Transportation Science

A Continuous Model for Designing Corridor Systems with Modular Autonomous Vehicles Enabling Station-wise Docking

Zhiwei Chen, Xiaopeng Li, Xiaobo Qu

Please scroll down for article—it is on subsequent pages

A Continuous Model for Designing Corridor Systems with Modular Autonomous Vehicles Enabling Station-wise Docking

Zhiwei Chen,[a] Xiaopeng Li,[a,]* Xiaobo Qu[b,]*

[a] Department of Civil and Environmental Engineering, University of South Florida, Florida 33620; [b] Department of Architecture and Civil Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden
*Corresponding authors
Contact: zhiweic@usf.edu (ZC); xiaopengli@usf.edu, https://orcid.org/0000-0002-5264-3775 (XL); xiaobo@chalmers.se (XQ)

Abstract. The "asymmetry" between spatiotemporally varying passenger demand and fixed-capacity transportation supply has been a long-standing problem in urban mass transportation (UMT) systems around the world. The emerging modular autonomous vehicle (MAV) technology offers us an opportunity to close the substantial gap between passenger demand and vehicle capacity through station-wise docking and undocking operations. However, there still lacks an appropriate approach that can solve the operational design problem for UMT corridor systems with MAVs efficiently. To bridge this methodological gap, this paper proposes a continuum approximation (CA) model that can offer near-optimal solutions to the operational design for MAV-based transit corridors very efficiently. We investigate the theoretical properties of the optimal solutions to the investigated problem in a certain (yet not uncommon) case. These theoretical properties allow us to estimate the seat demand of each time neighborhood with the arrival demand curves, which recover the "local impact" property of the investigated problem. With the property, a CA model is properly formulated to decompose the original problem into a finite number of subproblems that can be analytically solved. A discretization heuristic is then proposed to convert the analytical solution from the CA model to feasible solutions to the original problem. With two sets of numerical experiments, we show that the proposed CA model can achieve near-optimal solutions (with gaps less than 4% for most cases) to the investigated problem in almost no time (less than 10 ms) for large-scale instances with a wide range of parameter settings (a commercial solver may even not obtain a feasible solution in several hours). The theoretical properties are verified, and managerial insights regarding how input parameters affect system performance are provided through these numerical results. Additionally, results also reveal that, although the CA model does not incorporate vehicle repositioning decisions, the timetabling decisions obtained by solving the CA model can be easily applied to obtain near-optimal repositioning decisions (with gaps less than 5% in most instances) very efficiently (within 10 ms). Thus, the proposed CA model provides a foundation for developing solution approaches for other problems (e.g., MAV repositioning) with more complex system operation constraints whose exact optimal solution can hardly be found with discrete modeling methods.

Keywords: transportation corridors • operational design • modular autonomous vehicles • station-wise docking • continuum approximation

1. Introduction

The "asymmetry" between spatiotemporally varying passenger demand and fixed-capacity transportation supply has been a long-standing problem in urban mass transportation (UMT) systems around the world, for example, the MTR system in Singapore (Sun et al. 2014), the Beijing Subway system (Shi et al. 2018), and the urban transit system in The Hague, Netherlands (Luo et al. 2018). Specifically, the passenger demand at one UMT station fluctuates substantially over an operational day with overly crowded passengers flooding the station during peak hours yet sparse passenger arrivals during off-peak hours, leaving vehicles underutilized. On the other hand, the passenger demand is clustered at stations in central business areas, and suburban and rural areas usually witness relatively low passenger demand during the same period or vice versa. A typical solution to this asymmetry in the state

of the practice is to adjust vehicle dispatch frequencies to make the transportation capacity better aligned with the passenger demand level through transit timetable design (Caimi, Kroon, and Liebchen 2017). This topic, that is, demand-responsive transit timetable design, is also extensively investigated in the literature, from timetables with constant headway (e.g., Ceder 2001) and periodic timetables (e.g., Liebchen 2007) to non-periodic timetables (e.g., Sun et al. 2014).

However, time-varying vehicle dispatch frequency alone cannot bridge the huge gap between passenger demand and transit capacity. As a result, in many UMT systems, excessive passenger queues at stations in central business areas during peak hours (because demand is much greater than supply) and empty vehicle units in suburban and rural areas during off-peak hours (because supply is far greater than demand) can still be commonly observed. To further address these issues, as discovered in our preceding papers focusing on UMT shuttle systems with one origin and one destination (Chen, Li, and Zhou 2019, 2020), the emerging modular autonomous vehicles (MAV) can be used as an elastic medium reconciling the asymmetry between demand and supply. Specifically, MAV technology allows modular pods to be flexibly docked and undocked into modular vehicles with different formations (or lengths, capacities) at any station in a UMT corridor system (NEXT 2019). With this, long vehicles can be dispatched at times and stations with intensive demand and short vehicles at those with relatively sparse demand, thereby potentially diminishing both the vehicle operational cost and passenger waiting cost.

Simple though this concept seems, designing operational plans for MAV-based corridor systems with multiple origins and destinations enabling station-wise docking is challenging. It has been well acknowledged that conventional transit timetable design for UMT corridors (in which mainly vehicle dispatch times are decided) is NP-hard and, thus, usually has to be solved by heuristics (Caprara, Fischetti, and Toth 2002; Niu and Zhou 2013). The investigated problem in this paper, obviously, is as hard as the conventional ones at its minimum because it involves the joint design of dispatch time and capacity. Indeed, several pioneering studies have considered the option of multiple vehicle capacity in transit timetable design (e.g., Ceder 2001; Albrecht 2009; Hassold and Ceder 2014; Guo, Chow, and Schonfeld 2017; Chen, Li, and Zhou 2019, 2020). Different from these studies, we focus on corridor systems in which station-wise docking and undocking is allowed; that is, MAVs can dock and undock at any station along the corridor. This modification requires a decision on vehicle length to be made at each single station along the corridor and, thus, at least increases the number of decision variables by a factor of the number of the stations. Thus, the solution approaches

from previous studies cannot be directly applied to the investigated problem in this paper. Further, none of the aforementioned studies provides analytical results to the investigated problem for MAV-based UMT corridor systems. Such an analytical solution offers managerial insights for UMT operators and, more importantly, enables the development of efficient solution algorithms to address large-scale instances. Hence, there still lacks an appropriate approach that can solve the joint design problem in MAV-based UMT corridor systems in real-world contexts efficiently and offer analytical insights in the meanwhile.

To enable such a solution approach, the classical continuum approximation (CA) framework proposed by Newell (1971) to design dispatching policies for a transportation route should be considered. The CA approach approximates discrete decision variables with continuous smooth functions and decomposes the original problem into a finite number of separable subproblems in relatively homogeneous (time) neighborhoods. The subproblems can usually be formulated with simple multivariate functions that can be analytically solved independently, rendering the CA approach promising in tackling large-scale instances. Thus, it has been applied to solve more complicated transit design problems (e.g., Hurdle 1973; Newell 1974; Wirasinghe 1990; Daganzo 2009, 2010; Chen, Li, and Zhou 2020). However, two unique challenges in the investigated problem are not fully addressed in these studies. First, the investigated corridor system allows passengers to board and alight at any station, and thus, the seat demand (i.e., the number of passengers who want to board a vehicle) at a station includes passengers boarding at its upstream stations yet remaining on the vehicle (i.e., with downstream destinations) and those who board at the current station. Therefore, the seat demand at a station is not simply an exogenous input, but related to the dispatch decisions at its upstream stations, which couples decisions at different stations. Also, the passenger boarding order at one station affects the dispatch decisions at its downstream stations. Consider a simple example consisting of three stations and with the passenger demand at each station shown in Figure 1 with the destination of each passenger marked above the passenger's head. Assume that the capacity of one modular pod is one passenger. If a vehicle with two modular pods is dispatched at station 1, all passengers can board this vehicle in one dispatch (solution 1). However, if a vehicle with one modular pod is dispatched at station 1, then the number of modular pods dispatched at station 2 is dependent on the passenger boarding order at the first station (solutions 2 and 3). To address this issue, Wirasinghe (1990) extended Newell's (1971) policy in a many-to-many demand setting by replacing in Newell's (1971) analytical

Figure 1. (Color online) An Illustrative Example for the Spatiotemporal Correlation Decisions in a UMT Corridor System
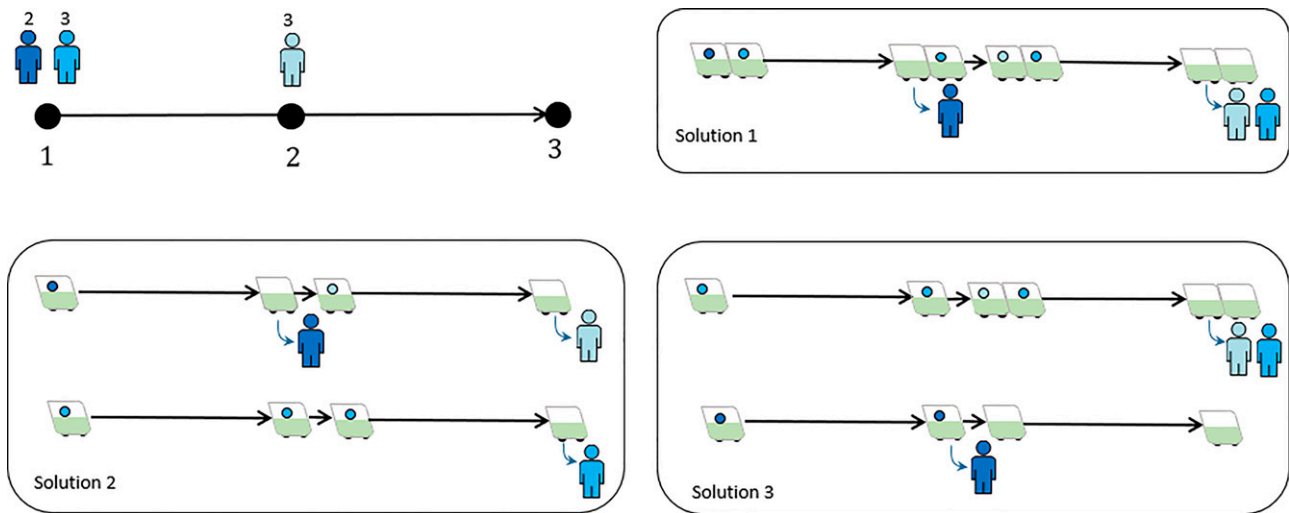
formulas the arrival demand rate by a seat demand rate. However, the seat demand rate was estimated with historic counts of boarding and alighting in Wirasinghe's (1990) study, which actually ignores the fact that the seat demand is interrelated with the current dispatch decisions. Further, Wirasinghe's (1990) method cannot be applied in systems in which historic boarding and alighting data are not available, for example, the proposed MAV-based UMT corridors, new UMT corridors, or subway systems with only passenger origin–destination (OD) information. The other challenge in modeling the investigated problem under the CA framework is the consideration of time-varying vehicle formations and station-wise docking. This consideration makes the vehicle dispatch cost a variable related to the capacity of the dispatched vehicles, which has not been considered in many CA studies except for Chen, Li, and Zhou (2020). Yet the station-wise docking operation postulates decisions on vehicle formations at each station along the corridor, and the model in Chen, Li, and Zhou (2020) applies to shuttle systems in which these decisions are made only at the start station. Thus, fundamental methodological innovation is needed to develop a CA model that can jointly design dispatch headway and capacity for transportation corridors with MAVs enabling station-wise docking.

To bridge this methodological gap, this paper proposes a CA model to jointly design the dispatch headway and capacity for many-to-many transit corridors in which MAVs can change their capacities through docking and undocking operations at any station along the corridor. One fundamental premise of the CA framework to be applicable is the "local-impact property"—that is, decisions in a local neighborhood are mainly affected by settings in this and nearby neighborhoods but not much by distant ones. Yet, in the investigated corridor systems, the dispatch decisions are spatiotemporally correlated. Specifically, dispatch decisions at a station affect the passenger arrival process at its downstream stations, therefore linking together decisions at different stations. On the other hand, decisions at one time affect the number of passengers waiting for boarding at the following times, therefore coupling decisions at different times together. To address this issue, we investigate theoretical properties of the optimal solution(s) to the investigated problem in a special (yet not uncommon) case, which enables us to approximately estimate the seat demand simply with the passenger arrival counts. These properties also reveal an approximate relation between the dispatch headway and vehicle formation, allowing the elimination of the vehicle formation decisions from the model. With these, the original problem regains the local impact property and can be decomposed into a finite number of separable problems with only one decision variable and two constraints. Overall, this paper makes contributions to the literature from the following three aspects:

i. We investigate the operational design of MAV-based transit corridor systems enabling station-wise docking and undocking under the CA framework. This proposed CA model presents a macroscopic view of the system and yields a simple analytical solution approach. The analytical results enable efficient solution methods for relevant large-scale problem instances and offer managerial insights to system operators.

ii. We prove some theoretical properties of the optimal solution(s) to the investigated problem that

break the spatiotemporal correlation between dispatch decisions and, thus, constitute a theoretical foundation for the CA model. With this, we advance the CA methodology to the joint design of dispatch headway and vehicle capacity that considers many-to-many demand patterns, station-wise vehicle capacity adjustment, and other factors (e.g., passenger boarding order, minimum dispatch headway).

iii. We conduct two sets of numerical experiments with realistic traffic data of various scales to test the solution quality and computation speed of the CA model. Results show that the proposed CA model can produce near-optimal solutions to the investigated problem very efficiently in extensive parameter settings, and its discrete counterpart can hardly be solved to optimality by an off-the-shelf commercial solver, Gurobi. Also, the investigated theoretical properties are verified and interesting managerial insights are offered through these numerical experiments.

Overall, this paper fills the methodological void of operational design for MAV-based transportation corridors enabling station-wise docking. It provides analytical and numerical methods for solving realistic problem instances and offers managerial insights to UMT operators. Additionally, although vehicle repositioning decisions are not considered in the proposed CA model, we show that the CA model can be easily extended to obtain highly accurate, near-optimal repositioning decisions in almost no time. Thus, it could serve as a foundation for the development of efficient solution approaches for more complex MAV-based system design problems.

The remainder of this paper is organized as follows. Section 2 introduces the problem setting and general problem formulation. Section 3 reviews several classical CA models and analyzes the theoretical properties of the investigated problem. Section 4 presents the CA model, the analytical solution approach, and the discretization method. Section 5 verifies the computation performance of the proposed CA model and the theoretical properties as well as reveals some managerial insights with numerical experiments. Finally, Section 6 concludes the paper and briefly discusses future research directions.

2. Problem Statement and General Problem Formulation

This section presents the investigated problem and general problem formulation. For the convenience of the readers, the key notation used throughout the paper is summarized in Online Appendix A.

2.1. Problem Statement

We consider a unidirectional UMT corridor system in which a set of stations indexed as $s, u \in \mathcal{S} := \{1, 2, \ldots, S\}$

Figure 2. (Color online) The Investigated Corridor System (Top) and Cumulative Passenger Counts (Bottom)
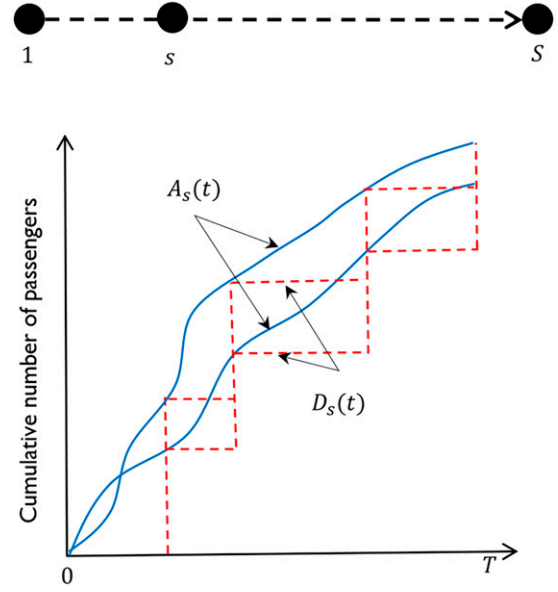
is placed along an urban transportation corridor with 1 and S being the start and terminal stations as shown in Figure 2. The sets of upstream and downstream stations of $s \in \mathcal{S}$ are denoted as $\mathcal{S}_s^-$ and $\mathcal{S}_s^+$, respectively. Passengers arrive at each station continuously during an operational horizon $[0, T]$. We describe the passenger arrival process at station $s \in \mathcal{S}$ with cumulative arrival demand curves denoted as $A_s(t)$, $\forall t \in [0, T]$ (the solid curves in Figure 2).[1] The corresponding passenger arrival rate $\forall s \in \mathcal{S}$ is defined as $a_s(t) := A'_s(t)$. To capture the distribution of passenger destinations from origin station $s \in \mathcal{S}$ among downstream station $u \in \mathcal{S}_s^+$, we denote the proportion of passengers arriving at station s and heading to station u at time t (among all passengers arriving at s at t) as $p_{su}(t)$, $\forall s \in \mathcal{S}, u \in \mathcal{S}_s^+, t \in [0, T]$. To serve the passengers, a set of K MAVs are dispatched across the operational horizon, indexed as $k \in \mathcal{K} := [1, 2, \cdots K]$, where the index increases with the dispatch time. These dispatches result in a cumulative passenger departure curve at each $s \in \mathcal{S}$ denoted as $D_s(t)$, $\forall t \in [0, T]$. The time at the start station and vehicle formation (defined in assumption (ii)) at station s for dispatch k are denoted as t_k, $\forall k \in \mathcal{K}$ and i_{ks}, $\forall k \in \mathcal{K}, s \in \mathcal{S}$, respectively. For the convenience of notation, define $t_0 := 0$ and $t_{K+1} := T$. We illustrate the operational concept and basic assumptions of the proposed MAV-based transportation corridor systems as follows.

i. The dispatched vehicles move along the corridor from the start to the terminal stations, separated by a minimum dispatch headway $\underline{h}$. The vehicles' dwell time at each station is relatively constant, and the travel time between two consecutive stations is time dependent. To simplify the model formulation, we

shift the time coordinate at stations $s > 1$ such that the departure time of any vehicle is the same across all stations along the corridor (Sun et al. 2014). The passenger demand arrival curves are shifted along with the time coordinate. Please refer to Online Appendix B for details about the time coordinate shifting process.

ii. Different from vehicles in existing UMT systems, MAVs can flexibly change their formations (i.e., length or capacity) at any station along the corridor by docking and undocking modular pods. We index vehicle formations as $i \in \mathcal{I} := \{1, 2, \cdots, I\}$, where I is the maximum allowable number of pods in one MAV. Each vehicle formation $i \in \mathcal{I}$ consists of i identical modular pods and, thus, has a capacity of ic, where c is the capacity of a single pod. Following Newell (1971), we assume that the number of modular pods is always sufficient so that there are always some vehicles in each formation for dispatching at each station.[2]

iii. Following previous studies (e.g., Niu, Zhou, and Gao 2015; Yin et al. 2017), we further assume that oversaturated traffic is not allowed at each station. However, note that multiple vehicle formations are considered in the investigated problem. Thus, different from previous studies, this assumption incorporates more general cases in which passengers waiting at the stations are not cleared after each dispatch.

iv. At each station $s \in \mathcal{S}$, passengers board a vehicle following a first-in, first-served (FIFS) principle. In reality, the passenger boarding order is extremely difficult to determine (especially when there are multiple destinations), but in general, passengers who arrive early have more chances to board a vehicle than those arriving later. Therefore, this assumption is not too distant from reality (Niu and Zhou 2013).

