1-2019

Routing and scheduling for a last-mile transportation system

Hai WANG
Singapore Management University, haiwang@smu.edu.sg

DOI: https://doi.org/10.1287/trsc.2017.0753

Published in Transportation Science, July 2017, Advance online.
http://doi.org/10.1287/trsc.2017.0753

Routing and Scheduling for a Last-Mile Transportation System

Hai Wang[a]

[a] School of Information Systems, Singapore Management University, Singapore 178902

Contact: haiwang@smu.edu.sg, http://orcid.org/0000-0002-9260-3428 (HW)

Abstract. The last-mile problem concerns the provision of travel services from the nearest public transportation node to a passenger's home or other destination. We study the operation of an emerging last-mile transportation system (LMTS) with batch demands that result from the arrival of groups of passengers who desire last-mile service at urban metro stations or bus stops. Routes and schedules are determined for a multivehicle fleet of delivery vehicles, with the objective of minimizing passenger waiting time and riding time. An exact mixed-integer programming (MIP) model for LMTS operations is presented first, which is difficult to solve optimally within acceptable computational times. Computationally feasible heuristic approaches are then developed: a myopic operating strategy that uses only demand information from trains that have already arrived, a metaheuristic approach based on a tabu search that employs demand information over the entire service horizon, and a two-stage method that solves the MIP model approximately over the entire service horizon. These approaches are implemented in a number of computational experiments to evaluate the system's performance, and demonstrate that LMTS is notably preferable to a conventional service system under certain conditions.

History: This paper has been accepted for the *Transportation Science* Special Issue on Recent Advances in Urban Transportation Through Optimization and Analytics.

Funding: This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

Keywords: last mile • batch demand • routing and scheduling • mixed-integer programming • myopic strategy • tabu search

1. Background and Literature Survey

The last-mile problem (LMP) has attracted growing attention in the past decade. Generally, it refers to the design and provision of travel services from a public transportation node to a passenger's final destination. Several developments have boosted interest in the problem. First, many cities and governments are under pressure to increase public transport's share of urban trips to reduce road congestion and air pollution. Urban transport planners have increasingly recognized that the unavailability of last-mile services—which offer alternatives to walking to or from the nearest metro station—is one of the main deterrents to the use of public transport. Second, an aging population in many cities has greatly increased the demand for such services. Third, legal requirements to ensure adequate mobility for certain demographic groups, such as people with physical disabilities or schoolchildren, are becoming more common; these are also the groups most likely to need last-mile services.

This paper focuses on a generic context, which is outlined schematically in Figure 1. A last-mile transportation system (LMTS) serves a public transportation node, such as a rapid-transit metro station, where trains discharge passengers in large or small groups (batches). Passengers' final destinations (homes, workplaces, public institutions, etc.) are spatially distributed in the urban area served by the node. A set of last-mile stops (LM stops) has also been specified and distributed spatially within this urban area. While a passenger's final destination can be anywhere within the area, LM stops are limited to a finite number of locations that are convenient for loading and unloading passengers, such as existing public transit stops, entrances to hotels, crossroads near office buildings, and loading areas adjacent to residential buildings or complexes. A fleet of vehicles is available to transport each passenger to the LM stop closest to her final destination. Empty vehicles then return to the station to pick up waiting or newly arrived passengers. The routes and schedules of LMTS vehicles are flexible and can be changed based on specific last-mile service requests. Essentially, LMTS is an on-demand urban transportation system with batch demands. We describe the setting in more detail in Section 2.

Many issues must be addressed when designing and operating an LMTS. In an earlier paper addressing the planning side (Wang and Odoni 2016), we focused on an LMTS from a stochastic and planning perspective, and provided closed-form approximations for its performance as a function of the system's fundamental design parameters. These approximations can be used for the preliminary planning and design of an LMTS, and in particular for determining approximate resource

Figure 1. Schematic of a Last-Mile Transportation System

requirements, such as the number of vehicles/servers needed to achieve a specified level of service (LOS).

A large body of literature has generated various models related to the LMP, with early work dating to the 1960s and expanding rapidly over the last 15 years. There are two last-mile contexts: freight LMTS and passenger LMTS. A freight LMTS is usually referred to as a "last-mile supply chain." Many studies have addressed this aspect, given the burgeoning role of e-commerce. By contrast to operations optimization, which we examine here, these studies focus more on concept discussions and simulations (e.g., Punakivi, Yrjölä, and Holmström 2001); general fulfillment strategies (e.g., Lee and Whang 2001); delivery options effects (e.g., Esper et al. 2003); demand characteristics effects (e.g., Boyer, Prud'homme, and Chung 2009, Song et al. 2009); and specific application contexts, such as Balcik, Beamon, and Smilowitz's (2008) study of last-mile distribution geared to the needs of humanitarian relief chains. As for passenger LMTS, various case studies have analyzed LMTS in different contexts, such as a bicycle-sharing program for an LMTS in Beijing (Liu, Jia, and Cheng 2012). Personal rapid transit (PRT), which refers to a variety of transportation systems with characteristics that are similar, in some respects, to the passenger LMTS we consider here, has also attracted significant attention in the past decade. Research has been conducted on PRT system control frameworks (e.g., Anderson 1998), financial assessments (e.g., Bly and Teychenne 2005, Berger et al. 2011), performance approximations (e.g., Lees-Miller, Hammersley, and Davenport 2009, Lees-Miller, Hammersley, and Wilson 2010), and case studies (e.g., Mueller and Sgouridis 2011). None of these papers addresses the detailed LMTS operating issues that are the subject of this paper, in which we focus on a generic passenger LMTS from an optimization and operational perspective.

A large body of research also concerns demand responsive transit (DRT), which is another type of on-demand service. Some papers focus on DRT concept discussions, practical implementation, and assessment of simulations in case studies, such as Horn (2002a), Mageean and Nelson (2003), Brake, Nelson, and Wright (2004), Enoch et al. (2004), Engles and Iacometti (2004), Palmer, Dessouky, and Abdelmaguid (2004), and Quadrifoglio, Dessouky, and Ordóñez (2008), among others. For example, Mageean and Nelson (2003) introduce the concept of telematics-based DRT services and present the results of their evaluation of a set of DRT technologies and operations at urban and rural sites across Europe. Palmer, Dessouky, and Abdelmaguid (2004) present the results of a nationwide study of DRT involving 62 transit agencies, which suggests that the use of a paratransit computer-aided dispatching system and agency service delivery is beneficial for productivity. Other papers focus on approximations of the performance of DRT from a stochastic viewpoint. Models have been developed to assist in system design and regulation; for example, Diana, Dessouky, and Xia (2006) study how to determine the number of vehicles needed to provide a DRT service of prespecified quality, in which service quality is evaluated in terms of waiting times at stops and maximum allowed detours. Daganzo (1978) presents an analytic model to predict average waiting and riding times in urban transportation systems, such as dial-a-ride buses and taxis, and Wilson and Hendrickson (1980) critically review models to predict the performance of such flexibly routed transportation services. A few papers also discuss DRT routing options in specific contexts, such as an insertion algorithm with the objective of minimizing total vehicle travel time or maximizing total ridership (Horn 2002b) and an evolutionary heuristics approach with a mixed objective (Chevrier et al. 2012). To some extent, the LMTS studied in this paper is a specific variant of a broadly defined DRT concept—namely, a demand responsive transportation system that addresses last-mile service requests with batch passenger demand and a shared passenger origin. Unlike most papers in the literature, we also focus on LMTS routing and scheduling from an optimization and operational perspective, and evaluate the performance of LMTS by applying alternative operational approaches.

Routing and scheduling problems have long been studied, and they comprise a large body of literature; we will mention here only a few of the most influential papers that are relevant to our problem. The vehicle routing problem with time windows (VRPTW) has been the subject of intensive study, using both heuristic and exact optimization approaches. A good review of the VRPTW literature can be found in Bräysy and Gendreau (2005a, b). The dial-a-ride problem (DARP) and related variations have also been extensively investigated by, e.g., Jaw et al. (1986) and Lei, Laporte, and Guo (2012). Cordeau and Laporte (2007) provide a good critical review of the DARP literature. In addition, scheduling for multiserver vehicle systems is an important problem that has been studied in diverse contexts; examples include Liu and Liu (1998), Zee, Harten, and Schuur (2001), and Lee, Mazumdar, and Shroff (2006).

By contrast to the systems and problems that have been studied previously, routing and scheduling for LMTS has the following features:

(1) passengers requesting last-mile service arrive in batches at the metro station, instead of individually, and passengers in the same batch have the same service time window;

(2) passengers requesting last-mile service share a last-mile origin (which is also the vehicle depot)—i.e., the metro station at which they were discharged from trains;

(3) the objective is to improve LOS by minimizing passenger waiting time and riding time, rather than to reduce operating costs; the latter is the main goal, for instance, of VRPTW.

In summary, LMTS presents a general-capacity, multivehicle routing and scheduling problem, with successive batches of passengers having the same time windows and a shared origin. These features provide intuitive groupings of the relevant operational decisions and provide incentives to identify heuristic approaches. For example, batch demands inspire grouping of route decisions between successive batches, which we use in a tabu search metaheuristic (definition of solution attributes in Section 5) and a two-stage method (definition of first-stage decision variables in Section 6).

The paper's contributions are twofold: first, the development of routing and scheduling approaches for an innovative urban transportation system concept—the LMTS—and second, the assessment of the performance of these approaches and of the LMTS concept, in general. Specifically, from an operational and optimization perspective, we formulate an exact mixed-integer programming (MIP) model for optimizing routing and scheduling decisions for a generic LMTS. Next, from an operational and optimization perspective, we develop three computationally feasible heuristics that employ different types of demand information: a myopic approach for the case in which no advanced demand information is available; and a tabu search metaheuristic and a two-stage MIP heuristic, both of which employ full advance demand information. Last, we evaluate the performance of LMTS under these routing and scheduling approaches, and compare them with the performance of a conventional service system with fixed routes and schedule.

