \mathbf{q}^{[n-1]}; A^{[n]}\left(t^{[n]}; \lambda_n \delta T\right), P_r\right), \quad n = 1, \ldots, N, \end{aligned} \tag{17}$$

where $\mathbf{q}^{[0]} = 0$ and $\mathbf{T}_l^{[0]} = 0$ are defined.

The total delay for car platoons from the simulation is presented as follows:

$$d_t = d_t\left(B_1^{\mathrm{POF}}, B_1^{\mathrm{PAE}}, \ldots, B_N^{\mathrm{POF}}, B_N^{\mathrm{PAE}}; A^{[1]}\left(t^{[1]}; \lambda_1 \delta T\right), \ldots, A^{[N]}\left(t^{[N]}; \lambda_N \delta T\right), P_r\right). \tag{18}$$

2.2.3. Simulation Procedure. The nth subsimulation ($n = 1, \ldots, N$) is performed as follows (Figures 4 and 5; Table 1):

[SIM-n]

Step 0: initialization for environment. Give the sets of POF and PAE tollbooths in service, B_n^{POF} and B_n^{PAE}, respectively, the free-parking time, t_f, the no-extra-pay time after POF cars paying inside the car park, t_e, and the travel time from CS_1 to tollbooth j when there is no queue in front of tollbooth j, τ_j. Input $\mathbf{q}^{[n-1]}$ from the last subsimulation.

Step 1: arrival information generation. Generate the arrival time $t_K^{[n]}$ for each car $(K)^{[n]}$ and the total number of cars arriving, $K_l^{[n]}$ in the period $t^{[n]} \in [0, \delta T]$ at CS_1 via a stochastic process with arrival intensity λ_n, parking duration $c_K^{[n]}$ via empirical distribution in Figure 4(a), and the travel time from car park to CS_1, $(t_d)_K^{[n]}$ via empirical distribution in Figure 4(b). Set $(K)^{[n]} = 1$.

Step 2: car type generation. If $c_K^{[n]} < t_f$, the car is labelled as PAE one. If $c_K^{[n]} \geq t_f$, the type of car $(K)^{[n]}$ is determined using Bernoulli distribution according to the surveyed preference of drivers to the POF system;

Step 3: tollbooth selection. PAE car chooses the PAE tollbooth with the shortest estimated waiting time at time $t_K^{[n]}$. That is, find a PAE tollbooth label $j \in I_1 = \{i \in B_n^{\mathrm{PAE}} | \omega_i^{e[n]}(K) \leq \omega_l^{e[n]}(K), \forall l \in B_n^{\mathrm{PAE}}\}$ with the probability $1/|I_1|$. POF car could use either the PAE or POF tollbooth and will choose the tollbooth with the shortest estimated waiting time. That is, use equation (1) to find a tollbooth label $j \in I_2 = \left\{i \in B_n^{\mathrm{IN}} | \omega_i^{e[n]}(K) \leq \omega_l^{e[n]}(K), \forall l \in B_n^{\mathrm{IN}}\right\}$ with the probability $1/|I_2|$.

Step 4: waiting time computation. Once the car $(K)^{[n]}$ becomes the car k chooses the jth tollbooth, i.e., $(k)_j^{[n]}$, we can compute the waiting time $\omega_j^{B[n]}(k)$ using equation (A.8) and the waiting time $\omega_j^{Q[n]}(k)$ using equation (A.11).

Step 5: service time generation. Generate the time to be served, $s_j^{[n]}(k)$, at tollbooth j. The service time depends on the type of tollbooths, the type of cars, and the computed waiting time $\omega_j^{[n]}(k)$ in equation (9), as presented in Table 1.

Step 6: move to next arriving car. Obtain the delay of the car $(k)_j^{[n]}$, $d_j^{[n]}(k)$, in equation (10). And $(K)^{[n]} = (K)^{[n]} + 1$.

Step 7: simulation termination criterion. If $(K)^{[n]} > K_l^{[n]}$, simulation terminates and give the outputs $d_t^{[n]}$, $\mathbf{q}^{[n]}$, and $\mathbf{T}_l^{[n]}$; return to Step 2 otherwise.

2.2.4. Validation for Simulation Model

(1) Studied Car Park. Terminal 3A of Chongqing Jiangbei International Airport, China came into operation in August 2017. According to the planning document, there will be approximately 8000 and 12300 passengers arriving at the terminal during peak hours in years 2020 and 2040, respectively. In future, a large number of cars will come to pick up passengers in the car park especially when subway is out of service after 11:00 p.m. To accommodate the considerably heavy flows leaving the car park in the future, the toll plaza for the car park exit is designed with 16 tollbooths: 13 for cars, 2 for coaches, and 1 for motorcycles, as Figure 6(a) presents. However, there is no need for operators to keep all 13 tollbooths in service since the flows from the park are not high so far. Usually, 2 PAE and 2 POF tollbooths are in service.

Video observation was adopted to capture the arrival and service features of car platoons at the toll plaza for this car park exit. Travel times from car park to CS_1 are estimated via the distances between all parking spaces and CS_1. However, the car park did not provide the data of parking durations for us due to the lack of detectors for parking spaces. Fortunately, we investigated the operation of P7 car park for Shanghai Hongqiao International Airport in 2017, and large quantities of parking duration samples were manually obtained by videos at the parking spaces. The data were used in this study and assumed to be able to reflect the parking

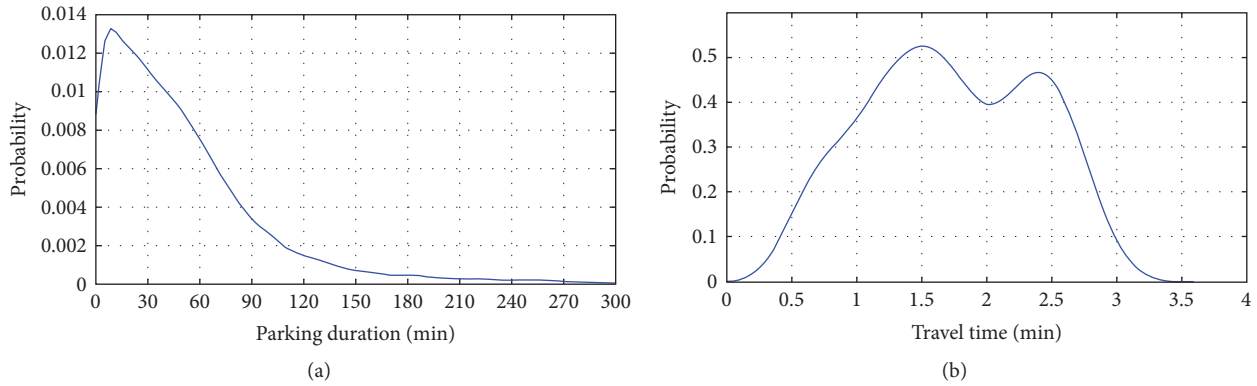


Figure 4: Probability densities of (a) parking durations and (b) travel times from car park to CS_1.

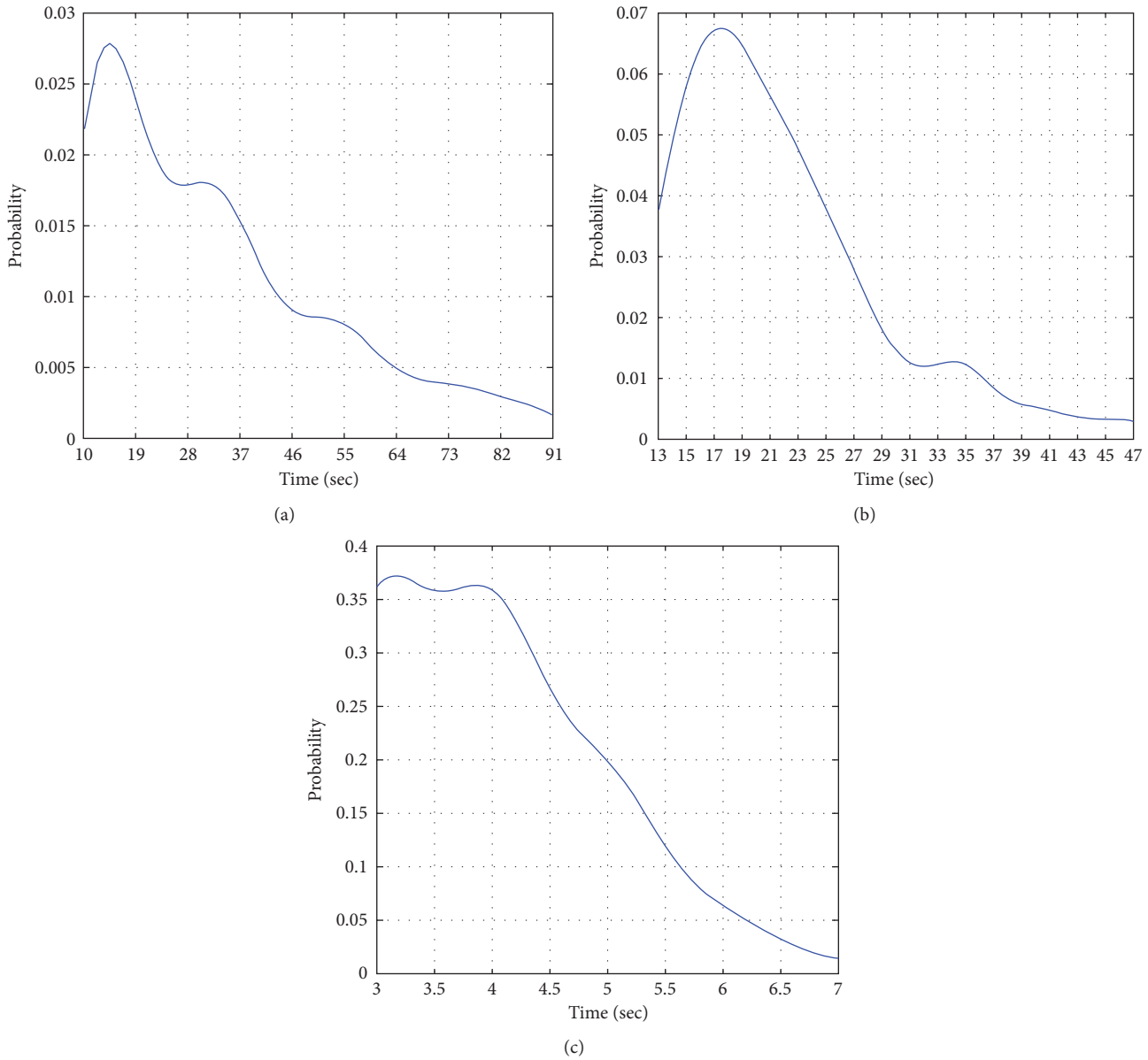
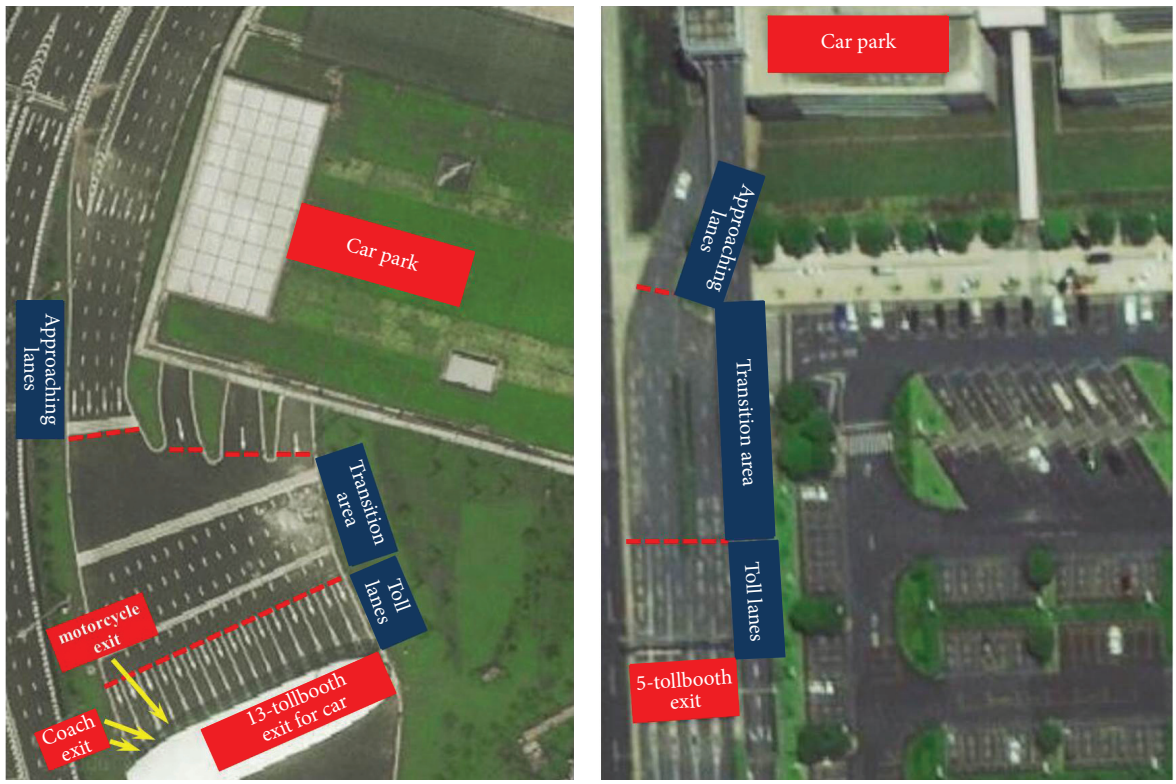


Figure 5: Probability densities of service times at toll plaza. (a) Probability density of car plate automatically verified plus QR code paid time. (b) Probability density of car plate automatically verified plus cash paid time. (c) Probability density of car plate automatically time.