These operations result in two cost components in the system, that is, the operational cost and the passenger waiting cost (Yin et al. 2017). The operational cost primarily consists of the energy cost spent on operating the vehicles. Other cost components, such as the crew salary, can be easily incorporated into this general cost structure if needed. According to Daganzo (2005), the operational cost of transportation systems is subadditive. It can be divided into a constant term regardless of the vehicle capacity and a component dependent on the vehicle capacity. Therefore, we define the operational cost as a concave function over the vehicle formation i to account for its economics of scale (Holmberg and Tuy 1999, Daganzo 2005). Specifically, the operational cost of a vehicle in formation i, denoted as f_i, satisfies

$$\lambda f_i + (1 - \lambda)f_j \leq f_k, \forall i, j, k := \lambda i + (1 - \lambda)j \in \mathcal{I} \cup \{0\}, \\ \lambda \in [0, 1], \quad (1)$$

where $f_0 := 0$. The passenger waiting cost refers to product of a homogeneous unit-time waiting cost w for passengers and the total passenger waiting time.

The objective of this study is to find an optimal schedule for corridor systems satisfying the aforementioned operational concepts such that the desired objective of the system operator can be achieved. The optimal schedule is subject to four groups of constraints (see Section 2.2 for details). The major decisions include the number of dispatches K, dispatch time t_k, vehicle formation i_{ks}, and cumulative departure curve $D_s(t)$. Following previous studies (e.g., Yin et al. 2017; Chen, Li, and Zhou 2020), we aim to search for an optimal trade-off between the operational cost and the passenger waiting cost in the system.

2.2. General Problem Formulation

With Equation (1), the total operational cost in the system can be formulated as

$$C_f(i_{ks}) := \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} f_{i_{ks}}. \quad (2)$$

According to Newell (1971), the total passenger waiting time at station $s \in \mathcal{S}$ can be quantified as the area between its cumulative arrival demand curve and departure curve, that is, $\int_0^T ((A_s(t) - D_s(t))dt$. Because the passenger boarding order needs to be considered in this study, we decompose the cumulative arrival curve $A_s(t)$ as T time-dependent arrival rate curves

$$A_s(t', t) := \begin{cases} 0 & t \in [0, t') \\ a_s(t') & t \in [t', T], \end{cases} \forall s \in \mathcal{S}, t' \in [0, T],$$

each of which corresponds to the arrival demand at time $t' \in [0, T]$, respectively. Likewise, the cumulative departure curve at station s, $D_s(t)$, can be decomposed as T departure rate curves, denoted as $\mathcal{D}_s(t', t), \forall s \in \mathcal{S}, t', t \in [0, T]$. To facilitate the illustration of these definitions, consider a simple example with $A_s(t)$ and $D_s(t)$ shown in Figure 3(a). We see from Figure 3(a) that two, two, and one passengers arrive at station s at times 1, 2, and 3, respectively. Three MAVs are dispatched to serve the passengers. The decomposed curves are shown in Figure 3(b), and the projections of these curves for different t' are plotted in Figure 3, (c)–(e). Take passengers arriving at time 2 (i.e., $t' = 2$) as an example. The time-dependent arrival rate curve $A_s(2, t) = 0, \forall t \in [0, 2)$ and $A_s(2, t) = a_s(2) = 2, \forall t \in [2, 3]$, and the time-dependent departure rate curve $D_s(2, t) = 0, \forall t \in [0, 2), D_s(2, t) = 1, \forall t \in [2, 3)$ (one passenger arriving at $t' = 2$ has been transported), and $D_s(2, 3) = 2$ (all passengers arriving at $t' = 2$ have been transported). The decomposed curves for other times can be illustrated in a similar way.

Figure 3. (Color online) An Example for Cumulative Arrival and Departure Curve Decomposition
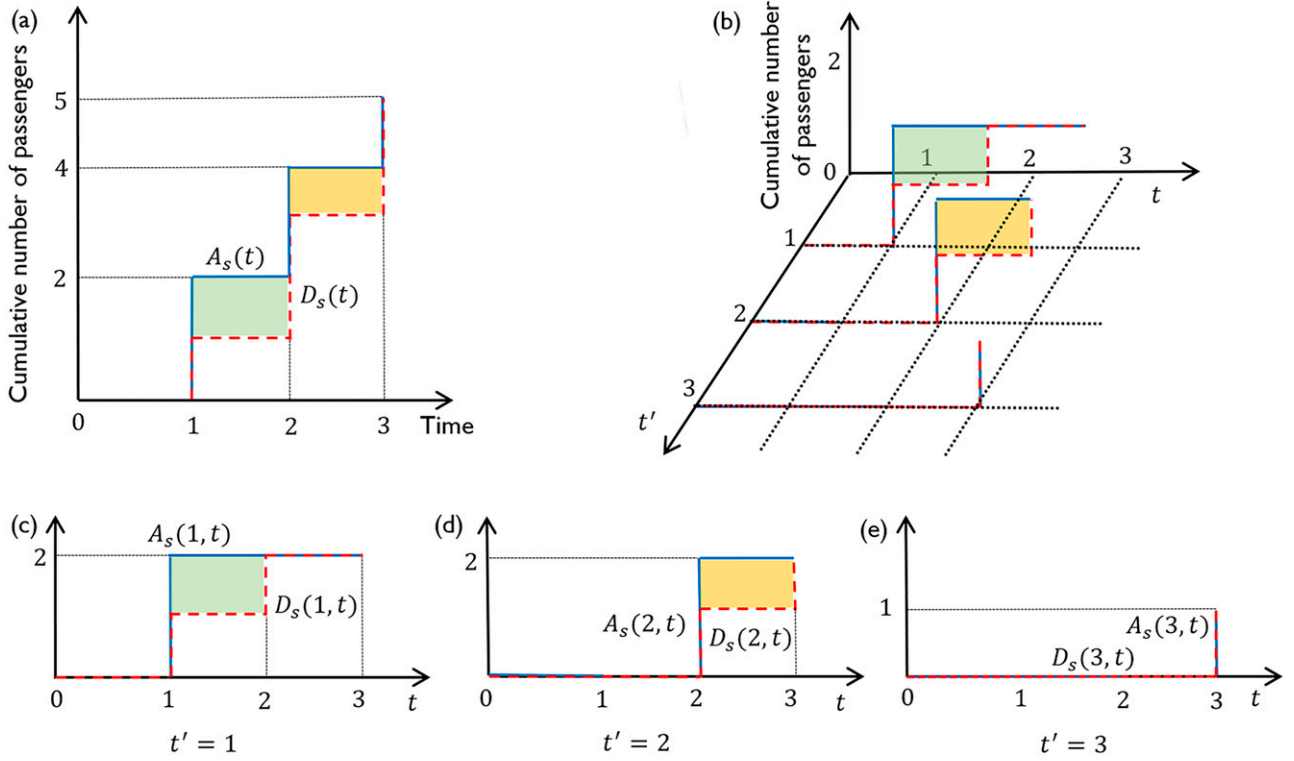
Notes. (a) The original cumulative arrival and departure curves. (b) The decomposed time-dependent cumulative arrival and departure curves. (c) The decomposed curves corresponding to arrival demand at time 1. (d) The decomposed curves corresponding to arrival demand at time 2. (e) The decomposed curves corresponding to arrival demand at time 3. Here the time horizon is discretized into three time intervals for illustrative purposes.

We further define the number of passengers who arrive at station s at time t' and board dispatch k as $d_{ks}(t') := D_s(t', t_k) - D_s(t', t_{k-1})$, $\forall k \in \mathcal{K}, s \in \mathcal{S}$, $t' \in [0, t_k]$. For dispatch k, all passengers arriving at station s no later than t_k can get on board as long as there is still enough capacity. Thus, $\int_0^{t_k} d_{ks}(t') dt'$ represents the number of passengers boarding dispatch k at station s. Note that the right derivative is used at points that are not differentiable (this applies throughout the paper). For example, for the second dispatch in Figure 3, we obtain $d_{2s}(1) = 1$, $d_{2s}(2) = 1$, and $d_{2s}(3) = 0$. Summing over all these values yields the number of passengers boarding at the second dispatch at stations, that is, two. Here, summation is used because the time horizon is discretized although, in a continuous time horizon, an integral of $d_{ks}(t')$ with respect to t' over zero to t_k is needed.

With these definitions, $\int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) dt$ is the total waiting time for passengers who arrive at station s at t'. Then, the integral of this term over t' from zero to T, that is,

$$\int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) dt dt',$$

is the total waiting time for all passengers arriving at station s during time horizon $[0, T]$. Thus, the total passenger waiting cost can then be formulated as

$$C_w(D_{st'}(t', t)) := w \sum_{s \in \mathcal{S}} \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) dt dt'. \tag{3}$$

Further, define e_{ks}, $\forall k \in \mathcal{K}, s \in \mathcal{S}$ as the number of passengers on board at dispatch k after passenger boarding and alighting at station s. With this, we formulate the objective function of the investigated problem as

$$\min_{K, [t_k, i_{ks}, e_{ks}, \forall k \in \mathcal{K}, s \in \mathcal{S}], \{d_{ks}(t'), D_s(t', t), \forall s \in \mathcal{S}, k \in \mathcal{K}, t' \in [0, T]\}} C_f(i_{ks})$$
$$+ C_w(D_s(t', t)), \tag{4}$$

where the first term represents the total vehicle operational cost across all dispatches and the second term the total passenger waiting cost across the operational horizon. To reflect the general operational details of UMT systems, the following four groups of constraints are considered.

i. Minimum headway requirement: These constraints impose a least-time separation between two

consecutive dispatched vehicles because of safety considerations and limited resources.

$$t_k - t_{k-1} \geq \underline{h}, \quad \forall k \in \mathcal{K} \setminus \{1\}, \tag{5}$$

ii. Departure curve conservation: These constraints formulate the cumulative departure conservation process mathematically with the time-dependent departure rate curve $\mathcal{D}_s(t', t)$. Constraints (6) define that $\forall t \in [t', T]$, $\mathcal{D}_s(t', t)$ increases by the number of boarding passengers if a vehicle is dispatched at t (i.e., $t = t_k$) and remains the value at the previous dispatch otherwise (i.e., $t \in (t_k, t_{k+1})$). Constraints (7) require $\mathcal{D}_s(t', t)$ to equal $A_s(t', t)$ at T, indicating that all passengers arriving at station s during the investigated time horizon are transported at the end of the operational horizon.

$$D_s(t', t) = \begin{cases} D_s(t', t_{k-1}) + d_{ks}(t') & \forall t = t_k, \\ D_s(t', t_k) & \forall t \in (t_k, t_{k+1}), \end{cases}$$
$$\forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T], \tag{6}$$

$$D_s(t', t) = A_s(t', t), \quad \forall t = T, s \in \mathcal{S}, t' \in [0, T]. \tag{7}$$

iii. Passenger boarding dynamics: These constraints model the passenger boarding process. Constraints (8) indicate passengers boarding at dispatch k at the start station may arrive at this station at any t' no later than t_k. At the other stations, the number of passengers on dispatched vehicle k equals the number of passengers on vehicle k at the previous station $s - 1$ plus the number of passengers boarding at station s minus the number of passengers alighting at station s. Constraints (9) are imposed because of the capacity limit of each vehicle formation. Constraints (10) require that the number of passengers boarding dispatch k at station s at time t should be no more than the passenger demand at station s at time t'. Constraints (11) describe the FIFS rule for passenger boarding. Note that $D_s(t', t) \leq A_s(t', t), \forall s \in \mathcal{S}, t', t \in [0, T]$, so $0 \leq \frac{D_s(t', t)}{A_s(t', t)} \leq 1, \quad \forall s \in \mathcal{S}, t', t \in [0, T]$. Therefore, for any $t' < t'' \in [0, T]$ at station $s \in \mathcal{S}$ when $\frac{D_s(t', t)}{A_s(t', t)} < 1, \forall t \in [0, T]$ (which means that passengers arriving at station s at time t' have not all boarded), $\frac{D_s(t'', t)}{A_s(t'', t)}, \forall t \in [0, T]$ must be zero as required by $\frac{D_s(t'', t)}{A_s(t'', t)} \leq \left\lfloor \frac{D_s(t', t)}{A_s(t', t)} \right\rfloor = 0$ (which means that no passengers arriving at station s at time $t'' > t'$ can board).

$$e_{ks} = \begin{cases} \displaystyle\int_{t'=0}^{t_k} d_{ks}(t')dt' & \text{if } s = 1, \forall k \in \mathcal{K} \\ \displaystyle e_{k(s-1)} + \int_{t'=0}^{t_k} d_{ks}(t')dt' - \sum_{u \in \mathcal{S}_s^-} \int_{t'=0}^{t_k} d_{ku}(t')p_{us}(t')dt' \\ \hspace{4cm} \forall s \in \mathcal{S} \setminus \{1\}, \forall k \in \mathcal{K} \end{cases} \tag{8}$$

$$e_{ks} \leq i_{ks}c, \quad \forall k \in \mathcal{K}, s \in \mathcal{S}, \tag{9}$$

$$d_{ks}(t') \leq A_s(t', t_k) - D_s(t', t_{k-1}), \quad \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, t_k], \tag{10}$$

$$\frac{D_s(t'', t)}{A_s(t'', t)} \leq \left\lfloor \frac{D_s(t', t)}{A_s(t', t)} \right\rfloor, \quad \forall t' < t'' \in [0, T], t \in [0, T], s \in \mathcal{S}, \tag{11}$$

iv. Variable domains: These constraints define the domains of all decision variables.

$$K \in \mathbb{Z}^+; t_k \in [0, T], \quad \forall k \in \mathcal{K}; i_{ks} \in \mathcal{I}, e_{ks} \in \mathbb{R}^+,$$
$$d_{ks}(t') \in \mathbb{R}^+, \quad \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T]. \tag{12}$$

For a realistic urban transportation corridor system, the defined problem is a very complex optimization problem that is difficult to solve to its exact optimum (optima) because of a large number of decision variables and its nonlinear problem structure. The purpose of the original model formulation presented here is to illustrate the complexity of the investigated problem and the formulation we try to approximate in the next section. However, there are other ways to formulate the investigated problem. For example, the problem can be formulated as a mixed integer programming model if the operational time horizon is discretized. A discrete mixed integer programming model, even in its linear form, involves a large amount of decision variables and requires excessive computational resources for solving an optimal or even a near-optimal solution (see Section 5.1.2 for numerical results). Thus, this study aims to tackle the investigated problem via a parsimonious continuous model with a continuous time representation. We focus on situations in which the passenger arrival demand rate varies relatively slowly (within a time scale comparable to the optimal vehicle dispatch headways) across the operational horizon at each station. This slowly varying property of passenger demand indicates that the total cost in a time neighborhood only depends on the dispatch decisions (i.e., dispatch headway and vehicle capacity) around this and nearby neighborhoods and is not much affected by the distant settings. This feature enables us to develop a continuum approximation (CA) approach to solve near-optimal solution(s) to the investigated problem very efficiently.

3. Theoretical Properties
In this section, we study the theoretical properties of the optimal solution(s) to the investigated problem. These properties offer critical insights into the (near-) optimal solution structure of the investigated problem and a theoretical foundation for solving the problem with the CA framework.

3.1. Review of Classic CA Models

We first review several classic CA models that are highly pertinent to the investigated problem to help the readers better understand the challenges in applying the CA method to the investigated problem. As pointed out previously, the first challenge in modeling the investigated problem with the CA method is the spatiotemporal correlation of dispatch decisions. To address this challenge, Newell (1971) aggregated the passenger demand at each station $s \in \mathcal{S}$ along a transportation route at any time $t \in [0, T]$ as

$$a(t) := \sum_{s \in \mathcal{S}} a_s(t), \; \forall t \in [0, T].$$

This way, the corridor system was converted to a one-to-one shuttle system in which passenger boarding and alighting dynamics are ignored. Then, Newell (1971) approximated the cost component in the converted shuttle system as

$$\sum_{k \in \mathcal{K}} \int_{t_{k-1}}^{t_k} \left(\frac{f_{Ic}}{h(t)} + \frac{wh(t)a(t)}{2} \right) dt = \int_0^T \left(\frac{f_{Ic}}{h(t)} + \frac{wh(t)a(t)}{2} \right) dt,$$

where $h(t)$ is the vehicle dispatch headway at $t \in [0, T]$ and the first and second terms represent the operational cost and passenger waiting cost, respectively. Note that only one vehicle formation is considered in Newell's (1971) model, and thus, without loss of generality, we use formation I in the preceding equation. To minimize this integral function is equivalent to minimizing its integrand, so the optimal headway, denoted as $h^*(t)$, can be analytically solved as

$$h^*(t) = \max \left\{ \sqrt{\frac{2f_{Ic}}{wa(t)}}, \frac{Ic}{a(t)} \right\}, \; \forall t \in [0, T].$$

However, this approximation implicitly ignores the passenger alighting pattern along the corridor so that $a(t)$ is usually greater than the actual seat demand. Taking the case in Figure 1 as an example, with Newell's (1971) model, we would use three as the passenger demand in the converted shuttle system, but the actual seat demand is two and one at stations 1 and 2, respectively. Thus, Newell's (1971) policy tends to provide a lower bound for the optimal dispatch headway (and upper bound for vehicle formation in this study) in a many-to-many corridor system. In light of this issue, Wirasinghe (1990) estimates the seat demand at the maximum load point, denoted as $s(t)$, with historical boarding and alighting counts from each dispatch in the existing schedule and revises Newell's (1971) policy as

$$h^*(t) = \max \left\{ \sqrt{\frac{2f_{Ic}}{wa(t)}}, \frac{Ic}{s(t)} \right\}, \; \forall t \in [0, T].$$

Simple though Wirasinghe's (1990) method is, it ignores the fact that the seat demand at a station is related to the current dispatch decisions rather than some historic schedules. For example, assume that solution 2 is used as the existing schedule for the system shown in Figure 1, which produces a seat demand of one at station 2. If the current decision at station 1 is to dispatch a vehicle with two modular pods, the actual seat demand at station 2 is two. Therefore, the actual seat demand values estimated with historic boarding and alighting counts are not necessarily true. Also note that using the estimated seat demand at the maximum load point, Wirasinghe's (1990) method essentially solves the problem for a transportation corridor by transforming it into a one-to-one shuttle system. Thus, the decision correlation between different stations in a transportation corridor is not effectively addressed in the existing CA literature.

The other challenge, adopting multiple vehicle formations at any station along the corridor, has not been considered in most CA studies. In the analytical formulas of Newell's (1971) and Wirasinghe's (1990) models, the passenger waiting cost is not associated with the vehicle capacity. These formulas hold when only unsaturated traffic and one vehicle formation are considered because all passengers are cleared after a vehicle leaves a station. However, in a system with unsaturated traffic and multiple vehicle formations as investigated in this paper, the number of passengers left behind at a station after a dispatch and the associated waiting cost are dependent on the vehicle formation decisions. Further, the operational cost is a given input independent of the dispatch decisions on vehicle formations in previous studies, rendering it inapplicable to design the optimal vehicle formations at each station for each dispatch. To the best of the research team's knowledge, the only CA model that has considered multiple vehicle formations in the operational design for transit systems is proposed in Chen, Li, and Zhou (2020). However, this preceding study focuses on shuttle systems in which decisions on vehicle formations only need to be made at the start station.