The paper is organized as follows. In Section 2, we describe in detail the operational LMTS problem and discuss the associated fundamental assumptions. In Section 3, we propose an exact MIP model for the LMTS routing and scheduling problem. Section 4 describes a myopic operating strategy that uses only demand information from trains that have already arrived, and could easily be implemented and used in practice; the strategy provides a default solution, which is particularly valuable when advance demand information is not available. Section 5 presents a fast tabu search metaheuristic that employs demand information over the entire service horizon and, in most contexts, is superior to the solution provided by the myopic operating strategy. Section 6 proposes a two-stage heuristic method for solving the exact MIP model presented in Section 3 over the entire service horizon. Section 7 defines a set of test instances and performs a number of computational experiments to evaluate the performance of LMTS and demonstrate the approaches described in Sections 3–6 in a number of settings. Finally, Section 8 contains a summary and concluding remarks.

2. Problem Description

We now describe in more detail the problem depicted in Figure 1. The LMTS, which serves a transit metro station, operates as follows: Any passenger who needs last-mile service is required to register in a service-reservation system (either via a smartphone application or website), by selecting the LM stop closest to her final destination. The passenger can provide advance notice of her arrival time at the metro station—i.e., the time at which she will need last-mile service. In practical terms, advance notice could be generated in a number of ways, with each entailing different length of advance notices and service horizons. For example, at one extreme, consider the case in which all passengers are regular subscribers. Each passenger follows a known schedule every day (e.g., Request last-mile service from metro station X to LM Stop 1 at 6:00 P.M. from Monday to Friday), and the metro service is punctual and reliable. This offers the LMTS operator lengthy advance notice of each passenger's service requirement and a service horizon that could span a sequence of many metro arrivals (e.g., We will serve about 65 passengers every afternoon, with known destinations, who will arrive in a sequence of six metro trains that reach the station between 5:00 and 6:00 P.M.). In this environment, the LMTS operator aims to optimize service to the entire (known) set of passengers over the entire service horizon (in this instance, 5:00–6:00 P.M.).

By contrast, if service subscribers have schedules that vary from day to day (or if the metro system is crowded and unreliable), service requests may be known only a short time before passengers arrive at the station. In that case, a passenger could use a smartphone application or tap a smart card on a special-purpose screen to send the service request when she enters any station for the purpose of traveling to the specific station with last-mile service or when she enters her train. The resulting message to the LMTS includes her arrival time at the metro station with last-mile service (which is easy to predict once the passenger boards a train) and her ultimate LM stop. Thus, the advance notice is on the order of 10–20 minutes, and the LMTS operator can plan the service only for passengers arriving on one of the incoming (e.g., one to three) metro trains. In an extreme case, if passengers cannot provide advance notice and cannot reserve LMTS service in advance (due to a lack of reservation systems or because of highly uncertain schedules for both passengers and metros), the LMTS operator is forced to plan service using solely demand information from trains that have already arrived at the station.

Since the number of LM stops served by the LMTS is finite, the number of possible vehicle routes (i.e., sequences of LM stops in a delivery trip, or route) is also finite (although potentially large). We assume that, based on the service region's geometry and historical demand data, we can preselect a set of feasible routes that are practical in the sense of satisfying some typical constraints, such as limits on the maximum number of LM stops on a single route or the route's maximum travel distance or travel time. For each feasible route, a traveling salesman problem (TSP) heuristic or exact algorithm can be used to obtain the optimal sequence for LM stops (with shortest total travel distance/time as the criterion) and the corresponding travel distance/time to each LM stop.

Once the set of preselected routes is specified, the LMTS operations problem consists of determining, for each problem instance, the vehicle routes and schedules that will actually be used to deliver passengers to their LM stops. Operational decisions include assignment of each passenger to a vehicle, route selection, and each vehicle's schedule. If a route is selected in the final operational decision, a vehicle will be dispatched to visit all of the LM stops specified on that route, according to a corresponding schedule; if there are no passengers destined for a particular LM stop, the system will choose a route that does not include that stop. Drivers are given a detailed plan that specifies the route, schedule, and number of passengers for each LM stop on every service trip (e.g., Depart the station at 6:36 P.M. on the Stop 3–Stop 1–Stop 2 route; deliver 4 passengers to Stop 3, 2 passengers to Stop 1, and 3 passengers to Stop 2). Each passenger will receive a message (on her smartphone or by tapping her smart card on a screen when she arrives at the station) with the vehicle she has been assigned to, the scheduled departure time from the station, the route, and the scheduled arrival time at her LM stop (e.g., Board Vehicle #123, which will depart the station at 6:36 P.M.; the route will be Stop 3–Stop 1–Stop 2; you will arrive at your destination, LM Stop 1, at 6:41 P.M.). After completing each trip, the vehicle returns to the station to pick up passengers for its next trip.

We summarize our assumptions: (1) LM stops are prespecified; (2) a sufficiently large set of feasible routes for LMTS vehicles is preselected based on geometry, demand patterns, and some practical constraints; (3) with advance notice of demand, each passenger's arrival time and destination LM stop (i.e., demand information) will be known for a prespecified time period; and (4) the delivery fleet consists of m vehicles, each with integer capacity c. We aim to create a detailed plan for fleet operations, with the objective of minimizing a weighted sum of passenger's waiting time before boarding a vehicle and in-vehicle riding time. We will then evaluate the system's performance compared to a conventional service.

As noted earlier, aside from demands that have already arrived at the station, the length of advance notice depends on the practical implementation and service-reservation requirements of the LMTS. In this paper, we assume that if a demand becomes known through advance notice (whether 10 minutes or hours before arrival at the station), that demand will materialize exactly as expected. In some contexts, this "deterministic" version of the LMTS operations problem is a reasonable approximation to reality. Its solution can also serve as a benchmark for contexts that contain stochastic variability.

3. MIP Formulation

In this section, we present an exact MIP model for the LMTS operations problem described in Section 2, assuming that the LMTS operator receives advance notice of every passenger's service requirement and has a service horizon that spans a sequence of several metro arrivals. The LMTS operator aims to optimize service to the entire (known) set of passengers over the entire service horizon. We introduce the following notation in Table 1.

In this formulation, we discretize the time into intervals of one minute so that we can approximate what will happen in practice. The objective function (1) is defined as minimizing the weighted sum of the time spent by all passengers in the LMTS—i.e., waiting time before boarding a vehicle and in-vehicle riding time

$$\text{minimize} \quad \beta_w \sum_t \sum_j r_j^t + \beta_r \sum_t \sum_j \sum_k t_{jk} z_{jk}^t. \qquad (1)$$

For example, since we discretize the time into one-minute intervals and r_j^t counts the unserved passengers at the end of time t, if $r_j^t = 20$ for some t, this

Table 1. Notation for the Exact MIP Model

Parameters	
J:	number of prespecified LM stops
K:	number of preselected routes
n_j^t:	number of passengers with destination at LM stop j arriving at the station at time t; obtained from train schedules and the last-mile's service-reservation system
ϕ_{jk}:	1 if LM stop j is served by route k; zero otherwise
t_k:	total service time of route k
t_{jk}:	travel time to LM stop j on route k
m:	number of vehicles in the fleet (fleet size)
c:	maximum number of passengers served by a vehicle (vehicle capacity)
β_w:	weight of passenger waiting time before boarding in the objective function
β_r:	weight of passenger in-vehicle riding time in the objective function
Decision variables	
z_{jk}^t:	number of passengers with destination at LM stop j assigned to route k at time t
w_k^t:	number of trips on route k initiated at time t
Intermediate variables	
r_j^t:	number of unserved passengers with destination at LM stop j waiting at the station at the end of time t
v^t:	number of available vehicles at the station at the end of time t

means that 20 minutes of waiting were added during the minute t for passengers going to stop j. If $\beta_w = \beta_r$, the objective is to minimize the sum of total passenger waiting and riding times. This, of course, is equivalent to minimizing the average elapsed time from a passenger's arrival at the station to her final delivery at the destination LM stop.