Table 1: Service times in all possible scenarios.

Parking duration	Car type	Condition of free-parking (for PAE car) or no-extra-pay (for POF car)	Tollbooth type	Service time
$c_K^{[n]} < t_f$	PAE	$c_K^{[n]} + (t_d)_K^{[n]} + \tau_j + w_j^{[n]}(k) > t_f$	PAE	Figure 5(b)
		$c_K^{[n]} + (t_d)_K^{[n]} + \tau_j + w_j^{[n]}(k) \le t_f$		Figure 5(c)
$c_K^{[n]} \ge t_f$	PAE	—	PAE	Figure 5(b)
	POF	$(t_d)_K^{[n]} + \tau_j + w_j^{[n]}(k) > t_e$	PAE	Figure 5(b)
			POF	Figure 5(a)
		$(t_d)_K^{[n]} + \tau_j + w_j^{[n]}(k) \le t_e$	PAE or POF	Figure 5(c)



(a) (b)

Figure 6: Two car parks for data sources. (a) Car park for Terminal 3A of Chongqing Jiangbei International Airport. (b) P7 car park for Shanghai Hongqiao International Airport.

condition in the car park for Terminal 3A of Chongqing Jiangbei International Airport.

In addition, the P7 car park for Shanghai Hongqiao International Airport has 4 PAE tollbooths and 1 staff only one in 2015, as Figure 6(b) illustrates. The billing rules between two parks are the same. Hence, the service times for PAE tollbooths in this park can be supplemented to the database of this study.

In sum, to develop and calibrate the simulation model in our study, the arrival profiles were collected from the car park exit for Terminal 3A of Chongqing Jiangbei International Airport. The parking durations were obtained in the P7 car park for Shanghai Hongqiao International Airport. The data for service times in POF and PAE tollbooths were observed in both two car park exits. The data are presented in Appendix B.

We use queue length to validate the simulation model via hypothesis test. The queue length evolutions are recorded at the same time we collected cars' arrival data by video at 23:00–01:00 on fourteen consecutive days in November 2018, as presented in the previous section. During the observation, 1 PAE and 3 POF tollbooths are in service where the sets for PAE and POF tollbooths in service are $\{10\}$ and $\{1, 2, 5\}$, respectively. The data of arrival and queues are divided by one hour, and then, 28 samples are captured.

(2) Validation. We validate the simulation model via hypothesis test. The queue length evolutions are recorded at the same time we collected cars' arrival data by video at 23:00–01:00 on fourteen consecutive days in November 2018, as presented in Appendix. During the observation, 1 PAE and 3 POF tollbooths are in service where the sets for PAE and

POF tollbooths in service are {10} and {1, 2, 5}, respectively. The data of arrival and queues are divided by one hour, and then, 28 samples are captured.

In each sample, the arrival profile is fixed as presented in Figure 7. We can adopt queue length at per second in the evolution or the average queue length to compare the results from 1000 parallel simulations and the one from observation. However, it is time-consuming and unnecessary for the validation to compare the queue length at every second in the evolution (the case of 432 cars arriving is shown in Figure 8). Hence, we use the average queue length (AQ) as the key measure for validation in 28 samples.

The hypothesis testing for each sample (or experiment) σ ($\sigma = 1, \ldots, 28$) is presented as $H_{0\sigma}$: $\mu_{1\sigma} = \mu_{2\sigma}$, where $\mu_{1\sigma}$ and $\mu_{2\sigma}$ denote the AQ from observation and the mean of AQs from 1000 parallel simulations, respectively. The boxplots for simulation results and corresponding observed results are shown in Figure 9. Given the 0.95 confidence level, the significance, lower bound (LB), and upper bound (UB) of confidence interval (CI) for each sample are listed in Table 2.

According to Table 2, only four hypotheses $H_{0\sigma}$ ($\sigma = 3, 5, 7, 12$) are rejected while others are accepted. The simulation model is valid for the description of the queueing system at car park exit. Admittedly, the average queue length as the key measure is sensitive since it only can be integer. It is nevertheless easy to observe in experiments and is an acceptable alternative for validation.

3. Solution Algorithm

To solve the simulation-based integer programming model for discrete-time DTAP, SIMIP-DTAP, we propose a solution algorithm with two steps. The first step is to decompose the discrete-time DTAP into subproblems by periods where the arrival intensity and allocation scheme are static. The second one is to solve each period-based subproblem. The two steps are detailed in the next two sections, respectively.

3.1. Rolling Horizon Approach to Decomposition of Discrete-Time DTAP. This section presents a rolling horizon (RH) approach to decompose the discrete-time DTAP, handling scheme dynamics and output stochasticity from simulation. The complexity of discrete-time DTAP is mainly given by the number of integer variables, which increases with the number of periods and subsimulations [29]. In addition, it is difficult to obtain an optimal or even a near-optimal solution, since a long time horizon needs to be considered [30]. Furthermore, car park operators usually estimate the exiting intensity of cars in a period before the beginning of the period. Thus, they divide the whole tollbooth operation problem into several stages, where each stage is only for exiting cars in the coming period. Only the plan for the near future is updated and executed. Based on this practice, an RH approach is utilized to decompose our problem, which is helpful in decreasing the computation time. Note that considering the required computation time, the model in a real-world tollbooth operation needs to be executed a few minutes before the onset of each roll period.

The RH approach is a decomposition method where the allocation problem at a stage is developed conditionally on the optimal allocation scheme from the last stage. Therefore, the discrete-time DTAP is decomposed into static subproblems on the stage-by-stage basis. According to Assumption 1, this decomposition coincides with discretization for the simulation on the period-by-period basis. The RH approach is presented in Figure 10, where only three stages are shown for instance. The parameter $(T_l)^{[n]}$ is the stage horizon which is stochastic, i.e., the time elapsed from when the period n starts to when the last car leaves the toll plaza in the nth subsimulation. The parameter r is the roll period for each stage and $r = \delta T$; The parameter T_{n-1} denotes the start time for the stage s ($s = n$). Hereinafter, we use s and n interchangeably.

The computational steps of RH approach to solving discrete-time DTAP are presented as follows.

[RH-DTAP]

Step 0: initialization. Give the stage $s = 1$, subsimulation $n = 1$, time $t = 0$, optimal allocation scheme at stage $(s-1)$, α^*_{n-1}, and optimal output from stage $(s-1)$, $\mathbf{q}^{[n-1]*}$.

Step 1: solve the **n**th static tollbooth allocation problem (STAP-**n**) based on inputs α^*_{n-1} and $\mathbf{q}^{[n-1]*}$, and nth subsimulation at stage s. Obtain the optimal allocation scheme α^*_n and output $\mathbf{q}^{[n]*}$.

Step 2: update the stage, subsimulation, and time, $s = s + 1$, $n = n + 1$, and $t = t + \delta T$.

Step 3: iteration termination criterion. If $s > N$, iteration terminates and give the optimal allocation scheme sequence $\alpha^* = (\alpha^*_1, \ldots, \alpha^*_N)$; return to Step 1 otherwise.

In RH, the subproblem STAP-n ($n = 1, \ldots, N$) is also described as a SIMIP model, presented as follows:

[SIMIP-STAP-n]

$$\boldsymbol{\alpha}^*_n \in \arg\min_{\alpha_n} p^{[n]}_{\mathrm{cf}}(\boldsymbol{\alpha}_n) = \mathbb{E}\left[(1-w)\beta d^{[n]}_t\left(\boldsymbol{\alpha}_n; q^{[n-1]*}\right) + w c^{[n]}_p(\boldsymbol{\alpha}_n)\right], \tag{19}$$

$$\text{subject to } \alpha^{[n]}_j = -1, 0, 1, \quad j \in B, \tag{20}$$

$$\prod_{j \in B} \alpha^{[n]}_j = 0, \tag{21}$$

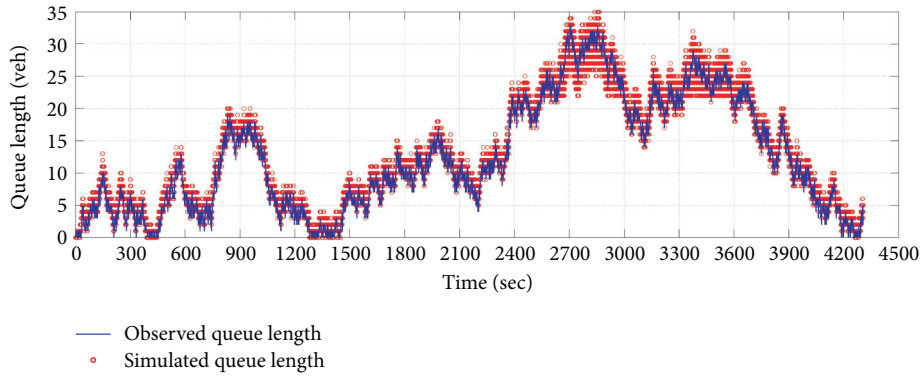


Figure 7: Comparison of queue lengths at every second in the evolution between 1000 simulations and observation in one case.

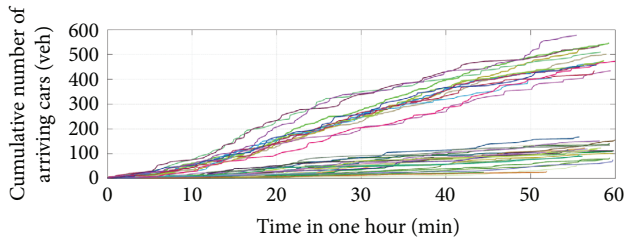


Figure 8: Observed arrival profiles for low-flow case (the lines with different colors represent the different samples of arrivals we observed).

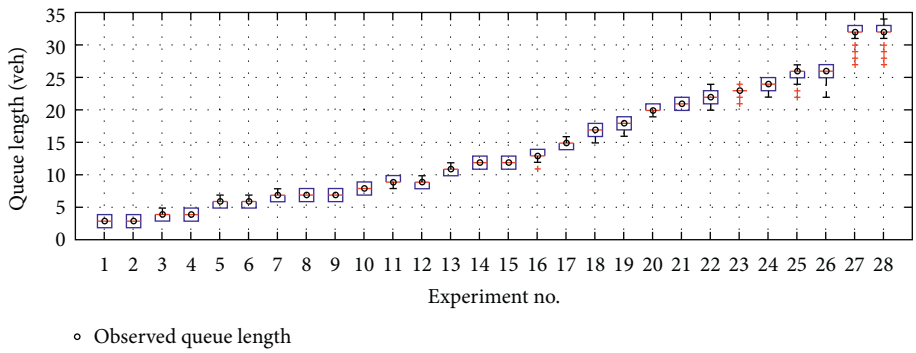


Figure 9: Comparison of queue lengths at every second in the evolution between simulations and observation in one case.

Table 2: Significances and 0.95 confidence intervals for validation samples.