3.2. Theoretical Properties

To address the challenges in applying the CA model to tackle the investigated problem, this section investigates analytical properties of the optimal solution(s). We first rewrite Constraints (8) and (10). Integrating both sides of Constraints (10) over t' yields

$$\int_{t'=0}^{t_k} d_{ks}(t')dt' \leq \int_{t'=0}^{t_k} (A_s(t', t_k) - D_s(t', t_{k-1}))dt',$$
$$\forall k \in \mathcal{K}, s \in \mathcal{S}.$$

Applying this inequality to Constraints (8) yields

$$
e_{ks} \leq \begin{cases} \int_{t'=0}^{t_k} (A_s(t', t_k) - D_s(t', t_{k-1}))dt' & s = 1 \\ \qquad\qquad \cdot \\ e_{k(s-1)} - \sum_{u \in S_s^-} \int_{t'=0}^{t_k} d_{ku}(t') p_{us}(t')dt' \\ + \int_{t'=0}^{t_k} (A_s(t', t_k) - D_s(t', t_{k-1}))dt' & s > 1, \ \forall k \in \mathcal{K}. \end{cases}
$$
(13)

Further, we define the number of passengers left at station $s \in S$ right after a dispatch k as the passenger queue at station s, that is,

$$
q_s(t_k) := \left(\int_{t'=0}^{t_k} (A_s(t', t_k) - D_s(t', t_k))dt' \right), \ \forall s \in S,
$$

and the number of passengers who are left at any upstream station of s (inclusive) and destined to any station in the downstream direction of station s as the cross-sectional passenger queue at station s, that is,

$$
\tilde{q}_s(t_k) := \sum_{u \in S_{s+1}^-} \sum_{v \in S_s^+} \left(\int_{t'=0}^{t_k} (A_u(t', t_k) - D_u(t', t_k)) p_{uv}(t')dt' \right),
$$
$$
\forall s \in S.
$$

Define $\tilde{q}_s(t_0) = 0, \ \forall s \in S$ for the convenience of notation. With these, we next investigate properties of queued passengers after each dispatch in the optimal solution(s) in the following three propositions.

Proposition 1. *An optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\}$ to Problems (4)–(12) with time-dependent arrival rate curves $A_s(t', t), \ \forall s \in S$ and time-dependent departure rate curves $D_s(t', t), \ \forall s \in S$ must satisfy one of the following conditions for each dispatch $k \in \mathcal{K}$: (i) $t_k - t_{k-1} = \underline{h}$ or (ii) there exists at least one $s \in S$ such that $q_s(t_k) = 0$.*

Proof. This proposition is proven by contradiction. If the condition in this proposition does not hold, then, in this optimal solution, there exists a $\bar{k} \in \mathcal{K}$ with $t_{\bar{k}} - t_{\bar{k}-1} > \underline{h}$ and $q_s(t_{\bar{k}}) > 0, \ \forall s \in S$. Then, we can construct an alternate solution $\{\hat{K}, \hat{t}_k, \hat{i}_{ks}, \hat{e}_{ks}, \hat{d}_{ks}(t'), \hat{D}_s(t', t)\}$ with the following steps: (i) keep the number of dispatches and vehicle formations the same as those in the optimal solution, that is, $\hat{K} = K, \ \hat{i}_{ks} = i_{ks}, \ \forall k \in \mathcal{K}, s \in S$; and (ii) push the dispatch time of $\bar{k}$ forward until either $\hat{t}_{\bar{k}} - \hat{t}_{\bar{k}-1} = \underline{h}$ or we find an $\bar{s} \in S$ such that $\hat{q}_{\bar{s}}(t_{\bar{k}}) = 0$ while keeping $\hat{t}_k = t_k$ for all other k indexes. Note that, with this construction process, the resultant alternate solution obviously satisfies Constraints (5).

Because all decisions before dispatch $\bar{k}$ in the alternate solution are the same as those in the optimal solution, we have $\hat{e}_{ks} = e_{ks}, \ \forall k < \bar{k}, s \in S$, $\hat{d}_{ks}(t') = d_{ks}(t'), \ \forall k < \bar{k}, s \in S, t' \in [0, T]$, and

$$
\hat{D}_s(t', t) = D_s(t', t), \ \forall s \in S, t' \in [0, T], t \in [0, \hat{t}_{\bar{k}}).
$$

Next, we solve $\hat{e}_{ks}, \hat{d}_{ks}(t')$, and $\hat{D}_s(t', t)$ for time indexes since dispatch $\bar{k}$ in the alternate solution. Based on how we construct the alternate solution, we have

$$
\int_{t'=0}^{T} (A_1(t', \hat{t}_{\bar{k}}) - D_1(t', \hat{t}_{\bar{k}-1}))dt'
$$
$$
= \int_{t'=0}^{\hat{t}_{\bar{k}}} (A_1(t', \hat{t}_{\bar{k}}) - D_1(t', \hat{t}_{\bar{k}-1}))dt' \geq i_{\bar{k}1}c.
$$

This inequality together with Equations (9) and (13) (k replaced by $\bar{k}$) yield $\hat{e}_{\bar{k}1} = i_{\bar{k}1}c$. We apply this result into Equation (8) and then solve it with $\hat{D}_s(t', t) = D_s(t', t), \ \forall s \in S, t' \in [0, T], t \in [0, \hat{t}_{\bar{k}})$ (proved earlier) and Equations (10) and (11) (with $k = \bar{k}$), which results in $\hat{d}_{\bar{k}1}(t') = d_{\bar{k}1}(t'), \ \forall t' \in [0, T]$. Applying this equation into Equation (6) with $k = \bar{k}$, we further obtain $\hat{D}_1(t', \hat{t}_{\bar{k}}) = D_1(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}1}(t'), \ \forall t' \in [0, T]$. Next, at station 2, again, based on the way we construct the alternate solution, we obtain

$$
\hat{e}_{\bar{k}1} - \int_{t'=0}^{\hat{t}_{\bar{k}}} \hat{d}_{\bar{k}1}(t') p_{12}(t')dt'
$$
$$
+ \int_{t'=0}^{\hat{t}_{\bar{k}}} (A_2(t', \hat{t}_{\bar{k}}) - D_2(t', \hat{t}_{\bar{k}-1}))dt'
$$
$$
\geq i_{\bar{k}2}c.
$$

With the same analysis used earlier, we obtain $\hat{e}_{\bar{k}2} = i_{\bar{k}2}c, \ \hat{d}_{\bar{k}2}(t') = d_{\bar{k}2}(t'), \ \forall t' \in [0, T]$, and $\hat{D}_2(t', \hat{t}_{\bar{k}}) = D_2(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}2}(t'), \ \forall t' \in [0, T]$. Applying this analysis over stations 3 to S, we obtain $\hat{e}_{\bar{k}s} = i_{\bar{k}s}c, \ \forall s \in S$, $\hat{d}_{\bar{k}s}(t') = d_{\bar{k}s}(t'), \ \forall t' \in [0, T], s \in S$, and $\hat{D}_s(t', \hat{t}_{\bar{k}}) = D_s(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}s}(t'), \ \forall t' \in [0, T], s \in S$. Further, because no MAVs are dispatched until $\hat{t}_{\bar{k}+1}$, we obtain from Constraints (6) that

$$
\hat{D}_s(t', t) = \hat{D}_s(t', \hat{t}_{\bar{k}}) = D_s(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}s}(t') = D_s(t', t_{\bar{k}}), \ \forall s
$$
$$
\in S, t' \in [0, T], t \in [\hat{t}_{\bar{k}}, t_{\bar{k}}].
$$

Then, because $\hat{t}_k = t_k, \hat{i}_{ks} = i_{ks}, \ \forall k > \bar{k}, s \in S$, we obtain $\hat{e}_{ks} = e_{ks}, \ \forall k > \bar{k}, s \in S$, $\hat{d}_{ks}(t') = d_{ks}(t'), \ \forall k > \bar{k}, s \in S, t' \in [0, T]$, and

$$
\hat{D}_s(t', t) = D_s(t', t), \ \forall s \in S, t' \in [0, T], t \in (t_{\bar{k}}, T].
$$

These results indicate that the alternate solution also satisfies Constraints (6)–(12). Therefore, the alternate solution is feasible to the investigated problem.

Regarding the change in the objective value, the operational cost of the alternate solution is obviously the same as that of the optimal solution because $\hat{i}_{ks} = i_{ks}, \ \forall k \in \mathcal{K}, s \in S$. Further, the difference between

the passenger waiting cost of the optimal solution and that of the alternate solution is

$$
w\Bigg[\sum_{s\in\mathcal{S}}\int_{t'=0}^{T}\int_{t=0}^{T}(A_s(t',t)-D_s(t',t))dtdt'
$$

$$
-\sum_{s\in\mathcal{S}}\int_{t'=0}^{T}\int_{t=0}^{T}(A_s(t',t)-\hat{D}_s(t',t))dtdt'\Bigg]
$$

$$
=w\Bigg[\sum_{s\in\mathcal{S}}\int_{t'=0}^{T}\int_{t=0}^{T}(\hat{D}_s(t',t)-D_s(t',t))dtdt'\Bigg]
$$

$$
=w\Bigg[\sum_{s\in\mathcal{S}}\int_{t'=0}^{T}\int_{t=\hat{t}_{\tilde{k}}}^{t_{\tilde{k}}}(d_{\tilde{k}s}(t'))dtdt'\Bigg].
$$

$$
=w(t_{\tilde{k}}-\hat{t}_{\tilde{k}})\Bigg[\sum_{s\in\mathcal{S}}\int_{t'=0}^{T}(d_{\tilde{k}s}(t'))dt'\Bigg]>0.
$$

Thus, the objective value of the alternate solution is always strictly smaller than that of the optimal solution, which is a contradiction. This completes the proof. □

Proposition 2. *Any dispatch k in an optimal solution $\{K,t_k,i_{ks},e_{ks},d_{ks}(t'),D_s(t',t)\}$ to Problems (4)–(12) with time-dependent arrival rate curves $A_s(t',t)$, $\forall s\in\mathcal{S}$ and time-dependent departure rate curves $D_s(t',t)$, $\forall s\in\mathcal{S}$ must satisfy $\tilde{q}_s(t_k)<c$, $\forall s\in\mathcal{S}$ if the following conditions hold:*
  i. $S(f_2-f_1)<cw\underline{h}$.
  ii. $\sum_{u\in\mathcal{S}_{s+1}^-}\sum_{v\in\mathcal{S}_s^-}(\int_{t'=t_{k-1}}^{t_k}(A_u(t',t'))p_{uv}(t')dt')\le(I-1)c$, $\forall s\in\mathcal{S}$.

Proof. This proposition is proven with induction.
   Base case: First, investigate the case with $k=1$. If there exists a $\bar{s}\in\mathcal{S}$ such that $\tilde{q}_{\bar{s}}(t_1)\ge c$, we define the first origin station at which there are passengers in cross-sectional queue $\tilde{q}_{\bar{s}}(t_1)$ as

$$
s_1:=\operatorname*{argmin}_{u<\bar{s}\in\mathcal{S}}\Bigg\{\sum_{v\in\mathcal{S}_{\bar{s}}^+}\Bigg(\int_{t'=0}^{t_1}(A_u(t',t_1)-D_u(t',t_1))p_{uv}(t')\Bigg)>0\Bigg\},
$$

and the last destination station to which passengers in cross-sectional queue $\tilde{q}_{\bar{s}}(t_1)$ travel as

$$
s_2:=\operatorname*{argmax}_{v>\bar{s}\in\mathcal{S}}\Bigg\{\sum_{u\in\mathcal{S}_{\bar{s}+1}^-}\Bigg(\int_{t'=0}^{t_1}(A_u(t',t_1)-D_u(t',t_1))p_{uv}(t')dt'\Bigg)>0\Bigg\}.
$$

   Then, we construct an alternate solution $\{\hat{K},\hat{t}_k,\hat{i}_{ks},\hat{e}_{ks},\hat{d}_{ks}(t'),\hat{D}_s(t',t)\}$, where $\hat{K}=K$, $\hat{t}_k=t_k$, $\forall k\in\mathcal{K}$, $\hat{i}_{ks}=i_{ks}$, $\forall k>1,s\in\mathcal{S}$, and

$$
\hat{i}_{1s}=\begin{cases}i_{1s}+1 & \text{if }s\in[s_1,s_2]\\ i_{1s} & \text{otherwise}\end{cases},\ \forall s\in\mathcal{S}.
$$

That is, at the first dispatch, we increase the number of modular pods by one between s_1 and s_2 (inclusive) while keeping the vehicle formations at other

stations the same as those in the optimal solution. With condition (ii), we obtain $\hat{i}_{1s}\le I-1$, $\forall s\in\mathcal{S}$. Further, because the optimal solution satisfies Constraints (7) (i.e., all passengers are transported at T), the alternate solution increases the vehicle's capacity at the first dispatch and, thus, obviously also satisfies Constraints (7). Thus, the constructed vehicle formations are always feasible to the investigated problem.

   To analyze the change in the objective value, we first solve variables $\hat{e}_{1s}$, $\forall s\in\mathcal{S}$, $\hat{d}_{1s}(t')$, $\forall s\in\mathcal{S},t'\in[0,T]$, and $D_s(t',t)$, $\forall s\in\mathcal{S},t'\in[0,T],t\in[0,t_2)$. Based on the way we construct the alternate solution, these variables can be computed for three types of stations separately as follows.

   Type 1: $s\in[1,s_1)\cup(s_2,S]$. Because $\hat{t}_k=t_k,\hat{i}_{ks}=i_{ks}$, $\forall k\in\mathcal{K},s\in[1,s_1)\cup(s_2,S]$ and the cross-sectional queue at station $\bar{s}$ does not affect these stations, it is obvious that $\hat{e}_{1s}=e_{1s}$, $\forall s\in[1,s_1)\cup(s_2,S]$, $\hat{d}_{1s}(t')=d_{1s}(t')$, $\forall s\in[1,s_1)\cup(s_2,S],t'\in[0,T]$, and

$$
\hat{D}_s(t',t)=D_s(t',t),\ \forall s\in[1,s_1)\cup(s_2,S],t'\in[0,T],
$$
$$
t\in[0,t_2).
$$

   Type 2: $s\in[s_1,\bar{s}]$. Because $\tilde{q}_s(t_1)<c,\forall s\in[s_1,\bar{s})$ in the optimal solution, all passengers in the cross-sectional queue at station $s\in[s_1,\bar{s})$ can be accommodated by increasing the number of modular pods at this station by one, that is, $\hat{e}_{1s}=e_{1s}+\tilde{q}_s(t_1)$, $\forall s\in[s_1,\bar{s})$ (Constraints (9) and (13)). For $\bar{s}$, we obtain $\hat{e}_{1\bar{s}}=e_{1\bar{s}}+c$. To further compute $d_{ks}(t')$ and $D_s(t',t)$, we define

$$
t_{1s}^1:=\operatorname*{arginf}_{t'\in[0,T]}\{A_s(t',t_1)-D_s(t',t_1)>0\},\ \forall s\in\mathcal{S}
$$

to represent the smallest arrival time index of the passengers at station s who are not served by the first dispatch in the optimal solution (also the arrival time index of the first passenger boarding the additional modular pod at dispatch 1 in the alternate solution based on Constraints (11)). With this, we solve the arrival time index of the last passenger who boards the additional modular pods at dispatch 1 at station s in the alternate solution as

$$
t_{1s}^2:=\operatorname*{arginf}_{t''\in[t_{1s}^1,T]}\Bigg\{\int_{t'=t_{1s}^1}^{t''}(A_s(t',t_1)-D_s(t',t_1))dt'\ge o_{1s}\Bigg\},\ \forall s\in\mathcal{S},
$$

where

$$
o_{1s}:=\hat{e}_{1s}-\sum_{u\in\mathcal{S}_s^-}\sum_{v\in\mathcal{S}_s^+}\Bigg(\int_{t'=0}^{t_1}\Big(\hat{d}_{1u}(t')p_{uv}(t')\Big)dt'\Bigg)
$$

$$
-\Bigg(e_{1s}-\sum_{u\in\mathcal{S}_s^-}\sum_{v\in\mathcal{S}_s^+}\Bigg(\int_{t'=0}^{t_1}(d_{1u}(t')p_{uv}(t'))dt'\Bigg)\Bigg),\ \forall s\in\mathcal{S}
$$

represents the number of passengers boarding the additional modular pod at dispatch 1 at s. Then, with similar analysis as in Proposition 1, we obtain

$$\hat{d}_{1s}(t') = \begin{cases} d_{2s}(t') & \text{if } t' \in [t^1_{1s}, t^2_{1s}] \\ d_{1s}(t') & \text{otherwise} \end{cases}, \ \forall s \in [s_1, \bar{s}], t' \in [0, T].$$

$$\hat{D}_s(t', t) = \begin{cases} D_s(t', t_2) & \text{if } t' \in [t^1_{1s}, t^2_{1s}], t \in [t_1, t_2] \\ D_s(t', t) & \text{otherwise}, \end{cases}$$

$$\forall s \in [s_1, \bar{s}], t' \in [0, T], t \in [0, t_2].$$

Type 3: $s \in (\bar{s}, s_2]$. Similar to type 2 stations, the cross-sectional queue at this type of station can also be cleared in the alternate solution by a single modular pod. Thus, we have

$$\hat{e}_{1s} = e_{1s} + \sum_{u \in \mathcal{S}_s^-} \sum_{v \in \mathcal{S}_s^+} \left(\int_{t'=0}^{t_1} \left(\hat{d}_{1u}(t') p_{uv}(t') \right) dt' \right) + q_s(t_1),$$

$$\forall s \in (\bar{s}, s_2].$$

Then, with the same analysis as at type 2 stations, we obtain that the analytical formulas of $\hat{d}_{1s}(t')$ and $\hat{D}_s(t', t)$ for $s \in [s_1, \bar{s}]$ also hold for $s \in (\bar{s}, s_2]$.

From this discussion, we have that $\hat{D}_s(t', t_1) \geq D_s(t', t_1), \forall s \in \mathcal{S}, t' \in [0, T]$. Because $\hat{t}_k = t_k, \hat{i}_{ks} = i_{ks}, \forall k > 1, s \in \mathcal{S}$, we obtain $\hat{D}_s(t', t) \geq D_s(t', t), \forall s \in \mathcal{S}, t' \in [0, T], t \in (t_1, T]$ based on Constraints (8)–(10) (i.e., the same amount of passengers are transported at each dispatch, if not more, in the alternate solution because of the increased capacity at dispatch 1).

With this, the difference between the operational cost between the optimal solution and the alternate solution is

$$\triangle C_f := \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \left(f_{i_{ks}} - f_{\hat{i}_{ks}} \right) = \sum_{s \in [s_1, s_2]} \left(f_{(i_{1s})} - f_{(i_{1s}+1)} \right)$$

$$\geq S(f_1 - f_2).$$

The difference of the passenger waiting cost between the optimal and alternate solution is

$$\triangle C_w := w \left[\sum_{s \in \mathcal{S}} \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) \, dt \, dt' \right.$$

$$\left. - \sum_{s \in \mathcal{S}} \int_{t'=0}^{T} \int_{t=0}^{T} \left(A_s(t', t) - \hat{D}_s(t', t) \right) dt \, dt' \right]$$

$$\geq w \left[\sum_{s \in \mathcal{S}} \int_{t'=t^1_{1s}}^{t^2_{1s}} \int_{t=t_1}^{t_2} \left(\hat{D}_s(t', t) - D_s(t', t) \right) dt \, dt' \right]$$

$$= w \left[\sum_{s \in \mathcal{S}} \int_{t'=t^1_{1s}}^{t^2_{1s}} \int_{t=t_1}^{t_2} (D_s(t', t_2) - D_s(t', t_1)) \, dt \, dt' \right]$$

$$= w(t_2 - t_1) \left[\sum_{s \in \mathcal{S}} \int_{t'=t^1_{1s}}^{t^2_{1s}} (d_{2s}(t')) dt' \right] \geq wc(t_2 - t_1) \geq \triangle i_{1\bar{s}} cw \underline{h}.$$

Then, based on condition (i) in this proposition, we obtain that the difference between the objective function of the optimal and alternate solutions is

$$\triangle C_f + \triangle C_w \geq S(f_1 - f_2) + cw \underline{h} > 0.$$

This forms a contradiction. We repeat this process until $\tilde{q}_{\bar{s}}(t_1) < c$ is satisfied. Thus, $\tilde{q}_{\bar{s}}(t_1) < c, \forall s \in \mathcal{S}$ must hold for the base case.