The formulation has the following constraints:

(a) Passenger flow constraints: Since all passengers have one shared origin (the metro station), expressions (2)–(4) define and constrain the number of unserved passengers with destinations at each LM stop waiting at the metro station at the end of each time t

$$r_j^0 = n_j^0 - \sum_k z_{jk}^0 \cdot \phi_{jk}, \quad \forall\, j, \tag{2}$$

$$r_j^t = r_j^{t-1} + n_j^t - \sum_k z_{jk}^t \cdot \phi_{jk}, \quad \forall\, j, t \geq 1, \tag{3}$$

$$r_j^t \geq 0, \quad \forall\, j, t. \tag{4}$$

(b) Vehicle flow constraints: Since all vehicles have one shared origin (the metro station), expressions (5)–(7) define and constrain the number of available vehicles waiting at the metro station at the end of each time t

$$v^0 = m - \sum_k w_k^0, \tag{5}$$

$$v^t = v^{t-1} + \sum_k w_k^{t-t_k} - \sum_k w_k^t, \quad \forall\, t \geq 1, \tag{6}$$

$$v^t \geq 0, \quad \forall\, t. \tag{7}$$

(c) Service capacity constraints: Expression (8) guarantees the vehicle service capacity restriction

$$\sum_j z_{jk}^t \cdot \phi_{jk} \leq c \cdot w_k^t, \quad \forall\, k, t. \tag{8}$$

(d) Domains of decision variables

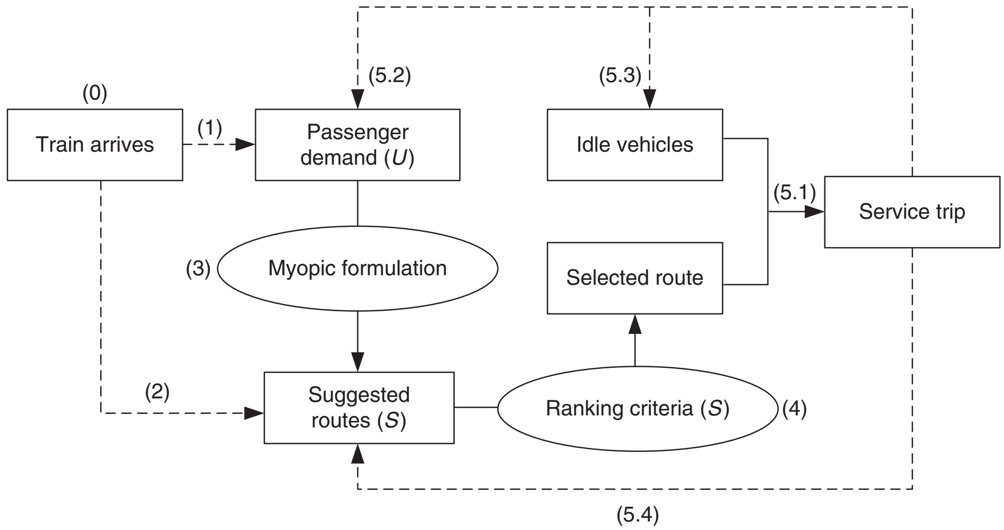
$$w_k^t \in \mathbf{Z}^*, \quad z_{jk}^t \in \mathbf{R}^* \quad \forall\, j, k, t. \tag{9}$$

The model described above deviates from a class of traditional vehicle routing and scheduling problems by combining the following features in a single formulation: (1) the routing and scheduling of a multivehicle fleet with general capacity; (2) sequences of batch demands over a period of time; (3) the same service release time (arrival time at the metro station) and the same service time window for all passengers in a single batch; (4) one shared origin (which is also the vehicle depot) for all service requests; and (5) a performance metric that measures the waiting time and riding time of passengers, instead of the travel distance/time of vehicles. The problem that these features represent can be denoted as the $(P, cap\mathbf{c}|S, r_j|Graph|\sum_j C_j)$ problem of de Paepe et al. (2004), which has been proved in that paper to be NP-hard. Therefore, it is not surprising that it is difficult to obtain optimal, or even near-optimal, solutions for large-scale instances of our MIP model. We can explore the features of this problem to develop computationally feasible heuristic approaches for application in different contexts, as described in Sections 4–6.

4. Myopic Operation

In this section, we describe a myopic approach to solving the LMTS operations problem using only demand information from trains that have already arrived at the station. When a train (batch of passengers) arrives at the station, we consider (1) the new passengers arriving on that train, and (2) any previously arrived passengers who are already waiting for last-mile service. On arrival of each train, the LMTS operator specifies assignments and delivery routes for passengers in both classes. The myopic approach is easy to implement in practice and can provide quick solutions. It is also important to note that the myopic approach may offer a default solution to the LMTS operations problem, especially when no advance demand information is available to the operator—i.e., a passenger's last-mile service request becomes known only when she actually arrives at the station. This is currently the case in most urban transportation systems, and thus myopic models can be of practical interest.

Figure 2. Procedure for a Myopic Approach



4.1. Procedure for Myopic Operation

Under the myopic approach, when a train arrives at the station, we solve the LMTS operations problem for the current state of the system without any look-ahead. Let u_j be the number of passengers waiting at the station destined for LM stop j after a train arrives. Note that u_j includes both any newly arrived passengers with destination j and any previously unserved passengers destined for j. Based on demand $U = \{u_1, u_2, \ldots, u_J\}$ following a train's arrival, we can determine a set of passenger assignments to vehicles and the set of routes $S = \{k_1, k_2, \ldots\}$ that these vehicles will travel. It is important to note that (1) S is based on service requests that have already arrived at the station; (2) S is updated (emptied and redetermined) after the arrival of each train; and (3) we could have multiple identical routes in S—i.e., there exists $i \neq j$, such that $k_i = k_j, k_i \in S, k_j \in S$. For example, if a large number of passengers from one train request last-mile service to LM stop j, several identical routes could serve that single stop within a short period of time. The procedure used to perform the myopic operations is illustrated in Figure 2 and described in Table 2.

The process is repeated every time a train (demand batch) arrives until the end of the service horizon for the LMTS. The arrival of a train is a natural trigger for the LMTS operator to redetermine a new set of routes given the new batch of passengers.

4.2. Myopic Formulation

In step 3 of the myopic approach, the set of routes S is updated (emptied and redetermined) after the arrival of each train. Updated solutions are based on the service requests of the passengers from the train that has just arrived at the station and the previously unserved passengers waiting at the station. An MIP model, myopic formulation, is proposed to make route suggestions after the arrival of each train. Let MF_i denote the myopic formulation for the decision epoch after the arrival of train i. Additional notation for MF_i is introduced in Table 3.

The Myopic Formulation MF_i is defined as follows:

$$\text{minimize} \quad \beta_1 \cdot g + \beta_2 \cdot \sum_k t_k \cdot w_k + \beta_3 \cdot \sum_j \sum_k t_{jk} \cdot z_{jk} \tag{10}$$

$$\sum_k z_{jk} \cdot \phi_{jk} = n_j^{i,U} + n_j^i, \quad \forall j, \tag{11}$$

$$\sum_j z_{jk} \cdot \phi_{jk} \le c \cdot w_k, \quad \forall k, \tag{12}$$

$$\sum_k w_k = g, \tag{13}$$

$$g, w_k \in \mathbf{Z}^*, \quad z_{jk} \in \mathbf{R}^*, \quad \forall j, k. \tag{14}$$

The first part of objective function (10) aims to reduce passenger waiting time before boarding (a smaller g means more ride sharing and earlier boarding); the second part of the objective function also aims to reduce

Table 2. Procedure for a Myopic Approach

(0) Whenever a new train arrives at the station:
- (1) Update demand U using information about the newly arrived passengers.
- (2) Empty the set of suggested routes S.
- (3) Determine a new set of routes S (use an MIP model, Myopic Formulation, Section 4.2).
- (4) Determine the priority of the routes in S (use ranking criteria, Section 4.3).
- (5) Whenever there are idle vehicles:
  - (5.1) Dispatch an idle vehicle to serve the route with the highest priority in S.
  - (5.2) Update the status of passengers.
  - (5.3) Update the status of vehicles.
  - (5.4) Delete the selected route from S.

Table 3. (Additional) Notation for Myopic Formulation MF_i

Parameters:	
$n_j^{i,U}$:	number of unserved passengers with destination at LM stop j before the arrival of train i
n_j^i:	number of passengers with destination at LM stop j arriving on train i
$\beta_{\{1,2,3\}}$:	coefficients in the objective function
Decision variables:	
z_{jk}:	number of passengers with destination at LM stop j assigned to a route k in the decision epoch
w_k:	number of trips on route k suggested in the decision epoch
g:	total number of trips needed in the decision epoch

passenger waiting time (from a queueing perspective, less total service time in a service system means a lower "utilization ratio," and therefore less waiting time for customers); the third part of the objective function aims to reduce passenger in-vehicle riding time. In practice, the values for β_1, β_2, and β_3 can be adjusted to incorporate different preferences. In the numerical experiments presented in Section 7, we test different combinations of β_1, β_2, and β_3, and find that a general relation of $\beta_1 > \beta_2 > \beta_3$ will provide high-quality solutions (measured by the sum of passenger waiting time and riding time). Constraint (11) ensures that every passenger is assigned to a route; (12) guarantees that the vehicle capacity is not exceeded; (13) captures the total number of trips needed; and (14) defines decision-variable domains, in which the integrality of z_{jk} can be relaxed.

Note that, in the myopic approach, we do not plug the limited information about arrived demand directly into the MIP model in Section 3 for route selection. When there is no future demand, the MIP model in Section 3 favors dispatching direct service to an LM stop (i.e., a route with a single LM stop) for the last trip of each vehicle. This, in turn, generates vehicle schedules with late completion times and makes it difficult for vehicles to return to the station early, even when the next train will arrive soon. Essentially, the advantage of the MIP model in Section 3 is its ability to consider interactions between demands from different batches. This advantage is diminished if information about future demand is not available. To some extent, the myopic formulation proposed in this section views the LMTS as a queue. The larger weights of $\beta_1 g$ and $\beta_2 \sum_k t_k w_k$ in its objective function will generate a relatively lower long-term utilization ratio for the LMTS from a queueing perspective, and provide a good LOS. In computational experiments, we test and find that the myopic formulation (with certain combinations of β_1, β_2, and β_3) performs well.

When g is set to a fixed value, the formulation MF_i is exactly the same as the traditional capacitated facility location problem (CFLP) if ϕ_{jk} is not taken into consideration. The decision variable w_k is analogous to the choice of location for the facility, and z_{jk} is analogous to the assignment of demand to the chosen facility. CFLP is a well-studied problem: Sridharan (1995) reviews various heuristic and exact methods for CFLP. When solving MF_i, we begin by setting the decision variable $g = \lceil (\sum_j (n_j^{i,U} + n_j^i))/c \rceil$. Because of the topological relations between preselected routes and LM stops, MF_i may be infeasible with this initial value of g. We increase g by one whenever MF_i is infeasible. When the number of LM stops is not large (e.g., J does not exceed 20), the corresponding MF_i can be solved directly and quickly with common commercial optimization software, such as CPLEX.