Experiment no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$H_{0\sigma}$	0	0	1	0	1	0	1	0	0	0	0	1	0	0
Significance	0.67	0.43	0.00	0.90	0.00	0.22	0.02	0.28	0.22	0.69	0.42	0.00	0.16	0.93
LB of CI	2.94	2.97	3.84	3.95	5.78	5.93	6.90	6.93	6.98	7.96	8.97	8.71	10.92	11.96
UB of CI	3.04	3.08	3.93	4.04	5.85	6.02	6.99	7.02	7.07	8.05	9.06	8.78	11.01	12.05
Experiment no.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
$H_{0\sigma}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Significance	0.97	0.24	0.47	0.54	0.94	0.62	0.20	0.63	0.36	0.51	0.06	0.09	0.79	0.30
LB of CI	11.95	12.98	14.94	16.94	17.95	19.97	20.91	21.97	22.94	23.94	25.91	25.91	31.93	31.90
UB of CI	12.05	13.07	15.03	17.03	18.06	20.05	21.02	22.06	23.02	24.03	26.00	26.01	32.05	32.03

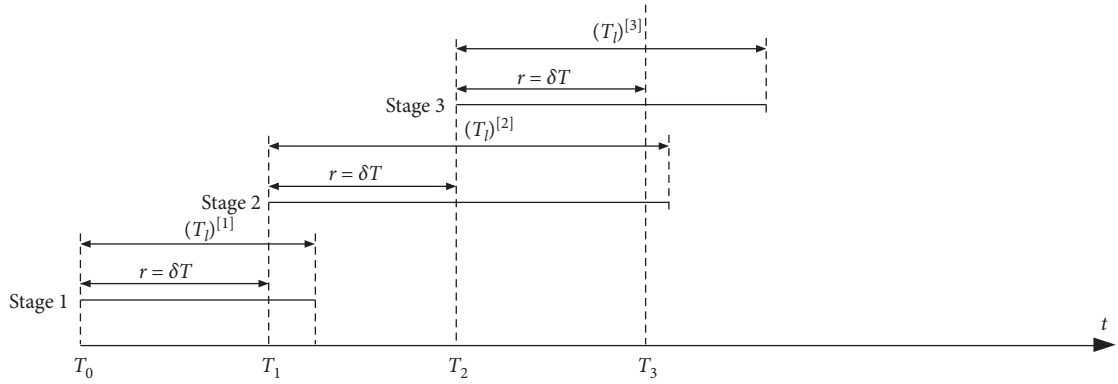


Figure 10: The rolling horizon approach.

$$\mathbb{E}\left[q_j^{[n-1]*}(\delta T)\right]\left(\alpha_j^{[n]} - \alpha_j^{[n-1]*}\right) = 0, \quad j \in B, \tag{22}$$

where $\mathbf{q}^{[n-1]*}$ and $\alpha_j^{[n-1]*}$ are the optimal results from [SIMIP-STAP-$(n-1)$] as well as the constants in [STAP-n]. We define that $\mathbf{q}^{[0]*} = 0$ and $\alpha_j^{[0]*} = 0$. $c_p^{[n]}(\alpha_n)$ is obtained using the following equation:

$$c_p^{[n]}(\boldsymbol{\alpha}_n) = c_m\left|B_n^{\mathrm{PAE}}\right|\max\left(\delta T, (T_l)^{[n]}\right) + c_e\left(\left|B_n^{\mathrm{PAE}}\right| + \left|B_n^{\mathrm{POF}}\right|\right)\max\left(\delta T, (T_l)^{[n]}\right), \quad n = 1, \ldots, N. \tag{23}$$

Note that $c_p^{[n]}(\alpha_n)$ here is different from but close to that in equation (2). The solution of the subproblem STAP-n is explicated in the next section.

3.2. Kriging Metamodel Algorithm for Solving Period-Based Subproblem. Due to the stochasticity of the outputs from the simulation, the SIMIP-STAP-n is computationally difficult to yield the exact solution. An approximation to the solution can be derived by existing algorithms. In this paper, the Kriging metamodel is used as a surrogate. The metamodel method is favored in many literature studies due to the determinacy of the response, thus making it computationally efficient. The metamodel is an approximation that is inexpensive to compute to describe the original model with the complicated and computationally expensive response.

3.2.1. Framework. The integer (global) optimization problem can be written in the following form:

$$\text{minimize} \quad G(\mathbf{x}), \tag{24}$$

$$\text{subject to } x_j^l \le x_j \le x_j^u, \quad \forall j = 1, \ldots, J, \tag{25}$$

where $G(\mathbf{x})$ denotes the objective function defined in equation (19). x_j^l and x_j^u denote the lower and upper bounds on the variable x_j, respectively. The optimization model defined in (24) is replaced by the sum of a constant value and a Gaussian random error term as follows:

$$G(\mathbf{x}) = \rho + Z(\mathbf{x}), \tag{26}$$

where ρ is the mean value of $G(\mathbf{x})$ and$Z(\mathbf{x})$ is a stationary Gaussian random process with mean zero, variance σ^2, and nonzero covariance. The covariance is expressed as

$$\mathrm{cov}\left[Z(\mathbf{x}^m), Z\left(\mathbf{x}^h\right)\right] = \sigma^2\psi\left(\mathbf{x}^m, \mathbf{x}^h\right), \tag{27}$$

where $\mathrm{cov}[Z(\mathbf{x}^m), Z(\mathbf{x}^h)]$ is the covariance between two sample points $\mathbf{x}^m$ and $\mathbf{x}^h$. $\psi(\mathbf{x}^m, \mathbf{x}^h)$ is the Kriging function, or Gaussian exponential correlation function, equivalently. The Kriging function is defined as follows:

$$\psi\left(\mathbf{x}^m, \mathbf{x}^h\right) = \exp\left[-\sum_{j=1}^{J}\lambda_j\left(x_j^m - x_j^h\right)^2\right], \tag{28}$$

where $\lambda = [\lambda_1, \ldots, \lambda_j, \ldots, \lambda_J]$ denotes the vector of unknown correlation parameters and x_j^m and x_j^h denote the jth variable in $\mathbf{x}^m$ and $\mathbf{x}^h$, respectively. J denotes the dimension of $\mathbf{x}$ and here $J = |B|$ in this paper. Note that the parameter λ_j can be regarded as measuring the importance of the variables x_j^m and x_j^h. The larger of λ_j, the greater the Euclidean distance between x_j^m and x_j^h, hence the lower the correlation between them.

Let n_s denote the number of the known sample points for variable $\mathbf{x}$. The predictive estimate $\widetilde{G}(\mathbf{x})$ of the response function $G(\mathbf{x})$ at unknown point $\mathbf{x}$ is given by

$$\widetilde{G}(\mathbf{x}) = \widetilde{\rho} + \widetilde{R}^T(x)\widetilde{\Psi}^{-1}(\mathbf{G} - 1\widetilde{\rho}), \tag{29}$$

where $\mathbf{G} = (G(\mathbf{x}^1), G(\mathbf{x}^2), \ldots, G(\mathbf{x}^{n_s}))^T$, 1 is the $n_s \times 1$ unit column, and $\widetilde{\Psi}$ is the estimated correlation matrix and is diagonally symmetric where

$$\widetilde{\Psi} = \begin{bmatrix} \widetilde{\psi}\left(\mathbf{x}^1, \mathbf{x}^1\right) & \widetilde{\psi}\left(\mathbf{x}^1, \mathbf{x}^2\right) & \cdots & \widetilde{\psi}\left(\mathbf{x}^1, \mathbf{x}^{n_s}\right) \\ \widetilde{\psi}\left(\mathbf{x}^2, \mathbf{x}^1\right) & \widetilde{\psi}\left(\mathbf{x}^2, \mathbf{x}^2\right) & \cdots & \widetilde{\psi}\left(\mathbf{x}^2, \mathbf{x}^{n_s}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\psi}\left(\mathbf{x}^{n_s}, \mathbf{x}^1\right) & \widetilde{\psi}\left(\mathbf{x}^{n_s}, \mathbf{x}^2\right) & \cdots & \widetilde{\psi}\left(\mathbf{x}^{n_s}, \mathbf{x}^{n_s}\right) \end{bmatrix} \in \mathscr{R}^{n_s \times n_s}, \tag{30}$$

where

$$\widetilde{\psi}\left(\mathbf{x}^m, \mathbf{x}^h\right) = \exp\left[-\sum_{j=1}^{J} \widetilde{\lambda}_j \left(x_j^m - x_j^h\right)^2\right]. \tag{31}$$

$\widetilde{R}(\mathbf{x})$ is the correlation vector between the unknown point $\mathbf{x}$ and sample points where

$$\widetilde{R}(\mathbf{x}) = \left(\widetilde{\psi}\left(\mathbf{x}, \mathbf{x}^1\right), \widetilde{\psi}\left(\mathbf{x}, \mathbf{x}^2\right), \ldots, \widetilde{\psi}\left(\mathbf{x}, \mathbf{x}^{n_s}\right)\right)^T. \tag{32}$$

Therefore, in order to yield the value of $\widetilde{G}(\mathbf{x})$, the estimators $\widetilde{\rho}$ and $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_K, \widetilde{\lambda}_2, \widetilde{\lambda}_K$ need to be obtained, given by

$$\widetilde{\rho} = \left(1^T \widetilde{\Psi}^{-1} 1\right)^{-1} \left(1^T \widetilde{\Psi}^{-1} \mathbf{G}\right), \tag{33}$$

$$\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J\right) = \underset{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J > 0}{\arg\max} -\frac{1}{2}\left\{n_s \ln\left[\frac{1}{n_s}(\mathbf{G} - 1\widetilde{\rho})^T \widetilde{\Psi}^{-1} (\mathbf{G} - 1\widetilde{\rho})\right] + \ln\left(\det \widetilde{\Psi}\right)\right\}, \tag{34}$$

where det is the determinant of matrix [35]. According to equation (33), $\widetilde{\rho}$ is obtained based on $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_J$ while $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_J$ are estimated via $\widetilde{\rho}$ in terms of equation (34). Hence, as claimed in Kleijnen [36], this estimation is a difficult mathematical problem, and the divide-and-conquer methodology [23] is used here for estimation. The algorithm to estimate the above parameters PE is presented as follows:

[PE]

Step 0: initialize the parameters of $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_K, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_K$, and obtain the $\widetilde{\psi}(\mathbf{x}^m, \mathbf{x}^h)$ defined in equation (31) and $\widetilde{\Psi}$ defined in equation (30)

Step 1: compute the value of $\widetilde{\rho}$ in terms of equation (33)

Step 2: substitute $\widetilde{\Psi}$ obtained in Step 0 and $\widetilde{\rho}$ obtained in Step 1 into equation (34), solve equation (34), and update $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_K, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_K$

Step 3: update $\widetilde{\psi}(\mathbf{x}^m, \mathbf{x}^h)$ and $\widetilde{\Psi}$ with the updated $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_J$, and substitute this updated into equation (33)

Step 4: if the convergence criteria have not been satisfied, then return to Step 1; otherwise, stop and output the estimators of $\widetilde{\rho}$ and $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_J, \widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\lambda}_J$

If the problem has other constraints in addition to the bound constraint (25), it can be formulated as the following form:

$$\text{minimize} \quad G(\mathbf{x}), \tag{35}$$

subject to

$$H_j(\mathbf{x}) \le 0, \quad \forall j = 1, \ldots, J, \tag{36}$$

$$x_j^l \le x_j \le x_j^u, \quad \forall j = 1, \ldots, J. \tag{37}$$

In order to use Kriging metamodel, the nonpositive constraint (36) is added as the penalty term to the objective function value at every candidate point $\mathbf{x}^m$, [37, 38] which is rewritten as

$$G_P\left(\mathbf{x}^m\right) = \begin{cases} G_{\max} + w_p z\left(\mathbf{x}^m\right), & \text{if } z\left(\mathbf{x}^m\right) > 0, \\ G\left(\mathbf{x}^m\right), & \text{otherwise,} \end{cases} \tag{38}$$

where $G_P(\cdot)$ denotes the penalty augmented objective function (PAOF). $G_{\max}$ represents the worst feasible objective function value so far. $z(\mathbf{x}^{[k]}) = \sum_{j=1}^{m} \max\left\{0, H_j(\mathbf{x}^{[k]})\right\}^2$ is the constraint violation function, and w_p denotes the penalty factor which is positive and adjusted. This definition guarantees that the penalty augmented function values of the infeasible points are larger than the worst feasible objective function value.

3.2.2. Kriging Metamodel Algorithm. Based on the principle interpreted in Section 3.2.1, SIMIP-STAP-n can be transformed into an integer (global) optimization problem through approximating the outputs of simulation by metamodels.