Induction step: Assume that, for a $k \in \mathcal{K} \setminus \{K\}$, $\tilde{q}_s(t_k) < c, \forall s \in \mathcal{S}$. At time t_{k+1}, if there exists a $\bar{s} \in \mathcal{S}$ such that $\tilde{q}_{\bar{s}}(t_{k+1}) \geq c$, then similar to the base case, we can construct an alternate solution $\{\hat{K}, \hat{t}_k, \hat{i}_{ks}, \hat{e}_{ks}, \hat{d}_{ks}(t'), \hat{D}_s(t', t)\}$ by raising the number of modular pods at dispatch k to $i_{(k+1)s} + 1$ at the origin and destination stations of passengers in $\tilde{q}_s(t_k)$, and K and $\hat{t}_k, \forall k \in \mathcal{K}, \hat{i}_{ks}$ for other k, s indexes remain the same as those in the optimal solution. Then, with a similar analysis as that to prove the base case, we can show that the alternate solution is feasible to the investigated problem and that its objective value is always strictly lower than that of the optimal solution, which forms a contradiction. Therefore, $\tilde{q}_s(t_{k+1}) < c, \forall s \in \mathcal{S}$ also holds. This completes the proof. $\square$

Proposition 3. *Any dispatch k in an optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\}$ to Problems (4)–(12) with time-dependent arrival rate curves $A_s(t', t), \forall s \in \mathcal{S}$ and time-dependent departure rate curves $D_s(t', t), \forall s \in \mathcal{S}$ must satisfy $q_s(t_k) < c, \forall s \in \mathcal{S}$ if the following conditions hold:*

i. $S(f_2 - f_1) < cw\underline{h}$.

ii. $\sum_{u \in \mathcal{S}_{s+1}^-} \sum_{v \in \mathcal{S}_s^+} \left(\int_{t'=t_{k-1}}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (I-1)c$, $\forall s \in \mathcal{S}$.

Proof. Based on Proposition 2, $q_1(t_k) = \tilde{q}_1(t_k) < c$. Further, with the definitions of $\tilde{q}_s(t_k)$ and $q_s(t_k)$, we obtain

$$\tilde{q}_2(t_k) = \int_{t'=0}^{t_k} (A_1(t', t_k) - D_1(t', t_k)) p_{uv}(t') dt' + q_2(t_k)$$

$$< q_1(t_k) + q_2(t_k) < c,$$

which, together with $0 \leq q_1(t_k) < c$, yields $q_2(t_k) < c$. Then, this proposition can be easily proved by iteratively applying this relationship; that is, $\tilde{q}_s(t_k) < q_1(t_k) + q_2(t_k) + \cdots + q_s(t_k) < c$. This completes the proof. $\square$

Propositions 1–3 give necessary condition(s) for the optimal solution(s) to the investigated problem. Proposition 1 indicates that the dispatch headway should be the minimum usable value unless the passenger queue is zero at at least one station. Note that no special conditions are needed for Proposition 1, and thus, it is a general property for the investigated problem. Propositions 2 and 3 indicate that the number of passengers after each dispatch would be relatively small values (i.e., less than the capacity of a single modular pod) at each station. Although the

conditions in Propositions 2 and 3 are not universal, they enable us to draw analytical insights into the optimal solution structure that would otherwise be impossible. Further, the applicability domain of this theoretical result is quite wide in real-world cases. Condition (i) usually holds for common UMT systems because the passenger waiting cost is usually of a higher magnitude compared with the associated vehicle operational cost (Huang et al. 2017; Yin et al. 2017; Chen, Li, and Zhou 2019, 2020). Condition (ii) is generally approximately true under the unsaturated demand scenario. Even for cases in which the conditions are not satisfied, we found that the resultant approximation error of a continuous approximation model guided by these theoretical findings are relatively small. Indeed, the passenger queue at each station can be cleared (not just bounded by c) after each dispatch most of the time in the optimal solution(s) (see Section 5.1.3 for numerical results). Hence, the conditions stated in these propositions, albeit nonexhaustive, provide important analytical insights into the nature of the (near-)optimal solutions and a theoretical foundation for breaking the spatiotemporal correlation between dispatch decisions.

With these results, we further investigate properties of the optimal vehicle formations. Define the seat demand at time t at station s, denoted as $\tilde{a}_s(t)$, as those who would travel through the segment between station s and $s + 1$ at this time, that is,

$$\tilde{a}_s(t) := \sum_{u \in \mathcal{S}_{s+1}^-} \sum_{v \in \mathcal{S}_s^+} (A_u(t',t')p_{uv}(t)) = \sum_{u \in \mathcal{S}_{s+1}^-} \sum_{v \in \mathcal{S}_s^+} (a_u(t)p_{uv}(t)),$$

$$\forall s \in \mathcal{S}, t \in [0,T].$$

Note that this is different from the one defined in Wirasinghe (1990) because historic cumulative departure curves are not involved. Let $\tilde{a}_{ks}$, $\underline{i}_{ks}$, and $\bar{i}_{ks}$ be the seat demand, conditional lower-bound vehicle formation, and upper-bound vehicle formation for dispatch k at station s, respectively. For dispatch k, the actual seat demand at station s also includes the cross-sectional passenger queue of this station from the previous dispatch. Thus, the seat demand for dispatch k at station s can be formulated as $\tilde{a}_{ks} := \tilde{q}_s(t_{k-1}) + \int_{t_{k-1}}^{t_k} \tilde{a}_s(t)dt, \forall k \in \mathcal{K}, s \in \mathcal{S}$. With this, we define the conditional lower-bound vehicle formation as $\underline{i}_{ks} := \min\{\lfloor \frac{\tilde{a}_{ks}}{c} \rfloor, I\}$, which represents the longest vehicle formation whose capacity is smaller than the seat demand (i.e., the shortest one that results in $\tilde{q}_s(t_k) < c$). Likewise, we define the upper-bound vehicle formation as $\bar{i}_{ks} := \min\{\lceil \frac{\tilde{a}_{ks}}{c} \rceil, I\}$, which represents the shortest vehicle formation whose capacity is greater than the seat demand (i.e., the shortest one that results in $\tilde{q}_s(t_k) = 0$). With these definitions, the following result regarding the lower

bound to the optimal vehicle formation must hold because, otherwise, Proposition 2 is violated.

Proposition 4. *Any dispatch k in an optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t',t)\}$ to Problems (4)–(12) with time-dependent arrival rate curves $A_s(t',t)$, $\forall s \in \mathcal{S}$ and time-dependent departure rate curves $D_s(t',t)$, $\forall s \in \mathcal{S}$ must satisfy $\underline{i}_{ks} \leq i_{ks}$, $\forall s \in \mathcal{S}$ if the following conditions hold:*

i. $S(f_2 - f_1) < cw\underline{h}$.

ii. $\sum_{u \in \mathcal{S}_{s+1}^-} \sum_{v \in \mathcal{S}_s^+} \left(\int_{t'=t_{k-1}}^{t_k} (A_u(t',t'))p_{uv}(t')dt' \right) \leq (I-1)c$, $\forall s \in \mathcal{S}$.

Finally, the following proposition reveals the upper bound to the optimal vehicle formation in general.

Proposition 5. *An optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t',t)\}$ to Problems (4)–(12) must satisfy $i_{ks} \leq \bar{i}_{ks}$, $\forall k \in \mathcal{K}, s \in \mathcal{S}$.*

Proof. This proposition is also proven by contradiction. If the condition in this proposition does not hold, then there exist $\bar{k} \in \mathcal{K}$ and $\bar{s} \in \mathcal{S}$ such that $i_{\bar{k}\bar{s}} > \bar{i}_{\bar{k}\bar{s}}$ in this optimal solution. Note that $\bar{i}_{\bar{k}\bar{s}} < I$ (i.e., $\bar{i}_{\bar{k}\bar{s}} = \lceil \frac{\tilde{a}_{\bar{k}\bar{s}}}{c} \rceil$) because, otherwise, this optimal solution is not feasible. Then, we can construct an alternate solution $\{\hat{K}, \hat{t}_k, \hat{i}_{ks}, \hat{e}_{ks}, \hat{d}_{ks}(t'), \hat{D}_s(t',t)\}$, where $\hat{K} = K$, $\hat{t}_k = t_k$, $\forall k \in \mathcal{K}, \hat{i}_{\bar{k}\bar{s}} = \bar{i}_{\bar{k}\bar{s}}$, and $\hat{i}_{ks} = i_{ks}$ for all other k and s indexes. Because $\bar{i}_{\bar{k}\bar{s}} = \lceil \frac{\tilde{a}_{\bar{k}\bar{s}}}{c} \rceil$, we obtain $i_{\bar{k}\bar{s}}c > \hat{i}_{\bar{k}\bar{s}}c \geq \tilde{a}_{\bar{k}\bar{s}}$, meaning that both $\hat{i}_{\bar{k}\bar{s}}$ and $i_{\bar{k}\bar{s}}$ can transport all passengers who want to board dispatch $\bar{k}$ at station $\bar{s}$. Therefore, $\hat{e}_{ks} = e_{ks}$, $\forall k \in \mathcal{K}, s \in \mathcal{S}$, $\hat{d}_{ks}(t') = d_{ks}(t')$, $\forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0,T]$, $\hat{D}_s(t',t) = D_s(t',t)$, $\forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0,T]$, $t \in [0,T]$. Thus, the constructed alternate solution is feasible to the investigated problem. Further, this result also indicates that the passenger waiting costs are the same between the alternate and optimal solutions. Thus, the difference between the objective value of the optimal solution and that of the alternate solution is

$$\sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \left(f_{i_{ks}} - f_{\hat{i}_{ks}} \right) = f_{(i_{\bar{k}\bar{s}})} - f_{(\hat{i}_{\bar{k}\bar{s}})} > 0,$$

which forms a contradiction. This completes the proof. □

These two propositions offer an analytical solution to the upper bound and the conditional lower bound for the optimal vehicle formations and, thus, constitute a theoretical foundation for designing efficient discretization methods for the CA approach.

4. CA

This section presents the CA approach to obtaining near-optimal solutions to the investigated problem efficiently. We first present how we approximate the original problem in a relatively homogeneous setting and the resulting problem decomposition. Then,

analytical solutions to the decomposed problem and the discretization approach are discussed.

4.1. Model Formulation

With the theoretical findings in the previous section, we can reformulate the original Problems (4)–(12) under the CA framework. Given an optimal solution $o :=\{t_k, i_{ks}\}_{\forall k \in \mathcal{K}, s \in \mathcal{S}}$ to Problems (4)–(12) with time-dependent arrival rate curves $A_s(t', t)$ and time-dependent departure rate curves $D_s(t', t)$, $\forall s \in \mathcal{S}$, we define the dispatch headway at time $t \in [0, T]$ as $\hat{h}(t) := t_k - t_{k-1}$, (s.t.) $t \in [t_{k-1}, t_k), \exists k \in \mathcal{K}$, and the vehicle formation at station $s \in \mathcal{S}$ at time $t \in [0, T]$ as $\hat{i}_s(t) := i_{ks}$, s.t. $t \in [t_{k-1}, t_k), \exists k \in \mathcal{K}, \forall s \in \mathcal{S}$. For convenience of notation, we define $\hat{h}(t_K) := t_K - t_{K-1}$ and $\hat{i}_s(t_K) := i_{Ks}$, $\forall s \in \mathcal{S}$. With this, the objective value corresponding to optimal solution o can be formulated as

$$OP(o) := \sum_{s \in \mathcal{S}} \left(\sum_{k \in \mathcal{K}} f_{i_{ks}} + w \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) \, dt \, dt' \right)$$

$$= \sum_{s \in \mathcal{S}} \left(\int_0^T \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + w \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) \, dt \, dt' \right). \quad (14)$$

s.t. (5)–(12).

Propositions 1–3 and empirical experiments show that the passenger queue at each station is relatively small after each dispatch most of the time in the optimal solution(s) when only unsaturated traffic is present. This finding enables us to relax the FIFS rule when describing passenger boarding behavior. Therefore, we can replace the time-dependent arrival and departure rate curves by the original ones to compute the passenger waiting cost. This results in a first approximation to the original objective function (4) as

$$OP(o) \approx \sum_{s \in \mathcal{S}} \left(\int_0^T \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + w(A_s(t) - D_s(t)) \, dt \right)$$

$$= \sum_{s \in \mathcal{S}} \left(\sum_{k \in \mathcal{K}} \int_{t_{k-1}}^{t_k} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} + w(A_s(t) - D_s(t)) \right) dt \right).$$

Note that $A_s(t)$ is very close to $D_s(t)$ at each $t = t_k$, $\forall k \in \mathcal{K}$ (Propositions 1–3). Hence, we have $a_s(t) = \frac{A_s(t) - A_s(t_{k-1})}{t - t_{k-1}} \approx \frac{A_s(t) - D_s(t_{k-1})}{t - t_{k-1}}, t \in [t_{k-1}, t_k)$, $\forall k \in \mathcal{K}$, which yields

$$OP(o) \approx \sum_{s \in \mathcal{S}} \left[\sum_{k \in \mathcal{K}} \left(\int_{t_{k-1}}^{t_k} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + w \int_{t_{k-1}}^{t_k} (A_s(t) - D_s(t)) \, dt \right) \right]$$

$$\approx \sum_{s \in \mathcal{S}} \left[\sum_{k \in \mathcal{K}} \left(\int_{t_{k-1}}^{t_k} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + w \int_{t_{k-1}}^{t_k} (a_s(t)(t - t_{k-1})) \, dt \right) \right]. \quad (15)$$

Because $a_s(t)$ changes relatively slowly between two consecutive dispatches, we treat it as constant and pull it out of the integral, which yields

$$OP(o) \approx \sum_{s \in \mathcal{S}} \left[\sum_{k \in \mathcal{K}} \left(\int_{t_{k-1}}^{t_k} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + \frac{w a_s(t)}{2} (t_k - t_{k-1})^2 \right) \right]$$

$$= \sum_{s \in \mathcal{S}} \left[\sum_{k \in \mathcal{K}} \left(\int_{t_{k-1}}^{t_k} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} \right) dt + \frac{w a_s(t)}{2} \int_{t_{k-1}}^{t_k} (t_k - t_{k-1}) \, dt \right) \right]$$

$$\approx \int_0^T \left[\sum_{s \in \mathcal{S}} \left(\frac{f(\hat{i}_s(t))}{\hat{h}(t)} + \frac{w a_s(t) \hat{h}(t)}{2} \right) \right] dt. \quad (16)$$

The difference between (16) and (15) is

$$\sum_{s \in \mathcal{S}} \left[\sum_{k \in \mathcal{K}} \left(\int_{t_{k-1}}^{t_k} (A_s(t_{k-1}) - D_s(t_{k-1})) \, dt \right) \right] < cTS,$$

which is a relatively small value compared with $OP(o)$. Additionally, numerical experiments with exact solution methods show that $A_s(t_k) - D_s(t_k) = 0$ most of the time. Thus, we feel safe to say that this approximation mostly holds, particularly for large instances. Solving the investigated problem with an approximated objective function (16), however, is not much different from solving it with the original one because both $\hat{h}(t)$ and $\hat{i}_s(t)$, $\forall s \in \mathcal{S}$ are step functions. To further simplify this, we replace $\hat{h}(t)$ and $\hat{i}_s(t)$, $\forall s \in \mathcal{S}$ with continuous functions $h(t)$ and $i_s(t)$, $\forall s \in \mathcal{S}$, respectively, and obtain

$$OP(o) \approx \int_0^T \left(\sum_{s \in \mathcal{S}} \left(\frac{f(i_s(t))}{h(t)} + \frac{w a_s(t) h(t)}{2} \right) \right) dt, \quad (17)$$

where the operational cost of continuous vehicle formation $i_s(t) \in [0, I]$ is defined as $f(i_s(t)) := f_i$, s.t. $i_s(t) \in (i-1, i], \exists i \in \mathcal{I}, \forall t \in [0, T], s \in \mathcal{S}$. For convenience of notation, define $f(0) := 0$. To ensure that the approximation is accurate, the following relationship should hold:

$$\int_0^T \left(\hat{h}(t) \right) dt \approx \int_0^T (h(t)) \, dt,$$

$$\int_0^T \left(\hat{i}_s(t) \right) dt \approx \int_0^T (i_s(t)) \, dt, \ \forall s \in \mathcal{S}.$$

Until now, the spatiotemporal correlation between dispatch decisions is eliminated, and we can solve the problem for each time point independently. However, the approximated objective function (17) still has $S + 1$ number of decision variables for which the solution approach might not be very simple. Thus, we need to further simplify the model formulation. Specifically, we obtain that, for $t \in [t_{k-1}, t_k), \forall k \in \mathcal{K}, i_s(t)c \approx \hat{i}_s(t)c = \hat{i}_s(t_k)c \approx \int_{t_{k-1}}^{t_k} \tilde{a}_s(t) dt \approx (t_k - t_{k-1})\tilde{a}_s(t) = h(t)\tilde{a}_s(t)$. Applying

this relationship to Equation (17) yields the final approximate objective function with only one decision variable as

$$OP(o) \approx \int_0^T \left[\sum_{s \in \mathcal{S}} \left(\frac{f\left(\frac{h(t)\tilde{a}_s(t)}{c}\right)}{h(t)} + \frac{wa_s(t)h(t)}{2} \right) \right] dt. \quad (18)$$

Clearly, the $h(t)$ value that minimizes Equation (18) also minimizes the integrand at every $t \in [0, T]$, so the original problem can be decomposed across the operational horizon as the following set of subproblems, each for one time point t.

$$c^*(t) := \min_{h(t)} c_t(h(t)) := \sum_{s \in \mathcal{S}} \left(\frac{f\left(\frac{h(t)\tilde{a}_s(t)}{c}\right)}{h(t)} + \frac{wa_s(t)h(t)}{2} \right),$$

$$\forall t \in [0, T].$$

This is s.t.

$$\underline{h} \leq h(t) \leq \frac{Ic}{\tilde{a}_s(t)}, \quad \forall s \in \mathcal{S}, \ t \in [0, T]. \quad (20)$$

Constraints (20) are imposed because of the minimum headway constraints and the vehicle capacity constraints in the original problem. These unit-time problems at each time point $t \in [0, T]$ have only one decision variable $h(t)$ and two constraints, and they can be independently solved analytically at each time $t \in [0, T]$. Thus, these single-variable optimization problems are much simpler to solve than the original problem. After solving the unit-time problems across the entire operational horizon, we can apply $c^*(t)$ to the following equation to obtain an approximate objective value to optimal solution o:

$$OP(o) \approx \int_0^T c^*(t) dt. \quad (21)$$

4.2. Analytical Solution

This section presents an analytical solution approach to the unit-time problems (19) and (20). The proposed solution approach first divides the feasible region of $h(t)$ into a finite number of subregions, in each of which a local minimum to the objective function (19) exists. Then, the local minimum in each subregion can be analytically solved. Finally, a comparison between all local minima can easily lead us to the global minimum of the unit-time problems. The following discussion applies to each $t \in [0, T]$, but we omit the index $\forall t \in [0, T]$ in the formulas for convenience of notation.

