European Journal of Operational Research 299 (2022) 263–282

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Vehicle dispatching plan for minimizing passenger waiting time in a corridor with buses of different sizes: Model formulation and solution approaches

Mohammad Sadrani[a,*], Alejandro Tirachini[b,c], Constantinos Antoniou[a]

[a] Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich 80333, Germany
[b] Transport Engineering Division, Civil Engineering Department, Universidad de Chile, Chile
[c] Instituto Sistemas Complejos de Ingeniería, Chile

ARTICLE INFO

Article history:
Received 6 March 2021
Accepted 29 July 2021
Available online 8 August 2021

Keywords:
Transportation
Heterogeneous fleet
Stochastic travel times
Mixed-integer nonlinear programming
Simulated annealing

ABSTRACT

Urban public transportation agencies sometimes have to operate mixing vehicles of different sizes on their routes, due to resource limitations or historical reasons. Services with different passenger-carrying capacities are provided to passengers during a mixed-fleet operation. A fundamental question arising here is how to optimally deploy a given fleet of different bus sizes to provide services that minimize passenger waiting time. We formulate a mixed-fleet vehicle dispatching problem as a Mixed-Integer Nonlinear Programming (MINLP) model to optimize dispatching schemes (dispatching orders and times) when a given set of buses of different sizes are available to serve demand along a route. The objective is to minimize the average passenger waiting time under time-dependent demand volumes. Stochastic travel times between stops and vehicle capacity constraints (i.e., introducing extra waiting time due to denied boarding) are explicitly modeled. A Simulated Annealing (SA) algorithm coupled with a Monte Carlo simulation framework is developed to solve large real-world instances in the presence of stochastic travel times. Results show that, in addition to dispatching headway, bus dispatching sequence can strongly affect waiting times under a mixed-fleet operation. Indeed, with an optimal dispatching sequence, a more accurate adjustment of supply to demand is possible in accordance with time-dependent demand conditions, and the total savings in waiting time are mainly driven by a further reduction in the number of passengers left behind. The optimality of uneven dispatching headways stems from two elements: having a mixed fleet and having localized peaks on demand that make buses run full.

1. Introduction

In some cities, urban public transportation agencies have to operate combining vehicles of different sizes on their routes, due to specific demand patterns, historical reasons or resource limitations, e.g., when different sizes of buses are purchased at different times through different contracts. Such a situation takes place for example in peak periods if the total fleet size required is larger than the number of articulated buses at disposal; therefore, the agency uses standard and articulated buses at the same time. This type of arrangement is seen in several large cities around the world. For example, in Santiago, Chile, a city in which 317 bus routes operate on a working day, 116 routes (37%) are operated with mixed fleets during the morning peak period, either combining small (8-meter long) with standard (12-meter long) buses, or standard with articulated (18-meter long) buses in one single route[1]. The key question arising here is how to optimally utilize a given mixed fleet (heterogeneous fleet) with buses of different sizes to provide services that minimize passenger waiting time, considering the fact that bus operators can provide services with different passenger-carrying capacities under a mixed-fleet operation. Hence, there appears a need to formulate and solve a mixed-fleet dispatching problem for a more efficient utilization of available resources in terms of vehicle capacity, particularly for high-demand cases in which vehicles operate with larger occupancy levels and thus crowding-related problems due to in-vehicle capacity constraints are more noticeable for public transport commuters. For example, it would be

* Corresponding author.
E-mail addresses: m.sadrani@tum.de (M. Sadrani), Alejandro.tirachini@ing.uchile.cl (A. Tirachini), c.antoniou@tum.de (C. Antoniou).

[1] Information based on Santiagoś bus route schedules available at http://www.dtpm.cl/index.php/programas-de-operacion, accessed 15 Feb 2021.

https://doi.org/10.1016/j.ejor.2021.07.054
0377-2217/

important to avoid a further exacerbation of denied boardings in this situation.

Crowding in public transport systems has become a perennial problem during peak hours with the rapid growth in demand. Nowadays, crowding in public transport services is known as one of the major phenomena affecting passengers' travel experience, mode choice, and system performance (Drabicki, Kucharski, Cats & Szarata, 2021; Hörcher & Tirachini, 2021). For example, public transport crowding phenomena can significantly increase the number of passengers failing to board due to a lack of capacity [i.e., extra waiting times for passengers being left behind by one or more vehicles due to overcrowding, who have to wait for a longer period of time to board the next coming vehicle(s)] (Yap, Cats & van Arem, 2020). Moreover, regarding the effects of in-vehicle crowding on the design and appraisal of public transport projects, Cats, West and Eliasson (2016) have revealed that neglecting crowding matters and in-vehicle capacity constraints can lead to an unrealistic estimation (overestimation) of the actual benefits of a public transport system, especially in transport systems with a high volume of travelers during peak periods. Hence, to offer a precise representation of reality, many researchers in the field of timetable design and vehicle scheduling have explicitly considered vehicle capacity constraints and the passengers left behind due to an overcrowded situation in their service supply optimization models (e.g., Cats et al., 2016; Dai, Liu, Chen & Ma, 2020; Gao, Kroon, Schmidt & Yang, 2016; Sánchez-Martínez, Koutsopoulos & Wilson, 2016; Wang, Tang, Ning, van den Boom & De Schutter, 2015).

Regarding the value of waiting time savings, several studies have revealed that the value of waiting time is higher than that of in-vehicle time, as passengers feel more dissatisfied with their waiting times at stops (e.g., Cats et al., 2016; Wardman, 2004; Xumei, Qiaoxian & Guang, 2011). Hence, waiting time is considered as one of the most striking travel time components for evaluating the level of service from a passenger's point of view (Niu, Zhou & Gao, 2015). Particularly, in high-demand public transport corridors during peak periods, waiting times can be too long for passengers who are unable to board a service due to a lack of capacity (see Fig. 1), thereby reducing the attractiveness and reliability of a public transport system substantially. With the rapid growth of public transportation ridership, left behind passengers due to overcrowding is becoming a main concern for many transit agencies (Sun & Xu, 2012; Zhu, Koutsopoulos & Wilson, 2017). This challenging problem is particularly prevalent on some crowded public transport systems across the globe (e.g., Beijing, Moscow, Sao Paulo, Santiago, Hong Kong), in which it is not unusual to operate vehicles at (or near) crush capacity[2] during the peak hours (Tirachini, Hensher & Rose, 2014). In this paper, we show that the above-mentioned problem is exacerbated in the operation with a given fleet of buses of heterogeneous sizes, if vehicles are not dispatched at the right order and times to properly meet time-dependent passenger demand along a route.

To date, there are a large number of published studies in the field of timetabling and vehicle scheduling that have endeavored to improve the quality of services provided to users by minimizing passenger waiting time at stops (e.g., Abdolmaleki, Masoud & Yin, 2020; Altazin, Dauzère-Pérès, Ramond & Trefond, 2020; Barrena, Canca, Coelho & Laporte, 2014a,b; Hassannayebi & Zegordi, 2017; Luo, Liu, Yu, Tang & Li, 2019; Nachtigall & Voget, 1997; Newell, 1971; Niu & Zhou, 2013; Niu et al., 2015; Sánchez-Martínez et al., 2016; among many others). Passenger waiting time can be strongly influenced by bus dispatching headways (Ceder & Marguier, 1985).

Fig. 1. Passengers left behind due to capacity constraints[3].

Up to now, the problem of setting dispatching headways has attracted considerable scholarly attention. Szeto and Wu (2011) proposed a joint optimization model for the route design and frequency setting problems. The main objective of the model was to minimize the number of transfers and the total travel time. An integrated solution method made up of a genetic algorithm and a neighborhood search heuristic was developed to solve the problem. Hadas and Shnaiderman (2012) proposed a new method for determining the dispatching headways and vehicle sizes based on the stochastic characteristics of Automatic Passenger Counting (APC) and Automatic Vehicle Location (AVL) data within a supply chain optimization model. The objective of the model was to minimize the total cost due to empty seats and un-served demand. Li, Xu and He (2013) developed a stochastic optimization model involving random passenger demand, boarding/alighting times and bus travel times, to find the optimal frequency with the aim of minimizing the waiting time for users and maximizing the expected bus company profits. The model was compared to the traditional frequency setting models of Newell (1971) and Ceder (1984). Martínez, Mauttone and Urquhart (2014) proposed a mixed integer linear programming (MILP) formulation for a transit frequency optimization problem. Since the model was intractable for large instances on a real transit network in Uruguay, a Tabu Search approach was adopted to solve the problem.

Berrebi, Watkins and Laval (2015) proposed a real-time bus dispatching policy to minimize passenger waiting time on a high-frequency bus route. For a general railway network, Meng, Luan and Zhou (2016) developed a cumulative flow variables-based integer programming model for dispatching trains under a stochastic environment, including stochastic capacity breakdown durations, segment running times and station dwell times. Gkiotsalitis and Cats (2018) developed a mathematical model for setting the dispatching headways of bus lines in a city network. The model explicitly included bus capacity and fleet size constraints. Moreover, demand, headway and travel time variations at different time periods were taken into account. The results showed that the dispatching headways are particularly susceptible to changes in some factors, such as demand volumes and bus running costs. Moreover, Gkiotsalitis and Alesiani (2019) developed a bus movement optimization model which can deal with travel time and passenger demand uncertainty to generate robust dispatching times for all bus trips in a timetable.

Furthermore, in a more recent strand of literature dealing with bus dispatching problems, Zhang and Liu (2019) formulated a time-dependent bus dispatching problem in a multi-modal context. In lieu of explicitly optimizing the size of dispatched bus fleet, the authors developed an adaptive fleet size adjustment approach with a target level of bus loading factor. Gkiotsalitis (2020) extended a

[2] Vehicles are operated with (near)full capacity, e.g., a high density of standees is observed inside vehicles.

[3] The photo has been taken from https://citylimits.org/2018/03/05/while-subways-get-the-spotlight-bus-riders-frustration-grows-as-numbers-dwindle/.

mathematical model for a periodic bus dispatching control of high-frequency services. An iterative gradient approximation solution method was also designed to reduce the computing burden of the proposed periodic dispatching control. Moreover, Gkiotsalitis and Van Berkum (2020a) introduced a novel rolling-horizon optimization model for adjusting the dispatching times of buses in rolling horizons. The proposed strategy outperforms myopic methods that determine the dispatching time of each bus trip in isolation. In the above-mentioned studies, vehicle dispatching schemes are designed for the operation with uniform fleets of buses (uniform-fleet operation), i.e., public transport providers do not need to deal with a heterogeneous fleet dispatching problem, in which vehicles of different sizes are concurrently used to meet the passenger demand.