The Kriging metamodel algorithm (KMA) starts with an initial experimental design. The initial experimental design is to produce n_s points which are repeatedly generated via a symmetric Latin hypercube design [39]. In Latin hypercube design, each variable of $\mathbf{x}$ is stratified into n_s equal intervals, and in each subinterval, a sample point is randomly generated.

Then, the optimization works iteratively and, in each iteration, new points for calculating the next expensive function are added to the set of already sampled points. A candidate point-based approach is applied to determine the new sample site. Candidates are generated by (a) uniformly

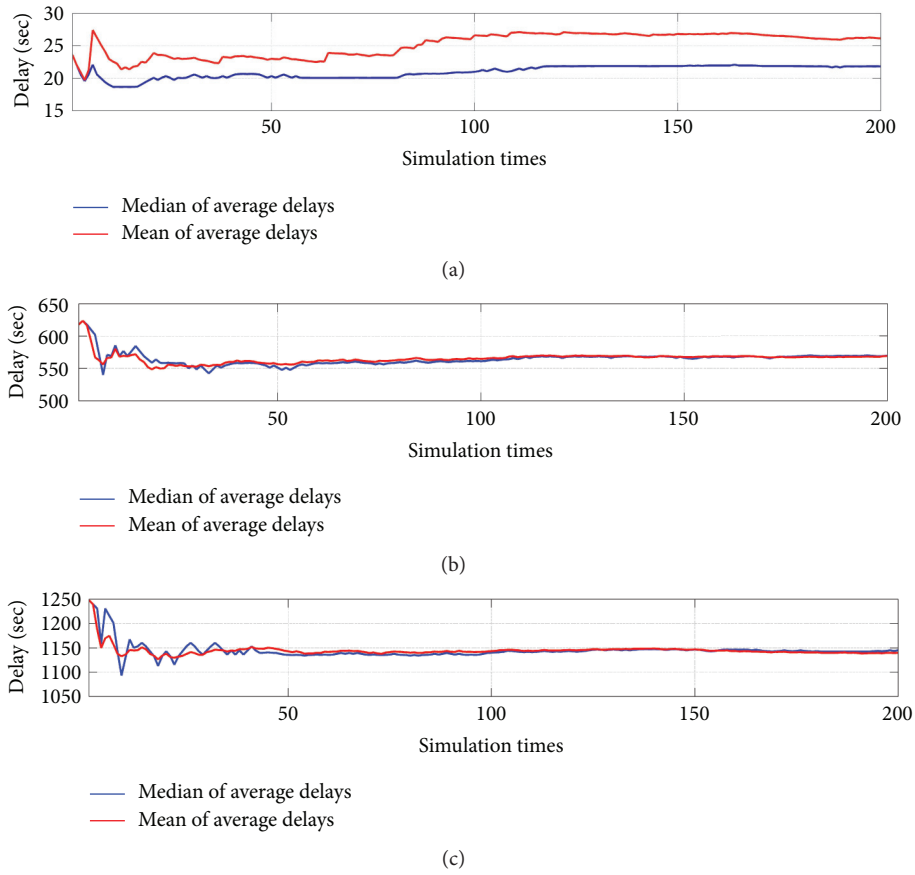


FIGURE 11: Median and mean of the average delays with different simulation times (a) when 1000 cars arrive, (b) when 2000 cars arrive, and (c) when 3000 cars arrive.

selecting points and (b) perturbing the best point found so far. The rth candidate point in the iteration is denoted by χ^r, $r = 1, \ldots, R$.

In the given set of randomly generated candidate points, the parameters of the Kriging metamodel are computed by the algorithm PE and the $\tilde{G}(\chi^r)$ is then used to approximate the $G(\chi^r)$ at every candidate point χ^r. A good candidate point for function evaluation ideally should have a low estimated function value (since the goal is to minimize it) and should be far away from previously evaluated points (since this promotes a more global search). Hence, the selected point has the best weighted score based on two criteria: (1) estimated function value obtained from the response surface model (response surface criterion V_R) and (2) minimum distance from previously evaluated points (distance criterion V_D) [40]. The weighted score is computed as

$$W(r) = w_R V_R(\boldsymbol{\chi}^r) + w_D V_D(\boldsymbol{\chi}^r), \tag{39}$$

where $w_R + w_D = 1$, $w_R \geq 0$ is the weight for the response surface criterion, and $w_D \geq 0$ is the weight for the distance criterion. The response surface criterion for every candidate point χ^r is given by

$$V_R(\boldsymbol{\chi}^r) = \begin{cases} \dfrac{\tilde{G}(\boldsymbol{\chi}^r) - G_{\min}}{G_{\max} - G_{\min}}, & \text{if } G_{\max} \neq G_{\min}, \\ 1, & \text{otherwise}, \end{cases} \tag{40}$$

where $G_{\max} = \max_{r=1,\ldots,R} \tilde{G}(\chi^r)$ and $s_{\min} = \min_{r=1,\ldots,R} \tilde{G}(\chi^r)$. The distance criterion for every candidate point χ^r is given by

$$V_D(\boldsymbol{\chi}^r) = \begin{cases} \dfrac{\Delta_{\max} - \Delta(\boldsymbol{\chi}^r)}{\Delta_{\max} - \Delta_{\min}}, & \text{if } \Delta_{\max} \neq \Delta_{\min}, \\ 1, & \text{otherwise}, \end{cases} \tag{41}$$

where $\Delta(\chi^r) = \min_{m=1,\ldots,n} \chi^r - \mathbf{x}^m$, $\Delta_{\max} = \max_{t=1,\ldots,T} \Delta(\chi^r)$, and $\Delta_{\min} = \min_{t=1,\ldots,T} \Delta(\chi^r)$.

The candidate point with the lowest weighted score is chosen as the new sample site in the next iteration until the maximum number of allowed function evaluations has been reached.

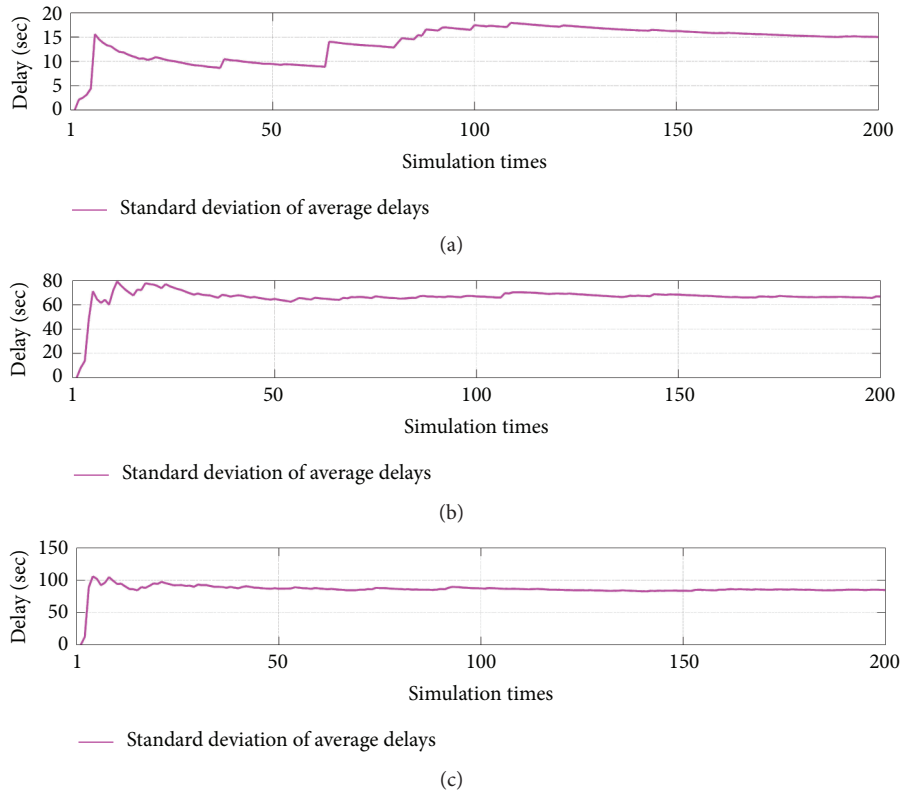


Figure 12: Standard deviation of the average delays with different simulation times (a) when 1000 cars arrive, (b) when 2000 cars arrive, and (c) when 3000 cars arrive.

4. Numerical Example

In this section, a numerical example is presented for the car park exit for Terminal 3A of Chongqing Jiangbei International Airport, China, as Figure 6(a) demonstrates. The toll plaza for the car park exit has 7-lane approach and 13 tollbooths for cars where $N_A = 7$, $N_R = 0$, and $N_L = 6$. We number the tollbooths from No. 1 to No. 13 consecutively (from the left- to right-hand side in the toll plaza). q_i^B for the toll lanes from No. 7 to No. 12 are 7, 6, 5, 5, 4, and 4 cars, respectively. According to our investigation, nearly 70% of drivers prefer the POF system.

The datasets for the random parameters in the simulation are established according to Appendix B. The algorithm is coded in MATLAB and run on a personal computer with Intel (R) Core (TM) i7-8700 3.20 GHz CPU.

4.1. Convergence Analysis for Simulation Model. More parallel evaluations for the simulation model will lead to more accurate distribution of results, but they are computationally expensive. Therefore, in the second part, we need to perform convergence analysis such that an appropriate number of evaluations are found which balances the accuracy and computation performance.

In the convergence analysis, average and total delays are chosen as the key measures under the arrival intensities of 1000, 2000, and 3000 cars in an hour. To accommodate the arrival intensity, 2 PAE and 5 POF tollbooths are assumed to be in service where the sets for PAE and POF tollbooths in service are $\{10, 11\}$ and $\{1, 2, 5, 6, 9\}$, respectively. The median, mean, and standard deviation of the key measures are obtained with different parallel evaluations for the simulations, as demonstrated in Figures 11–13.

As Figures 11–14 illustrate, after 150 parallel evaluations of simulations, the convergence patterns for the median, mean, and standard deviation of key measures are smooth. Thus, 150 parallel simulations are adequate for the subsequent performance analysis and optimization for the operation of toll plaza. Since the distributions of average and total delays represent either skewed or normal patterns, the median rather than mean is considered as the statistical quantity in the study when we develop SIMIP-DTAP.

4.2. Appropriate Operation Schemes for Toll Plaza. In this part, we find the solution for SIMIP-DTAP as the recommended tollbooth operation schemes, with different weights of operational costs. The horizon is 8 hours. The sequence of arrival intensities is $[1500, 2500, 3500, 4000, 4000, 3500, 3000, 2000]$ cars per hour. These intensities are higher than the current situation shown in Section 2.2 and are for future consideration. The manpower and electricity expense for each tollbooth are ¥ 20.00 and ¥ 1.00 per hour, respectively, where ¥ denotes the Chinese currency unit. The VOT for all people in one car is ¥ 50.00 per hour. In KMA for each stage of RH-DTAP, 200 function evaluations are allowed at the

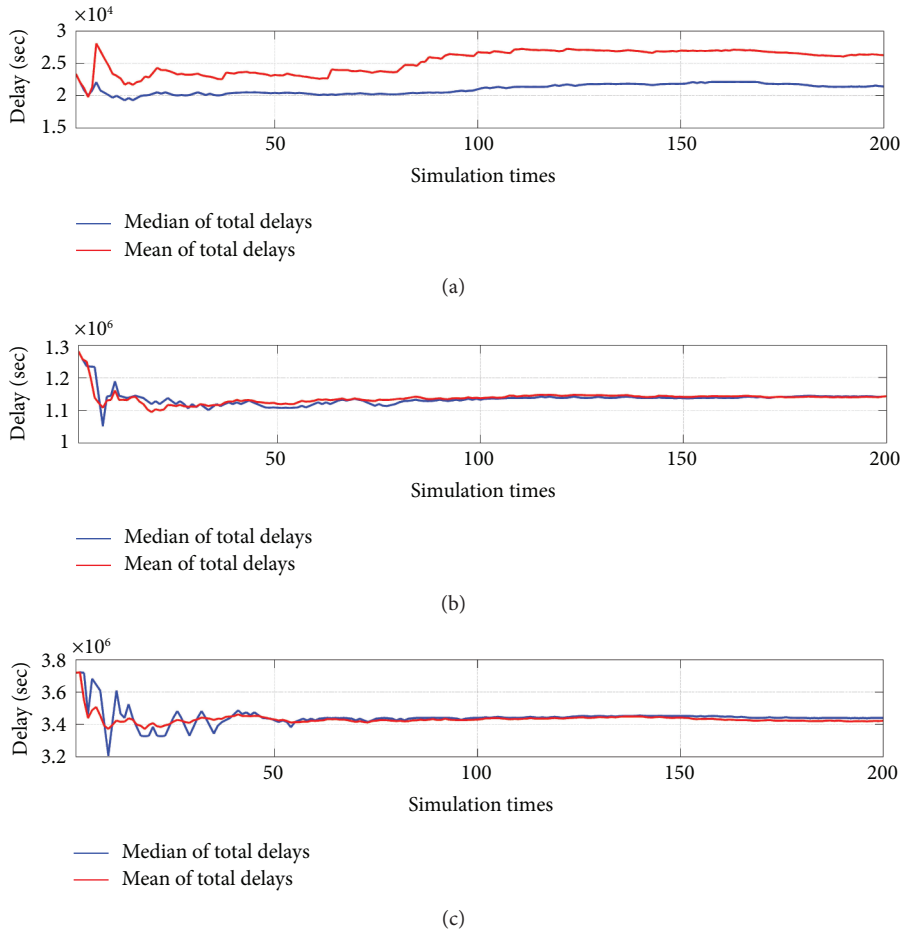


Figure 13: Median and mean of the total delays with different simulation times (a) when 1000 cars arrive, (b) when 2000 cars arrive, and (c) when 3000 cars arrive.