We first present how to divide the feasible region into a finite number of subregions in which $c_t(h(t))$ is strictly unimodal as shown in Figure 4. According to the definition of $f(i_s(t))$ in the previous section, for each station, $s \in \mathcal{S}, f\left(\frac{h(t)\tilde{a}_s(t)}{c}\right)$ is a step function specified by a set of discrete points (joints between pairs of

Figure 4. (Color online) Subregion Division for the Feasible Region of $h(t)$ for Unit-Time Problems
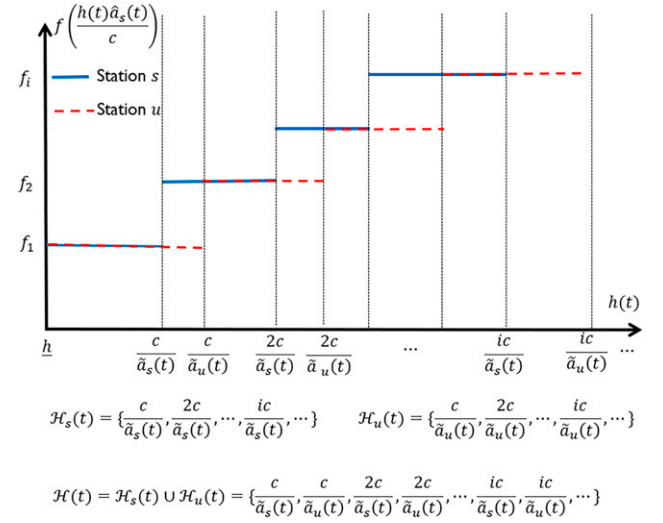
consecutive pieces of the function) denoted as $\mathcal{H}_s(t) := \left\{ \frac{ic}{\tilde{a}_s(t)} \mid \underline{h} \leq \frac{ic}{\tilde{a}_s(t)} \leq \min\left\{ \frac{Ic}{\tilde{a}_s(t)}, T \right\}, \forall i \in \mathcal{I} \right\}$, where the elements are ascendingly ordered. With this, we can divide the feasible region for all possible headway values into a finite number of subregions with a set of segregation points denoted as an ascendingly ordered set $\mathcal{H}(t) := \cup_{s \in \mathcal{S}} \mathcal{H}_s(t)$. For convenience of notation, denote the lth element in $\mathcal{H}(t)$ as $h_l(t)$ in the following analysis. Define $X_{l-1}(t) := \sum_{s \in \mathcal{S}} f\left(\frac{h(t)\tilde{a}_s(t)}{c}\right) > 0$, $Y(t) := \sum_{s \in \mathcal{S}} \frac{w\tilde{a}_s(t)}{2} \geq 0$, which are both constant over $h(t) \in (h_{l-1}(t), h_l(t)]$ based on the definition of the step function. Then, $c_t(h(t)) = X_{l-1}(t)/h(t) + Y(t)h(t)$ is a strictly convex economic order quantity (EOQ) function of $h(t)$. Solving $c_t'(h(t)) = 0$ yields $h(t) = \sqrt{\frac{X_{l-1}(t)}{Y(t)}}$, and thus, there are three cases regarding the monotonicity of $c_t(h(t))$ in subregion $(h_{l-1}(t), h_l(t)], \forall l \in [2, \cdots, |\mathcal{H}(t)|]$ as follows:

Case 1. If $h_l(t) < \sqrt{\frac{X_{l-1}(t)}{Y(t)}}$, then $c_t(h(t))$ is strictly decreasing over $(h_{l-1}(t), h_l(t)]$. In this case, the local minimum is achieved at $h(t) = h_l(t)$.

Case 2. If $h_{l-1}(t) > \sqrt{\frac{X_{l-1}(t)}{Y(t)}}$, then $c_t(h(t))$ is strictly increasing over $(h_{l-1}(t), h_l(t)]$. In this case, the local minimum is obtained at $h(t) = h_{l-1}(t)$ (which is rolled back to the previous piece of the step function with a slight abuse of notation).

Case 3. If $h_{l-1}(t) \leq \sqrt{\frac{X_{l-1}(t)}{Y(t)}} \leq h_l(t)$, $c_t(h(t))$ is strictly decreasing $\forall h(t) \in \left(h_{l-1}(t), \sqrt{\frac{X_{l-1}(t)}{Y(t)}} \right)$ and then strictly increasing $\forall h(t) \in [\sqrt{\frac{X_{l-1}(t)}{Y(t)}}, h_l(t)]$. In this case, the local minimum is obtained at $h(t) = \sqrt{\frac{X_{l-1}(t)}{Y(t)}}$.

This yields an analytical solution to the optimal $h(t)$ in subregion $(h_{l-1}(t), h_l(t)]$ denoted as $h_{l-1}^*(t)$ as follows

$$h_{l-1}^*(t) = \begin{cases} h_{l-1}(t), & \text{if } \sqrt{\dfrac{X_{l-1}(t)}{Y(t)}} < h_{l-1}(t); \\[2ex] \sqrt{\dfrac{X_{l-1}(t)}{Y(t)}}, & \text{if } h_{l-1}(t) \le \sqrt{\dfrac{X_{l-1}(t)}{Y(t)}} \le h_l(t), \\ & \forall l \in [2, \ldots, |\mathcal{H}(t)|]; \\[2ex] h_l(t), & \text{if } \sqrt{\dfrac{X_{l-1}(t)}{Y(t)}} > h_l(t). \end{cases} \quad (22)$$

Then, we can obtain the local minimum in subregion $[h_{l-1}(t), h_l(t)]$, denoted as $c_{l-1}^*(t)$, by applying Equation (22) to $c_t(h(t))$, that is, $c_{l-1}^*(t) = c_t(h_{l-1}^*(t))$. With the local minima in all subregions $[h_{l-1}(t), h_l(t)]$, $\forall l \in [2, \cdots, |\mathcal{H}(t)|]$, the global minimum over the entire feasible region can be obtained as

$$c^*(t) = \min_{l \in [2, \ldots, |\mathcal{H}(t)|]} c_{l-1}^*(t),$$

with the minimizer as

$$l^*(t) = \operatorname*{argmin}_{l \in [2, \ldots, |\mathcal{H}(t)|]} c_{l-1}^*(t).$$

This yields the optimal headway as

$$h^*(t) = h_{l^*(t)}^*(t),$$

and the optimal vehicle formation as

$$i_s^*(t) = \frac{h^*(t)\tilde{a}_s(t)}{c}, \quad \forall s \in \mathcal{S}.$$

Finally, after solving the unit-time problems $\forall t \in [0, T]$, we can plug $c^*(t)$, $\forall t \in [0, T]$ into Equation (21) to obtain an approximate value of $OP(\boldsymbol{o})$.

4.3. Discretization Issues

The continuous solution $\{h^*(t), i_s^*(t)\}$, $\forall s \in \mathcal{S}, t \in [0, T]$ obtained from the analytical solution approach approximates the optimal dispatch headways and vehicle formations and allows us to find an near-optimal total cost without explicitly solving the original problem. However, this continuous solution cannot be directly applied for discrete vehicle dispatching decisions. A discretization method is needed to convert $h^*(t)$ into discrete time points for each dispatch t_k^* and $i_s^*(t)$ into discrete vehicle formation i_{ks}^*. We first discuss the method to discretize $h^*(t)$. Several methods are proposed to find an approximate step function to the continuous headway function $h^*(t)$ in the literature, such as Daganzo (2005) and Chen, Li, and Zhou (2020). Here, we opt to adopt the one presented in Chen, Li, and Zhou (2020) because of its simplicity and computational efficiency. This method essentially iteratively applies the definition of headway, that is, $h^*(t_{k-1}^*) = t_k^* - t_{k-1}^*$, $\forall k = K, K-1, \cdots, 1$. Specifically, because a vehicle must be dispatched at T (otherwise Constraints (7) are violated), this algorithm sets T as

the time for the last dispatch, that is, $t_K^* = T$. Starting from the point $(t = T, h^*(t) = 0)$, this algorithm draws a 45° line backward in time and find its intersection with $h^*(t)$, which locates the time for the previous dispatch, that is, t_{K-1}^*, as the abscissa of this intersection point. This process is then repeated from t_{K-1}^* to locate t_{K-2}^* from t_{K-2}^* to t_{K-3}^* until we reach a point $t_k^* < \underline{h}$. Note that this algorithm also determines the values of K^* and $\mathcal{K}^*$.

Once the discrete time points t_k^*, $\forall k \in \mathcal{K}^*$ are obtained, we can move on to discretize $i_s^*(t)$, $\forall s \in \mathcal{S}$. Chen, Li, and Zhou (2020) compute the weighted average of all vehicle formations in time interval $[t_{k-1}^*, t_k^*]$ and round it to an integer as the vehicle formation for dispatch k. For shuttle systems investigated in Chen, Li, and Zhou (2020), the discretized results from this approach are not distant from the continuous values because discrete vehicle formations are used in solving the unit-time problems. However, in this study, discrete vehicle formations $\hat{i}_s(t)$, $\forall s \in \mathcal{S}$ are approximated by a continuous function $i_s(t)$, $\forall s \in \mathcal{S}$, whose original step function (i.e., $\hat{i}_s(t)$) cannot be well estimated by the weighted average method. Thus, here, we propose a greedy heuristic to discretize $i_s^*(t)$, $\forall s \in \mathcal{S}$ to improve the discretization accuracy. Note that, as long as the dispatch times are determined, the vehicle formation of each station can be determined independently. Thus, the following analysis can be applied to all $s \in \mathcal{S}$, but we omit this index for convenience of notation. Propositions 4 and 5 indicate that, for each dispatch k at station s, the optimal vehicle formation should be selected between the conditional lower bound vehicle formation $\underline{i}_{ks}$ and upper bound vehicle formation $\bar{i}_{ks}$. For $k \in \mathcal{K}^* \setminus \{K^*\}$, if $(\tilde{a}_{ks} - \underline{i}_{ks}c)w(t_k^* - t_{k-1}^*) > f_{\bar{i}_{ks}} - f_{\underline{i}_{ks}}$, indicating that the increase in the passenger waiting cost by using the conditional lower bound vehicle formation cannot be offset by the decrease in the operational cost, we select the upper bound vehicle formation, that is, $i_{ks}^* = \bar{i}_{ks}$. Otherwise, the conditional lower bound vehicle formation is used, that is, $i_{ks}^* = \underline{i}_{ks}$. For the last dispatch K, the upper bound should always be used, that is, $i_{Ks}^* = \bar{i}_{Ks}$, because Constraints (7) are violated otherwise.

These discretization methods can be formally stated as the pseudocode in Algorithm 1.

Algorithm 1 (Discretization)
  Input: $h^*(t), i_s^*(t)$, $\forall s \in \mathcal{S}, t \in [0, T], \underline{h}$
  Output: t_k^*, i_{ks}^*, $\forall k \in \mathcal{K}$, $\forall s \in \mathcal{S}$
  1: $K \leftarrow 0$
  2: $t_K^* \leftarrow T$
  3: $\mathcal{K} \leftarrow \{K\}$
  4: **while** $t_K^* \ge \underline{h}$ **do**
  5:    $K \leftarrow K + 1$
  6:    $\mathcal{K} \leftarrow \mathcal{K} \cup \{K\}$
  7:    $t_k^* \leftarrow \operatorname*{argsup}_{t \in [0, t_{k-1}^*]} \{h^*(t) = t_{K-1}^* - t\}$

8: **end while**

9: $i^*_{Ks} \leftarrow \left\lceil \dfrac{\int_{t^*_{K-1}}^{t^*_K} \tilde{a}_s(t)\,dt}{c} \right\rceil, \; \forall s \in \mathcal{S}$

10: **for** k in $\mathcal{K} \setminus \{K\}$ **do**

11:    **for** s in $\mathcal{S}$ **do**

12:     $\tilde{a}_{ks} \leftarrow \int_{t^*_{k-1}}^{t^*_k} \tilde{a}_s(t)\,dt, \underline{i}_{ks} \leftarrow \lfloor \frac{\tilde{a}_{ks}}{c} \rfloor, \bar{i}_{ks} \leftarrow \lceil \frac{\tilde{a}_{ks}}{c} \rceil$

13:     **if** $(\tilde{a}_{ks} - \underline{i}_{ks}c)w(t^*_k - t^*_{k-1}) > (f_{\bar{i}_{ks}} - f_{\underline{i}_{ks}})$ **then**

14:      $i^*_{ks} \leftarrow \bar{i}_{ks}$

15:     **else**

16:      $i^*_{ks} \leftarrow \underline{i}_{ks}$

17:     **end if**

18:    **end for**

19:   **end for**

20: Reverse $(i^*_{ks}, t^*_k) = (i^*_{K-k,s}, t^*_{K-k}), \; \forall k \in \mathcal{K} := \{0, 1, \cdots, K\}$

5. Numerical Experiments

This section assesses the proposed CA model with two sets of numerical experiments. The first set of experiments is built on a hypothetical transit corridor in Mandl's (1980) network to evaluate the computational performance of the proposed CA model, verify the theoretical properties, and analyze the system performance with varying input parameters. The second set of experiments applies the CA model to design future MAV service with realistic travel demand data collected from the Batong line in the Beijing Subway system. The purpose of these experiments is to investigate the applicability of the CA model to solve real-world problems and to further reveal some interesting managerial insights. All experiments are run on a DELL Studio PC with 3.60 GHz of Intel Core 17-7700 CPU and 16 GB RAM in a Windows environment. All the algorithms are implemented in Visual C++ 2015.

5.1. Case Study 1: A Hypothetical Transit Corridor in Mandl's (1980) Network

5.1.1. Experimental Settings. We first explain how we set up the numerical experiments based on a hypothetical transit corridor in Mandl's (1980) network. As shown in Figure 5, Mandl's (1980) network includes a group of cities in Switzerland, consisting of 15 nodes and 21 bidirectional links. Without loss of generality, we select 11 nodes (the nodes with shaded circles) to construct a unidirectional transportation corridor starting at node 1 and terminating at node 5 (with the direction of each link indicated by the arrows). Because we do not have access to the historic spatiotemporal travel demand data for this network, we use the transit OD demand matrix during the peak hour in Arbex and da Cunha (2015) to simulate multiple time-dependent OD matrices as follows.

i. For an OD pair, we compute the arrival demand rate per minute based on the peak-hour demand in Arbex and da Cunha (2015) and use it as an input to

Figure 5. (Color online) Mandl's Network and the Investigated Corridor System
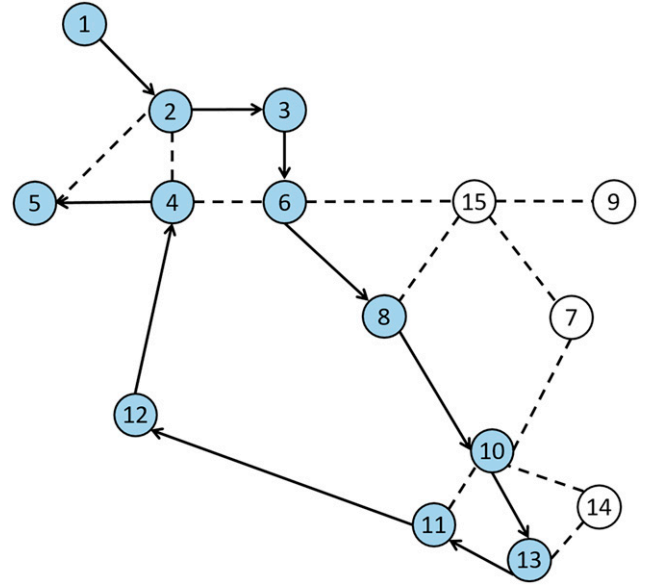
simulate the passenger arrival process as a Poisson process. The simulation process determines the arrival time of each passenger, based on which we construct the time-dependent OD matrices for this OD pair during the peak hour.

ii. To estimate the travel demand for other periods within the day, we multiply the OD demand matrix in Arbex and da Cunha (2015) by the ratio of the time-of-day factor of the target period to that of the peak hour. Then, we repeat the first step to construct the time-dependent OD matrices for this target period.

We assume that the constructed corridor is served by MAVs that can dock and undock at any station along the corridor. To make sure all passengers can be served at the end of the operational horizon with a minimum dispatch headway of $\underline{h} = 3$ minutes, we set $\mathcal{I} := [1, 2, 3]$ and $c = 50$ passengers/pod. Further, because no empirical data regarding the operational cost of MAVs are available, we simply adopt the calibrated function from Chen, Li, and Zhou (2020), that is, $f_i = C^F + C^V(i)^\alpha = 1.912 + 29.50i, \; \forall i \in \mathcal{I}$. Besides this, the unit-time waiting cost per passenger is set as $w = \$0.8/\text{minute}$.

5.1.2. Computation Performance of the CA Model. This section examines the computation performance of the proposed CA model. To investigate how the computation performance varies with the instance size, we design 36 instances with different numbers of origin stations (i.e., 2–10) and lengths of the operational horizon (i.e., 0.5, 1, 1.5, and 2 hours). The proposed CA approach is compared with a discrete model

solved by an existing commercial solver, Gurobi (Chen and Li 2021). In the CA model, we compute the objective values in two ways. The first approach, called CA-I, obtains the objective value directly from the continuous analytical solutions. The second approach, CA-D, feeds the discretized solutions into the objective function of the discrete model to compute the objective value. To quantitatively measure the difference between the solution from the CA model (denoted as V_{CA}) and that from the discrete model (denoted as V_{GU}), we compute the relative gap between their objective values as $\frac{V_{CA}-V_{GU}}{V_{GU}} \times 100$. We also record the computation time of each solution method to compare their computation speed. Note that we set the solution time limit of Gurobi as one hour; that is, if Gurobi cannot solve the model to optimality within an hour, the program is terminated. The relative gaps and computation time of all instances are summarized in Table 1, in which C, C_f, and C_w represent the total system cost, operational cost, and total passenger waiting cost, respectively. In the first column in Table 1, we denote an instance as "IN–number of origin stations–length of the operational horizon." For example, IN-2-0.5 represents an instance with two origin stations and an operational horizon of 0.5 hour. Also, note that there are two values in the eighth column in Table 1 with the ones outside the brackets representing the time for Gurobi to construct and solve the model and those inside the brackets representing the solution time only. Finally, when Gurobi cannot solve the model to optimality, either a feasible solution is returned, indicated by a solution time of 3,600 seconds, or not even a feasible solution can be obtained, indicated by a slash in the eighth column in Table 1.

We see from Table 1 that the relative gaps in the total system cost from CA-I can be positive or negative, indicating that CA-I can either overestimate or underestimate the optimal objective value. This is because CA-I produces the results directly from the continuous analytical solutions, which can be either feasible or infeasible to the investigated problem. Despite the signs of the relative gaps in total system costs from CA-I, their absolute values are all under 4%, and only two of them are above 3%. These results indicate that the CA-I estimates the total cost of transit corridors with good accuracy even without discretization. As for the two cost components, the estimation errors are a bit higher in some instances (e.g., instances with two origin stations), but most of them are still under 4%. For those with a relatively large relative gap in one component, the other cost component usually has a relatively small gap, and thus, the two cost components offset each other so that the estimated total system cost is still accurate. For example, in instance IN-2-0.5 (the instance with two origin stations and an operational horizon of 0.5 hour), the relative error in the

operational cost is 9.17%, but this deviation is balanced with a small relative error in the waiting cost of −1.44%, resulting in a relative error in the total cost of only 3.53%. For CA-D, all relative gaps are positive except the one from IN-2-2 because the discretization method ensures that all solutions are feasible. Surprisingly, in instance IN-2-2 (the instance with two stations and an operational horizon of two hours), CA-D produces an objective value lower than that from Gurobi (note that neither is an optimal solution), indicating that sometimes CA-D produces near-optimal solutions with higher accuracy than Gurobi does. Further, the relative gaps in the total cost are all below 5%, and most of them are under 2%, indicating the CA-D can produce near-optimal solutions to the investigated problem with high accuracy. The relative gaps in the two cost components show a similar trend as those from CA-I. Overall, these results verify the near-optimality of the CA model. Also, we observe that the relative gap (i.e., the accuracy of the CA model) generally decreases as the instance size increases, meaning the CA model is appropriate for large-scale problem instances in the real world.