In the context of operation with a heterogeneous fleet of buses, dell'Olio, Ibeas and Ruisánchez (2012) constructed an optimization model with constraints on bus capacity to optimize bus size and headway. A series of numerical experiments were conducted using different fleet configurations, including homogenous fleets made up of buses of the same size, and heterogeneous fleets composed of different bus sizes. The findings demonstrated that a better service can be provided by the use of heterogeneous fleets. In another study which set out to create a bus scheduling timetable based on multiple bus sizes, Ceder, Hassold and Dano (2013) formulated a bi-objective mathematical model, in which the first objective was to minimize the deviation of the headways from a desired even headway, and the second one was to minimize the deviation of the observed passenger loads from a desired even-load level of the vehicles at the maximum-load point. Duran-Micco, Vermeir and Vansteenwegen (2020) formulated a transit network design and frequency setting problem while considering a heterogeneous fleet (buses of different sizes and technologies). The authors take two objectives into account, including the minimization of the total travel time and CO2 emissions. Results show that the heterogeneous fleet can reduce travel times and emissions simultaneously, compared to scenarios without a heterogeneous fleet.

In the literature of public transport services, there has been a growing number of publications focusing on the topic of designing timetables with dynamic passenger demand (e.g., Canca, Barrena, Algaba & Zarzo, 2014; Meng & Zhou, 2019; Robenek, Azadeh, Maknoon, de Lapparent & Bierlaire, 2018; Zhang, Li & Qiao, 2018). For instance, to achieve an efficient train timetable that can fully utilize the limited infrastructure and rolling stock resources, Meng and Zhou (2019) developed an integrated train service plan optimization model with variable passenger demand. The authors introduced a team-based scheduling approach to coordinate demand assignment, routing, and timetabling tasks. The proposed method could efficiently increase operators' profits and passenger travel demand satisfaction. Nowadays, with the rapid development of monitoring technologies, there are several tools to obtain detailed demand information through on-vehicle equipment, such as APC and AVL devices and smartcard payment systems. For example, Munizaga and Palma (2012) proposed a new method to estimate a public transport Origin–Destination matrix at a high level of accuracy from smartcard data in Santiago, Chile. In another work undertaken by Aguiléra, Allio, Benezech, Combes and Milion (2014) in Paris, the authors developed a new method to measure passenger flows in an underground transit system using Cell-phone data. They showed that the measures are consistent with those inferred from automated fare collection data. The knowledge of precise demand information is assumed by Niu et al. (2015) to develop a nonlinear integer programming model that finds the optimal skip-stop pattern on a rail transit corridor. The objective of the model was to minimize the total passenger waiting time under both high and medium-resolution time-varying demand data. Moreover, Luo et al. (2019) proposed an optimization model for dynamic bus dispatching problem, taking several types of real-time information such as dynamic passenger flows and road traffic conditions into consideration. The principal objective of the model was to minimize the total passenger waiting time. The authors developed a memory-based genetic algorithm to solve the model.

It is important to point out that public transport planning decisions are typically categorized into three levels, namely strategic, tactical, and operational decisions. For example, at the strategic planning level (related to long-term decisions), the set of routes and stops are determined. At the tactical planning level (related to medium-term decisions), fleet size requirement is determined, including the number and type of vehicles purchased to operate. Finally, at the operational planning level, short-term decisions (e.g., vehicle scheduling, driver scheduling, and driver rostering) are made. We refer the interested readers to Desaulniers and Hickman (2007) and Farahani, Miandoabchi, Szeto and Rashidi (2013) for a comprehensive review of these aspects. For example, Desaulniers and Hickman (2007) stated that the common way is to consider strategic and tactical planning decisions as input, and then determine a better way of using the agencies' resources in order to improve the level of service provided to users in operational planning decisions. In this study, we focus on the operational planning level, and we assume that the fleet size (the number and type of vehicles) is already determined at the tactical planning stage, which is a well-established and reasonable assumption according to the usual classification of public transport planning decisions (Desaulniers & Hickman, 2007; Farahani et al., 2013). Indeed, under a mixed-fleet operation, we attempt to optimize vehicle dispatching schemes for a more efficient utilization of available resources (improving capacity utilization).

Moreover, public transport operators may be able to perform interlining between different bus routes (i.e., the allocation of one vehicle to a different line at the end of one round), and interlining decisions could be adopted in real-time cases, for instance, to deal with sudden changes on demand conditions. In such a case, even though the total bus fleet is fixed, the fleet per route does not need to be fixed. Our model does not deal with this situation, we rather tackle the optimization problem of single routes, or the case of multiple routes when the interlining decisions have been made at an upper decision level, therefore without introducing changes to fleet allocation between routes on a day-to-day basis. In other words, we are modeling a situation in which it is fair to assume that the fleet size per line is fixed. Resorting to the usual classification of public transport planning decisions, we are assuming that the decisions of fleet allocation per route per period were made on a tactical level (including possible interlining), and the decision of which bus to dispatch at a particular time and order is made on the operational level.

The main contributions of this work to the state-of-the-art are summarized here. While a considerable amount of literature has been published on the problem of setting bus dispatching headways, there have been no attempts to examine how bus dispatching policies can affect passenger waiting times in a mixed-fleet operation (i.e., in the operation with buses of different sizes), considering the fact that a mixed fleet of vehicles can provide services with different passenger-carrying capacities during real-world operations. To bridge this gap, we develop a Mixed-Integer Nonlinear Programming (MINLP) model to optimize vehicle dispatching schemes (in terms of both vehicle dispatching order and vehicle dispatching time) when a given set of buses with different sizes (capacities) are operated to serve demand along a route. The objective of the proposed model is to minimize the average passenger waiting time when the passengers' demand changes along the time.

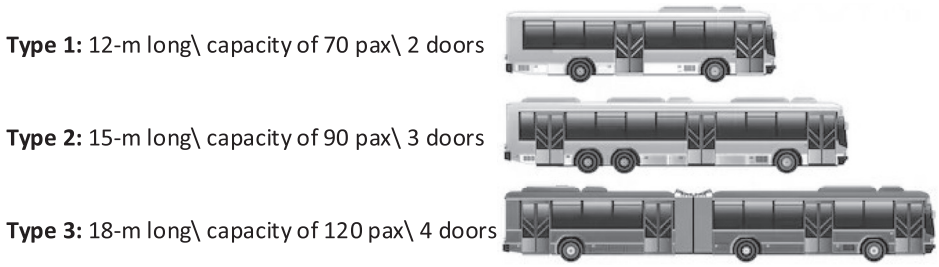

Fig. 2. Three different available bus types used for mixed-fleet operations in this work.

Binary decision variables on dispatching order (size assignment to each bus service)

x_{mi} $m = 1, 2, 3;\quad i = 1, 2, \ldots, N_v$

$m \setminus i$	$i = 1$	$i = 2$	...	$i = N_v - 1$	$i = N_v$
$m = 1$	x_{11}	x_{12}	...	$x_{1(N_v-1)}$	x_{1N_v}
$m = 2$	x_{21}	x_{22}	...	$x_{2(N_v-1)}$	x_{2N_v}
$m = 3$	x_{31}	x_{32}	...	$x_{3(N_v-1)}$	x_{3N_v}
	$T^d_{1,1}$	$T^d_{2,1}$	...	$T^d_{(N_v-1),1}$	$T^d_{N_v,1}$

Continuous decision variables on dispatching time

$T^d_{i,j}$ $i = 1, 2, \ldots, N_v;\quad j = 1$

Fig. 3. Structure of one initial solution.

To reflect realistic operating conditions in the proposed mixed-fleet dispatching problem, vehicle capacity constraints are explicitly modeled to account for the actual impacts of full capacity operations during peak hours, i.e., introducing extra waiting times for passengers failing to board due to a lack of capacity, who have to wait for the next vehicle(s). Particularly, we examine how a more accurate adjustment of the dispatching order for a given set of vehicles with different sizes can prevent denied boarding problems from further exacerbation. Moreover, to provide real-world characteristics of urban bus systems, we model stochastic travel times between stops, which pose additional challenges when solving the problem. Given the existence of a discrete set of dispatching sequences that can be prescribed for buses in our mixed-fleet operation, the proposed dispatching problem is a permutation-based combinatorial optimization problem with huge discrete search spaces in real-world cases. It would be practically challenging to solve the problem, due to the NP-hardness of this problem leading to heavy computational complexity for large real-world instances. A Simulated Annealing (SA) algorithm with state-of-the-art operators, which can efficiently produce feasible neighboring solutions, is designed to solve real-world instances within a reasonable computing time. A Monte Carlo simulation method is also embedded as a subroutine into the SA to handle travel time uncertainty in the presence of stochastic travel times. Moreover, we propose a novel full integer space enumeration method, by which the master MINLP problem is decomposed into a certain number of continuous NLP subproblems, to provide a direction towards the global optimal solutions for small and medium-sized instances. Accordingly, to obtain explicit ideas about the quality of the solutions found by the SA algorithm, the performance of the SA is assessed by comparing its results to the optimal solutions obtained through the GAMS software in several small and medium-sized test problems. Finally, the optimization model and the solution algorithm are tested through performing a broad range of numerical experiments based on the actual data from a real bus corridor in Sydney, Australia. As another contribution, we also analyze the differences in optimal bus dispatching patterns due to having different degrees of demand resolution available; for this purpose, a comparison between knowing passenger arrival rates every 15 minutes (high-resolution demand) vs. every 60 minutes (low-resolution demand) is performed. Indeed, in the design of dispatching schemes, we attempt to accentuate the importance of considering specific times, at which demand surges to the highest level during the peak hours (peak inside the peak), to avoid a further exacerbation of denied boarding problem at maximum-load points during real-world operations.

This study demonstrates that, in addition to dispatching headways, bus dispatching order can strongly affect passenger waiting time under a mixed-fleet operation. This is because depending on the dispatching order of each vehicle, bus operators can provide services with different capacities to travelers during a mixed-fleet operation. Hence, with an optimal adjustment of the dispatching order, a more accurate adjustment of supply (vehicle capacity) to demand is possible in line with time-dependent demand volumes, thus leading to a better utilization of the existing resources and reducing the number of passengers left behind due to binding capacity constraints (i.e., avoiding a further deterioration of denied boarding problems). Obviously, in the operation with buses of one size, the setting of bus dispatching order has no meaning to bus operators anymore. Moreover, we show that the optimality

of uneven dispatching headways stems from two elements: having a mixed fleet and having localized peaks on demand that make buses run full.

The remainder of this paper is organized as follows. In Section 2, an optimization model is formulated for a mixed-fleet vehicle dispatching problem. Solution algorithms are provided to solve the problem in Section 3. In Section 4, a series of numerical experiments are carried out to assess the efficiency and effectiveness of the mathematical model and the solution approaches. Finally, Section 5 provides conclusions and potential directions for future research.