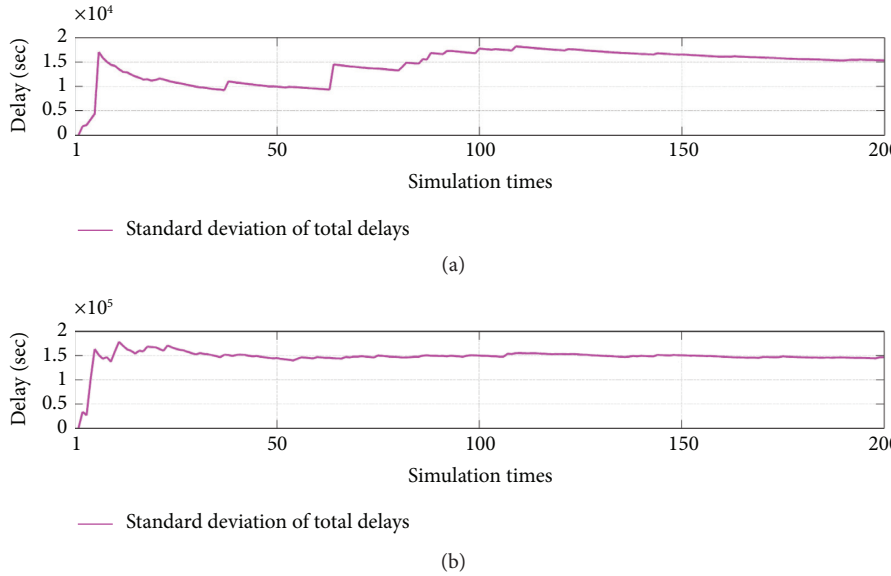
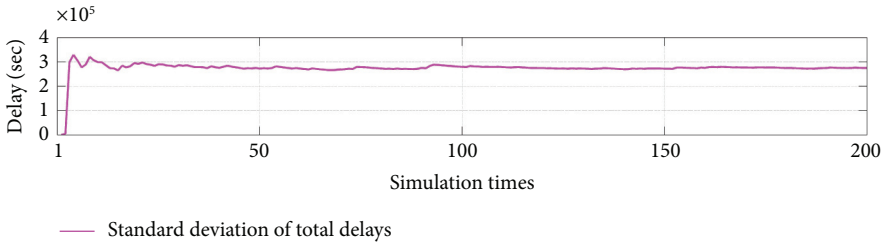
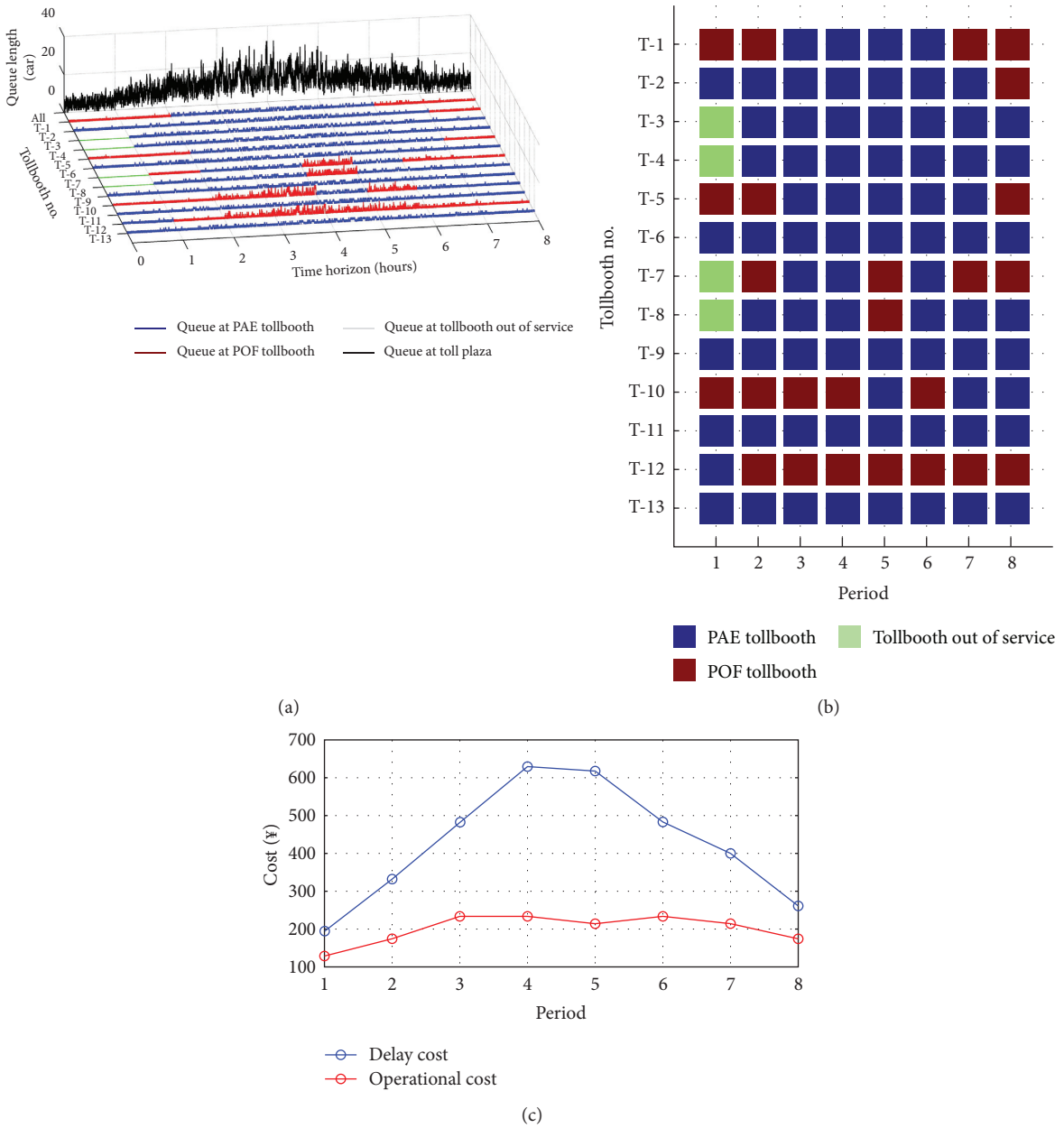


Figure 14: Continued.



(c)

FIGURE 14: Standard deviation of the total delays with different simulation times (a) when 1000 cars arrive, (b) when 2000 cars arrive, and (c) when 3000 cars arrive.



(a) (b) (c)

FIGURE 15: Recommended tollbooth operation scheme and its performance with weight of operational cost = 0.2. (a) Queue length evolution with time. (b) Best tollbooth operation scheme. (c) Costs in each period.

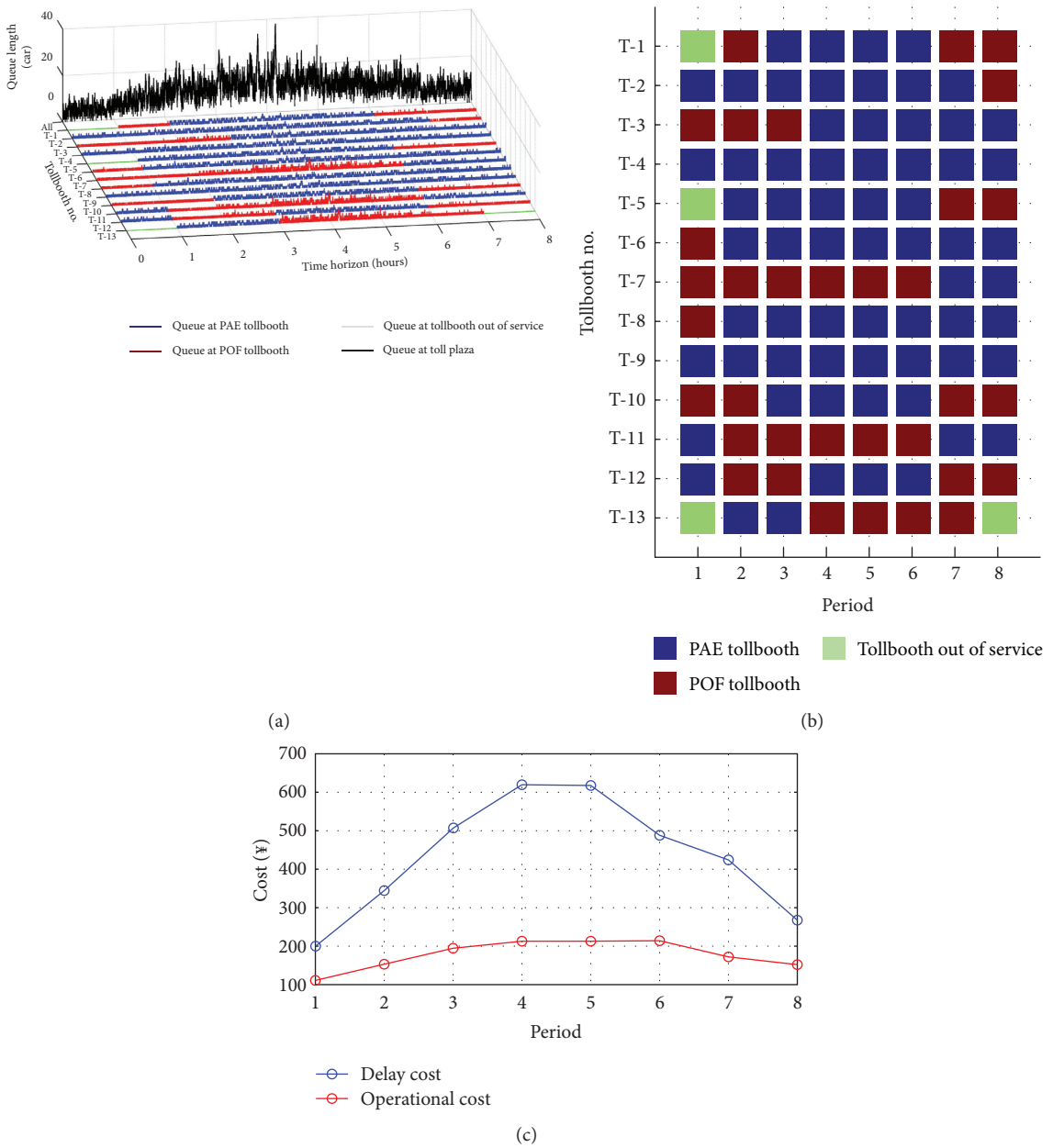


Figure 16: Recommended tollbooth operation scheme and its performance with weight of operational cost = 0.5. (a) Queue length evolution with time. (b) Best tollbooth operation scheme. (c) Costs in each period.

maximum, with 20 initial points and 20 new sample points in each iteration.