Regarding the computation speed, we can see that Gurobi cannot solve most of the instances to optimality and could not even provide a feasible solution in many instances given the solution time limit of an hour. Further, it took Gurobi hours to construct the model before solving it (e.g., slightly over an hour in IN-3-1.5). However, the CA models could tackle all instances with astonishing efficiency, that is, within 10 ms. Besides, we can also see that the computation time of Gurobi increases dramatically as the instance size increases although that of the CA models are relatively stable. This is because the investigated problem is NP-hard (as its special case, the classic timetabling problem, is proven NP-hard; Caprara, Fischetti, and Toth 2002) and the discrete model involves a large number of decision variables and constraints as the problem size grows. Whereas the CA model only needs to solve a finite number of unit-time problems, and thus, has only linear time complexity. These results show that the CA models are also attractive from the perspective of computational efficiency.

Next, we investigate how the computational performance of the CA model varies when the relationship between the passenger waiting cost and operational cost changes. To this end, we change the value of unit-time waiting cost per passenger w in IN-2-0.5 and IN-4-0.5 and solve the model with Gurobi, CA-I, and CA-D again. These two instances are selected because they produce the largest and smallest relative gaps by CA-D among all instances whose discrete models can be solved to optimality in the preceding experiments. Note that the operational cost f_i can be changed for experiments as well, but here, we simply select w for

illustrative purposes. The results from these experiments are summarized in Table 2. As we can see from Table 2, the relative gaps of the two cost components C_f and C_w are more evident (e.g., 20.41% from CA-D in IN-4-0.5), but the gap of the total system cost C still remains relatively small (less than 4% from CA-I and less than 6% from CA-D, respectively). The average of the absolute values of the relative gaps from CA-I and CA-D are 2.24% and 2.67%, respectively, which are much smaller than the relevant deviation of w (62.5%). These results indicate that the proposed CA approaches still produce near-optimal solutions with high accuracy even when the parameter variation is drastic. This finding is consistent with those from several previous studies using the CA model (e.g., Li et al. 2016). Further, we also observe from Table 2 that the computation time of Gurobi changes

substantially as w changes because the number of nodes that Gurobi has to explore is case sensitive (i.e., dependent on values of the input parameters). In contrast, the number of unit-time problems that the CA approaches has to solve is fixed for one particular instance regardless of the parameter variation. Thus, the solution time of the CA model does not change with the parameter variation, verifying the robustness of the computation speed of the CA model.

5.1.3. Results Verification. This section reports some numerical results to further verify the validity of the CA model as well as the theoretical properties. We first present the optimal operational plans from Gurobi and near-optimal operational plans from CA-D of four selected instances in Figure 6. Next, the number of passengers left at the stations after each dispatch in

Table 1. Computational Performance of the CA Model Under the Default Parameter Setting

	Relative gap, %/Objective value (10^3)						Computation time, s		
	CA-I			CA-D					
Instance	C	C_f	C_w	C	C_f	C_w	Gurobi	CA-I	CA-D
IN-2-0.5	3.53	9.17	−1.44	4.12	5.67	2.76	71.77 (45.60)	0.001	0.001
IN-3-0.5	3.12	0.14	6.44	3.26	3.57	2.90	584.45 (526.63)	0.001	0.001
IN-4-0.5	2.15	2.73	1.53	1.69	4.84	−1.64	280.21 (177.21)	0.001	0.001
IN-5-0.5	1.39	0.82	2.09	1.83	5.39	−2.55	298.02 (123.18)	0.001	0.001
IN-6-0.5	1.21	0.02	2.56	1.73	4.23	−1.08	417.18 (163.45)	0.001	0.001
IN-7-0.5	0.58	−0.32	1.74	1.15	2.98	−1.21	1,772.18 (1,433.07)	0.001	0.002
IN-8-0.5	0.52	−1.00	2.65	1.24	2.67	−0.75	2,858.44 (2,432.32)	0.001	0.002
IN-9-0.5	0.23	−2.03	3.63	1.19	1.47	0.76	4,136.26 (3,600)	0.001	0.002
IN-10-0.5	0.02	−2.90	4.61	1.69	0.00	4.36	4,259.64 (3,600)	0.002	0.002
IN-2-1	2.68	8.16	−2.14	3.14	2.87	3.38	3,955.82 (3,600)	0.001	0.001
IN-3-1	1.85	−1.18	5.22	2.26	1.80	2.76	4,394.16 (3,600)	0.001	0.002
IN-4-1	2.26	2.98	1.49	1.99	4.96	−1.14	5,025.16 (3,600)	0.001	0.002
IN-5-1	0.31	−0.76	1.62	0.67	2.71	−1.85	5,833.16 (3,600)	0.001	0.002
IN-6-1	−0.52	−0.37	−0.68	0.29	2.85	−2.50	6,976.23 (3,600)	0.001	0.003
IN-7-1	**17.51**	**9.68**	**7.84**	**17.75**	**10.03**	**7.72**	/	0.001	0.003
IN-8-1	**18.83**	**10.71**	**8.12**	**19.15**	**10.03**	**7.93**	/	0.001	0.003
IN-9-1	**19.82**	**11.55**	**8.27**	**20.52**	**11.22**	**8.56**	/	0.001	0.003
IN-10-1	**20.46**	**12.06**	**8.40**	**21.21**	**11.97**	**8.62**	/	0.001	0.004
IN-2-1.5	2.53	7.63	−1.91	2.04	2.18	1.92	5,301.36 (3,600)	0.001	0.002
IN-3-1.5	0.47	1.65	−0.74	0.16	2.92	−2.66	7,307.53 (3,600)	0.002	0.002
IN-4-1.5	0.75	1.32	0.14	0.02	1.38	−1.42	10,161.30 (3,600)	0.002	0.004
IN-5-1.5	**15.79**	**8.66**	**7.13**	**15.74**	**8.75**	**6.99**	/	0.002	0.004
IN-6-1.5	**21.00**	**11.05**	**9.95**	**20.99**	**11.14**	**9.85**	/	0.002	0.004
IN-7-1.5	**23.33**	**12.99**	**10.34**	**23.50**	**13.18**	**10.32**	/	0.002	0.006
IN-8-1.5	**25.11**	**14.45**	**10.66**	**25.36**	**17.74**	**10.63**	/	0.002	0.006
IN-9-1.5	**26.41**	**15.58**	**10.82**	**27.09**	**15.86**	**11.23**	/	0.002	0.007
IN-10-1.5	**27.35**	**16.38**	**10.96**	**28.10**	**16.80**	**11.30**	/	0.001	0.009
IN-2-2	2.50	13.57	−6.24	−0.20	4.66	−4.03	8,941.71 (3,600)	0.002	0.002
IN-3-2	1.80	2.40	1.17	1.94	0.67	3.27	15,525.20 (3,600)	0.002	0.003
IN-4-2	**15.79**	**8.19**	**7.60**	**15.54**	**8.03**	**7.50**	/	0.002	0.005
IN-5-2	**19.76**	**10.87**	**8.88**	**19.70**	**10.97**	**8.73**	/	0.002	0.005
IN-6-2	**26.23**	**13.92**	**12.31**	**26.14**	**13.88**	**12.26**	/	0.001	0.006
IN-7-2	**29.17**	**16.37**	**12.80**	**29.43**	**16.52**	**12.90**	/	0.002	0.008
IN-8-2	**31.40**	**18.24**	**13.16**	**31.79**	**18.40**	**13.39**	/	0.002	0.009
IN-9-2	**33.03**	**19.69**	**13.34**	**34.01**	**19.83**	**14.17**	/	0.002	0.009
IN-10-2	**34.28**	**20.75**	**13.53**	**34.96**	**21.09**	**13.87**	/	0.004	0.010

Note. Bold fonts represent objective value.

Table 2. Computational Performance of the CA Model Under Varying Values of w

| | Relative gap, % | | | | | | Computation time, s | | |
| | CA-I | | | CA-D | | | | | |
w	C	C_f	C_w	C	C_f	C_w	Gurobi	CA-I	CA-D
IN-2-0.5									
0.3	0.76	−0.06	2.21	3.17	0.42	7.90	140.514 (113.660)	0.001	0.001
0.4	1.79	−5.83	13.53	2.20	0.00	5.60	122.065 (113.660)	0.001	0.001
0.5	2.76	−5.44	13.63	1.85	0.00	4.30	150.467 (122.950)	0.001	0.001
0.6	3.42	−3.20	11.20	2.07	−2.83	7.82	68.530 (41.040)	0.001	0.001
0.7	3.71	3.51	3.92	2.74	−5.67	11.20	104.203 (76.820)	0.001	0.001
0.8	3.53	9.17	−1.44	4.12	5.67	2.76	74.810 (47.440)	0.001	0.001
0.9	3.28	8.76	−1.25	4.16	5.51	3.05	68.802 (41.380)	0.001	0.001
1.0	2.99	9.87	−2.13	4.10	5.51	3.05	62.022 (34.630)	0.001	0.001
1.1	2.70	10.05	−2.26	4.04	5.51	3.05	64.035 (35.380)	0.001	0.001
1.2	2.45	10.10	−2.29	3.99	5.51	3.05	60.682 (32.600)	0.001	0.001
1.3	2.22	10.10	−2.29	3.73	5.51	2.71	51.786 (23.500)	0.001	0.001
IN-4-0.5									
0.3	2.19	4.48	−3.07	5.29	−1.31	20.41	976.485 (866.610)	0.001	0.001
0.4	2.16	3.12	0.36	5.04	−1.27	16.90	450.632 (339.020)	0.001	0.001
0.5	2.07	2.70	1.10	1.23	1.26	1.20	254.632 (147.390)	0.001	0.001
0.6	2.00	0.41	4.18	1.23	−1.23	4.62	248.309 (147.390)	0.001	0.001
0.7	2.10	1.73	2.54	2.19	0.00	4.77	297.036 (189.820)	0.001	0.001
0.8	2.15	2.73	1.53	1.69	4.84	−1.64	287.487 (179.400)	0.001	0.001
0.9	2.01	4.91	−1.50	1.50	4.79	−2.38	182.355 (74.790)	0.001	0.001
1.0	1.72	9.47	−4.85	1.27	6.05	−2.79	218.819 (110.750)	0.001	0.001
1.1	1.34	10.92	−6.04	1.06	6.05	−2.79	170.782 (62.430)	0.001	0.001
1.2	1.01	7.31	−3.72	0.91	2.34	−0.16	192.991 (78.910)	0.001	0.001
1.3	0.80	7.40	−3.78	1.14	3.50	−0.50	164.861 (55.860)	0.001	0.001

the optimal solutions from Gurobi to four selected instances are plotted in Figure 7. Finally, the lower bound, upper bound, and optimal vehicle formations (obtained from Gurobi) of two selected instances are presented in Figure 8.

From Figure 6, we see that the optimal vehicle formations vary dramatically across time and space in all four instances, verifying that it is necessary to adjust the vehicle capacity across all stations over time in UMT systems. Also, the near-optimal dispatch times from the CA model and the optimal ones from the discrete model are exactly the same. Note that minor variations may be observed in larger instances, but comparisons cannot be provided here because Gurobi can only solve certain small instances to optimality given the limited computational resources. The vehicle formations from both approaches are the same for the majority of dispatches across all stations (over 75%) with some minor local variations in the instances shown in Figure 6, (b)–(d). For the instance shown in Figure 6(a), the discrepancy between the vehicle formations obtained from Gurobi and CA-D is more evident with only 16 out of the 40 dispatches being the same. However, the relative gap between these two solutions is only 5.28%, which still satisfies the requirements of most engineering applications. The excellent performance of the CA model partially lies in the EOQ structure of the investigated

problem, whose objective value remains relatively stable when the variations in the solutions remain within a certain range (Daganzo 2005). Based on these observations, we conclude that, though discrepancies can be witnessed for certain dispatches between solutions from the CA model and those from the discrete model, the objective values do not deviate substantially. Thus, the CA model can produce near-optimal solutions very close to those from the discrete model across a wide range of input parameter values. These results further demonstrate the solution accuracy of the CA model.

As can be seen from Figure 7, the passenger queue drops to zero at at least one station after each dispatch in the four instances except the fifth dispatch in IN-4-0.5 with $w = 0.3$. For this exception, the vehicle dispatch headway equals the minimum dispatch headway (i.e., three minutes). These results verify Proposition 1. Also, we can see that the number of passengers left at the stations after a dispatch reaches zero in most cases in the optimal solutions to the selected instances. For a few cases in which the passenger queue is not cleared, the number of passengers left behind is still less than the capacity of a single vehicle (i.e., 50). We also observe that the larger the value of w, the smaller the passenger queue and the lower the possibility that the passenger queue is not zero after a dispatch. These results verify Propositions 2 and 3. Further, Figure 8

Figure 6. Optimal Solutions from the Discrete Model (Left) and Near-Optimal Solutions from CA-D (Right) of Four Selected Instances

(a)

IN-4-0.5, w=0.3

(b)

IN-4-0.5, w=0.7

(c)

IN-4-0.5, w=0.9
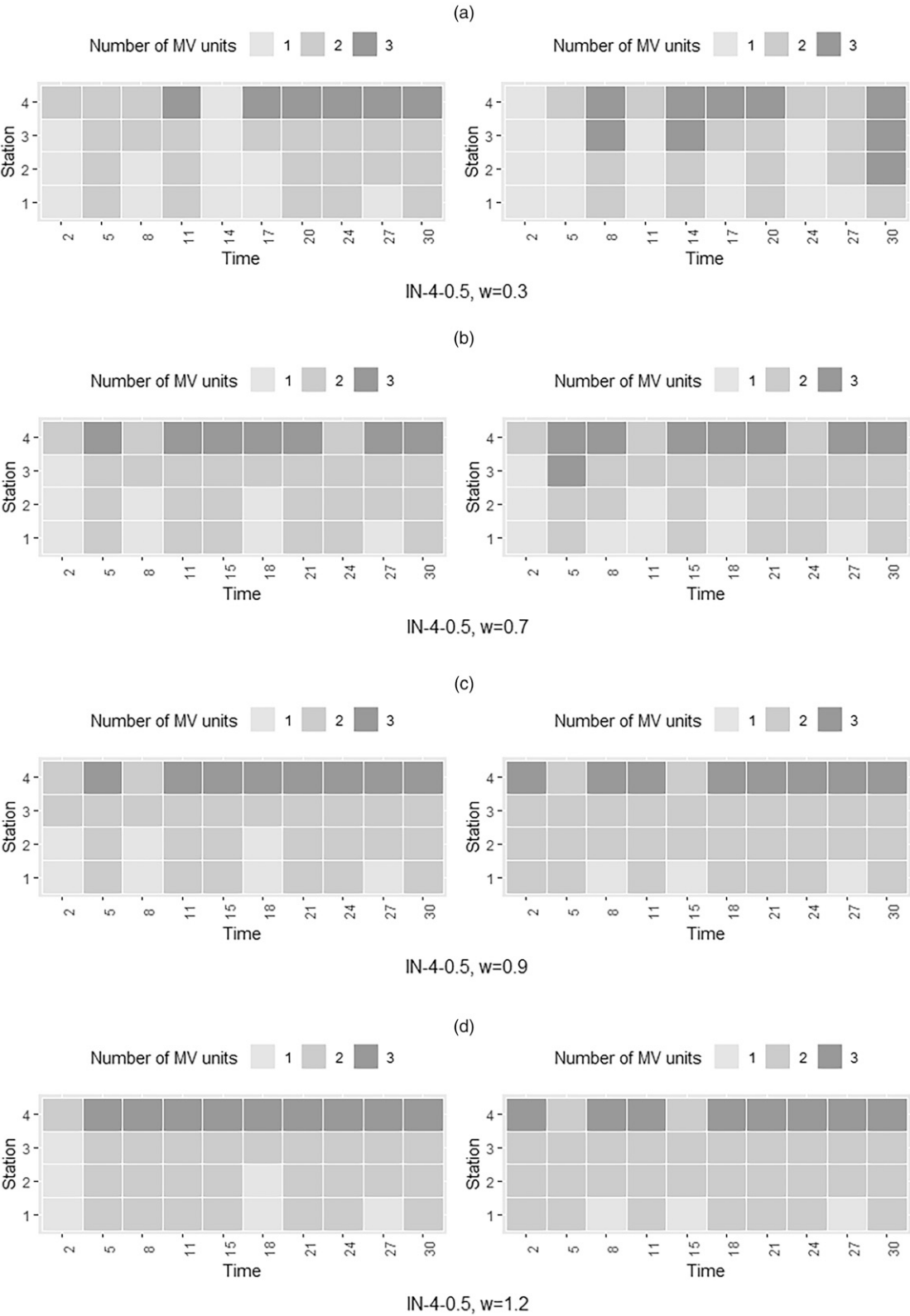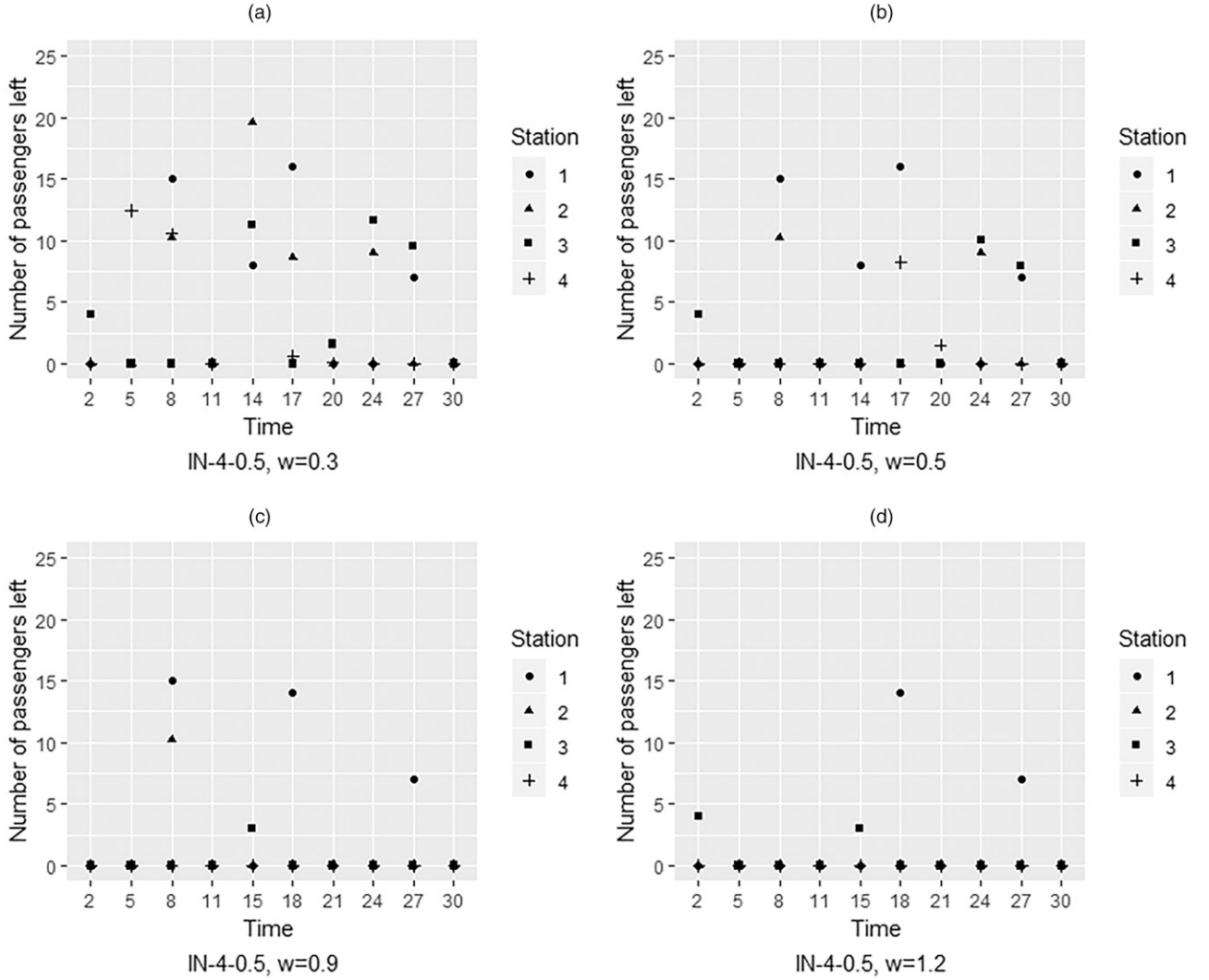
(d)