2. Mathematical formulation

In this section, we formulate a heterogeneous fleet vehicle dispatching problem as a Mixed-Integer Nonlinear Programming (MINLP) model to optimize vehicle dispatching patterns (in terms of dispatching order and dispatching times). The objective of the model is to minimize the average passenger waiting time under time-dependent demand volumes, when a heterogeneous fleet of buses (with different passenger-carrying capacities) is available to meet demand along a bus corridor. Hence, we formulate a detailed objective function that is accurately able to compute passenger waiting times even in the case of failing to board due to a lack of capacity (extra waiting time due to denied boarding). Moreover, several different constraints that need to be considered in real-world urban bus operations are formulated, such as passenger flow constraints, vehicle movement constraints, resource availability constraints, and vehicle capacity constraints in the presence of different sizes of buses which can provide services with different capacities during a mixed-fleet operation. The notation used in the model formulation is listed in Table 1. Before presenting the model, the main assumptions and aspects considered in the mathematical formulation of a mixed-fleet operation are summarized as follows:

- We focus on high-frequency bus services, and thus headways are so short that passengers do not need to plan for arriving at stops (i.e., we assume that passengers arrive randomly at stops).
- There exists a given mixed fleet composed of three different bus sizes: 12-meter (standard) bus with a capacity of 70 passengers and 2 doors, 15-meter (rigid) bus with a capacity of 90 passengers and 3 doors, and 18-meter (articulated) bus with a capacity of 120 passengers and 4 doors (see Fig. 2).
- We focus on the operational planning level, and we assume that the fleet size (the number and type of vehicles) is already determined at the tactical planning level.
- The number of buses of each size is given, i.e., for a certain bus size, the number of associated available vehicles is already given. As an illustrative example, suppose a given mixed fleet: {12, 12, 12, 12, 12, 15, 15, 15, 18, 18}, in which the resource limitations on buses of each size are known, which are at 5, 3, and 2 for 12-, 15-, and 18-m long buses respectively.
- We consider bus operations during a predefined planning horizon on a general bi-directional bus corridor.
- We assume that the first and last buses are dispatched at certain times, which are the beginning and end of the planning horizon for the sake of simplicity.
- We model varying dwell times, which can vary depending on the bus type (i.e., the number of bus doors, see Fig. 2) and on the number of passengers alighting and boarding at each stop.
- We assume that vehicles are not permitted to overtake each other.
- We assume that travel times of vehicles between stops are stochastic, drawn from a log-normal distribution.
- We explicitly model vehicle capacity constraints, and therefore:
  - (a) the number of passengers who can successfully board a vehicle at a stop cannot exceed the remaining capacity inside the vehicle at that stop;
  - (b) in the case of failing to board due to capacity constraints for any number of times, the model is able to continue computing the actual waiting time of passengers (even if some passengers are being left behind by two or more successive services due to an oversaturated condition) until they successfully board a bus service with enough room, i.e., introducing extra waiting times for passengers left behind due to a lack of capacity, who have to wait for the next coming vehicle(s);
  - (c) if there is not enough capacity inside a bus to carry all the passengers waiting for it at a stop, we assume that all the waiting passengers, irrespective of their destinations, have the same chance to board[4].

The mixed-fleet vehicle dispatching problem is formulated as follows:

$$\underset{x_{mi},T^d_{i,1}}{Min}\ \mathrm{AWT} = \frac{1}{P}\cdot \sum_{i\in V, i\geq 2}\sum_{j\in S}\sum_{k\in S, k>j}\left(\underbrace{\lambda_j\left[T^d_{i-1,j}\right].OD_{j,k}\left[T^d_{i-1,j}\right].H_{i,j}.\frac{H_{i,j}}{2}}_{\text{(i)}} + \underbrace{N^f_{i-1,j,k}.H_{i,j}}_{\text{(ii)}}\right) \tag{1}$$

s.t:

$$x_{1i}+x_{2i}+x_{3i}=1 \quad \forall i\in V \tag{2}$$

$$C_i = c_1x_{1i}+c_2x_{2i}+c_3x_{3i} \quad \forall i\in V \tag{3}$$

$$\sum_{i\in V} x_{1i} = A \tag{4}$$

$$\sum_{i\in V} x_{2i} = B \tag{5}$$

$$\sum_{i\in V} x_{3i} = C \tag{6}$$

$$H_{i,j} = T^d_{i,j} - T^d_{i-1,j} \quad \forall i\in V-\{1\}, \forall j\in S \tag{7}$$

$$T^a_{i,j} = T^d_{i,j-1} + \delta_a + T^r_{i,j} + \delta_d \quad \forall i\in V, \forall j\in S-\{1\} \tag{8}$$

$$T^r_{i,j} \sim \text{lognormal}(r_j, \sigma_j) \quad \forall i\in V, \forall j\in S-\{1\} \tag{9}$$

[4] This is indeed a well-known assumption among researchers, under which it is assumed that passengers with different destinations arrive randomly and are well mixed together at each stop. In this situation, if the boarding demand exceeds the remaining capacity inside a vehicle, all the waiting passengers are assumed to have the same chance to board regardless of their destinations (e.g., Wang et al., 2015; Gao et al., 2016; Sánchez-Martínez et al., 2016; Dai et al., 2020).

[5] Three different bus types are available in our mixed-fleet operations (see Fig. 2): type 1 is a 12-m long bus; type 2 is a 15-m long bus, and type 3 is an 18-m long bus, hence:

Binary variable x_{1i} would be 1 if a 12-m vehicle is allocated to i^{th} bus service, otherwise 0;

Binary variable x_{2i} would be 1 if a 15-m vehicle is allocated to i^{th} bus service, otherwise 0;

Binary variable x_{3i} would be 1 if an 18-m vehicle is allocated to i^{th} bus service, otherwise 0.

Table 1
List of notations.[5]

Symbol	Description	Unit
Sets		
V	Set of vehicles, V= {1, 2, …, N_v}	
M	Set of vehicle types, M= {1, 2, 3} for three available bus sizes	
S	Set of stops, $S = \{1, 2, \ldots, N_s\}$	
Indices		
i	Index of buses	
m	Index of bus type	
j, k	Index of stops	
Input parameters		
N_v	Total number of buses in a given fleet of mixed bus sizes ($N_v = A + B + C$)	
A	Available number of 12-m long buses in the given mixed fleet	
B	Available number of 15-m long buses in the given mixed fleet	
C	Available number of 18-m long buses in the given mixed fleet	
N_s	Number of stops along the bus route	
$\lambda_j[t]$	Passenger arrival rate at stop j at time t	pax/min
$OD_{j,k}[t]$	O-D matrix (the percentage of waiting passengers at stop j, who aim to travel from stop j to destination stop k) at time t	%
P	Total demand (the total number of waiting passengers who arrived at stops), which is a constant value during the entire study period	pax
δ_a	Acceleration time	s
δ_d	Deceleration time	s
r_j	Mean running times between stops $j-1$ and j	min
σ_j	Standard deviation of running times between stops $j-1$ and j	min
α_0	Time required for opening and closing bus doors	s
α_1	Average alighting time per passenger	s/pax
α_2	Average boarding time per passenger	s/pax
P_i^a	Proportion of passengers alighting through the busiest door of bus i	%
P_i^b	Proportion of passengers boarding through the busiest door of bus i	%
T_1	First dispatching time (beginning of the planning horizon)	min
T_2	Last dispatching time (end of the planning horizon)	min
$h_{\min}$	Minimum dispatching headway	min
$h_{\max}$	Maximum dispatching headway	min
c_1	Capacity of a 12-m long bus	pax/veh
c_2	Capacity of a 15-m long bus	pax/veh
c_3	Capacity of an 18-m long bus	pax/veh
Auxiliary variables		
C_i	Total capacity of bus i	pax/veh
$T_{i,j}^a$	Arrival time of bus i at stop j	min
$T_{i,j}^d$	Departure time of bus i from stop j, $j \geq 2$	min
$T_{i,j}^s$	Dwell time of bus i at stop j	min
$T_{i,j}^r$	Running time of bus i between stops $j-1$ and j	min
$H_{i,j}$	Headway between buses $i-1$ and i at stop j	min
$N_{i,j,k}^w$	Number of passengers with trip $j \to k$ waiting for bus i at stop j	pax
$N_{i,j}^w$	Total number of passengers waiting for bus i at stop j	pax
$N_{i,j}^{on}$	Number of passengers on bus i between stops $j-1$ and j	pax
$N_{i,j}^b$	Number of passengers boarding bus i at stop j	pax
$N_{i,j}^a$	Number of passengers alighting bus i at stop j	pax
$N_{i,j,k}^s$	Number of passengers with trip $j \to k$, who can board bus i at stop j	pax
$N_{i,j,k}^f$	Number of passengers with trip $j \to k$, who fail to board bus i at stop j	pax
Decision variables		
x_{mi}	Binary variable which is 1 if a type m vehicle is dispatched as i-th service; otherwise 0 ($m = 1, 2, 3$ for the three different bus)	
$T_{i,1}^d$	Dispatching time of bus i from the first stop	min

$$T_{i,j}^d = T_{i,j}^a + T_{i,j}^s \quad \forall i \in V, \forall j \in S - \{1\} \tag{10}$$

$$h_{\min} \leq T_{i,1}^d - T_{i-1,1}^d \leq h_{\max} \quad \forall i \in V - \{1\} \tag{11}$$

$$T_{i,j}^s = \alpha_0 + P_i^a \alpha_1 N_{i,j}^a + P_i^b \alpha_2 N_{i,j}^b \quad \forall i \in V, \forall j \in S \tag{12}$$

$$N_{i,j,k}^w = \lambda_j\left[T_{i-1,j}^d\right].OD_{j,k}\left[T_{i-1,j}^d\right].(\overbrace{T_{i,j}^d - T_{i-1,j}^d}^{H_{i,j}}) + N_{i-1,j,k}^f \quad \forall i \in V - \{1\}, \forall j,k \in S, \quad j < k \tag{13}$$

$$N_{i,j}^w = \sum_{k \in S, k > j} N_{i,j,k}^w \quad \forall i \in V, \forall j \in S \tag{14}$$

$$N_{i,j}^{on} = N_{i,j-1}^{on} - N_{i,j-1}^a + N_{i,j-1}^b \quad \forall i \in V, \forall j \in S - \{1\} \tag{15}$$

$$N_{i,j}^b = \min\left\{N_{i,j}^w, C_i - N_{i,j}^{on} + N_{i,j}^a\right\} \quad \forall i \in V, \forall j \in S \tag{16}$$

$$N_{i,j}^a = \sum_{j' \in S, j' < j} N_{i,j',j}^s \quad \forall i \in V, \forall j \in S - \{1\} \tag{17}$$

$$N_{i,j,k}^s = \frac{N_{i,j}^b}{N_{i,j}^w} N_{i,j,k}^w \quad \forall i \in V, \forall j,k \in S, \quad j < k \tag{18}$$

$$N_{i,j,k}^f = N_{i,j,k}^w - N_{i,j,k}^s \quad \forall i \in V, \forall j,k \in S, \quad j < k \tag{19}$$

$$T_{i,1}^d \geq 0 \quad \forall i \in V \tag{20}$$

$$x_{mi} \in \{0,1\} \quad \forall m \in M, \forall i \in V \tag{21}$$

As can be seen in expression (1), the objective of the problem is to minimize the average waiting time (AWT), obtained through