4.3. Ranking Criterion

The myopic formulation MF_i suggests a set of routes, S, that can be traveled before the next train arrives. Because the number of idle vehicles right after the arrival of train i might be less than the number of routes in S, we may not, in some cases, be able to offer all of the routes in S immediately. Therefore, we need a ranking criterion to determine the order in which the suggested routes in S will be traveled. The routes will then be traveled sequentially, according to the priority specified by the ranking criterion. With the suggested routes prioritized, the only remaining task is to assign passengers to the routes. Generally, a trip with shorter travel time that serves more passengers should be given a higher priority than a trip with longer travel time that serves fewer passengers. Therefore, as a simple criterion, for a route with $w_k = 1$, we use the value $\sum_j z_{jk}\phi_{jk}/t_k$ to set the priority, where $\sum_j z_{jk}\phi_{jk}$ is the total number of passengers the trip on route k will serve, and t_k is the total service time on route k. The suggested routes in S are then ranked and selected in a descending order of $\sum_j z_{jk}\phi_{jk}/t_k$. When $w_k > 1$, we can apply a similar criterion.

This approach is indeed myopic. In the formulation MF_i, any passengers (last-mile service requests) that appear after the arrival of train i have no influence on MF_i decisions. The myopic formulation MF_i essentially suggests a set of routes S based only on demand information from trains that have already arrived, without considering passengers who will arrive in the future. Assuming a situation in which train i brings only one passenger destined for a particular LM stop j and train $i+1$ brings a large number of passengers destined for the same LM stop j, the myopic approach could dispatch a route that serves LM stop j before the arrival of train $i+1$, while more advanced approaches could hold the vehicle until the arrival of train $i+1$. With the holding strategy, passengers destined for LM stop j from both train i and train $i+1$ could share the ride, which might reduce the waiting time of passengers from train $i+1$ due to the immediate availability of the vehicle being held for them. Because of the heuristic ranking criterion, the myopic approach is not myopic optimal. It is, however, easy to implement in practice and can provide quick solutions.

5. Tabu Search

In this section—and by contrast to the absence of information about future demand under the myopic approach—we describe a method based on a tabu search, in which we assume that all demand information over the entire service horizon, which spans a sequence of several metro arrivals, is known and used. In this environment, the LMTS operator aims to optimize service to the entire (known) set of passengers over the entire period of interest.

The tabu search, which was proposed and developed by Glover (1986, 1989, 1990a, b), is a local search metaheuristic that explores the solution space by moving, at each iteration, from the current solution to the best solution in its neighborhood. The main concepts include attributes, neighborhood, moves, a tabu list, aspiration criteria, and termination conditions. The tabu search has been applied, with good results, to various types of routing and scheduling problems. Examples include vehicle routing (Gendreau, Hertz, and Laporte 1994, Cordeau and Maischberger 2012), job-shop scheduling (Hurink, Jurisch, and Thole 1994), nurse scheduling (Dowsland 1998), real-time vehicle routing and dispatching (Gendreau et al. 1999), vehicle routing with time windows (Cordeau, Laporte, and Mercier 2001), split-delivery vehicle routing (Archetti, Speranza, and Hertz 2006), vehicle routing with simultaneous pick-up and delivery service (Montané and Galvao 2006), and dynamic dial-a-ride (Berbeglia, Cordeau, and Laporte 2012).

In what follows, we first introduce the notation and attributes used in the method. We then provide detailed descriptions of tabu search concepts for the LMTS operations problem, assuming that all demand information is known and used.

5.1. Notation and Attributes

Because of the batch demand with the same service release time in the LMTS, it is intuitive and reasonable to view route trips between successive demand batches as a "decision group." Specifically, in this tabu search metaheuristic, we can define solution attributes as the routes of trips initiated during each interarrival time period of trains (batches of passengers). Let T_i denote the arrival time of train i, and let $h_i = [T_i, T_{i+1})$ denote the time period between the arrival of train i and the arrival of train $i+1$. We denote the operation solution s as $(R_1, R_2, \ldots, R_I)$, where R_i is the set of routes initiated during time period h_i (R_I is the set of routes initiated after the arrival of the last train; i.e., train I). Note that, in this approach, we make decisions at the beginning of the period of interest by considering all of the requests over the entire service horizon. The set of routes R_i is therefore not the decision we make at time point T_i, as in the myopic approach, but the decision we make for operations during the time period $h_i = [T_i, T_{i+1})$, assuming that demand information over the entire service horizon is known and used. For example, in an LMTS operations problem with three trains, a solution s denoted by $(R_1 = \{k_1, k_3\}, R_2 = \{k_2\}, R_3 = \{k_1, k_4\})$ represents the operation plan in which the vehicle fleet initiates two service trips during the time period h_1, one on route k_1 and the other on route k_3; one service trip on route k_2 during h_2; and two service trips during h_3, one on route k_1 and the other on route k_4.

Note that the solution $s = (R_1, R_2, \ldots, R_I)$ represents only the routes of trips initiated during each time period h_i, while the sequence/priorities of routes within each time period should be determined by some ranking criteria. We can use the same ranking criterion used in the myopic approach described in Section 4.3.

5.2. Neighborhood and Moves

It is important to consider the space (neighborhood) in which the search will be conducted. In this paper, we define two natural neighborhoods. The first involves changing the routes of trips within a single time period h_i: (1) swap LM stops between two trips, (2) shift an LM stop from one trip to another, (3) split a trip into two trips so that one serves a single LM stop and the other the remaining LM stop(s), (4) add an LM stop to a trip, (5) add a trip serving a single LM stop, and (6) eliminate an LM stop from a trip. The second neighborhood involves changing the routes of trips during two consecutive time periods, h_i and $h_{i'}$ ($i' = i+1$ or $i-1$): (1) swap LM stops between a trip in h_i and a trip in $h_{i'}$, (2) shift an LM stop from a trip in h_i to an existing trip in $h_{i'}$, and (3) shift an LM stop from a trip in h_i to a new single LM stop trip in $h_{i'}$. The moves are valid if and only if the new routes generated are feasible according to route preselection requirements. It is obvious that any solution, including the optimal solution, can be obtained by imposing a limited number of moves, as described above, on any other solution.

5.3. Tabu List

Each solution in the neighborhoods described in Section 5.2 contains one or several route changes. In this paper, a move is tabu if any move that reverses a change of route in recent iterations (as recorded in the tabu list) is forbidden. The best size of the tabu list for each kind of problem must be determined empirically, and computational tests must be implemented for different problems. Previous work on similar problems provides observations regarding good tabu list sizes. For example, Cordeau, Gendreau, and Laporte (1997) find that the best tabu list size for solving the periodic vehicle routing problem (PVRP) and multi-depot vehicle routing problem (MDVRP) is $7.5 \log_{10} n$, where n is the number of customers. Other work has shown experimentally that, for certain problems, a tabu list of variable size tends to give better results than a fixed one. For example, Taillard (1991) sets the size of the tabu list to a random number in a specified interval. In our problem, we have tested and compared several simple and common tabu list sizes in a number of computational experiments, in which a fixed size $1 + J/2$ works well—and better than other simple sizes—in most cases (J is the number of prespecified LM stops). Other simple tabu list sizes, either fixed or variable, can easily be tested and implemented.

5.4. Aspiration Criteria and Termination Conditions

With a similar criterion as in the myopic approach to prioritize the routes in each R_i of a solution $s = (R_1, R_2, \ldots, R_I)$, the only remaining task is passenger assignment. The objective value (sum of the passenger waiting time and riding time) can then be evaluated easily. We can use a solution obtained from other approaches, such as the myopic approach described in Section 4 or the two-stage method described in Section 6, as the initial solution to the tabu search metaheuristic. Aspiration criteria, if satisfied, allow moves that override tabus. In our problem, we allow a move that overrides a tabu, if that move results in an objective value that is better than the best known objective value identified so far.

Our termination rule is that the search terminates if a maximum number of total iterations (N_1) is reached or if the best solution so far has not been improved on for a certain number of iterations (N_2). In a number of computational experiments, we find that, with the initial solution from the myopic approach or the two-stage method, the tabu search improves the solutions quickly in the beginning (usually within 200 iterations) and, in most cases, converges rapidly afterward. We set $N_1 = 500$ and $N_2 = 50$ in the computational experiments, while other N_1 and N_2 can be easily implemented.

5.5. Tabu Search Algorithm

Based on the concepts discussed in Sections 5.1–5.4, the tabu search metaheuristic algorithm is described in Table 4.

Table 4. A Tabu Search Algorithm

(0) Obtain an initial solution s_0 from another approach.
  Set best solution $s^* = s_0$; current solution $s^c = s_0$; tabu list $TL = \varnothing$.
(1) REPEAT:
  IF termination condition is satisfied,
    STOP.
  ELSE
    (1.1) For each neighbor in the neighborhood of s^c, calculate the objective value.
    (1.2) Move to the best neighbor that is not tabu or satisfies the aspiration criteria.
    (1.3) Update s^*, s^c, and TL.

6. Two-Stage Method for Solving MIP

In this section, we describe a two-stage heuristic for solving approximately the MIP model presented in Section 3. The LMTS operator uses all demand information and aims to optimize service to the entire (known) set of passengers over the entire service horizon. In this method, we decompose the decisions into two stages. In the first stage, we modify the original exact MIP model proposed in Section 3 to make more aggregate decisions for every time period h_i. Batch demands provide us with intuitive groupings of the decisions. In the second stage, we implement the original exact MIP model with the decision variables (columns) generated using the information revealed from the aggregate decisions in the first-stage solution. Sections 6.1 and 6.2 describe the first and second stages, respectively, in detail.