IN-4-0.5, w=1.2

Figure 7. Number of Passengers Left at Each Station After Each Dispatch in the Optimal Solutions from Gurobi in Four Selected Instances
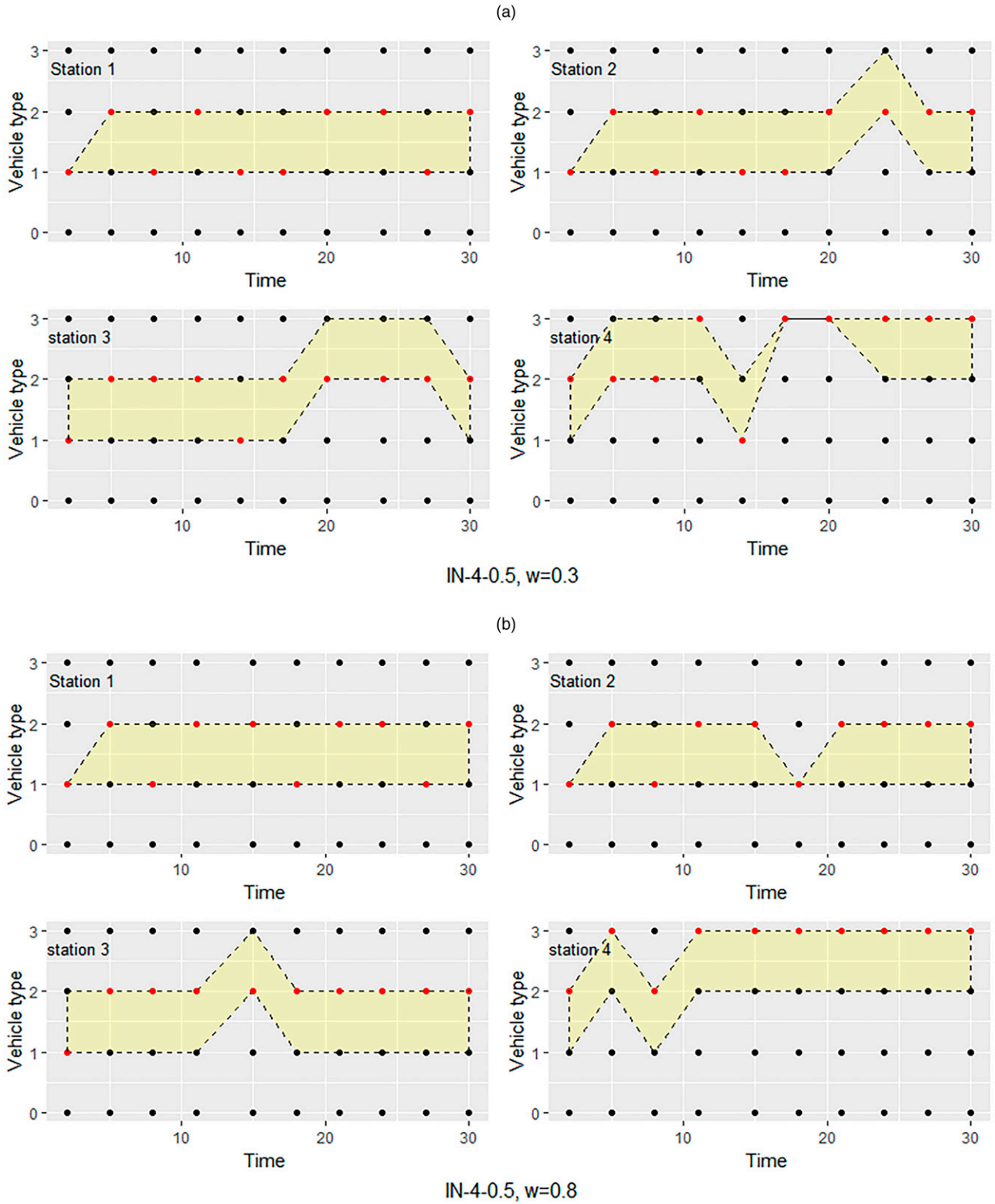

shows that the optimal vehicle formations are always selected between the conditional lower and upper bounds across all stations in the selected instances, verifying Propositions 4 and 5.

5.1.4. Sensitivity Analysis. This section evaluates the effectiveness of station-wise dynamic capacity adjustment (SDCA) and investigates how the system performance varies as key input parameters change. We select IN-10-2 for the analysis because it has the largest problem size. In each experiment, we vary one input parameter and keep the others the same as those defined in Section 5.1.1 unless stated otherwise. To demonstrate the effectiveness of SDCA in transit corridors, we compare the results of the proposed system with a benchmark system in which only the largest vehicle formation can be dispatched.

Further, following Chen, Li, and Zhou (2019, 2020), we adopt three metrics to quantitatively evaluate the system performance, namely, average load percentage (ALP), total operational cost (TOC), and total waiting cost (TWC). ALP is defined as the average of the ratios of the number of on-board passengers to the vehicle capacity of each dispatch across all stations; that is, $\frac{1}{KS}\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\frac{e_{ks}}{i_{ks}c}$, where e_{ks} and i_{ks} represent the number of on-board passengers and vehicle formation of dispatch k at station s, respectively. TOC and TWC refer to the first and second terms in the objective function (4), respectively. Results of the sensitivity analysis are summarized in Figure 9.

We first investigate the effectiveness of SDCA. As can be seen from Figure 9, introducing SDCA in UMT systems can improve the system performance for all the parameter values considered in this

Figure 8. (Color online) Lower Bound, Upper Bound, and Optimal Vehicle Formations from Gurubi of Two Selected Instances

Notes. All dots represent the entire feasible region. Dots along the top and bottom border of the shaded areas represent the upper and lower bound vehicle formations, respectively, after applying Propositions 4 and 5. Red dots are the optimal vehicle formations from Gurobi.

Figure 9. (Color online) Sensitivity Analysis of Case Study 1
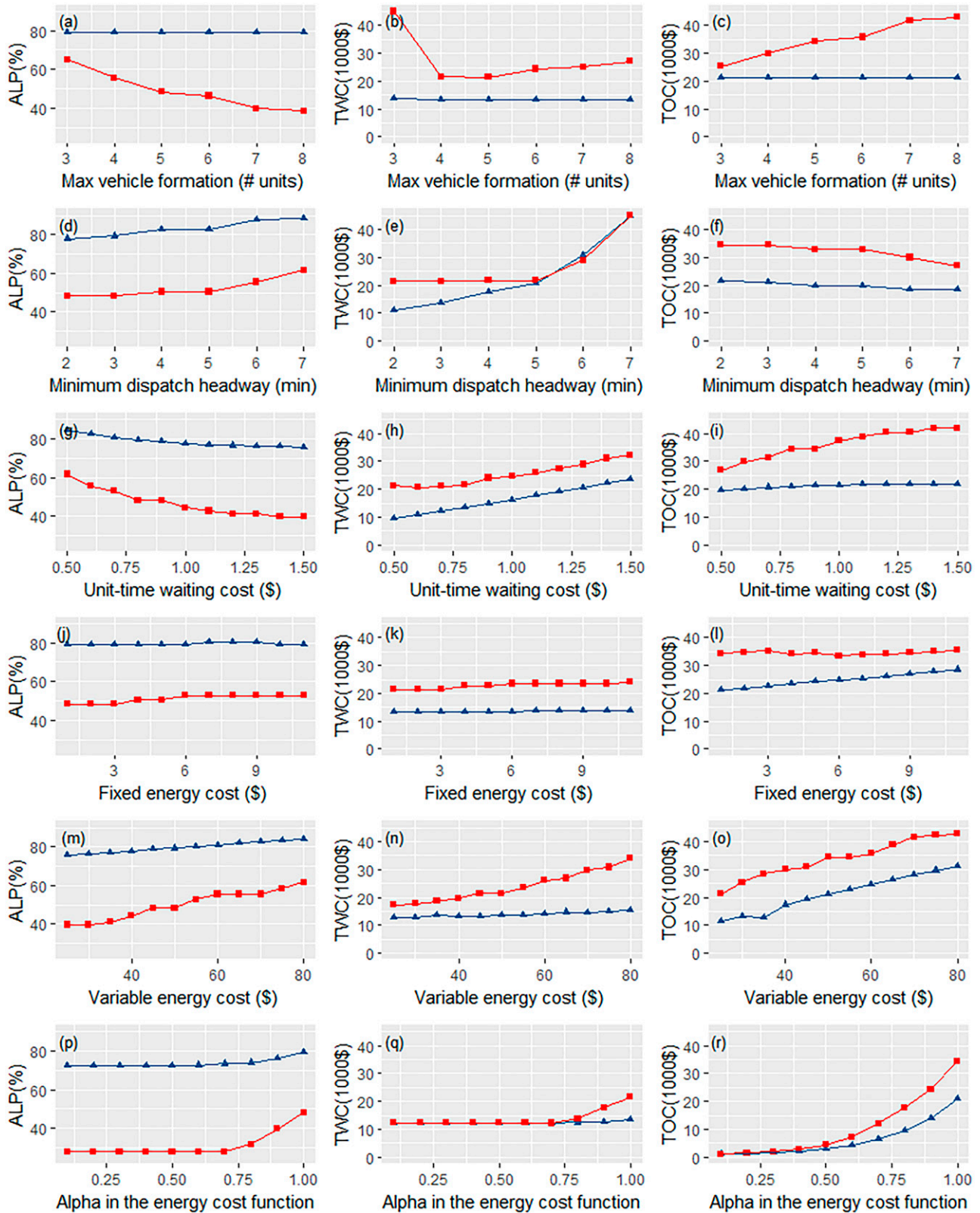
Notes. Blue lines with triangle dots represent results from the proposed MAV system. Red lines with square dots represent results from the benchmark system. Dots on the border of the graphs represent values larger than the maximum values.

section.[3] Specifically, by allowing SDCA along the corridor, we increase the average load percentage across all dispatches by up to around 40%, indicating a better resource utilization rate in the system. Besides this, the total passenger waiting cost and operational cost are both reduced in almost all instances. In one exception in which the total waiting cost increases, that is, Figure 9(e) when the minimum dispatch headway is six minutes, the decrease in the other cost component is larger so that the total system cost is still decreased. Thus, introducing station-wise dynamic capacity adjustment into UMTs can increase the vehicle utilization rate and decrease the total system cost.

Next, we analyze the effects of each input parameter on the effectiveness of SDCA. Figure 9, (a)–(c), shows that, as the maximum vehicle formation I increases, both the ALP gap (i.e., the vertical distance between the ALP curves) and the TOC gap (i.e., the vertical distance between the TOC curves) increase, and the TWC gap (i.e., the vertical distance between the TWC curves) first experiences a plummet and then a steady growth. These observations suggest that the improvement brought by SDCA is the least when $I = 4$ and gets better as I deviates from this value. When I is less than 4, the passenger waiting cost in the benchmark system is extremely high because the optimization model forces vehicles to be dispatched with longer headways. In the proposed MAV system, however, passengers are served with shorter vehicles and smaller headways, thus substantially reducing the passenger waiting cost. Once I reaches four, the near-optimal design of the proposed system remains the same, but the total system cost in the benchmark system keeps increasing, thus strengthening the effectiveness of the proposed SDCA operation. Because the total system cost in the proposed system reaches the minimum and remains the same after I reaches four, we set $I = 5$ in the following analysis to allow a larger feasible region as well as a better performance.

Figure 9, (d)–(f), reveals that the effectiveness of SDCA is less evident as the minimum dispatch headway $\underline{h}$ increases. Although the ALP gap remains relatively stable, both the TWC and TOC gaps (i.e., the vertical distance between the TOC curves) narrow down as $\underline{h}$ increases. The reason is that the number of passengers waiting for boarding at all stations for each dispatch increases with the minimum time difference between every two consecutive dispatches, therefore raising the probability of using longer vehicles in the proposed MAV system. As a result, the capability to flexibly adjust vehicle capacity based on the passenger demand level is limited as $\underline{h}$ increases, thus weakening the effectiveness of SDCA. Thus, reducing the minimum dispatch headway in transportation corridors, if possible, can improve the

effectiveness of SDCA. This result is consistent with the finding in Chen, Li, and Zhou (2019).

Figure 9, (g)–(i), implies that the increase in the unit-time passenger waiting cost w amplifies the effectiveness of the SDCA as indicated by the increasing ALP and TOC gaps. As w increases, the operational cost and the passenger waiting cost gradually increase in both the benchmark and MAV systems because vehicles are dispatched more frequently to serve passengers with higher unit-time waiting cost. The total waiting cost in both systems seems to increase at the same pace as indicated by the relatively stable TWC gap in Figure 9(h). Yet the increase in the operational cost is almost negligible in the MAV system because shorter vehicles can be used to keep the operational cost small even though vehicles are dispatched more frequently.

The variations in the parameter representing the fixed energy cost regardless of vehicle capacity, C^F, do not impose a substantial impact on the effectiveness of SDCA as shown in Figure 9, (i)–(l). This parameter increases the operational cost of each vehicle formation equally, and thus, the solutions are barely changed by the parameter variations. However, despite a slight decrease in the ALP gap, the SDCA saves more costs as the other two parameters increase. The increase in the parameter representing the variable energy cost dependent on the vehicle capacity, that is, C^V, results in a steady growth in the TWC gap while the TOC gap stays relatively stable (Figure 9, (n) and (o)), indicating a larger gap between the total system costs. Further, Figure 9, (q)–(r), shows that the cost saving is very limited when the unitless parameter in the operational cost function, α, is less than 0.4 (both the TWC and TOC gap are almost zero). This is because, when the value of α is small, the operational costs of different vehicle formations are almost the same. Thus, there is no substantial difference to dispatch long and short vehicles. As α increases, SDCA starts to bring observable savings in the operational cost because, in this case, it is more favorable to dispatch short vehicles when passenger demand is low. Moreover, when α reaches around 0.7, the passenger waiting cost is reduced as well, resulting in a more substantial saving in the total system cost. Thus, it is crucial to analyze the extent of the economics of scale in the operational cost in UMT systems before introducing the SDCA paradigm because, otherwise, system improvement may not be achieved.

5.2. Case Study 2: Future MAV Service on the Batong Line, Beijing Subway

In this section, we apply the CA model to design future MAV service for the Batong line in the Beijing Subway system. As shown in Figure 10, the Batong line is a bidirectional line with 13 stations. Following

Figure 10. (Color online) Batong Line in the Beijing Subway System
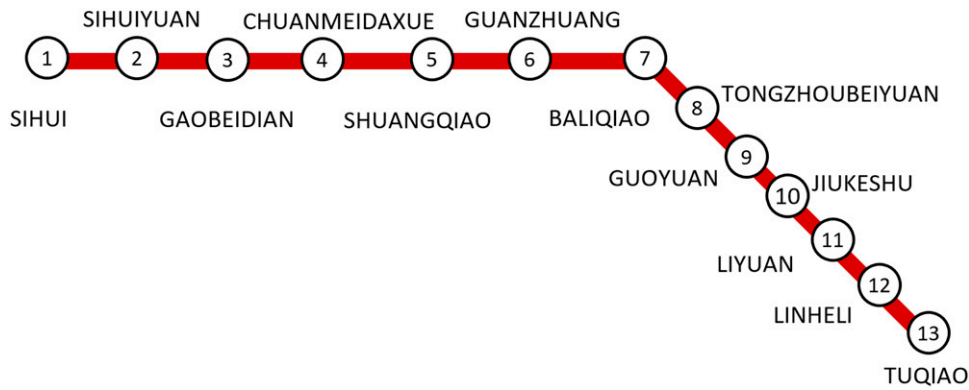
Chen, Li, and Zhou (2020), we simply select the direction from station 13 to 1 for the experiments in this case study. Because a large portion of passengers originating from these stations are destined to stations on other subway lines in the network, we treat all other stations in the network as a virtual destination station in the corridor. Further, oversaturated traffic is not the focus of this study, so we consider an operational horizon starting from 11:00 a.m. and terminating at 23:00 p.m., during which only unsaturated traffic is present. With this, we count the passenger demand between each OD pair per minute using smartcard data and obtain the passenger arrival demand passing through each station during each minute over the investigated operational horizon, as shown in Figure 11. In this case study, we envision a future scenario in which the operator introduces the MAV technology in the system so that vehicles can change their capacity flexibly at any station along the corridor to serve the passengers. Following the existing operational plan of the Batong line, we set $\mathcal{I} := [1, 2, 3, 4, 5, 6]$, $c = 226$ pax (i.e., passenger)/carriage, and $\underline{h} = 3$ mins. Besides this, w is set as \$0.10/minute based on the average monthly salary per capita in Beijing in 2017, and the

operational cost function is adapted from the one in Chen, Li, and Zhou (2020), resulting in $C^F = 0.39$, $C^V = 0.45$, and $\alpha = 0.5$. In each experiment, we vary only one parameter, and the others remain the same as the default values to investigate the sensitivity. To evaluate the effectiveness of SDCA, we compare the MAV system with a benchmark system in which only one vehicle formation can be dispatched.

We first plot the optimal design for three selected instances in Figure 12. We find that the optimal vehicle formations change dramatically across time and space in all three instances; that is, long vehicles are dispatched at stations and time periods with intensive passenger demand, and short vehicles are used for stations and time periods with relatively low passenger demand. Further, the larger the minimum dispatch headway, the higher the possibility of using long vehicles and, thus, the lower the flexibility to adjust vehicle capacity. These results are consistent with those in the previous hypothetical example.