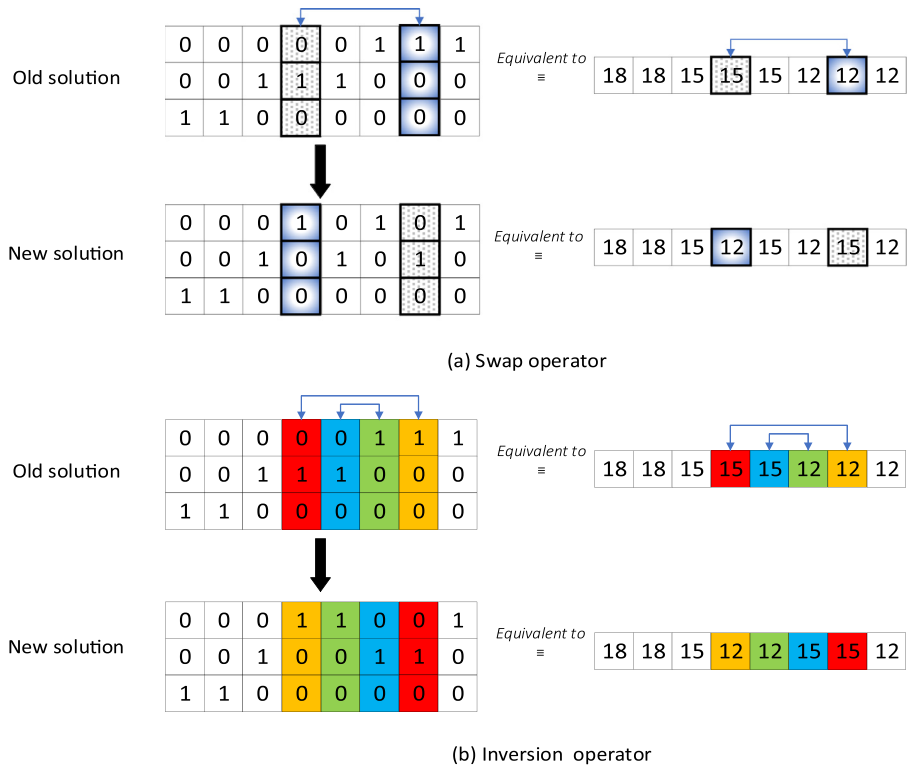

Fig. 4. An illustrative example for creating new dispatching sequences using two operators: (a) swap operator; and (b) inversion operator.

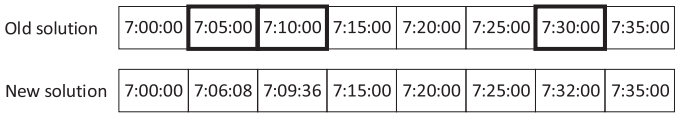

Fig. 5. An illustrative example for changing the departure times of the selected buses.

dividing the total waiting time by the total demand P (P is a fixed value during the entire analysis period). The total waiting time is composed of two parts: (i) waiting time for new passengers arriving at stops over the headway, and (ii) extra waiting time for passengers who were unable to board the preceding vehicle due to a lack of capacity, who have to wait for the next vehicle(s). Indeed, the total waiting time is derived from expression (13), in which the number of waiting passengers in both groups is computed (i.e., (i) the group of new arriving passengers, and (ii) the group of left-behind passengers). In high-frequency bus systems, headways are so short that passengers do not need to plan their arrival at stops, i.e., they arrive randomly at stops over the headway. Hence, for passengers in the group (i), the waiting time is averagely estimated as half of the headway ($H_{i,j}/2$), which is a well-known estimation in the literature of high-frequency bus services due to the random and unplanned arrival of passengers at stops (e.g., Dai et al., 2020; Furth & Wilson, 1981; Gkiotsalitis & Cats, 2018; Wu, Liu & Jin, 2017). On the other hand, for passengers in the group (ii) who were unable to board the previous service due to overcrowding, the extra waiting time is equal to the whole headway ($H_{i,j}$) because they have to wait for the next bus service. As a result, if a considerable number of passengers are left behind due to capacity constraints, passengers' total waiting time can climb dramatically, thereby declining the attractiveness of a public transport system substantially. Note that the formulation of $N^f_{i,j,k}$ in Eq. (19) enables us to compute the actual number of passengers being left behind by each vehicle. Accordingly, even if some passengers are left behind by two or more consecutive services due to an oversaturated condition (denied boarding for several cycles), the model is able to count them among the group of left-behind passengers ($N^f_{i,j,k}$). Hence, we can still correctly calculate their waiting times until they successfully board a service.

As can be seen in expression (1), dispatching headways of vehicles can directly affect the waiting times of passengers in both groups (i) and (ii). Moreover, we attempt to investigate how a proper decision on vehicle dispatching order with the consideration of time-dependent demand volumes can lead to a better utilization of vehicles' capacity, thereby reducing the number of passengers being left behind in group (ii) who need to wait for next arriving vehicle(s) (i.e., preventing denied boarding problems from further exacerbation). Hence, regarding the decision variables considered in the proposed mixed-fleet dispatching problem, we seek to find the optimal dispatching schemes, including the optimal dispatching order of vehicles (x_{mi}) and also the optimal dispatching times of vehicles from the original stop ($T^d_{i,1}$).

In light of the fact that bus operators can provide services with different capacities during a mixed-fleet operation depending on the dispatching order of each vehicle, expression (2) represents the size of vehicle allocated to i-th bus service. As can be seen in expression (3), the passenger-carrying capacity of bus service i depends on its size, e.g., the maximum number of passengers accommodated by a 12-m long bus is $c_1 = 70$ (pax). As explained in the model assumptions, we assume that there exists a given mixed fleet with buses of different sizes and the number of buses of each size is already given. Constraints (4–6) express resource limitations on the number of buses of each size for the three different bus sizes involved in our mixed-fleet operations. These constraints can make the problem more complicated, due to the combinatorial nature of the problem in terms of dispatching sequences. In essence, given the existence of a discrete set of dispatching sequences, which can be practically prescribed for buses in our

mixed-fleet operation, constraints (4-6) turn the proposed problem into a permutation-based combinatorial optimization problem (permutations with repetition due to the presence of several analogous buses in the given fleet). Further details on this issue are provided in Section 3.

Vehicle movement constraints that need to be considered in real-world urban bus operations are given in (7)–(12). Headway between two consecutive buses $i-1$ and i at stop j is calculated by (7). As can be seen in expression (8), the arrival time of bus i at stop j depends on four different time components: (1) the departure time of bus i from stop $j-1$, (2) the time required to accelerate from zero to cruise speed when bus i leaves stop $j-1$, (3) the running time between two adjacent bus stops $j-1$ and j, and (4) the time required to decelerate from cruise speed to zero when bus i wants to enter stop j.

In real-world operations, a broad range of external factors can affect bus running times, such as traffic conditions, traffic signals, bus drivers' behavior, weather, and so on (Wang & Haghani, 2020). Hence, buses might experience different running times between two adjacent stops $j-1$ and j. As can be seen in expression (9), to reflect realistic operating conditions, we assume that running times between stops are stochastic, drawn from a log-normal distribution with mean and standard deviation of r_j and σ_j respectively. It is worth noting that the log-normal distribution is a commonly-used distribution among researchers to model stochastic bus running times (e.g., Cats, Larijani, Koutsopoulos & Burghout, 2011; Delgado, Munoz & Giesen, 2012; Sánchez-Martínez et al., 2016; Zhao, Bukkapatnam & Dessouky, 2003), owing to a long right tail of the travel time distribution. As it is clear from (10), bus i leaves stop j after the completion of alighting and boarding processes (i.e., dwell time) at that stop.

The departure times of buses from the first stop ($T^d_{i,1}$) are considered as decision variables in this study. Indeed, buses can leave the first stop with varying dispatching headways during a predefined planning horizon denoted as $[T_1, T_2]$. Nevertheless, the dispatching headways are confined to the range of $[h_{\min}, h_{\max}]$ defined by policies [see constraint (11)]. Since this study focuses on high-frequency bus services, the upper bound is considered to be 12 minutes. The lower bound is also set to be 2 minutes in order to mitigate the bus bunching[6] phenomenon. Moreover, for the sake of problem simplicity, it is assumed that the first and last buses are dispatched at the beginning and end of the planning horizon respectively (i.e., $T^d_{1,1} = T_1$ and $T^d_{N_v,1} = T_2$).

As can be seen in expression (12), the dwell time of bus i at stop j depends on the number of passengers alighting and boarding at that stop through the busiest bus door, plus the fixed time spent opening and closing bus doors. With regard to the boarding and alighting policy, we assume that passengers use the same doors for alighting and boarding, however, the alighting process has priority over the boarding process (i.e., sequential boarding and alighting, in which boarding process is started after finishing the alighting process). Accordingly, the total dwell time at a stop will depend on the sum of the passengers' boarding and alighting times (Tirachini et al., 2014). Parameters α_1 and α_2 are the average alighting and boarding times per passenger respectively and depend on fare collection technology, bus floor height, platform layout, and so on. Moreover, parameters P^a_i and P^b_i represent respectively the proportions of passengers alighting and boarding through the busiest door of bus i and are dependent on the number of bus doors. For example, in the case of sequential boarding and alighting at all doors, the more doors a bus has the faster boarding and alighting is (Tirachini et al., 2014). In this study, we consider a heterogeneous bus fleet composed of three different bus sizes: 12-m long bus with 2 doors, 15-m long bus with 3 doors, and 18-m long bus with 4 doors (see Fig. 2).

In the following, passenger flow constraints given by (13)–(19) are described. Under time-dependent passenger demand, the actual number of passengers with trip $j \to k$ waiting for bus i at stop j ($N^w_{i,j,k}$) is essentially derived from the sum of two different groups of passengers

$$\left(i.e.,\ N^w_{i,j,k} = \underbrace{\left(\int_{T^d_{i-1,j}}^{T^d_{i,j}} \lambda_j[t].OD_{j,k}[t].dt \right)}_{i} + \underbrace{N^f_{i-1,j,k}}_{(ii)} \right).$$

The first group includes new passengers reaching their origin stops during the headway, whereas the second group includes those passengers who were unable to board the previous bus due to a dearth of capacity.