6.1. First Stage: Solve the MIP to the Level of Time Period h_i

In the first stage, we modify the original exact MIP model by replacing the time dimension t in the decision variables z^t_{jk} and w^t_k with the dimension of each train's ID i. In other words, instead of making detailed decisions for every discretized time instant t, we shift our focus to making more aggregate decisions for every time period h_i in the first stage, assuming that all demand information over the entire service horizon is known and used. This modified model can reduce the problem's scale and the computational time required. The (additional) notation is modified as in Table 5.

The objective function (15) captures part of the passenger waiting time before boarding a vehicle and all of the passenger in-vehicle riding time

$$\text{minimize} \quad \beta_w h_i \sum_i \sum_j r^i_j + \beta_r \sum_i \sum_j \sum_k t_{jk} z^i_{jk}. \tag{15}$$

The modified formulation has the following constraints:

(a) Passenger flow constraints: Expressions (16)–(18) are directly modified from expressions (2)–(4), respec-

Table 5. (Additional) Notation for the First-Stage Model

Parameters:
n^i_j: number of passengers with destination at LM stop j arriving on train i
Decision variables:
z^i_{jk}: number of passengers with destination at LM stop j assigned to route k during the time period h_i
w^i_k: number of trips on route k initiated during the time period h_i
Intermediate variables:
r^i_j: number of unserved passengers with destination at LM stop j waiting at the station at the end of time period h_i

tively. The dimension of time t is replaced with the train's ID i

$$r_j^1 = n_j^1 - \sum_k z_{jk}^1 \cdot \phi_{jk}, \quad \forall j, \tag{16}$$

$$r_j^i = r_j^{i-1} + n_j^i - \sum_k z_{jk}^i \cdot \phi_{jk}, \quad \forall j, i \geq 2, \tag{17}$$

$$r_j^i \geq 0, \quad \forall j, i. \tag{18}$$

(b) Vehicle flow constraints: Note that the decision variables w_k^i describe vehicle operations during the time period h_i. We cannot capture detailed vehicle schedules by replacing w_k^t with w_k^i in constraints (5)–(7), and we cannot guarantee the feasibility of vehicle schedules, as in the original MIP. Therefore, we use heuristic constraints (19)–(22) to replace (5)–(7). Constraints (19) and (20) limit the total number of trips that can be initiated within one and two consecutive intervals between trains, respectively, while constraints (21) and (22) limit the total service time of trips initiated within one and two consecutive intervals between trains, respectively. The values of the upper limits $m_{\max 1}$, $m_{\max 2}$, $t_{\max 1}$, and $t_{\max 2}$ can be set to roughly the values these quantities realize in solutions that use alternative approaches, such as the myopic approach

$$\sum_k w_k^i \leq m_{\max 1}, \quad \forall i = 1, 2, \ldots, I, \tag{19}$$

$$\sum_k (w_k^i + w_k^{i+1}) \leq m_{\max 2}, \quad \forall i = 1, 2, \ldots, I-1, \tag{20}$$

$$\sum_k t_k \cdot w_k^i \leq t_{\max 1}, \quad \forall i = 1, 2, \ldots, I, \tag{21}$$

$$\sum_k t_k \cdot (w_k^i + w_k^{i+1}) \leq t_{\max 2}, \quad \forall i = 1, 2, \ldots, I-1. \tag{22}$$

(c) Service capacity constraints: Expression (23) is modified from expression (8)

$$\sum_j z_{jk}^i \cdot \phi_{jk} \leq c \cdot w_k^i, \quad \forall k, i. \tag{23}$$

(d) Domains of decision variables

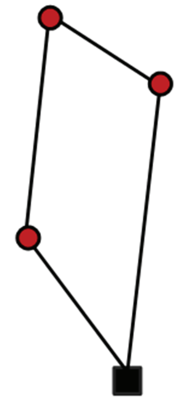
$$w_k^i \in \mathbf{Z}^*, \quad z_{jk}^i \in \mathbf{R}^* \quad \forall j, k, i. \tag{24}$$

Essentially, the modified formulation in the first stage uses the time period h_i as the smallest time unit to make decisions. However, unlike the myopic approach, the formulation does consider all of the service demands over the entire service horizon and the mutual interactions that occur among demands from all of the trains.

6.2. Second Stage: Column Generation in the Original Formulation

The solution of the modified formulation in the first stage highlights the routes that could be provided during each time period h_i. In the second stage, we implement the original exact MIP model proposed in Section 3 with the decision variables (columns) generated

Figure 3. (Color online) Route Selected in the First Stage



using the information revealed in the first-stage solution. Specifically, if $w_k^i > 0$ in the optimal solution of the first-stage problem, we generate vehicle decision variables $w_{k'}^t$ for (1) every time $t \in h_i$ and (2) every route k' that is a subtour of route k, including route k itself. For example, if route k, which serves three LM stops, is selected for the time period h_i in the first-stage solution, as shown in Figure 3, we generate decision variables $w_{k'}^t$ for (1) every time t in $h_i = [T_i, T_{i+1})$ and (2) the $2^3 - 1 = 7$ specific routes, where each route serves a subset of the three LM stops, as shown in Figure 4.

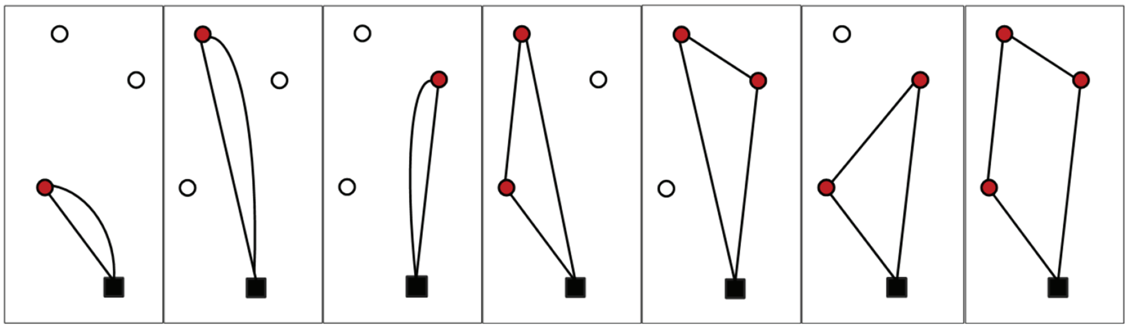
With the decision variables (columns) generated as described above, the exact MIP model in Section 3 can be solved in stage 2 in much less computational time.

7. Computational Study

We now present a computational study based on the approaches described in Sections 4–6. We compare the results of the myopic operating strategy using demand information only from trains that have already arrived at the station, the tabu search metaheuristic over the entire service horizon, and the MIP model over the entire service horizon, which is solved approximately in two stages. A conventional service with fixed routes and schedule is taken as a benchmark to evaluate the performance of LMTS using three operating strategies. Computational experiments were coded in Java and run on 64-bit computers with 3.6 GHz processors and 32 GB RAM. All of the corresponding MIP problems were solved using CPLEX 12, with a time limit of five minutes for each instance.

We first discuss the settings of test instances for the computational experiments in Section 7.1. We then describe, in Section 7.2, a common multivehicle conventional service with fixed routes and schedule that will serve as our benchmark for comparisons. Finally, Section 7.3 presents our computational results, followed by a brief discussion.

Figure 4. (Color online) Routes of Decision Variables Generated in the Second Stage



7.1. Settings of Test Instances

We consider an LMTS in a rectangular service region, as shown in Figure 1. To traverse the region's length and width, LMTS vehicles require 10 minutes and six minutes, respectively. The number of prespecified LM stops (which are assumed to be randomly and uniformly distributed throughout the region), J, takes values ranging from six to 12 in the experiments. We assume that vehicles travel according to the Euclidean metric, and the service time at each LM stop (e.g., for vehicle deceleration, loading/unloading of passengers, and vehicle acceleration) is set to one minute. Feasible routes are selected to satisfy the requirements that (1) the maximum number of LM stops on a route is three and (2) the maximum total service time (travel time + service time at stops) on a route is 14 minutes. The number of feasible routes, K, under such conditions will be in the region of 100–300.

Trains with passengers arrive at the metro station every 10 minutes. The size of a passenger batch from each train is assumed to be Poisson-distributed with intensity N taking values ranging from 10 to 30. Passenger destinations are assumed to be distributed among the LM stops either (1) uniformly (UN), (2) slightly heterogeneously (SH), or (3) extremely heterogeneously (EH). An example of demand intensity at LM stops in a case with $J = 8$ and $N = 16$ is shown in Table 6. Advance demand notices from 10 trains are assumed to be known (and used in the tabu search heuristic and two-stage method).

A fleet of vehicles with an identical capacity with values ranging from three to 12 serves the LMTS. Experiments with different fleet sizes are used to evaluate the performance of LMTS in situations of high and low vehicle utilization. For each combination of parameter settings (J, N, m, c) and passenger distribution type (UN, SH, or EH), we carry out 10 test instances. In each instance, we generate random locations for the LM stops and random sizes of passenger batches from each train based on Poisson-distributed demand.