The recommended tollbooth operation scheme derived from the solution of discrete-time DTAP and its performance are investigated with different weights of operational cost, as Figures 15–17 presented. It is worth noting that

(i) PAE tollbooths are usually more welcome than POF ones in spite of the different weights between traffic efficiency and cost-effectiveness

(ii) As the arrival intensity increases, the proportion of PAE tollbooths to POF ones grows

These two statements are drawn because PAE tollbooths are more flexible that they can serve both PAE and POF cars. When all POF tollbooths have long queues, POF cars can choose PAE tollbooths as alternatives. The average service times for POF and PAE tollbooths are 4 and 17 seconds, respectively. The ratio of POF to PAE cars is 7 : 3 if we do not take into account the cars who supposed to be free-parking and naturally become PAE cars. Therefore, the estimated balanced ratio (EBR) of POF to PAE tollbooths is roughly obtained as 1 : 1.8. When the arrival intensity is low, the ratio of POF to PAE tollbooths is close to EBR. When the arrival intensity is high, however, the ratio of POF to PAE tollbooths is far higher than EBR. The toll plaza still needs more PAE tollbooths than EBR to satisfy the heavy demand even if car operators want to control operational cost.

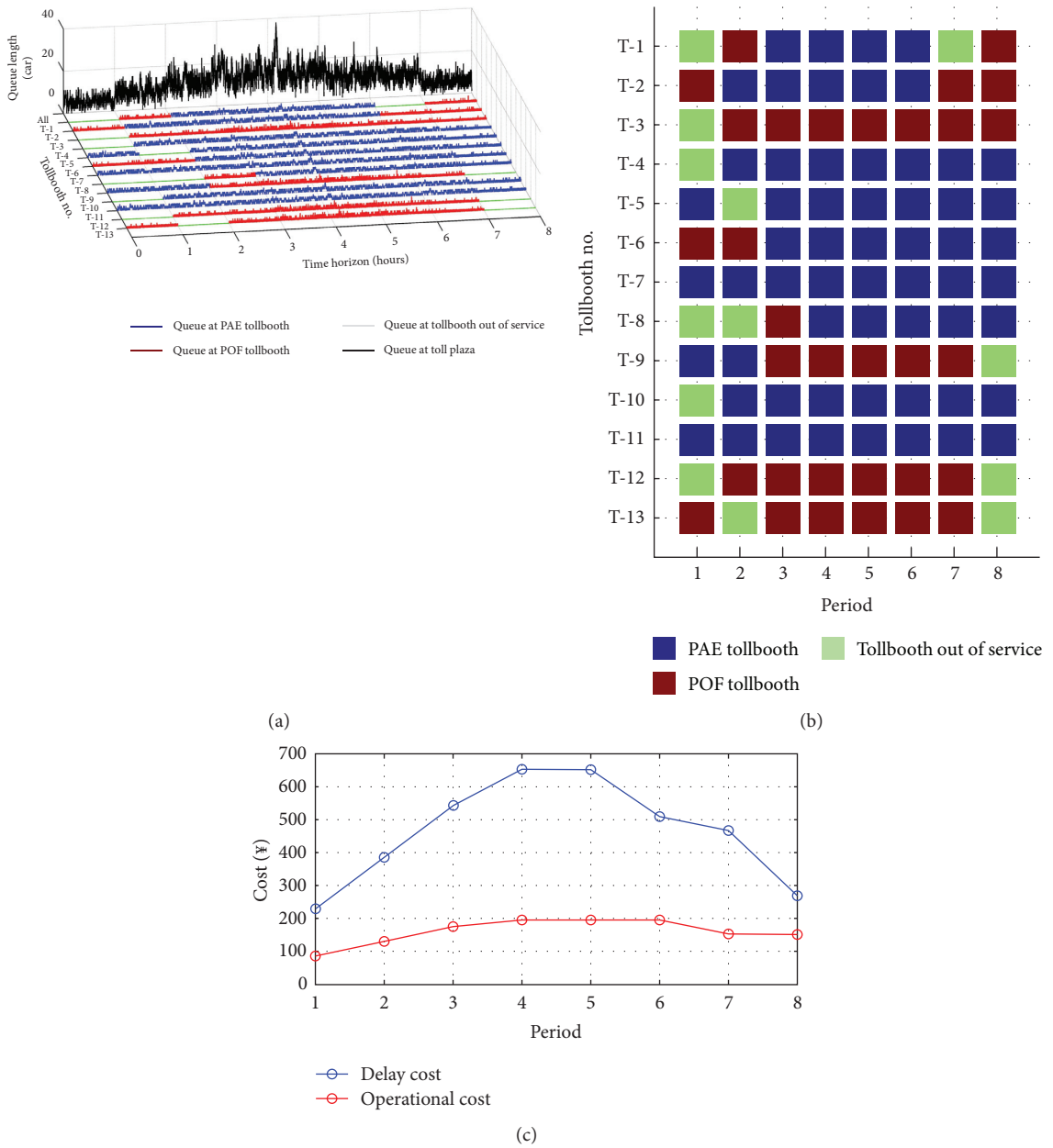


FIGURE 17: Recommended tollbooth operation scheme and its performance with weight of operational cost = 0.8. (a) Queue length evolution with time. (b) Best tollbooth operation scheme. (c) Costs in each period.

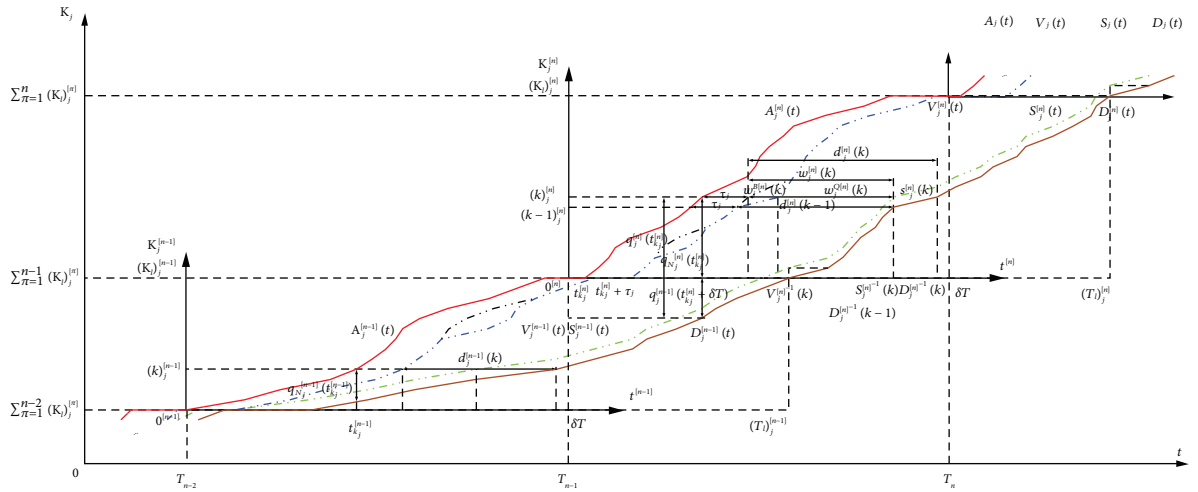


FIGURE 18: Profiles at a single tollbooth in the entire time horizon and divided periods.

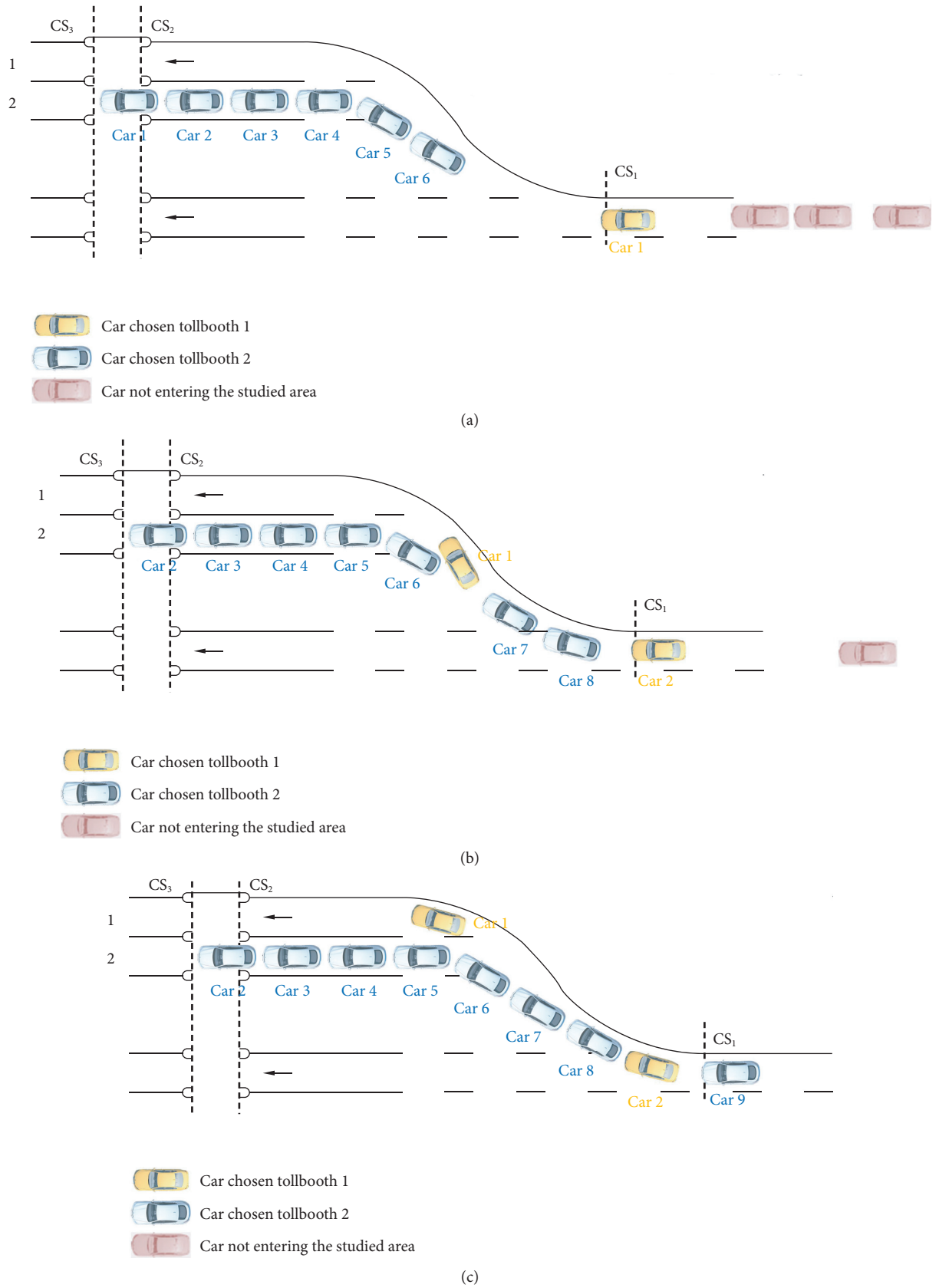


FIGURE 19: Formation and clearance of the blockage from queues of other tollbooths: (a) At time t_a, $q_2(t_a) \geq q_2^B$, newly arriving car 1 is blocked. (b) At time t_a, $q_2^{1_1}(t_b) < q_2^B$, newly arriving car 1 starts to move; $q_2(t_a) \geq q_2^B$, newly arriving car 2 is blocked. (c) At time t_c, $q_2^{2_1}(t_c) \geq q_2^B$, car 2 is still blocked.

In addition, the lengths of all periods are identical and the allocation of PAE and POF tollbooths is assumed to be changed only at the beginning of each period. Therefore, the proposed allocation schemes do not have immediate response to the event when a bursty flow arrives in the middle of the period. An event-based dynamic tollbooth allocation problem will be proposed in the further research.

5. Conclusions and Recommendations

This paper investigated traffic-efficient yet cost-effective operation of POF and PAE tollbooths at the toll plaza for the car park, formulated as a discrete-time dynamic tollbooth allocation problem (DTAP) where a simulation-based integer programming model is developed. To describe the traffic dynamic at the car park exit, a simulation model was first proposed to represent the queueing system of the toll plaza with the mixed-type customers (cars) and servers (tollbooths), where the service time is dependent on the waiting times of customers. The simulation also took into account the time-varying tollbooth allocations with respect to arrival intensities of car platoons, and the queueing due to the blockage from the queueing spillover on the adjacent servers in the transition area before the toll plaza. Then, the simulation model is validated at the studied car park exit. Based on the results from the simulation, an integer programming model was developed to minimize a weighted sum of the traffic efficiency measure and operational cost for operation schemes of toll plaza. It was decomposed by rolling horizon approach into period-based subproblems that are iteratively solved via the Kriging metamodel algorithm.