Since we are working with high-frequency bus systems, the headway (the time interval between the departures of two successive buses from one station, $H_{i,j} = T^d_{i,j} - T^d_{i-1,j}$) is a short enough time interval, during which the destination distribution vector and the passenger arrival rate $\lambda_j[t]$ do not fluctuate notably (Gao et al., 2016). Accordingly, following Gao et al., 2016, we assume that these parameters remain constant during the headway (from the departure time of vehicle $i-1$ to the departure time of vehicle i), and are equal to $OD_{j,k}[T^d_{i-1,j}]$ and $\lambda_j[T^d_{i-1,j}]$. Therefore, the proposed integral form in the case of group (i) can be approximately rewritten, as presented in Eq. (13).

According to the definitions of $N^w_{i,j,k}$ and $N^w_{i,j}$, expression (14) always holds. Indeed, the total number of passengers waiting for bus i at stop j is obtained through summing up across all the waiting passengers at origin stop j with different destinations ($k > j$). As can be seen in expression (15), the number of passengers traveling inside bus i between stops $j-1$ and j, $N^{on}_{i,j}$, is composed of those passengers remaining on bus i from the former segment ($j-2 \to j-1$) as their destination was not stop $j-1$, i.e., $(N^{on}_{i,j-1} - N^a_{i,j-1})$, plus passengers boarding bus i at stop $j-1$, i.e., $N^b_{i,j-1}$. Note that buses are empty when they arrive at the first stop (i.e., $N^{on}_{i,1} = 0,\ \forall i \epsilon V$). Expression (16) indicates that the number of passengers who can successfully get on bus i at stop j cannot be larger than the remaining capacity inside bus i at that stop.

As indicated in (17), the number of passengers who alight bus i at stop j will include the passengers who boarded bus i at the prior stops with the intention of traveling to stop j. Note that there is no demand for alighting at the first stop (i.e., $N^a_{i,1} = 0,\ \forall i \epsilon V$). As discussed in the model assumptions, if there is not enough room on bus i to carry all the passengers waiting for it at stop j, we assume that all the passengers, regardless of their destinations, have the same chance to get on [see expression (18)]. This is indeed a well-known assumption, under which it is assumed that passengers with different destinations arrive randomly and are well mixed together at each stop. Hence, if the boarding demand exceeds the residual capacity of a vehicle, all the waiting passengers are assumed to have the same chance to get on irrespectively of their destinations (e.g., Gao et al., 2016; Sánchez-Martínez et al., 2016; Wang et al., 2015). The number of passengers with trip $j \to k$, who are unable to board bus i at stop j due to a lack of capacity is obtained by (19). Indeed, Eq. (19) can account for the actual number of passengers left behind by each vehicle at any stop. Accordingly, even if denied boarding occurs for several cycles for some passengers due to an overcrowded situation, the model is able to count them among the left-behind passengers and to calculate their extra waiting time until they successfully board an

[6] Bus bunching phenomenon will happen when two or more buses on the same route arrive simultaneously at the same stop.

available service. Constraints (20) and (21) define the domain of decision variables.

3. Solution approaches

The proposed heterogeneous fleet vehicle dispatching problem was formulated as a Mixed-Integer Nonlinear Programming (MINLP) model, in which a number of constraints [e.g., constraints (16) and (18)] and the proposed objective function are nonlinear, thereby combining the difficulty of optimizing over integer variables with the handling of nonlinear functions.

For instance, in nonlinear constraint (18), all the terms (e.g., $N^b_{i,j}$ and $N^w_{i,j,k}$) are variables due to the presence of binary decision variables x_{mi} in the model. In essence, all the terms are dependent on the passenger-carrying capacity of bus services provided to users, and due to the possibility of dispatching services with varying capacities during a mixed-fleet operation [as presented in constraint (3)], those terms can change depending on the starting order of buses in a mixed-fleet operation. To be more precise, as explained for $N^b_{i,j}$ in Eq. (16), the actual number of passengers who can successfully board vehicle i at each stop depends on the capacity of vehicle i. Moreover, for $N^w_{i,j,k}$, as can be seen in Eq. (13), some of the passengers waiting for vehicle i at stop j are those passengers who were unsuccessful in boarding vehicle $i-1$ due to a shortage of capacity inside vehicle $i-1$, who have to wait for vehicle i. This indeed implies that $N^w_{i,j,k}$ can be dependent on the capacity of the preceding vehicle (vehicle $i-1$) during operations. In addition, $N^w_{i,j,k}$ is also affected by the dispatching times of vehicles (continuous decision variables in the model), as the number of new passengers arriving at stops is computed based on the headways between each two consecutive vehicles, as observed from Eq. (13).

As a further example in the objective function (1), according to part (ii) (i.e., the nonlinear expression of ($N^f_{i-1,j,k} \cdot H_{i,j}$), in which all the terms are variables as well), the extra waiting times due to denied boarding can depend on both binary and continuous decision variables of the model (i.e., dispatching order and times). More precisely, the extra waiting time in the case of failing to board, which is equal to the entire headway (i.e., $H_{i,j}$), is directly influenced by the dispatching times of vehicles (continuous decision variables of the model, $T^d_{i,1}$). On the other hand, as discussed earlier, the actual number of passengers being left behind [i.e., $N^f_{i-1,j,k}$ obtained through Eq. (19)] can be strongly dependent on the capacity of bus services provided to travelers during a mixed-fleet operation, which essentially depends on the dispatching sequence of vehicles (binary decision variables of the model, x_{mi}). Such arrangements further point to the difficulty of handling a mixed-fleet dispatching problem as long as the relevant binary variables on the dispatching order of vehicles are regarded as decision variables in our model. This aspect leads to the dependency of many passenger flow variables on the capacity of services, in such a detailed model that explicitly considers binding capacity constraints for any given bus size during a mixed-fleet operation.

Moreover, given the nature of the problem in terms of vehicle dispatching order (x_{mi}), constraints (4–6) (resource limitations on the buses of each size) transform the situation into a permutation-based combinatorial optimization problem (permutations with repetition due to the existence of several identical buses in the given set of vehicles), and the computational complexity of the problem will grow with a factorial rule depending on the fleet size and on the number of buses of each size. Given a set of N_v vehicles, such that there are A identical buses of type 1, B identical buses of type 2, and C identical buses of type 3, there are a total of $\frac{N_v!}{A! \times B! \times C!}$ distinct sequences for dispatching buses in our mixed-fleet operations. For example, in our real-world numerical experiments with a given fleet of 16 buses: {12, 12, 12, 12, 12, 12, 12, 12, 15, 15, 15, 15, 18, 18, 18}, vehicles can be dispatched in $\frac{16!}{9! \times 4! \times 3!} = 400,400$ possible sequences, which are computationally very expensive to be evaluated in real-world cases.

Overall, not only do MINLP problems amalgamate the two aspects of MILP and NLP problems, but also have some unique features. For example, although a strict convexity assumption can ensure the global uniqueness of an NLP solution, the same does not hold for MINLP problems (see Bonami, Kilinç & Linderoth, 2012 for further details on the scope and complex nature of MINLPs). Generally, MINLP problems turn out to be NP-hard by nature (Bonami et al., 2012; Burer & Letchford, 2012). Moreover, as it is obvious, typical combinatorial optimization problems with sequence-dependent setup are known to be strongly NP-hard (Alkaya & Duman, 2015; Bianco, Rinaldi & Sassano, 1987; Lin, Cheng, Pourhejazy & Ying, 2021; Osman & Potts, 1989; Ruiz & Stützle, 2007). It would be practically challenging to find the best dispatching sequence of vehicles from a huge discrete set of possible sequences in real-life cases. Meta-heuristic algorithms are known as one of the most efficient and frequently used search methods for solving such complex optimization problems, as they can find satisfactory suboptimal solutions within an acceptable computing time (Talbi, 2009).

In this research, we develop a Simulated Annealing (SA) algorithm with state-of-the-art features, taking the advantage of producing feasible neighboring solutions, to solve large real-world mixed-fleet vehicle dispatching problems within a reasonable computing time. Moreover, an additional complexity in our real-world problems is to deal with travel time stochasticity (due to the stochastic nature of the problem in terms of travel time uncertainty), which should be addressed when designing the solution algorithm. To tackle this issue, the SA algorithm is coupled with a Monte Carlo Simulation (MCS) method, which evaluates candidate solutions over several replications. Further details on the MCS are provided in the next subsection.

To obtain certain insights about the quality of solutions suggested by the SA algorithm, we offer a full integer space enumeration method (in Section 3.2), whereby the master MINLP problem is decomposed into a particular number of continuous NLP subproblems through fixing integer variables, to provide a direction towards the optimal solutions for small and medium-sized dispatching problems, measuring later the difference to the best solution found by the SA. To efficiently deal with the difficulty of handling binary variables and constraints (4–6) in the proposed mixed-fleet dispatching problem, we extensively describe these aspects in the design of our solution approaches. For example, we discuss how the proposed SA and its operators are properly designed to produce feasible neighborhood solutions that can satisfy these constraints, thereby enhancing the capability of the algorithm for a better exploitation of the best solutions within the feasible search space.

3.1. Simulated Annealing (SA) algorithm

SA is a metaheuristic algorithm known for its ability to avoid getting trapped into a local optimum by allowing for random neighborhood changes, which can be adapted to different optimization problems with discrete or continuous space states (Zhang, Qi, Lin & Miao, 2015). There has been a large amount of work where SA has been efficiently applied to various combinatorial optimization problems (Gomes & Oliveira, 2006; Karimi-Mamaghan, Mohammadi, Meyer, Karimi-Mamaghan & Talbi, 2021). In essence, SA is a single-solution based[7] algorithm in which the cooling

[7] Single-solution based algorithms manipulate and improve a single solution during the search process. On the other hand, in population-based algorithms (e.g., particle swarm, and evolutionary algorithms), a population of solutions is evolved (Talbi, 2009).

Algorithm 1
The main steps of the proposed SA algorithm.