Table 6. Demand Intensity at LM Stops

$J = 8$, $N = 16$	Highest demand over lowest demand	Demand intensity at LM stops
UN	2.0/2.0 = 1	2.0 for all
SH	3.0/0.5 = 6	0.5, 1.5, 1.5, 2.0, 2.0, 2.5, 3.0, 3.0
EH	4.0/0.2 = 20	0.2, 0.6, 1.0, 1.8, 2.2, 2.8, 3.4, 4.0

7.2. Conventional Service with Fixed Routes and Schedule

To study the potential advantages of on-demand service in LMTS, we introduce a multivehicle conventional transportation system with fixed routes and schedule as a benchmark for comparison. Ceder and Wilson (1986) have summarized various approaches to the design of conventional bus networks. For different settings and situations, the selection of bus routes and schedule normally considers multiple criteria, such as vehicle or bus route service time, passenger travel time, and transfer time. Many heuristics have been developed for this purpose; these typically share some common considerations, such as (a) limit on the maximum passenger detour time, (b) upper and lower bounds on route length, (c) demand/supply balance across all stops, and (d) a specified number of available routes. We have applied a simple integer programming model (25)–(30) that includes the considerations common to typical approaches. Given a vehicle fleet with fleet size m and vehicle capacity c, the model selects fixed bus routes from a set of possible routes and determines the number of buses that will serve each selected route. For possible bus routes, unlike the route preselection requirements in the LMTS, the limit on the maximum number of stops in a route is removed for the conventional service. The bound on the maximum route length is increased for the conventional service. The model with equal input values of m and c as in the LMTS represents a conventional bus system using the same vehicle fleet. However, in practice, the conventional bus system typically uses a vehicle fleet that has fewer and larger vehicles—i.e., smaller m and larger c—compared to on-demand systems. Table 7 introduces the relevant notation.

Table 7. (Additional) Notation for a Bus Route Design for Conventional Service

Parameters:	
t_d:	upper limit on the service time difference between any pair of bus routes
c_d:	upper limit on the demand/supply imbalance across all stops
d_k:	number of stops served on route k
λ_j:	passenger demand rate to LM stop j
M:	a large positive number
Decision variables:	
x_k:	integer variable to indicate the number of buses serving route k in the conventional service

The integer programming model is defined as follows:

$$\text{minimize} \quad \sum_k t_k \cdot x_k \tag{25}$$

$$\sum_k \phi_{jk} \cdot x_k \geq 1, \quad \forall j, \tag{26}$$

$$\frac{\sum_k c \cdot \phi_{jk} \cdot (x_k/(d_k \cdot t_k))}{\lambda_j} - \frac{\sum_k c \cdot \phi_{j'k} \cdot (x_k/(d_k \cdot t_k))}{\lambda_{j'}} \leq c_d, \quad \forall j, j' \in \{1, \ldots, J\}, \tag{27}$$

$$\sum_k x_k = m, \tag{28}$$

$$t_k - t_{k'} \leq M(2 - x_k - x_{k'}) + t_d, \quad \forall k, k' \in \{1, \ldots, K\}, \tag{29}$$

$$x_k \in \mathbf{Z}^*, \quad \forall k. \tag{30}$$

Objective function (25) minimizes the total service time of the selected bus routes, which is a common option in the bus design problem, and constraint (26) ensures that every LM stop is served by at least one bus route. The limit on the maximum route length partially takes care of points (a) and (b) mentioned above in connection with the bus design problem; constraint (27) ensures that demand/supply is balanced across all stops (point (c) above); constraint (28) sets the number of buses equal to the number of available vehicles in the fleet, which is a common option concerning point (d) in the bus design problem; constraint (29) ensures that the service time difference between any pair of selected bus routes does not exceed an upper limit, which also partially addresses points (a) and (b); and constraint (30) defines the domains of the integer decision variables.

As for the input values of the integer programming model, the passenger demand rate λ_j is obtained by calculating the average number of passengers per unit time requesting last-mile service to LM stop j using demand information over the entire service horizon. Different combinations of threshold parameters t_d and c_d are tested and compared. For any $x_k^* > 0$ in the optimal solution, we assume that x_k^* buses will serve route k evenly with a uniform service headway ($=t_k/x_k^*$); therefore, the frequency of buses serving route k is equal to x_k^*/t_k. We simulate the conventional bus system serving the selected routes with fixed routes and schedule and evaluate the system's LOS (in terms of passenger waiting time and riding time). We can then choose the values of t_d and c_d that maximize the LOS.

7.3. Results and Discussion

Tables 8–13 display the objective values (passenger waiting time + riding time) and the computational time associated with different operating strategies with diverse parameter settings. The myopic approach uses demand information for passengers that have already arrived at the station. The MIP two-stage method uses demand information over the entire service horizon. The tabu search method uses initial solutions obtained from the myopic approach and the MIP two-stage method, respectively. UN denotes uniform demand among LM stops; SH denotes slightly heterogeneous demand (the ratio of the highest demand over the lowest demand is six); and EH denotes extremely heterogeneous demand (the ratio is 20). The objective value is in minutes, and the running time (per instance) is in seconds.

Taking the conventional service with fixed routes and schedule as our benchmark, it is obvious that all of the nonnaïve methods for operating LMTS provide better service—i.e., with reduced passenger waiting time and riding time. The myopic operating strategy uses demand information from trains that have already arrived at the station, which can be easily implemented in LMTS; this method could reduce passenger waiting time and riding time significantly, compared to the conventional service. For example, in the UN case in Table 8 and Figure 5, total passenger waiting time and riding time in the LMTS myopic strategy is only

Table 8. Results for $J = 8$, $N = 16$, $c = 6$, and $m = 3$

	UN		SH		EH	
$J = 8$, $N = 16$ $c = 6$, $m = 3$	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	9.55 + 4.37		12.49 + 4.10		14.85 + 4.51	
Myopic	4.12 + 3.60	0.6	4.22 + 3.30	0.5	3.42 + 2.29	0.6
Myopic + tabu	2.11 + 3.37	0.6 + 5	2.39 + 3.07	0.5 + 4	1.97 + 3.08	0.6 + 4
Two-stage	**2.13** + **3.06**	121	**2.05** + **2.88**	81	**1.88** + **2.90**	107
Two-stage + tabu	2.00 + 3.20	121 + 5	2.10 + 2.99	81 + 4	1.87 + 2.95	107 + 4

Table 9. Results for $J = 8$, $N = 16$, $c = 6$, and $m = 7$

	UN		SH		EH	
$J = 8$, $N = 16$ $c = 6$, $m = 7$	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	1.61 + 4.83		1.24 + 3.17		1.40 + 3.90	
Myopic	0.06 + 4.00	0.5	0.01 + 3.60	0.5	0.01 + 3.81	0.5
Myopic + tabu	0.07 + 3.35	0.5 + 4	0.01 + 2.93	0.5 + 5	0.02 + 3.32	0.5 + 4
Two-stage	0.07 + 3.24	73	0.02 + 2.84	130	0.06 + 3.23	55
Two-stage + tabu	**0.06** + **3.24**	73 + 4	**0.01** + **2.84**	130 + 6	**0.03** + **3.23**	55 + 5

Table 10. Results for $J = 12$, $N = 30$, $c = 6$, and $m = 5$

	UN		SH		EH	
$J = 12$, $N = 30$ $c = 6$, $m = 5$	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	15.81 + 4.10		19.16 + 4.22		25.98 + 4.53	
Myopic	4.72 + 3.50	0.4	4.68 + 3.63	1.6	5.06 + 3.51	1.4
Myopic + tabu	2.97 + 3.30	0.4 + 36	2.99 + 3.45	1.6 + 24	3.30 + 3.45	1.4 + 25
Two-stage	**2.52** + **3.17**	512	**2.35** + **3.35**	452	**2.69** + **3.35**	435
Two-stage + tabu	2.72 + 3.21	512 + 28	2.52 + 3.41	452 + 19	2.86 + 3.42	435 + 20

Table 11. Results for $J = 12$, $N = 30$, $c = 6$, and $m = 7$

	UN		SH		EH	
$J = 12$, $N = 30$ $c = 6$, $m = 7$	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	5.55 + 4.12		11.68 + 3.66		13.88 + 3.88	
Myopic	1.05 + 3.80	1.8	1.24 + 3.55	1.5	1.02 + 3.89	1.3
Myopic + tabu	0.60 + 3.40	1.8 + 44	0.75 + 3.26	1.5 + 31	0.52 + 3.64	1.3 + 24
Two-stage	**0.60** + **3.29**	342	**0.71** + **3.13**	310	**0.51** + **3.56**	277
Two-stage + tabu	0.58 + 3.32	342 + 43	0.74 + 3.20	310 + 27	0.56 + 3.58	277 + 22

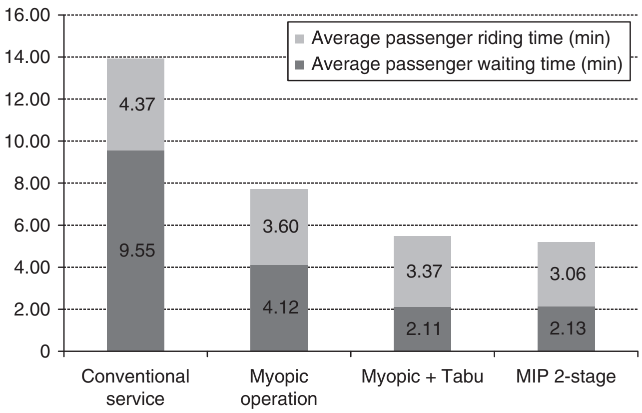
Table 12. Results for $J = 6$, $N = 24$, $c \times m = 24$ for UN

	$c = 12$, $m = 2$		$c = 8$, $m = 3$		$c = 4$, $m = 6$	
$J = 6$, $N = 24$ $c \times m = 24$, UN	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	19.78 + 5.74		11.90 + 5.00		7.64 + 3.11	
Myopic	9.43 + 3.52	0.4	5.81 + 3.53	0.3	2.26 + 3.29	0.3
Myopic + tabu	6.71 + 3.50	0.4 + 1	3.60 + 3.40	0.3 + 2	1.53 + 3.21	0.3 + 3
Two-stage	**5.73** + **3.38**	44	**3.07** + **3.27**	49	**1.33** + **3.15**	121
Two-stage + tabu	4.71 + 4.54	44 + 1	3.25 + 3.32	49 + 1	0.50 + 3.15	121 + 2