Next, we present the ALP, TWC, and TOC of all experiments in Figure 13. We see that introducing station-wise dynamic capacity adjustment can improve the performance of the Batong line in the Beijing Subway system across all instances with the

Figure 11. Passenger Demand Rate During the Unsaturated Period of the Batong Line in the Beijing Subway System
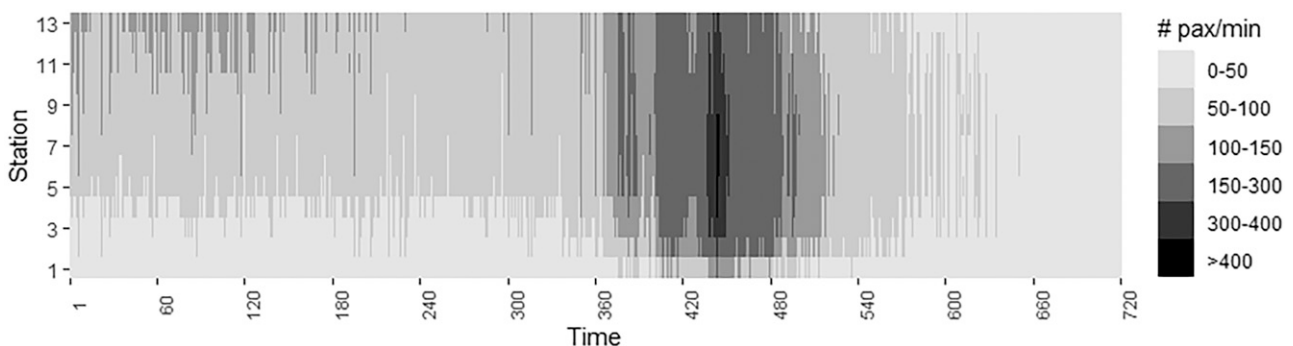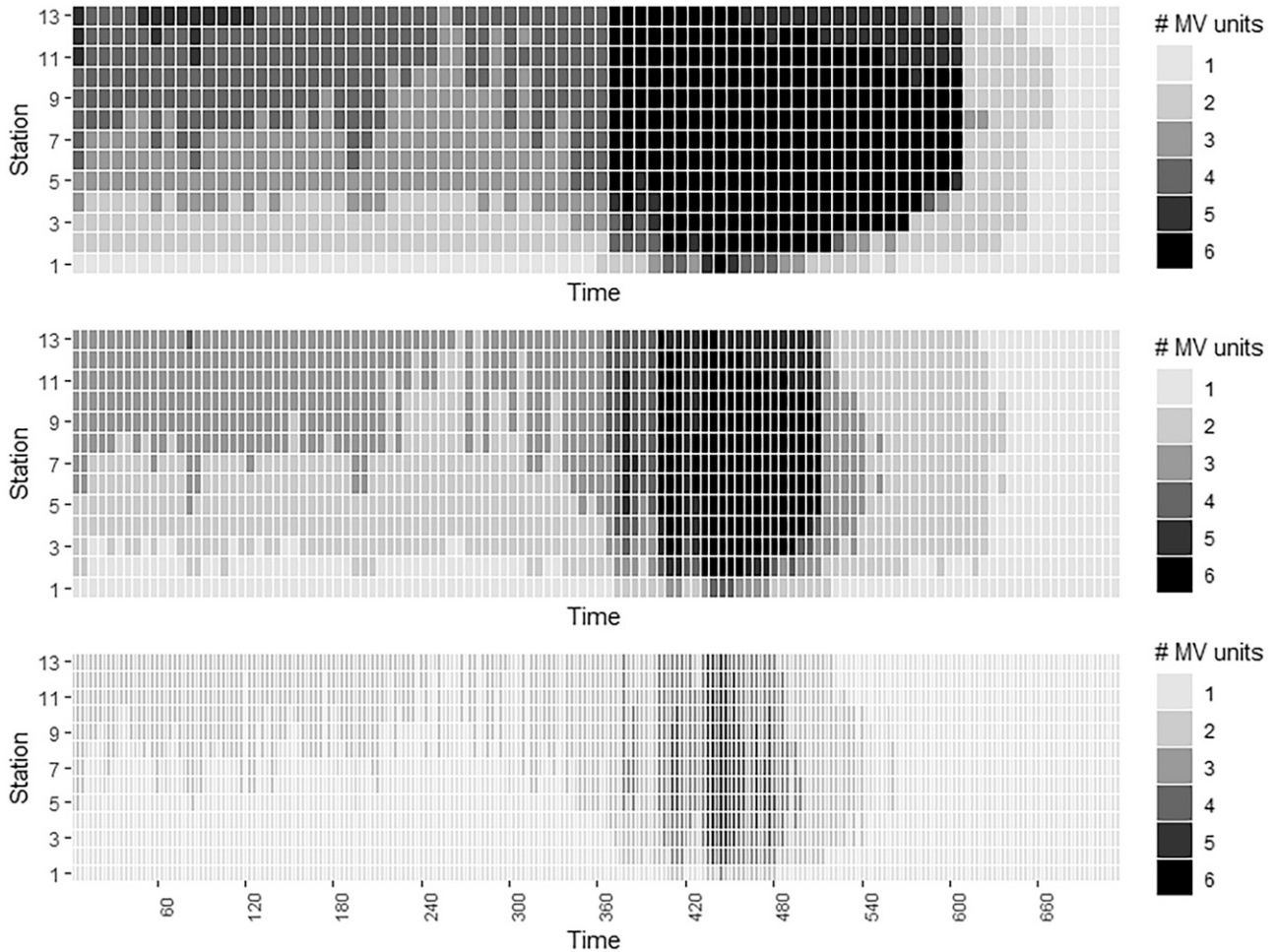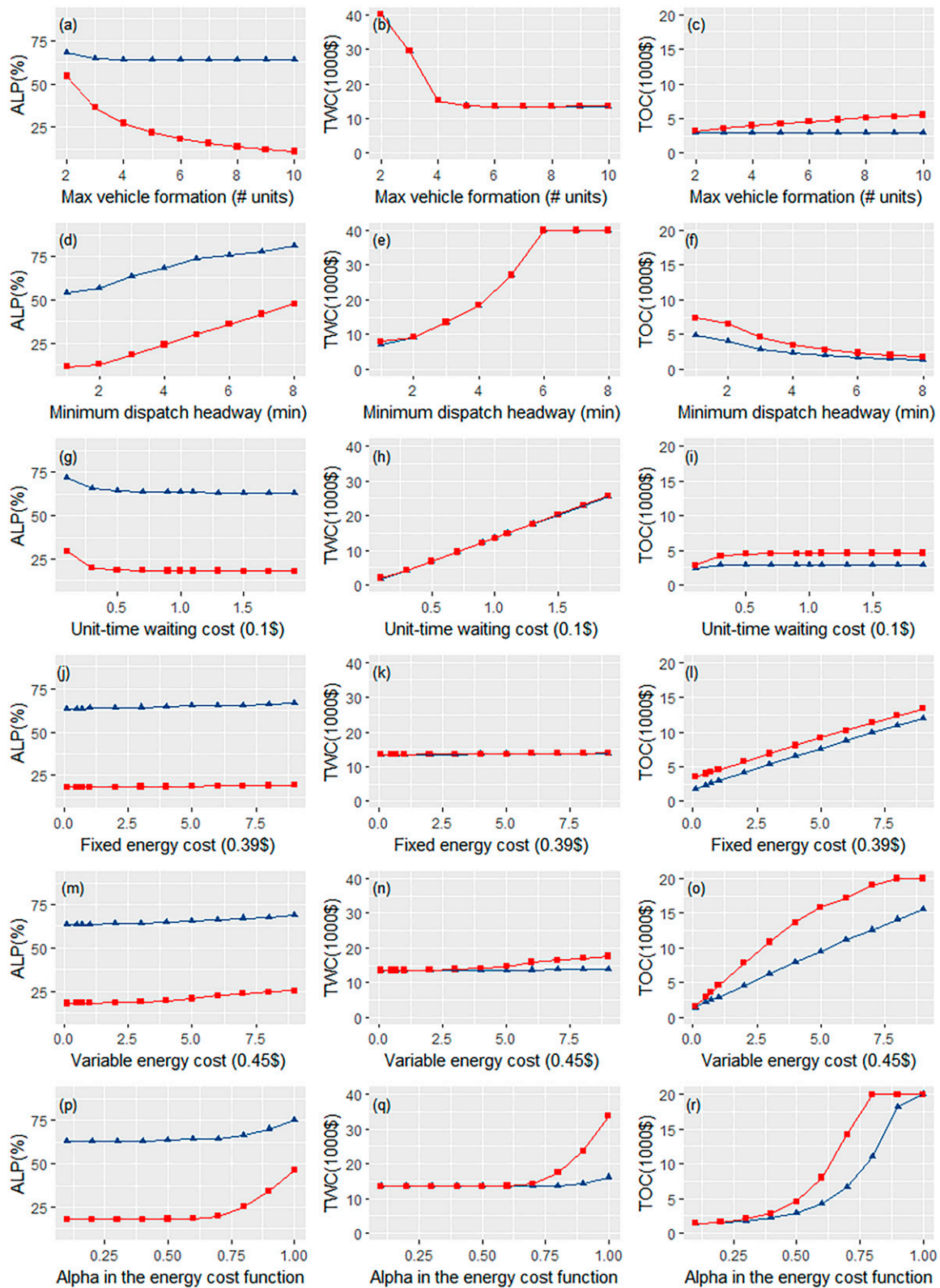
Figure 12. Optimal Design for the Batong Line in the Beijing Subway with a Minimum Dispatch Headway of Nine (Top Row), Six (Middle Row), and Three Minutes (Bottom Row)

average load percentage always being increased, the total passenger waiting cost being decreased or remaining the same, and the total operational cost always being decreased. The difference between this and the previous case is that, in addition to a few scenarios, the total passenger waiting costs are (almost) the same with and without SDCA. This is because the passenger demand in the Batong line is of a different magnitude compared with that in Mandl's (1980) network. As a result, the passenger waiting cost is usually substantially larger than the operational cost (as can be seen from values of the *y*-axes of the figures in the second and third columns in Figure 13), so the optimization model inherently adjusts the vehicle dispatch plan to maintain the passenger waiting cost at a relatively small value. This observation also reveals that the condition in the theoretical propositions may be satisfied in various UMT systems with similar cost structure as that in the Beijing Subway system, rendering a broad applicability domain of the proposed model.

As for each input parameter, we see that they influence the system performance metrics in a similar way as they influence the system performance in the previous case study. The exception is that, as the maximum vehicle formation increases, the effectiveness of SDCA is strictly increasing (the TWC is almost zero, and the TOC gap is widening), which can be probably attributed to the intensive passenger demand in the system. Also, we observe that, for the existing system, the total system cost reaches its minimum when $I = 4$, suggesting that reducing the number of carriages in each train to four units for the Batong line during the unsaturated hours can improve system performance. These results are also consistent with those in Chen, Li, and Zhou (2019). Finally, the mean computation times for CA-I and CA-D across all experiments are 0.027 and 1.129 seconds, respectively, which is very efficient for an instance with 13 stations and 720 time indexes (i.e., 12 hours with a discretization interval of one minute). The results verify the applicability of the proposed CA model in addressing real-world problems that are

Figure 13. (Color online) Sensitivity Analysis of Case Study 2

Notes. Blue lines with triangle dots represent results from the proposed MAV system. Red lines with square dots represent results from the benchmark system. Dots on the border of the graphs represent values larger than the maximum values.

extremely difficult to solve efficiently with exact modeling methods.

6. Conclusion

This paper proposes a CA model to solve the joint design of dispatch headway and vehicle formation for MAV-based transportation corridors. This model extends the traditional CA approach that designs dispatch policies for transportation systems to consider many-to-many demand patterns, station-wise vehicle capacity adjustment, and other factors (e.g., passenger boarding order, minimum dispatch headway). The challenges of modeling the investigated problem under the CA framework lie in the spatiotemporal correlation between the dispatch decisions, the complicated passenger boarding dynamics, and the consideration of multiple vehicle formations across all stations. To address these challenges, we investigate the theoretical properties of the investigated problem. These properties enable the problem to recover the local impact property and eliminate the vehicle formation decisions from the problem, based on which a macroscopic CA model is proposed to decompose the original problem into finite subproblems that can be analytically solved. With the analytical continuous solutions from the CA model, a greedy heuristic is designed to search for the discrete near-optimal solutions to the original problem. Numerical experiments were conducted for a hypothetical transportation corridor in Mandl's (1980) network and the Batong line in the Beijing Subway system to assess the computation performance of the CA model, verify the theoretical properties as well as reveal managerial insights for transit operators. The main findings are

1. The CA model can solve near-optimal solutions that can satisfy the requirements of most engineering applications very efficiently (less than 10 ms) compared with the discrete modeling method (which may not even yield a feasible solution in several hours). Its performance is robust under various parameter settings and is more accurate for large-scale instances. Thus, the proposed CA model is appropriate for solving large-scale operational design problem instances for transportation corridors in the real world, whose exact optimal solutions would take expensive computation resources or is impossible to solve.

2. The numerical experiments verify the analytical theoretical properties. These properties reveal the optimal number of passengers waiting for boarding is a relatively small value (bounded by the capacity of a single modular pod) after each dispatch across all stations along the corridor. Further, the optimal vehicle formations should always be either the upper or conditional lower bound vehicle formation. These

properties shed important managerial insights and constitute a theoretical foundation for constructing the CA model as well as the discretization heuristic. They can also facilitate the design of expedited solution algorithms to search for the exact optimal solutions to the investigated problem.

3. Introducing SDCA can improve the vehicle utilization rate and decrease the total system cost (i.e., the sum of passenger waiting and operational costs) for transportation corridors under a wide range of parameter settings. Different input parameters affect the effectiveness of SDCA differently. Specifically, the effectiveness of SDCA can be amplified by decreasing the minimum dispatch headway, the unit-time waiting cost per passenger, the variable operational cost dependent on the vehicle capacity, and the unitless parameter that reflects the economics of scale of the operational cost. The maximum number of modular pods allowed to be docked into one vehicle shows different influences on the performance of different systems. Finally, the parameter representing the fixed operational cost regardless of the vehicles' capacity does not affect the effectiveness of SDCA substantially.

4. Jointly optimizing the timetabling and vehicle repositioning in MAV-based transportation corridors results in a hardly tractable problem with existing commercial solvers. The microscopic details in MAV repositioning also render the development of analytical CA model very challenging. Thus, it is preferable to decompose the timetabling and vehicle repositioning as two separate problems. Online Appendix D shows that the timetabling decisions obtained by solving the proposed CA model could be easily extended to obtain near-optimal repositioning decisions (with a gap less than 5% in most cases) very efficiently (within 10 ms). Thus, the proposed CA model provides a foundation for developing solution approaches for other problems with more complex system operation constraints.

This study proposes a continuous model to solve the operational design for MAV-based transportation corridors enabling station-wise docking and can be extended in a number of directions. First, this model can be extended to the case in which oversaturated traffic may exist with modifications, such as adding a penalty factor or queue decomposition. Second, Online Appendix C shows that the performance of the CA model is robust under uncertain passenger demands. However, considering the demand uncertainty in the CA model and tackling it with a probabilistic approach will be an interesting and meaningful research direction. Finally, this study primarily focuses on the joint design of vehicle dispatch headway and capacity for a MAV-based corridor system without considering the rolling stock management (i.e., the

MAV repositioning activities). Rolling stock management is an important planning-level decision problem that can be considered in future research. Researchers can follow existing practice to solve the rolling stock management as a separate problem as we do in Online Appendix D, using results from this study as the inputs. Alternatively, they can consider introducing decision variables and constraints related to rolling stock management into the proposed model to search for further improvements compared with existing practice. The incorporation of these variables and constraints is expected to greatly increase the solution space of the model, which would render the problem to another level of difficulty. Future research efforts could be devoted to design more sophisticated solution approaches for such an integrated design.

Endnotes

[1] This study addresses a deterministic optimization problem in which the cumulative arrival curves $A_s(t)$ are assumed a priori rather than a stochastic or uncertain process. We acknowledge that this treatment does not consider potential uncertainties in the passenger demand during real-time operation. Nonetheless, this simplification remains acceptable for tactical design or planning that is mainly impacted by the expected demand values and is relatively robust to the minor stochasticity of actual demand realizations. In fact, many UMT operators generally design their timetables with historic or predicted demand counts at the planning stage. Please refer to Online Appendix C for numerical experiments on the performance of the proposed CA method under stochastic demands.

[2] The rolling stock management (in our case, determination of the MAV repositioning activities with a fleet size limit) is generally a separate problem from timetabling in urban mass transportation system design (Cadarso and Marín 2011; Caimi, Kroon, and Liebchen 2017). With assumption (ii), we can dispatch as many vehicles as we need at each station, and thus, vehicle repositioning activities are not necessary to be incorporated into the model. After solving the investigated problem, the results can be used as inputs for the rolling stock management planning model to obtain the optimal fleet size and MAV repositioning activities within the system. Please refer to Online Appendix D for an application of the proposed solution approach to the case in which vehicle repositioning activities (and, thus, the fleet size limit) are considered.

[3] The difference in some instances (i.e., Figure 9, (q) and (r)) is too small to be visually observed.

References

Albrecht T (2009) Automated timetable design for demand-oriented service on suburban railways. *Public Transportation* 1(1):5–20.

Arbex RO, da Cunha CB (2015) Efficient transit network design and frequencies setting multi-objective optimization by alternating objective genetic algorithm. *Transportation Res. Part B: Methodological* 81(2):355–376.

Cadarso L, Marín Á (2011) Robust rolling stock in rapid transit networks. *Comput. Oper. Res.* 38(8):1131–1142.

Caimi G, Kroon L, Liebchen C (2017) Models for railway timetable optimization: Applicability and applications in practice. *J. Rail Transport Planning Management* 6(4):285–312.

Caprara A, Fischetti M, Toth P (2002) Modeling and solving the train timetabling problem. *Oper. Res.* 50(5):851–861.

Ceder A (2001) Bus timetables with even passenger loads as opposed to even headways. *Transportation Res. Rec.* 1760(1):3–9.

Chen Z, Li X (2021) Designing corridor systems with modular autonomous vehicles enabling station-wise docking: Discrete modeling method. *Transportation Res. Part E: Logistics and Transportation Rev.* 152:102388.

Chen Z, Li X, Zhou X (2019) Operational design for shuttle systems with modular vehicles under oversaturated traffic: Discrete modeling method. *Transportation Res. Part B: Methodological* 122:1–19.

Chen Z, Li X, Zhou X (2020) Operational design for shuttle systems with modular vehicles under oversaturated traffic: Continuous modeling method. *Transportation Res. Part B: Methodological* 132:76–100.

Daganzo CF (2005) *Logistics Systems Analysis* (Springer Science & Business Media, New York).

Daganzo CF (2009) A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Res. Part B: Methodological* 43(10):913–921.

Daganzo CF (2010) Structure of competitive transit networks. *Transportation Res. Part B: Methodological* 44(4):434–446.

Guo QW, Chow JYJ, Schonfeld P (2017) Stochastic dynamic switching in fixed and flexible transit services as market entry-exit real options. *Transportation Res. Procedia* 23:380–399.

Hassold S, Ceder AA (2014) Public transport vehicle scheduling featuring multiple vehicle types. *Transportation Res. Part B: Methodological* 67:129–143.

Holmberg K, Tuy H (1999) A production-transportation problem with stochastic demand and concave production costs. *Math. Programming* 85(1):157–179.

Huang Y, Yang L, Tang T, Gao Z, Cao F (2017) Joint train scheduling optimization with service quality and energy efficiency in urban rail transit networks. *Energy* 138:1124–1147.

Hurdle VF (1973) Minimum cost schedules for a public transportation route—I. Theory. *Transportation Sci.* 7(2):109–137.

Li X, Ma J, Cui J, Ghiasi A, Zhou F (2016) Design framework of large-scale one-way electric vehicle sharing systems: A continuum approximation model. *Transportation Res. Part B: Methodological* 88:21–45.

Liebchen C (2007) Periodic timetable optimization in public transport. Waldmann KH, Stocker UM, eds. *Operations Research Proceedings 2006* (Springer, Karlsruhe), 29–36.

Luo D, Bonnetain L, Cats O, van Lint H (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Res. Rec.* 2672(8):175–186.

Mandl CE (1980) Evaluation and optimization of urban public transportation networks. *Eur. J. Oper. Res.* 5(6):396–404.

Newell GF (1971) Dispatching policies for a transportation route. *Transportation Sci.* 5(1):91–105.

Newell GF (1974) Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Sci.* 8(3):248–264.

NEXT (2019) Accessed October 13, 2020, https://www.next-future-mobility.com/.

Niu H, Zhou X (2013) Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. *Transportation Res. Part C: Emerging Tech.* 36:212–230.

Niu H, Zhou X, Gao R (2015) Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. *Transportation Res. Part B: Methodological* 76:117–135.

Shi J, Yang L, Yang J, Gao Z (2018) Service-oriented train timetabling with collaborative passenger flow control on an oversaturated

metro line: An integer linear optimization approach. *Transportation Res. Part B: Methodological* 110:26–59.

Sun L, Jian GJ, Lee DH, Axhausen KW, Erath A (2014) Demand-driven timetable design for metro services. *Transportation Res. Part C: Emerging Tech.* 46:284–299.

Wirasinghe SC (1990) Re-examination of Newell's dispatching policy and extension to a public bus route with many to many time-varying demand. *Internat. Sympos. Transportation Traffic Theory.*

Yin J, Yang L, Tang T, Gao Z, Ran B (2017) Dynamic passenger demand oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. *Transportation Res. Part B: Methodological* 97: 182–213.