- **Step (1):** Set T_0 and β. Let t $\leftarrow$ 0.
- **Step (2):** Generate a feasible initial solution y & evaluate the answer, i.e.,
  - **(2.1).** Generate a feasible initial solution y, in which the decision variables of the problem (dispatching orders and dispatching times) are randomly generated (the structure of one single solution is illustrated by Fig. 3, in which the first part of the figure is dedicated to the bus dispatching order and the second part indicates the dispatching times of buses from the first stop).
  - **(2.2).** Evaluate the value of the objective function for the initial solution, $f(y)$, through a Monte Carlo Simulation (MCS) method over several simulation-based trials to handle the uncertainty of stochastic travel times, drawn from a log-normal distribution (see Algorithm 2 for further information on the specific steps of the MCS method embedded into the SA). Let $y_{best} \leftarrow y$ and Go to Step 3.
- **Step (3):** Create a neighboring solution y' using various operators & evaluate the answer, i.e.,
  - **(3.1).** Create a neighboring solution y', wherein *bus dispatching order* and *bus dispatching times* are randomly changed into a new arrangement, while taking the advantages of producing feasible solutions, i.e.,
    Dispatching order schedule: To create a new bus dispatching order, as can be seen in Fig. 4, the dispatching order of vehicles in the former solution is changed into a new dispatching sequence through a random displacement by means of swapping or inversion operators. Such a permutation-based procedure in producing neighboring answers can ensure the new solutions will be feasible in terms of constraints (4–6), as the total fleet size and the number of buses of each size in the new solution will remain unchanged compared to the initial feasible solution, and merely the dispatching sequences of vehicles are updated in the new solutions.
    Dispatching time schedule: To create a new departure time schedule, the departure times in the previous solution are changed for some vehicles using a normal distribution. First, based on a given rate, a number of vehicles in one solution are randomly selected (e.g., vehicles 2, 3, and 7 in Fig. 5). Then, for each vehicle selected in turn, its departure time is changed by means of a normal distribution while considering the departure times of preceding and subsequent vehicles as certain bounds, i.e., $nT^d_{i,1} \sim N(T^d_{i,1}, \sigma^2) \sim T^d_{i,1} + \sigma\ N(0,1)$, where the standard deviation σ is defined as $\sigma = \mu \times (T^d_{i+1,1} - T^d_{i-1,1})$. Note that $T^d_{i,1}$ is the departure time of vehicle i in the previous solution and $nT^d_{i,1}$ is the new departure time generated for vehicle i. After performing several preliminary tests with different values, μ was set to be 0.1. In the meantime, the feasibility of the generated departure times is checked [to meet constraint (11)] and modified (regenerated), if needed.
  - **(3.2).** Evaluate the value of objective function for the generated neighboring solution, $f(y')$, through the MCS method in Algorithm 2.
- **Step (4):** If $f(y') \le f(y)$ or $r \le P_{ac}$ then $y \leftarrow y'$. If $f(y') \le f(y_{best})$ then $y_{best} \leftarrow y'$.
- **Step (5):** If the stopping criteria (I_{max}) is not met then $T_t = \beta \times T_{t-1}$, $t \leftarrow t+1$ and Go to Step 3; otherwise, stop and return y_{best}.

process of molten metals is simulated (Askarzadeh, dos Santos Coelho, Klein & Mariani, 2016). SA starts with a feasible initial solution and endeavors to ameliorate the current answer by generating a new solution in the vicinity of the current answer. Indeed, if the new solution leads to a lower objective function value, the current solution is replaced by the new solution; otherwise, SA rules decide whether the current solution is replaced by the new one or not. To be more precise, the algorithm starts with an initial positive temperature (T_0) and during the search, the temperature is steadily reduced. The probability of accepting a worse solution is also decreased as the temperature is reduced, i.e., although the algorithm might take a risk in accepting a worse solution in high temperatures, this risk-taking propensity will gradually decline with moving towards the end of the search process. In general, this strategy can help the algorithm to escape from local optimum solutions (Eglese, 1990; Meiri & Zahavi, 2006).

As discussed before, the proposed dispatching problem is a permutation-based problem in terms of vehicle dispatching order, i.e., a permutation of a given set of vehicles leads to a new ordering of those vehicles. As an illustrative example of adjusting vehicle dispatching order, Fig. 4 depicts a mixed-fleet dispatching problem with a given set of 8 vehicles: {12,12,12,15,15,15,18,18} that can be dispatched in $P(8;3,\ 3,\ 2) = \frac{8!}{3!\times 3!\times 2!} = 560$ different arrangements. In our SA algorithm, we employ efficient operators to produce diverse solutions in terms of dispatching sequence. In Algorithm 1, we describe the steps of the SA adopted to solve the proposed mixed-fleet vehicle dispatching problem. where:

$$P_{ac}(y, y', T_t) = \begin{cases} 1 & \text{if} f(y') \le f(y) \\ \exp(-\frac{f(y')-f(y)}{T_t}) & \text{Otherwise} \end{cases} \quad (22)$$

t	Iteration counter
I_{max}	Maximum number of iterations
T_0	Initial temperature
T_t	Temperature in iteration t
β	Cooling factor
y_{best}	Best found solution
$f(y)$	Objective function value for solution y
r	A uniform random number in [0, 1]
$P_{ac}(y, y', T_t)$	Probability function for accepting the non-improving solution y'.

As can be seen in expression (22), the probability of accepting a non-improving solution will depend on the difference between the corresponding objective function values, and also on the temperature at the relevant iteration. As discussed in Step 3, for the dispatching order of vehicles, neighborhood solutions are created through two different operators that are randomly used, including swap and inversion operators. In the swapping operator, two vehicles are randomly selected to be swapped in the same solution [see Fig. 4(a)]. In the inversion operator, a string of vehicles is randomly selected to be reversed in the same solution, as illustrated in Fig. 4(b).

To handle travel time uncertainty, each solution is repeatedly assessed over several simulation-based evaluations through a Monte Carlo Simulation (MCS) method which is a well-established method among researchers to cope with the uncertainty in stochastic programming problems to estimate expected values (Marseguerra, Zio & Podofillini, 2002), particularly among transportation researchers to handle the uncertainty of stochastic travel times in urban bus operations (e.g., Chen, Liu, Zhu & Wang, 2015; Gkiotsalitis & Van Berkum, 2020b; Liu, Yan, Qu & Zhang, 2013; Mou, Zhang & Liang, 2020; Wu et al., 2017; Zhang, Huang, Liu & Vu, 2020). For example, Liu et al. (2013) used a Genetic Algorithm (GA), combined with a MCS framework, to solve a stop-skipping service problem. In principle, the heuristic GA algorithm was employed to find the optimal stopping patterns, and the MCS method was employed to deal with travel time uncertainty in the process of solution evaluation. The same procedure is also executed in the studies of Chen et al. (2015) and Mou et al. (2020). Likewise, the MCS method is embedded as a subroutine of the SA algorithm in our study. Indeed, in Steps 2.2 and 3.2 of the SA algorithm, the evaluation of each solution is carried out by the MCS method over several replications due to the presence of stochastic travel times. The specific steps of the MCS scheme are summarized in Algorithm 2.

Algorithm 2
The steps of the Monte Carlo Simulation method incorporated into the SA.

(i) Set the MCS parameters: Set the counter of simulations m and its initial value as one; let $\bar{Z}^{(m)}$ denote the estimated value of the objective function (1); set the maximum number of simulations, $M_{max} = 1000$.
(ii) Perform travel time sampling: The vehicle travel time between each two successive stops is a random variable with predetermined mean and standard deviation. For each bus, sample the travel time between stops $j-1$ and j (i.e., $T^r_{i,j}$) using Eq. (9) based on its distribution function (log-normal distribution), where $i = 1, 2, \ldots, N_v$ and $j = 2, 3, \ldots, N_s$.
(iii) Calculate the variables: Based on the sampled travel time value, update the value of the relevant parameters in the solution using Eqs. (7)–(19), including bus movement and passenger flow calculations: bus travel time, headways, dwell time, arrival/departure time at each station, the number of passengers waiting, alighting and boarding the vehicle, and the number of passengers who are unable to board the vehicle.
(iv) Calculate the objective function value: Based on Eq. (1), calculate and update the objective value $Z^{(m)}$, and the final output of objective value is determined by the average value of simulation samples:

$$\bar{Z}^{(m)} = \frac{Z^{(m)} + (m-1)\,.\,\bar{Z}^{(m-1)}}{m} \tag{23}$$

(v) Check the stopping criterion: Increase the number of simulations by 1, i.e., $m = m+1$. If $m < M_{max}$, return to step ii; otherwise, stop and output the estimated objective function value $\bar{Z} = \bar{Z}^{(m)}$.

3.2. Full integer space enumeration

For the proposed mixed-fleet vehicle dispatching problem, we introduce a strategy to decompose the original MINLP problem into a certain series of continuous NLP subproblems with fixed binary variables for providing a direction towards the optimal solutions in the case of small and medium-sized instances. As discussed before, given the nature of the problem in terms of vehicle dispatching order, constraints (4–6) (resource limitations) transform the proposed mixed-fleet dispatching problem into a permutation-based combinatorial optimization problem (permutations with repetition due to the existence of several identical buses in the given set of vehicles), and the complexity of the problem will grow based on a factorial function depending on the fleet size and on the number of buses of each size. In principle, there exists a total of $\frac{N_v!}{A!\times B!\times C!}$ possible ways (in terms of dispatching sequence) for dispatching vehicles in a mixed-fleet operation. Accordingly, by fixing the dispatching sequences in $\frac{N_v!}{A!\times B!\times C!}$ different ways, we decompose the master MINLP problem into a certain number (equivalent to $\frac{N_v!}{A!\times B!\times C!}$) of continuous NLP subproblems, in each of which the optimal dispatching times of vehicles should be determined.

Indeed, each possible dispatching sequence is reflected by one of those subproblems. In other words, binary variables (x_{mi}) have already been fixed in each NLP subproblem and we just need to find the optimal dispatching times of vehicles in each subproblem (note that dispatching times are continuous decision variables in our model). Obviously, since the number of buses of each size will remain unchanged after carrying out a permutation on a given set of buses and merely the dispatching order of those buses are renewed in each subproblem, constraints (4–6) have been spontaneously satisfied in all the resulted subproblems. Finally, each NLP subproblem (for optimizing vehicles' dispatching times in that subproblem) is solved using the GAMS/CONOPT package that can determine that the solution is globally optimal in the NLP case and it will return Modelstat = 1 (Optimal). After solving all the NLP subproblems consecutively one after another, the best-found solution leading to the lowest passenger waiting time is identified. The specific steps of the proposed full integer space enumeration method are summarized in Algorithm 3. Indeed, the main aim of this method is to eliminate the difficulty of handling binary variables when solving small and medium dispatching instances.

Algorithm 3
The specific steps of the proposed full integer space enumeration method.

Step (1): Decompose the master MINLP problem into a certain number ($\frac{N_v!}{A!\times B!\times C!}$) of continuous NLP subproblems by fixing binary variables (vehicle dispatching order) in $\frac{N_v!}{A!\times B!\times C!}$ different ways, i.e., each subproblem is created based on one of those possible sequences, which can be prescribed for dispatching vehicles in a given mixed fleet.
Step (2): Solve the obtained NLP subproblems (with continuous decision variables of dispatching times) using the GAMS/CONOPT package for finding the optimal dispatching times of vehicles in each subproblem.
Step (3): Return the minimum-cost solution among the whole NLPs solved (return the obtained dispatching times together with the dispatching order already prescribed for that subproblem in Step 1).

Hence, we will employ this procedure for solving a set of small and medium-sized test problems (in Section 4.1), measuring later the difference from the best solution found by the SA to obtain certain insights about the quality of the attained solutions by the SA. Note that Algorithm 3 is not designed to handle travel time uncertainty and our test problems are solved with deterministic running times between stops to avoid further complexity and growth of computing times.