A numerical example from the studied car park exit was presented to illustrate the proposed simulation model and discrete-time DTAP. The recommended operation schemes are derived. In recommended schemes, PAE tollbooths are usually more welcome than POF ones in all recommended schemes with different weights between traffic efficiency and cost-effectiveness, since PAE tollbooths are more flexible that they can serve both PAE and POF cars. In particular, when the arrival intensity is high, the toll plaza needs far more PAE tollbooths to satisfy the heavy demand even if car operator wants to control operational cost.

Further research directions can be derived from the simulation methodology proposed. Either the decelerations or accelerations in the driving behaviors of car following, lane changing, and gap acceptance are not taken into account in the simulation. Neither are the competitions of right-of-way among all cars. The microscopic simulation needs to be more sophisticated. In addition, the lengths of all periods are identical, and the allocation of PAE and POF tollbooths is assumed to be changed only at the beginning of each period. Therefore, the proposed allocation schemes do not have immediate response to the event when a bursty flow arrives in the middle of the period. An event-based dynamic tollbooth allocation problem will be proposed in the further research.

Appendix

A. Evaluation of Queue Length in System and Waiting Time for Individual Car

In order to help understand the relationships of the inputs and outputs among all periods, we first establish the arrival and departure cumulative profiles of the cross sections for the entire time horizon $[0, T]$ and $[T_{n-1}, T_n]$ of each period n. Then, the relationship among profiles in periods is described, and the performance measures in each period are quantified, respectively.

The cumulative profiles for tollbooth j at CS_1, CS_2, and CS_3 are denoted as $A_j(t)$, $S_j(t)$, and $D_j(t)$, respectively. As demonstrated in Figure 18, these cumulative profiles are formulated in the rectangular coordinate $(\mathbf{t}, \mathbf{K}_j)$ where the $\mathbf{t}$ − axis represents the time and the $\mathbf{K}_j$ − axis represents the cumulative number of cars choosing tollbooth j in the simulation. These profiles are further divided by the period, as shown in Figure 6. The cumulative profiles for tollbooth j at CS_1, CS_2, and CS_3 are denoted as $A_j^{[n]}(t^{[n]})$, $S_j^{[n]}(t^{[n]})$, and $D_j^{[n]}(t^{[n]})$, respectively, and are formulated in the rectangular coordinate $(\mathbf{t}^{[n]}, \mathbf{K}_j^{[n]})$ where the $\mathbf{t}^{[n]}$-axis represents the time and the $\mathbf{K}_j^{[n]}$ − axis represents the cumulative number of cars choosing tollbooth j in the nth subsimulation, respectively. The relationships between profiles $A_j(t)$ and $A_j^{[n]}(t^{[n]})$, and $D_j(t^{[n]})$ and $D_j^{[n]}(t^{[n]})$ are described in equations (A.1) and (A.2), respectively:

$$A_j^{[n]}\left(t^{[n]}\right) = A_j\left(t^{[n]} + (n-1)\delta T\right) - \sum_{\pi=0}^{n-1} \left(K_l\right)_j^{[\pi]}, \quad t^{[n]} \in [0, \delta T], j \in B_n^{\mathrm{IN}}, n = 1, \ldots, N, \tag{A.1}$$

$$D_j^{[n]}\left(t^{[n]}\right) = D_j\left(t^{[n]} + (n-1)\delta T\right) - \sum_{\pi=0}^{n-1} \left(K_l\right)_j^{[\pi]}, \quad t^{[n]} \in \left[\min\left\{0, \left(T_l\right)_j^{[n-1]} - \delta T\right\}, \left(T_l\right)_j^{[n]}\right], j \in B_n^{\mathrm{IN}}, n = 1, \ldots, N, \tag{A.2}$$

where $(T_l)_j^{[n]}$ denotes the moment in the nth subsimulation when the last car leaves the tollbooth j, and $(T_l)_j^{[n]} = D_j^{[n]-1}((K_l)_j^{[n]})$ where the superscript -1 indicates the inverse function. In addition, we define that $(T_l)_j^{[0]} = 0$ and $(K_l)_j^{[0]} = 0$.

A.1. Queue Length with respect to Time. Now, we focus on evaluating the performance measures in each period. Let $q_j^{[n]}(\cdot)$ denote the queue length in front of tollbooth j with respect to the time in the nth subsimulation. Queue length in

a subsimulation is the summation of the newly formed queue length and the queue length in the last subsimulation at the same time, which is described as follows:

$$q_j^{[n]}\left(t_K^{[n]}\right) = q_j^{[n-1]}\left(t_K^{[n]} + \delta T\right) + q_{Nj}^{[n]}\left(t_K^{[n]}\right), \tag{A.3}$$

where $q_{Nj}^{[n]}(\cdot)$ denotes the newly formed queue length in front of tollbooth j by the arriving cars during the period n with respect to the time in the nth subsimulation. $q_j^{[n]}(\cdot)$ can be captured by iteratively obtaining $q_j^{[n-1]}(\cdot)$ and $q_{Nj}^{[n]}(\cdot)$. In addition, $q_j^{[0]}(\cdot) \equiv 0$ is defined. The newly formed queue length in front of tollbooth j by the arriving cars during the nth subsimulation at any time $t^{[n]}$, $q_{Nj}^{[n]}(t^{[n]})$, is given by equations (A.4), (A.5), or (A.6) with respect to different conditions as follows.

If $(T_l)_j^{[n-1]} < \delta T$, we have

$$q_{Nj}^{[n]}\left(t^{[n]}\right) = \begin{cases} A_j^{[n]}\left(t^{[n]}\right) - D_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in [0, \delta T], \\ (K_l)_j^{[n]} - D_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in \left[\delta T, (T_l)_j^{[n]}\right], \\ 0, & t^{[n]} \in \left[(T_l)_j^{[n]}, +\infty,\right) \end{cases} \quad j \in B_n^{\mathrm{IN}}, n = 1, \ldots, N. \tag{A.4}$$

If $\delta T \le (T_l)_j^{[n-1]} < 2\delta T$, we have

$$q_{Nj}^{[n]}\left(t^{[n]}\right) = \begin{cases} A_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in \left[0, \min\left\{0, (T_l)_j^{[n-1]} - \delta T\right\}\right], \\ A_j^{[n]}\left(t^{[n]}\right) - D_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in \left[\min\left\{0, (T_l)_j^{[n-1]} - \delta T\right\}, \delta T\right], \\ (K_l)_j^{[n]} - D_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in \left[\delta T, (T_l)_j^{[n]}\right], \\ 0, & t^{[n]} \in \left[(T_l)_j^{[n]}, +\infty,\right) \end{cases} \quad j \in B_n^{\mathrm{IN}}, n = 1, \ldots, N. \tag{A.5}$$

If $(T_l)_j^{[n-1]} \ge 2\delta T$, we have

$$q_{Nj}^{[n]}\left(t^{[n]}\right) = \begin{cases} A_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in [0, \delta T], \\ (K_l)_j^{[n]}, & t^{[n]} \in \left[\delta T, (T_l)_j^{[n-1]} - \delta T\right], \\ (K_l)_j^{[n]} - D_j^{[n]}\left(t^{[n]}\right), & t^{[n]} \in \left[(T_l)_j^{[n-1]} - \delta T, (T_l)_j^{[n]}\right], \\ 0, & t^{[n]} \in \left[(T_l)_j^{[n]}, +\infty,\right) \end{cases} \quad j \in B_n^{\mathrm{IN}}, n = 1, \ldots, N. \tag{A.6}$$

A.2. Waiting Time due to Blockage of the Queues from Other Servers. Let $\omega_j^{B[n]}(k)$ denote the waiting time for the car $(k)_j^{[n]}$ due to the blockage of the queues in front of other tollbooths. For the car $(k)_j^{[n]}$, the queueing blockage is presented in three scenarios as follows.

If there exists a toll lane $i \in B_n^{\mathrm{IN}}$ where $1 \le j < i \le N_R + 1$, such that $q_i\left(t_{k_j}^{[n]}\right) \ge q_i^B$, then the car is blocked by the queue in front of tollbooth i, as Figure 19 presents. q_i^B denotes the minimum queue length in front of tollbooth i which blocks the path for other cars reaching their objective toll lanes. The car has to wait until the queue in front of this car for any tollbooth i between toll lanes j and $(N_R + 1)$ is shorter than q_i^B. In other words, the queue behind the car $(k)_j^{[n]}$ (the car platoon arriving at CS_1 later than the car $(k)_j^{[n]}$) for any tollbooth i between toll lanes j and $(N_R + 1)$ does not affect the movement of this car. Let $q_i^{(k)_j^{[n]}}(t^{[n]})$ denote the queue length for tollbooth i in front of the car $(k)_j^{[n]}$ at time $t^{[n]}$ and we have

$$q_i^{k_j^n}\left(t^{[n]}\right) = A_i^{[n]}\left(t_{k_j}^{[n]}\right) - D_i^{[n]}\left(t^{[n]}\right), \quad t^{[n]} \in \left[t_{k_j}^{[n]}, +\infty\right). \tag{A.7}$$

The moment when the car $(k)_j^{[n]}$ blocked by the queue can restart to move to the objective toll lane, $t_{k_j^n}^d$, is given by $\min\left\{t^{[n]} > t_{k_j}^{[n]} \middle| \max_{j<i\le N_R+1}\left(q_i^{k_j^n}(t^{[n]}) - q_i^B\right) \le 0\right\}$. Then, $\omega_j^{B[n]}(k)$ is equal to $t_{k_j^n}^d - t_{k_j}^{[n]}$.

If $N_R + 1 \le j \le N_R + N_A$, then the car waits in the queue for the service and the blockage effect from other queues does not exist.

If there exists a toll lane $i \in B_n^{\mathrm{IN}}$ where $N_T - N_L \le i < j \le N_T$, such that $q_i(t_{k_j}^{[n]}) \ge q_i^B$, then the car is blocked by the queue in front of tollbooth i. The car has to wait until the queue for any tollbooth i between toll lanes $(N_T - N_L)$ and j is shorter than q_i^B. The moment when the car $(k)_j^{[n]}$ can restart to move, $t_{k^n}^d$, is given by $\min\left\{t^{[n]} > t_{k_j}^{[n]} \middle| \max_{N_T - N_L \le i < j}(q_i^{k_j}(t^{[n]}) - q_i^B) \le 0\right\}$.

To sum up, the waiting time for the car $(k)_j^{[n]}$ due to the blockage of the queues in front of the other tollbooths is given by

$$\omega_j^{B[n]}(k) = \begin{cases} \min\limits_{\max\limits_{j<i\le N_R+1}\left(q_i^{k_j^n}\left(t^{[n]}\right) - q_i^B\right) \le 0, t^{[n]} > t_{k_j}^{[n]}} \left(t^{[n]} - t_{k_j}^{[n]}\right), & \text{if } \exists 1 \le j < i \le N_R + 1 \\ & s.t.\ q_i\left(t_{k_j}^{[n]}\right) \ge q_i^B, \\ \min\limits_{\max\limits_{N_T - N_L \le i < j}\left(q_i^{k_j^n}\left(t^{[n]}\right) - q_i^B\right) \le 0, t^{[n]} > t_{k_j}^{[n]}} \left(t^{[n]} - t_{k_j}^{[n]}\right), & \text{if } \exists N_T - N_L \le i < j \le N_T \\ & s.t.\ q_i\left(t_{k_j}^{[n]}\right) \ge q_i^B, \\ 0, & \text{otherwise.} \end{cases} \tag{A.8}$$

A.3. Waiting Time due to Busyness of Selected Server. To help evaluate the waiting time due to the busyness of selected server, we first introduce a virtual profile for CS_2 at tollbooth j in the case that there is no queue in front of tollbooth j. The profile is denoted by $V_j^{[n]}(t)$, presented in Figure 1(a) and obtained using the following equation:

$$V_j^{[n]-1}(k) = A_j^{[n]-1}(k) + \tau_j + \omega_j^{B[n]}(k), \quad (k)_j^{[n]} \in K_j^{[n]}, \tag{A.9}$$

where τ_j represents the travel time from CS_1 to tollbooth j when there is no queue in front of tollbooth j.