Table 13. Results for $J = 12$, $N = 18$, $c \times m = 24$ for UN

	$c = 12$, $m = 2$		$c = 8$, $m = 3$		$c = 4$, $m = 6$	
$J = 12$, $N = 18$ $c \times m = 24$, UN	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)	Waiting time + Riding time (min)	Running time (sec)
Conventional	22.55 + 6.99		12.05 + 5.18		4.97 + 4.09	
Myopic	6.75 + 4.88	1.9	2.75 + 4.55	1.3	1.46 + 3.76	1.1
Myopic + tabu	4.06 + 4.27	1.9 + 96	1.58 + 3.95	1.3 + 93	0.96 + 3.45	1.1 + 27
Two-stage	3.95 + 3.89	612	**1.73** + **3.54**	605	**0.86** + **3.32**	271
Two-stage + tabu	**3.66** + **4.14**	612 + 79	1.63 + 3.66	605 + 89	0.89 + 3.37	271 + 22

Figure 5. Passenger Waiting Time and Riding Time for $J = 8$, $N = 16$, $c = 6$, and $m = 3$



55.5% of that in the conventional service. In addition, it can be seen that the advanced operating strategies that use demand information over the entire service horizon—i.e., the tabu search metaheuristic and the MIP two-stage method—provide even better routing and scheduling solutions. In the same instances, the tabu search, using the myopic solution as its initial solution, provides an objective value that is only 39.4% of that for the conventional service, and the MIP two-stage method provides an objective value that is only 37.3% of the conventional service. The tabu search that uses the MIP solution as its initial solution does not yield much further improvement. Since heuristics based on a tabu search have generated high-quality solutions in various similar problems (e.g., Cordeau, Laporte, and Mercier 2001, Archetti, Speranza, and Hertz 2006, Cordeau and Maischberger 2012), we can view this limited improvement as an encouraging indication of the MIP two-stage method's superior quality of solutions in general. Note that compared with the conventional service, LMTS waiting time is noticeably reduced, while riding time is reduced to a lesser degree.

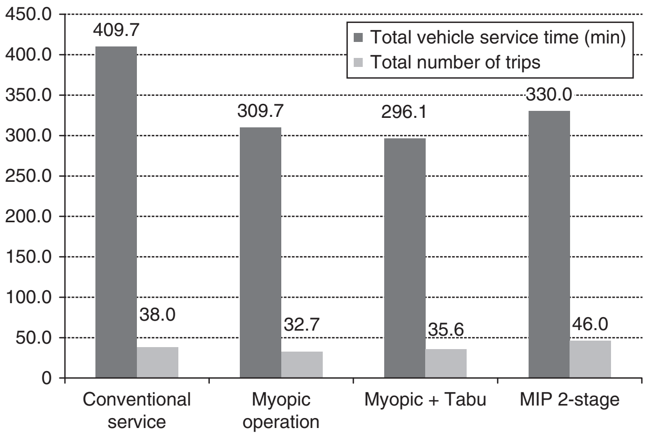
In terms of computational time, the myopic operating strategy can provide solutions in seconds, while the computational time of the tabu search metaheuristic depends on the parameters in the search termination conditions: if the maximum total number of iterations (N_1) is 500 and the maximum number of iterations without improvement (N_2) is 50, the tabu search improves the solutions quickly in the beginning and converges rapidly afterward. It usually terminates within 200 iterations with the parameters described above, and requires computational times that range from several seconds in small cases (small J and K) to one to two minutes in large cases (large J and K). The MIP two-stage method requires the longest computation time.

The advantage of a flexible LMTS is greater when vehicle capacity is small than when it is large. Tables 12 and 13 display results for systems with the same geometric configuration, same passenger demand, and an equal total vehicle capacity ($c \times m = 24$) in the fleet. We evaluate three cases: (1) $c = 12$, $m = 2$, minibuses; (2) $c = 8$, $m = 3$, minivans; and (3) $c = 4$, $m = 6$, normal taxis. We observe the following. First, if the two systems have exactly the same number m of vehicles and the same capacity c per vehicle, LMTS performs better than the conventional system in each of the three cases. Second, as vehicle capacity c gets smaller, both the LMTS and the conventional system will perform better (with shorter waiting time and riding time), due to the increased flexibility and more customized service. Third, if we only require that the total capacity ($c \times m$) of the two systems stays the same (but do not necessarily have the same c and m for the two systems), we would expect that, in practice, the advantage of LMTS over the conventional system will increase, as the conventional system is more likely to use larger vehicles and fewer vehicles than LMTS. Stated differently, if we compare an LMTS that uses small vehicles to a conventional service with large vehicles (with an equal $c \times m$), the advantages of the flexible LMTS become even greater.

In some cases with more heterogeneous demand distribution, such as EH in Tables 8 and 11, average passenger waiting time in the conventional service is greater than 10 minutes (the train headway), which means that passengers from one train will join the same queue with those from previous trains; after a couple of minutes, passengers who arrived at the station earlier can board a bus, while passengers who arrived later must wait for the next bus. When passenger destinations become more heterogeneous, passengers are more likely to suffer long waiting times (e.g., greater than 10 minutes). This is because conventional buses must visit every stop (even ones with zero passenger demand on a trip) on their predesigned routes. Therefore, with heterogeneous demand distribution, it is more likely that buses in a conventional system will visit a number of unnecessary stops.

In the case of low demand and low vehicle utilization, it is highly probable that a batch of passengers from any particular train will be served before the next train arrives. In these conditions, the performance of the myopic operating strategy (without using advance demand notice) is quite close to that of the tabu search or the MIP method (both of which use demand information over the entire service horizon). Stated differently, information about passengers on the next train does not help much to improve the system's performance, if most passengers on the current train can be served before the next train arrives. By contrast, when demand and vehicle utilization are high, it is better to employ advance demand information and consider demands from all trains. For example, Tables 8 and 9 have the same system parameters,

Figure 6. Vehicle Service Time and Number of Trips for $J = 8$, $N = 16$, $c = 6$, and $m = 3$



with the exception of the number of vehicles. Table 8 shows a high-utilization situation with three vehicles, while Table 9 presents a low-utilization case with seven vehicles. In the UN case, the tabu search metaheuristic (or the MIP two-stage method) uses demand information over the entire service horizon to reduce passenger waiting time and riding time, compared to the myopic operating strategy, by 15.8% (or 18.5%) in the low-utilization case, while the improvement is 29.0% (or 32.8%) in the high-utilization case. The same trend can be found in Tables 10 and 11: the improvement is 17.5% (or 19.8%) for low utilization and 23.7% (or 30.8%) for high utilization.

An LMTS can also have positive effects on vehicle utilization by reducing total vehicle service time (Figure 6) compared to a conventional system. Specifically, the LMTS that employs the tabu search metaheuristic, starting from myopic solutions, obtains operating plans with good service quality and the shortest vehicle service time. The LMTS that uses the MIP two-stage method obtains operating plans with the best service quality; this results from designing more customized routes, which typically make a larger number of trips. The MIP two-stage method can also reduce total vehicle service time, compared to the conventional service.

The results shown in Tables 8–13 also suggest that an LMTS may deliver significant cost savings for both service users (passengers) and service providers (e.g., municipalities, private companies). First, shorter waiting time and riding time translate into monetary savings for users. According to Gómez-Ibañez, Tye, and Winston (1999), for work trips in San Francisco, the monetary value of a unit of transfer waiting time is 195% of the user's after-tax wages, and the monetary value of a unit of in-vehicle riding time is 76% of the user's after-tax wages. The equivalent economic savings are large when we consider these monetary values of time. Second, the reduced vehicle service time achieved by LMTS also means monetary savings for its operators. Given a fixed vehicle fleet size, the transportation system's operating costs are proportional to the vehicle service time—e.g., the fuel cost for vehicles directly depends on vehicle travel time/distance, and the labor cost for drivers is positively correlated with vehicle service time. Third, the shorter vehicle service time also means less traffic congestion and reduced carbon emissions.

8. Conclusion

This paper develops routing and scheduling approaches for an innovative urban transportation system concept—the last-mile transportation system (LMTS)—and assesses the performance of these approaches and of the LMTS concept, in general. Specifically, from an operational and optimization perspective, the paper formulates an exact MIP model and develops several computationally feasible heuristics for optimizing routing and scheduling decisions for a generic LMTS. The system's performance is also evaluated in comparison to a conventional service system.

The LMTS routing and scheduling studied here has several features: (1) passengers requesting last-mile service arrive in batches at the metro station, instead of individually, and passengers in the same batch have the same service time window; (2) passengers requesting last-mile service share a common last-mile origin (which is also the vehicle depot); (3) the objective is to improve LOS. These features provide intuitive groupings of operational decisions and provide incentives for identifying heuristic approaches.

Given the service region's geometry (prespecified LM stops, feasible routes), the number and capacity of vehicles in the service fleet, and a set of known last-mile service requests (passenger arrival times and destinations), the operational strategies we have developed provide detailed routing and scheduling plans for the system's vehicle fleet, with the objective of minimizing the sum of passenger waiting time before boarding a vehicle and riding time. Computational experiments suggest that, compared to a conventional service system with fixed routes and schedule, an LMTS that operates in a variety of contexts—including the myopic operating strategy, which uses demand information from trains that have already arrived, and the tabu search metaheuristic and MIP two-stage method, which use demand information over the entire service horizon—performs better under a broad range of conditions. The myopic operating strategy can be implemented easily and quickly, and can provide a default solution, particularly when no advance demand information is available; the tabu search metaheuristic considers all demand information and offers good-quality solutions in a short computational time; and the MIP two-stage method provides the best solution over the entire service horizon, but with greater computational requirements. We believe that the strategies we have proposed will benefit LMTS operators by providing operating plans for these complex systems that are both cost-effective and offer a higher LOS for passengers.