In recent years, there has been a considerable progress within the field of MILP and NLP (Achterberg & Wunderling, 2013; Bazaraa, Sherali & Shetty, 2013), which also enriches the field of MINLP as decomposition techniques for MINLP problems rely often on solving these types of subproblems (Kronqvist, Bernal, Lundell & Grossmann, 2019). For example, there is a large number of different solvers available and the number is growing in the NLP case: solvers like CONOPT, SNOPT, Knitro, and Mosek are well-known commercial options, and IPOPT is a well-known opensource solver (see Kronqvist et al., 2019 for further details on the above-mentioned NLP solvers). Overall, there is a wide variety in the algorithms behind NLP solvers, e.g., CONOPT implements a generalized reduced gradient approach, whereas SNOPT employs a sequential quadratic programming method, and Knitro, Mosek, and IPOPT use an interior-point approach (see Biegler, 2010 for a comprehensive review of NLP).

4. Numerical experiments

4.1. Small and medium-sized test instances

To obtain certain insights about the quality of the solutions found by the SA algorithm, the performance of the SA is evaluated by comparing its results to the optimal solutions obtained through the GAMS software 24.7.1 in solving a set of test problems generated randomly. Indeed, 25 small and medium-sized test problems are randomly prepared with various sizes and features (see Table 2), including different number of vehicles (fleet size), buses of each size, and bus stops (other input parameters are given in Appendix A). The gaps between the best solutions found by the SA algorithm and the optimal solutions obtained by GAMS are computed using Eq. (24).

$$\text{GAP} = \frac{(\text{SA}_{\text{answer}} - \text{GAMS}_{\text{answer}})}{\text{GAMS}_{\text{answer}}} \times 100 \tag{24}$$

All the computational experiments are performed on a personal computer with Intel(R) Core(TM) i5–6500 CPU @ 3.20 GHz and 16.0 GB RAM. As can be seen in Table 2, the SA has found the optimal solutions in most of the test problems. Furthermore, the maximum gap (0.83%) is observed in instance #21 and is less than 1 percent. Note that the average passenger waiting time (i.e., objective function value) is measured in minutes and the results of the SA and GAMS are compared to each other with two decimals (i.e., 0.01 minute which is less than one second). Indeed, a neglectable gap

Table 2
Computational results of the SA vs. GAMS in solving small and medium instances.

Class	Instance number	Instance features						Objective value			Comp. time (*sec*)	
		N_s	N_v	A	B	C	NLP*	GAMS	SA	GAP (%)	GAMS	SA
Small	#1	6	3	1	1	1	6	0.24	0.24	0.00	270	7
	#2	6	4	2	1	1	12	0.32	0.32	0.00	1450	14
	#3	6	4	1	2	1	12	0.30	0.30	0.00	1461	13
	#4	6	4	1	1	2	12	0.29	0.29	0.00	1455	14
	#5	8	5	1	1	3	20	0.36	0.36	0.00	3020	16
	#6	8	5	2	2	1	30	0.40	0.40	0.00	3002	15
	#7	8	5	2	1	2	30	0.39	0.39	0.00	3014	15
	#8	10	6	1	4	1	30	0.55	0.55	0.00	4681	17
	#9	10	6	1	1	4	30	0.49	0.49	0.00	4570	15
	#10	10	6	1	3	2	60	0.53	0.53	0.00	9365	17
	#11	10	6	1	2	3	60	0.51	0.51	0.00	9359	16
	#12	10	6	2	2	2	90	0.55	0.55	0.00	14,256	17
Medium	#13	12	7	5	1	1	42	0.94	0.94	0.00	7560	21
	#14	12	7	1	1	5	42	0.77	0.77	0.00	7551	19
	#15	12	7	4	2	1	105	0.92	0.92	0.00	18,910	20
	#16	12	7	4	1	2	105	0.88	0.88	0.00	19,215	21
	#17	12	7	1	3	3	140	0.82	0.82	0.00	25,536	22
	#18	12	7	2	2	3	210	0.85	0.85	0.00	38,316	20
	#19	14	8	6	1	1	56	1.56	1.56	0.00	13,448	25
	#20	14	8	1	1	6	56	1.13	1.13	0.00	13,372	22
	#21	14	8	1	2	5	168	1.21	1.22	0.83	40,328	25
	#22	14	8	2	4	2	420	1.41	1.41	0.00	>86,400	26
	#23	14	8	2	2	4	420	1.29	1.29	0.00	>86,400	24
	#24	14	8	3	3	2	560	1.45	1.46	0.69	>86,400	25
	#25	14	8	2	3	3	560	1.35	1.35	0.00	>86,400	24
	Max. gap%									**0.83**		

* No. of continuous NLP subproblems (i.e., $\frac{N_v!}{A! \times B! \times C!}$) solved by GAMS.

of 0.83% (between the average waiting times of 1.21 and 1.22 minutes) is even less than one second and passengers do not notice from a practical viewpoint; nonetheless, we have merely provided such results with 2 decimals for research purposes. The computing time required by the SA is always less than 0.5 minutes even for medium test instances. By contrast, the computing times are much more expensive in the case of GAMS, where computation times will increase markedly with a growth of the fleet size. This is due to the fact that the total number of continuous NLP subproblems that needed to be solved using GAMS (in Algorithm 3) is increased substantially as a function of $\frac{N_v!}{A! \times B! \times C!}$ that indeed represents the number of possible sequences to dispatch buses in a mixed-fleet dispatching problem (e.g., in instance #25, 560 NLP subproblems are solved consecutively one after another without interruption, due to the existence of 560 possible sequences for dispatching vehicles). This is indeed a great challenge in solving real-life instances, in which a tremendous number of dispatching sequences can be prescribed for a mixed-fleet operation, e.g., there exist 400,400 possible arrangements for dispatching buses in our real-world example (with $N_v = 16$, $A = 9$, $B = 4$, $C = 3$) presented in the next section. This challenging issue further highlights the importance and application of heuristic optimization algorithms that enable practitioners to discover good suboptimal solutions within a rational computing time for such a complex problem, coping with the difficulty of handling binary variables in large practical instances.

It should be noted that, in all the test problems, the dispatching orders found by the SA are exactly the same as those obtained in the optimal solutions. Indeed, the only difference that leads to such an insignificant gap (0.83% in instance #21) is attributed to a very slight difference (about seconds) in some dispatching times suggested by the SA compared to the optimal results of GAMS. This shows that the capability of SA's operators with their special neighborhood search mechanisms is quite promising, as the designed swapping and inversion operators (in Fig. 4) can fruitfully generate a new feasible dispatching order of vehicles through a random displacement of vehicles within the same fleet, thereby enabling the algorithm for a better exploitation of the best solutions in the feasible search space. This prominent feature would be of paramount importance in finding a suitable dispatching arrangement for real-life instances, in which bus operators are practically confronted with numerous dispatching arrangements.

The performance of the SA can be sensitive to the user-defined parameters, including initial temperature T_0, and cooling factor β. Hence, several preliminary runs are carried out with different values of parameters to select the most suitable parameter values from a set of candidate values (the range of each parameter is given). Indeed, our initial experiments are performed under different combinations of parameters, including changes in T_0 (from 6 to 12 with a step value of 1) and in β (from 0.85 to 0.99 with a step value of 0.01), and the results are evaluated for each parameter combination through a maximum number of 100 iterations for each run. This is indeed a commonly-used procedure in the literature for tuning the parameters of metaheuristics, such as the SA algorithm (e.g., Pishvaee, Kianfar & Karimi, 2010; Shaabani & Kamalabadi 2016). Moreover, since the performance of metaheuristics can vary when solving instances with different sizes, the SA's parameters are separately adjusted for small, medium, and large-scale problems. Finally, the preferred values were set as $T_0 = 9$ and $\beta = 0.95$ for small instances, $T_0 = 9$ and $\beta = 0.99$ for medium ones, as well as $T_0 = 10$ and $\beta = 0.99$ for large real-life instances presented in the next section. The details of the fine-tuning process are not provided here to avoid a lengthy paper.

4.2. Application area and real-life case study (large-scale instance)

To assess the effectiveness and efficiency of the proposed optimization model and the solution approach, several numerical experiments are carried out based on data from a bi-directional bus route, Military Road in North Sydney, Australia, which consists of a total of $N_s = 24$ stops (12 stops in each direction) (see Tirachini et al. (2014) for more details of the bus route). We consider the planning horizon from 7:00 am to 8:30 am. In the base case scenario, it is assumed that the bus route is served by a given mixed

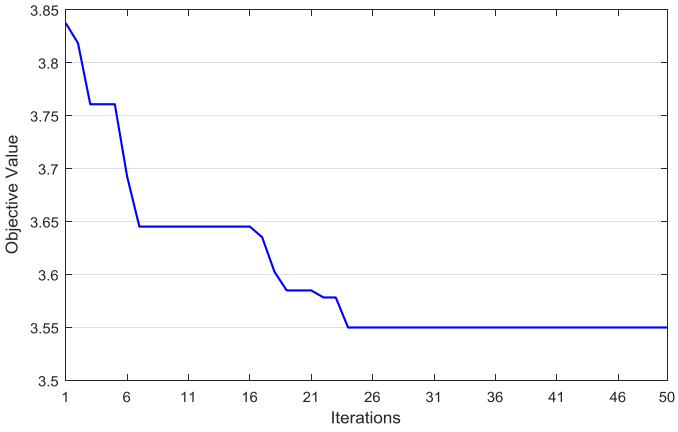

Fig. 6. Convergence trend of the simulated annealing algorithm in the large-scale problem.

fleet of 16 buses: {12, 12, 12, 12, 12, 12, 12, 12, 12, 15, 15, 15, 15, 18, 18, 18}. For example, under the assumption of even dispatching headways of 6 minutes (service frequency of 10 bus/h) and in a situation of constant passenger arrival rates at bus stops, regular bus headways and no passengers left behind (i.e., if buses never run at full capacity), the average waiting time would be 3 minutes.

The parameters used in this study are taken from Tirachini et al. (2014) and Tirachini (2014). For the sake of brevity, travel time distribution parameters and detailed information on demand rates are presented in Appendix A. In order to determine how the solution is sensitive to different degrees of demand availability, we compare the cases of low and high-resolution passenger arrival rates. As can be seen in Table A.6 and Fig. A.11, in the high-resolution demand case, the passenger arrival rates ($\lambda_j[t]$) are assumed to be constant during each 15-minute time interval, and they basically follow a bell-shaped pattern during the simulation time, peaking roughly at 7:45 am. In the low-resolution demand case; however, the passenger arrival rates remain constant during each one-hour period (see Table A.7), as commonly assumed in several bus supply optimization models (e.g., Hadas, Shnaiderman & Cedar, 2010; Niu et al., 2015; Tirachini et al., 2014). As discussed in Section 3, the proposed SA is coupled with a Monte Carlo simulation method to handle travel time uncertainty. Accordingly, the number of Monte Carlo simulations is set to be 1000.