Let $\omega_j^{Q[n]}(k)$ denote the waiting time for the car $(k)_j^{[n]}$ due to the busyness of tollbooth j, and we have $\omega_j^{Q[n]}(k) = S_j^{[n]-1}(k) - V_j^{[n]-1}(k)$. In the case that the car $(k)_j^{[n]}$ arrives at the tollbooth by following the tail of the queue, the service starts at the departure time of the car $(k-1)_j^{[n]}$. Otherwise, the service start time of the car $(k)_j^{[n]}$ is $V_j^{[n]-1}(k)$. Therefore, we have

$$S_j^{[n]-1}(k) = \begin{cases} D_j^{[n]-1}(k-1), & \text{if } V_j^{[n]-1}(k) \le D_j^{[n]-1}(k-1), \\ V_j^{[n]-1}(k), & \text{otherwise.} \end{cases} \tag{A.10}$$

$\omega_j^{Q[n]}(k)$ is given by

$$\omega_j^{Q[n]}(k) = \begin{cases} D_j^{[n]-1}(k-1) - V_j^{[n]-1}(k), & \text{if } V_j^{[n]-1}(k) \le D_j^{[n]-1}(k-1), \\ 0, & \text{otherwise.} \end{cases} \tag{A.11}$$

B. Data for Studied Car Park

B.1. Arrival Profile. Cars' arrival data were collected by video at 23:00–01:00 on fourteen consecutive days in November 2018. During the period, large numbers of flights arrived at the airport while subway was out of service. Many cars were expected to come and pick up passengers in the car park. The arrival samples cover from 20 to 650 cars per hour. Unfortunately, in these peak hours, no more than 650 cars exit the park per hour and the heavy-flow samples cannot be captured.

The main reason for underexpected car number is that many drivers illegally waited at the roadside lanes of the roadway around the airport rather than in the car park and then they pick up the passengers at the arrival floor in order to avoid the parking fees. When these waiting behaviors were strictly prohibited by the traffic police in future, the cars had to pour into the car park and the toll plaza at the exit would face the massive car flows. The car platoon's arrivals at the toll plaza are represented in Figure 8. They are approximated as Poisson processes in the case of low flow via Kolmogorov–Smirnov test. Then, we assume that Poisson arrival is still fit for moderate and high flows.

Indeed, flows departing airport's car park follow no typical stochastic process sometimes due to its high dependence on flight schedule. In addition, it is extremely hard to capture the stochastic processes with explicit mathematical forms at all levels of arrival intensities. Hence, the assumption of Poisson arrival for car platoons does not fully reflect reality; it nevertheless is an available alternative when we simulate the flow with time-varying arrival intensities.

B.2. Parking Duration and Travel Time from Car Park to CS_1. Both of the two car parks grant drivers certain time for free parking. If a car stays in the garage for less than 15 minutes, the one does not have to pay. Hence, parking duration and travel time from car park to CS_1 of a car affect whether the car itself has an opportunity to be free of charge when exiting the toll plaza. The distribution of parking durations is obtained from the parking occupancy data. The data provide the times of entering and leaving the parking space for each car. In the car park, nearly 15% of cars park for less than 10 minutes which have the chance to be free-parking. In addition, under 15% of cars stay for more than 5 hours in which the longest parking car parks for over 10 days. For the majority of cars parks no more than 5 hours and full-length parking duration figure cannot show the distribution features clearly, we demonstrate a part of the parking duration distribution (less than 5 hours), as presented in Figure 4(a). The travel time distribution from car park to CS_1 is estimated via the distances between all parking spaces and CS_1, as presented in Figure 4(b).

B.3. Service Time. At PAE tollbooths, free-parking cars wait for the car plates automatically verified by the cameras. The distribution of the service times for free-parking cars is presented in Figure 5(c). If the cars need to pay (i.e., PAE cars stay more than 15 minutes or POF cars stay more than 20 minutes after smartphone paying in the car park), the service time includes the time waiting for the car plates automatically verified by the cameras and the time paid at the toll plaza (i.e., QR code paid time in POF tollbooths or cash paid time in PAE tollbooths). The distributions of the service times when cars have to pay at POF and PAE tollbooths are presented in Figures 5(a) and 5(b), respectively. Usually, the QR code paid time is shorter than cash paid time, but some drivers spend far more time on QR scanning and smartphone payment than cash payment to tollmen.

Data Availability

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. Meanwhile, the code for the proposed model is also available.

Disclosure

The authors take sole responsibility for all views and opinions expressed in this paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by the National Natural Science Foundation for Young Scholars of China (Grant no. 71901190) and Special Project for Technology Innovation and Application Development of Chongqing, China (Grant no. cstc2019jscx-tjsbX0013).

References

[1] D.-w. Huang and W.-n. Huang, "The influence of tollbooths on highway traffic," *Physica A: Statistical Mechanics and its Applications*, vol. 312, no. 3-4, pp. 597–608, 2002.

[2] I. Sahin and G. Akyildiz, "Bosporus Bridge toll plaza in Istanbul, Turkey: upstream and downstream traffic features," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1910, pp. 99–107, 2005.

[3] K. Komada and T. Nagatani, "Traffic flow through multi-lane tollbooths on a toll highway," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 11, pp. 2268–2279, 2010.

[4] G. Aksoy, H. B. Celikoglu, and E. Gedizlioglu, "Analysis of toll queues by micro-simulation: results from a case study in Istanbul," *Procedia-Social and Behavioral Sciences*, vol. 111, pp. 614–623, 2014.

[5] Q.-M. He and X. Chao, "A tollbooth tandem queue with heterogeneous servers," *European Journal of Operational Research*, vol. 236, no. 1, pp. 177–189, 2014.

[6] T. V. Do, "A closed-form solution for a tollbooth tandem queue with two heterogeneous servers and exponential service times," *European Journal of Operational Research*, vol. 247, no. 2, pp. 672–675, 2015.

[7] P. Chakroborty, R. Gill, and P. Chakraborty, "Analysing queueing at toll plazas using a coupled, multiple-queue, queueing system model: application to toll plaza design," *Transportation Planning and Technology*, vol. 39, no. 7, pp. 675–692, 2016.

[8] D. Levinson and E. Chang, "A model for optimizing electronic toll collection systems," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 4, pp. 293–314, 2003.

[9] J. Klodzinski and H. Al-Deek, "Evaluation of toll plaza performance after addition of express toll lanes at mainline toll plaza," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1867, pp. 107–115, 2004.

[10] T. Ito and T. Hiramoto, "A general simulator approach to ETC toll traffic congestion," *Journal of Intelligent Manufacturing*, vol. 17, no. 5, pp. 597–607, 2006.

[11] K. Komada, S. Masukura, and T. Nagatani, "Traffic flow on a toll highway with electronic and traditional tollgates," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 24, pp. 4979–4990, 2009.

[12] G. Pang and Y. Zhou, "Two-parameter process limits for an infinite-server queue with arrival dependent service times," *Stochastic Processes and their Applications*, vol. 127, no. 5, pp. 1375–1416, 2017.

[13] R. Bekker, G. M. Koole, B. F. Nielsen, and T. B. Nielsen, "Queues with waiting time dependent service," *Queueing Systems*, vol. 68, no. 1, pp. 61–78, 2011.

[14] K. Y. Zhernovyi and Y. V. Zhernovyi, "M θ/G/1/m and M θ/G/1 systems with the service time dependent on the queue length," *Journal of Communications Technology and Electronics*, vol. 58, no. 12, pp. 1267–1275, 2013.

[15] J. Rodrigues, S. M. Prado, N. Balakrishnan, and F. Louzada, "Flexible M/G/1 queueing system with state dependent service rate," *Operations Research Letters*, vol. 44, no. 3, pp. 383–389, 2016.

[16] J. Lee and J. Kim, "A workload-dependent queue under a two-stage service policy," *Operations Research Letters*, vol. 34, no. 5, pp. 531–538, 2006.

[17] R. Bekker and S. C. Borst, "Optimal admission control in queues with workload-dependent service rates," *Probability in the Engineering and Informational Sciences*, vol. 20, no. 4, pp. 543–570, 2006.

[18] N. Nezamuddin and H. Al-Deek, "Developing microscopic toll plaza and toll road corridor model with PARAMICS," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2047, pp. 100–110, 2008.

[19] C. Russo, R. Harb, and E. Radwan, "Calibration and verification of SHAKER, a deterministic toll plaza simulation model," *Journal of Transportation Engineering*, vol. 136, no. 2, pp. 85–92, 2010.

[20] M. Fleming, N. Huynh, and Y. Xie, "Agent-based simulation tool for evaluating pooled queue performance at marine container terminals," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2330, pp. 103–112, 2013.

[21] S. Kim, "The toll plaza optimization problem: design, operations, and strategies," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 1, pp. 125–137, 2009.

[22] D. K. Hale, C. Antoniou, M. Brackstone, D. Michalaka, A. T. Moreno, and K. Parikh, "Optimization-based assisted calibration of traffic simulation models," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 100–115, 2015.

[23] Q. Cheng, S. Wang, Z. Liu, and Y. Yuan, "Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 422–438, 2019.

[24] S. Amaran, N. V. Sahinidis, B. Sharda, and S. J. Bury, "Simulation optimization: a review of algorithms and applications," *4OR*, vol. 12, no. 4, pp. 301–333, 2014.

[25] K. Kulkarni, K. Tran, H. Wang, and H. Lau, "Efficient gate system operations for a multipurpose port using simulation-optimization," in *Proceedings of the 2017 Winter Simulation Conference (WSC)*, Las Vegas, NV, USA, 2017.

[26] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Kevin Tucker, "Surrogate-based analysis and optimization," *Progress in Aerospace Sciences*, vol. 41, no. 1, pp. 1–28, 2005.

[27] J. Müller and C. A. Shoemaker, "Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems," *Journal of Global Optimization*, vol. 60, no. 2, pp. 123–144, 2014.

[28] D. Chen, "Research on traffic flow prediction in the big data environment based on the improved RBF neural network," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2000–2008, 2017.

[29] A. M. Kostin, G. Guillén-Gosálbez, F. D. Mele, M. J. Bagajewicz, and L. Jiménez, "A novel rolling horizon strategy for the strategic planning of supply chains. Application to the sugar cane industry of Argentina," *Computers & Chemical Engineering*, vol. 35, no. 11, pp. 2540–2563, 2011.

[30] S. Zhan, L. G. Kroon, J. Zhao, and Q. Peng, "A rolling horizon approach to the high speed train rescheduling problem in case of a partial segment blockage," *Transportation Research Part E: Logistics and Transportation Review*, vol. 95, pp. 32–61, 2016.

[31] S. Chand, V. N. Hsu, and S. Sethi, "Forecast, solution, and rolling horizons in operations management problems: a classified bibliography," *Manufacturing & Service Operations Management*, vol. 4, no. 1, pp. 25–43, 2002.

[32] L. Meng and X. Zhou, "Robust single-track train dispatching model under a dynamic and stochastic environment: a scenario-based rolling horizon solution approach," *Transportation Research Part B: Methodological*, vol. 45, no. 7, pp. 1080–1102, 2011.

[33] J. Silvente, G. M. Kopanos, E. N. Pistikopoulos, and A. Espuña, "A rolling horizon optimization framework for the simultaneous energy supply and demand planning in microgrids," *Applied Energy*, vol. 155, pp. 485–501, 2015.

[34] C.-C. Lu, K.-C. Ying, and H.-J. Chen, "Real-time relief distribution in the aftermath of disasters - a rolling horizon approach," *Transportation Research Part E: Logistics and Transportation Review*, vol. 93, pp. 1–20, 2016.

[35] Y. Xia, X. Liu, and G. Du, "Solving bi-level optimization problems in engineering design using kriging models," *Engineering Optimization*, vol. 50, no. 5, pp. 856–876, 2018.

[36] J. Kleijnen, *Design and Analysis of Monte Carlo Experiments*, Springer, Berlin, Germany, 2nd edition, 2015.

[37] Y. Du, S. Yu, Q. Meng, and S. Jiang, "Allocation of street parking facilities in a capacitated network with equilibrium constraints on drivers' traveling and cruising for parking," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 181–207, 2019.

[38] J. Müller, C. A. Shoemaker, and R. Piché, "SO-I: a surrogate model algorithm for expensive nonlinear integer programming problems including global optimization applications," *Journal of Global Optimization*, vol. 59, no. 4, pp. 865–889, 2014.

[39] K. Q. Ye, W. Li, and A. Sudjianto, "Algorithmic construction of optimal symmetric Latin hypercube designs," *Journal of Statistical Planning and Inference*, vol. 90, no. 1, pp. 145–159, 2000.

[40] R. G. Regis and C. A. Shoemaker, "A stochastic radial basis function method for the global optimization of expensive functions," *INFORMS*, vol. 19, pp. 497–509, 2007.