A natural extension for future research would be to consider stochastic versions of the LMTS operations problem; these would involve some combination of unreliable train schedules, probabilistic last-mile service requests, and uncertainty about vehicle service times due to traffic congestion. Another extension would be to consider combining the last-mile system described here with a first-mile system, in which vehicles would also pick up passengers at LM stops in the service region and transport them to the metro station; such an approach could potentially offer even greater social benefits.

Acknowledgments

The author would like to thank Professor Amedeo Odoni and Professor Cynthia Barnhart for valuable comments and suggestions concerning the paper.

References

Anderson JE (1998) Control of personal rapid transit systems. *J. Adv. Transportation* 32(1):57–74.

Archetti C, Speranza MG, Hertz A (2006) A tabu search algorithm for the split delivery vehicle routing problem. *Transportation Sci.* 40(1):64–73.

Balcik B, Beamon BM, Smilowitz K (2008) Last mile distribution in humanitarian relief. *J. Intelligent Transportation Systems* 12(2): 51–63.

Berbeglia G, Cordeau JF, Laporte G (2012) A hybrid tabu search and constraint programming algorithm for the dynamic dial-a-ride problem. *INFORMS J. Comput.* 24(3):343–355.

Berger T, Sallez Y, Raileanu S, Tahon C, Trentesaux D, Borangiu T (2011) Personal rapid transit in an open-control framework. *Comput. Indust. Engrg.* 61(2):300–312.

Bly PH, Teychenne R (2005) Three financial and socio-economic assessments of a personal rapid transit system. *Proc. 10th Internat. Conf. Automated People Movers*, 1–16.

Boyer KK, Prud'homme AM, Chung W (2009) The last mile challenge: Evaluating the effects of customer density and delivery window patterns. *J. Bus. Logist.* 30(1):185–201.

Brake J, Nelson JD, Wright S (2004) Demand responsive transport: Towards the emergence of a new market segment. *J. Transport Geography* 12(4):323–337.

Bräysy O, Gendreau M (2005a) Vehicle routing problem with time windows, Part I: Route construction and local search algorithms. *Transportation Sci.* 39(1):104–118.

Bräysy O, Gendreau M (2005b) Vehicle routing problem with time windows, Part II: Metaheuristics. *Transportation Sci.* 39(1): 119–139.

Ceder A, Wilson NH (1986) Bus network design. *Transportation Res. Part B: Methodological* 20(4):331–344.

Chevrier R, Liefooghe A, Jourdan L, Dhaenens C (2012) Solving a dial-a-ride problem with a hybrid evolutionary multi-objective approach: Application to demand responsive transport. *Appl. Soft Comput.* 12(4):1247–1258.

Cordeau JF, Laporte G (2007) The dial-a-ride problem: Models and algorithms. *Ann. Oper. Res.* 153(1):29–46.

Cordeau JF, Maischberger M (2012) A parallel iterated tabu search heuristic for vehicle routing problems. *Comput. Oper. Res.* 39(9):2033–2050.

Cordeau JF, Gendreau M, Laporte G (1997) A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks* 30(2):105–119.

Cordeau JF, Laporte G, Mercier A (2001) A unified tabu search heuristic for vehicle routing problems with time windows. *J. Oper. Res. Soc.* 52(8):928–936.

Daganzo CF (1978) An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Res.* 12(5):325–333.

de Paepe WE, Lenstra JK, Sgall J, Sitters RA, Stougie L (2004) Computer-aided complexity classification of dial-a-ride problems. *INFORMS J. Comput.* 16(2):120–132.

Diana M, Dessouky MM, Xia N (2006) A model for the fleet sizing of demand responsive transportation services with time windows. *Transportation Res. Part B: Methodological* 40(8):651–666.

Dowsland KA (1998) Nurse scheduling with tabu search and strategic oscillation. *Eur. J. Oper. Res.* 106(2):393–407.

Engles D, Iacometti A (2004) System architecture. Ambrosino G, Nelson JD, Romanazzo M, eds. *Demand Responsive Transport Services: Towards the Flexible Mobility Agency* (ENEA, Rome), 75–88.

Enoch M, Potter S, Parkhurst G, Smith M (2004) INTERMODE: Innovations in demand responsive transport. Report, Department for Transport and Greater Manchester Passenger Transport Executive, Leicester, UK. https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/3372/1/Intermode%20final%20June%202004.pdf.

Esper TL, Jensen TD, Turnipseed FL, Burton S (2003) The last mile: An examination of effects of online retail delivery strategies on consumers. *J. Bus. Logist.* 24(2):177–203.

Gendreau M, Hertz A, Laporte G (1994) A tabu search heuristic for the vehicle routing problem. *Management Sci.* 40(10):1276–1290.

Gendreau M, Guertin F, Potvin JY, Taillard E (1999) Parallel tabu search for real-time vehicle routing and dispatching. *Transportation Sci.* 33(4):381–390.

Glover F (1986) Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.* 13(5):533–549.

Glover F (1989) Tabu search—Part I. *ORSA J. Comput.* 1(3):190–206.

Glover F (1990a) Tabu search: A tutorial. *Interfaces* 20(4):74–94.

Glover F (1990b) Tabu search—Part II. *ORSA J. Comput.* 2(1):4–32.

Gómez-Ibañez J, Tye W, Winston C (1999) *Essays in Transportation Economics and Policy*, Vol. 42 (Brookings Institution Press, Washington, DC).

Horn ME (2002a) Multi-modal and demand-responsive passenger transport systems: A modelling framework with embedded control systems. *Transportation Res. Part A: Policy Practice* 36(2): 167–188.

Horn ME (2002b) Fleet scheduling and dispatching for demand-responsive passenger services. *Transportation Res. Part C: Emerging Tech.* 10(1):35–63.

Hurink J, Jurisch B, Thole M (1994) Tabu search for the job-shop scheduling problem with multi-purpose machines. *OR Spektrum* 15(4):205–215.

Jaw JJ, Odoni AR, Psaraftis HN, Wilson NH (1986) A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transportation Res. Part B: Methodological* 20(3):243–257.

Lee HL, Whang S (2001) Winning the last mile of e-commerce. *MIT Sloan Management Rev.* 42(4):54–62.

Lee JW, Mazumdar RR, Shroff NB (2006) Opportunistic power scheduling for dynamic multi-server wireless systems. *IEEE Trans. Wireless Comm.* 5(6):1506–1515.

Lees-Miller JD, Hammersley JC, Davenport N (2009) Ride sharing in personal rapid transit capacity planning. *12th Internat. Conf. Automated People Movers*, 321–332.

Lees-Miller JD, Hammersley JC, Wilson RE (2010) Theoretical maximum capacity as benchmark for empty vehicle redistribution in personal rapid transit. *Transportation Res. Record: J. Transportation Res. Board* 2146(1):76–83.

Lei H, Laporte G, Guo B (2012) Districting for routing with stochastic customers. *Eur. J. Transportation Logist.* 1(1–2):67–85.

Liu L, Liu X (1998) Dynamic and static job allocation for multi-server systems. *IIE Trans.* 30(9):845–854.

Liu Z, Jia X, Cheng W (2012) Solving the last mile problem: Ensure the success of public bicycle system in Beijing. *Procedia Soc. Behav. Sci.* 43:73–78.

Mageean J, Nelson JD (2003) The evaluation of demand responsive transport services in Europe. *J. Transport Geography* 11(4):255–270.

Montané FAT, Galvao RD (2006) A tabu search algorithm for the vehicle routing problem with simultaneous pick-up and delivery service. *Comput. Oper. Res.* 33(3):595–619.

Mueller K, Sgouridis SP (2011) Simulation-based analysis of personal rapid transit systems: Service and energy performance assessment of the Masdar City PRT case. *J. Adv. Transportation* 45(4):252–270.

Palmer K, Dessouky M, Abdelmaguid T (2004) Impacts of management practices and advanced technologies on demand responsive transit systems. *Transportation Res. Part A: Policy Practice* 38(7):495–509.

Punakivi M, Yrjölä H, Holmström J (2001) Solving the last mile issue: Reception box or delivery box? *Internat. J. Physical Distribution Logist. Management* 31(6):427–439.

Quadrifoglio L, Dessouky MM, Ordóñez F (2008) A simulation study of demand responsive transit system design. *Transportation Res. Part A: Policy Practice* 42(4):718–737.

Song L, Cherrett T, McLeod F, Guan W (2009) Addressing the last mile problem. *Transportation Res. Record: J. Transportation Res. Board* 2097(1):9–18.

Sridharan R (1995) The capacitated plant location problem. *Eur. J. Oper. Res.* 87(2):203–213.

Taillard E (1991) Robust taboo search for the quadratic assignment problem. *Parallel Comput.* 17(4):443–455.

Wang H, Odoni A (2016) Approximating the performance of a “last mile” transportation system. *Transportation Sci.* 50(2):659–675.

Wilson NH, Hendrickson C (1980) Performance models of flexibly routed transportation services. *Transportation Res. Part B: Methodological* 14(1):67–78.

Zee DJVD, Harten AV, Schuur P (2001) On-line scheduling of multi-server batch operations. *IIE Trans.* 33(7):569–586.