4.3. Optimal dispatching policy under high-resolution demand volumes

Fig. 6 shows the convergence trend of the SA algorithm. As can be seen, the SA experiences a sharp decline in average waiting time, from 3.84 to 3.55 (min/pax) in the first twenty-four iterations before tailing off, i.e., the SA reached a plateau after 24 iterations within 5.8 minutes. Note that each candidate solution is being evaluated over several (1000) replications, due to the implementation of the MCS method incorporated as a subroutine into the SA to accomplish the evaluation process. Obviously, the computing time would be much shorter if a bus motion model ignores real-life operating conditions, such as stochastic travel times between stops, for the sake of simplicity.

Fig. 7(a) gives information about the optimal dispatching headways and the optimal bus dispatching order found by the SA under the high-resolution demand case (15-minute-dependent demand volumes), by showing the bus dispatching order in a time scale. Passengers experience an average waiting time of 3.55 (min/pax) under this optimal dispatching strategy. In total, 9.9% of passengers are left behind and need to wait for a second bus to board, which explains that the average waiting time is larger than 3 minutes. Importantly, with the proposed strategy, buses of one size are not necessarily dispatched consecutively one after the other, because not doing so allows us to have a more precise adjustment of supply (vehicle capacity) to demand in accordance with time-dependent passenger demand, thereby leading to a better utilization of vehicles' capacity in a given fleet of heterogeneous buses. Indeed, due to the provision of services with varying passenger-carrying capacities under a mixed-fleet operation, vehicles' capacity should be supplied to public transport users in line with temporal changes in demand. Otherwise, if buses are not dispatched in an optimal sequence together with considering the passengers' demand that may fluctuate within the planning horizon (i.e., spatial and temporal demand unbalances), the capacity of vehicles might not be used in due course (non-optimal utilization of resources), thus increasing the average passenger waiting time due to an increase in the number of passengers left behind owing to capacity constraints. For example, 18-m long buses, having more capacity to accommodate passenger volumes at the maximum loading sections, are mostly dispatched to cover the 7:45 am spike in passenger volumes. Moreover, as it is clear from Fig. 7(a), larger buses are dispatched with a larger headway between vehicles compared to smaller buses. Indeed, when different sizes of buses are dispatched to serve a single route, due to their different capacities, the headway between them should be different, otherwise more passengers would be left behind by the smaller buses, resulting in greater delays. This aspect is further discussed in the next subsection. The values of mean, standard deviation, and coefficient of variation for dispatching headway are respectively equal to 5.96 minute, 1.85 minute, and 0.31 in the optimal solution.

4.4. Comparison to even headway solutions

We compare the optimal solution from Section 4.3 to the case in which buses are dispatched at a uniform headway of 6 minutes. As has been shown, as long as vehicle capacity constraints are not binding and passenger arrival rates at bus stops are uniform, an even headway minimizes waiting time (Osuna & Newell, 1972). We develop two even-headway dispatching scenarios:

(i) Same dispatching order as in Fig. 7(a), under the constraint of a fixed 6-minute dispatching headway [see Fig. 7(b)].
(ii) Optimal dispatching order, under the constraint of a fixed 6-minute dispatching headway [see Fig. 7(c)].

In case (i), we see that the number of passengers left behind, and consequently the average passenger waiting time increase by 55% and 11.5%, going from 309 to 480 (pax) and from 3.55 to 3.96 (min/pax) respectively, if buses in the optimal solution are dispatched at an even headway of 6 minutes, while maintaining their dispatching order. In case (ii), we assume that vehicles are operated with a fixed 6-minute dispatching headway and only the dispatching order of each vehicle is optimized in this situation. We see that the percentage of passengers left behind is 14% and passenger waiting time increases by 9%, reaching 3.87 (min/pax). This shows the benefits of dispatching buses at uneven headways in a situation with different bus sizes and binding vehicle capacity constraints.

4.5. Comparing the optimal dispatching order with other predefined orders

In this section, we conduct further comparisons between the optimal solution and alternative dispatching schemes. Here we test the case of different patterns in which buses of the same size are dispatched consecutively. We test six different dispatching scenarios: $D_{12-15-18}$, $D_{12-18-15}$, $D_{15-12-18}$, $D_{15-18-12}$, $D_{18-12-15}$, $D_{18-15-12}$, in which buses are dispatched with a predetermined order and only

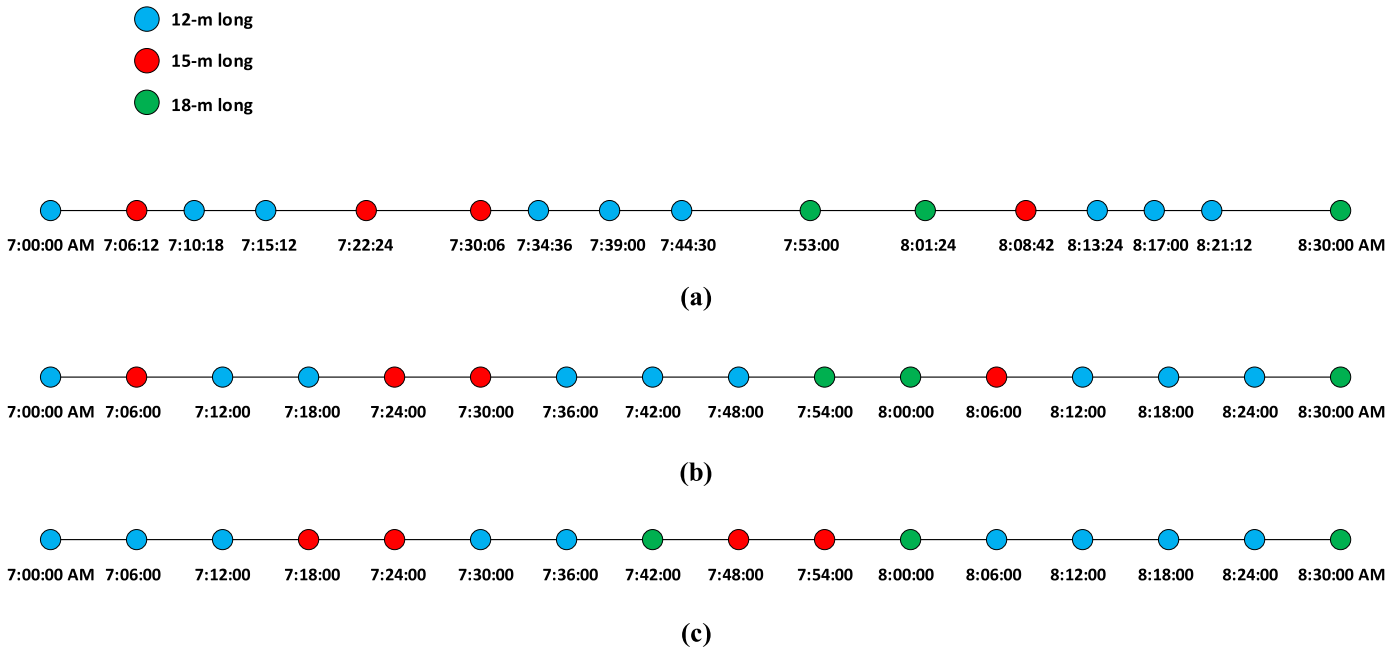

Fig. 7. Dispatching patterns under high-resolution demand data: (a) optimal dispatching pattern; (b) same dispatching order in (a) with the constraint of a fixed 6-minute dispatching headway; and (c) optimal dispatching order with the constraint of a fixed 6-minute dispatching headway.

Table 3
Comparing the optimal dispatching order with other predefined orders.

Scenario	The average passenger waiting time (min/pax)	Percentage of passengers left behind (%)
SA solution	3.55	9.9
$D_{12-15-18}$	4.45	20.6
$D_{12-18-15}$	4.14	17.1
$D_{15-12-18}$	4.04	15.9
$D_{15-18-12}$	4.42	19.9
$D_{18-12-15}$	4.78	23.6
$D_{18-15-12}$	4.24	18.2

the dispatching time of each bus is optimized. For example, in scenario $D_{12-15-18}$, 12 meters long buses are firstly dispatched, then 15 meters long buses, and finally 18 meters long buses are dispatched. The results are compared with the optimal scenario (see Table 3). Overall, the average passenger waiting time increases broadly in line with the percentage of passengers left behind. In the optimal scenario, the average passenger waiting time is 3.55 (min/pax), followed by a value of 4.04 (min/pax) in scenario $D_{15-12-18}$. Indeed, by comparing these two dispatching scenarios, we see that the optimal scenario leads to a decrease of 12.1% in the average passenger waiting, mainly caused by a further reduction in the percentage of passengers left behind, declining from 15.9% to 9.9%. Furthermore, using the optimal dispatching pattern instead of scenarios $D_{18-12-15}$, $D_{12-15-18}$, $D_{15-18-12}$, $D_{18-15-12}$, and $D_{12-18-15}$ can produce savings in the average passenger waiting time by 25.7, 20.2, 19.7, 16.8, and 14.3 percent, respectively. Therefore, in a mixed-fleet operation, it is relevant that bus agencies not only set bus dispatching headways, but also correctly assign vehicles of specific sizes at the right time, in order to minimize the unwanted effects of large peak demands that temporally use all the available vehicle capacities.

As can be seen in Fig. 7(a), 18-meter long buses are not dispatched early in the optimal scenario, showing that if larger buses (which have more room to carry passenger volumes at the maximum-load point of a route) are dispatched in an appropriate time to improve capacity utilization, they can reduce the number of passengers left behind; otherwise, it is probable that bus capacity is not used efficiently due to temporal and spatial differences in passenger volumes, thereby increasing passenger waiting times.

Fig. 8 gives information regarding the number of passengers left behind by each bus during the simulation time (a) in scenario $D_{15-12-18}$, and (b) in the optimal scenario. Looking firstly at Fig. 8(a), we see that the number of passengers who fail to board increases steadily when 12-m long buses are dispatched sequentially. Indeed, these buses have no enough room to accommodate passengers who missed the previous buses due to a shortage of capacity, and consequently this situation will continue to deteriorate when they are dispatched sequentially. As Fig. 8(b) shows, to optimize the capacity utilization of vehicles under the optimal scenario, buses of different capacities can be properly dispatched at specific times in accordance with demand conditions, and therefore the total number of passengers left behind by 12-m long buses reduces dramatically, dropping from 514 to 295 (pax).

4.6. Comparison between low and high-resolution demand cases

To understand how having high-resolution demand data instead of low-resolution demand information (one-hour-dependent demand volumes) can affect the optimal solution, a comparison between these two demand cases is made in this section. The relevance of this comparison rests on the fact that demand fixed on an hourly basis is common in most public transport frequency or dispatching setting models, e.g., Hadas et al. (2010), Tirachini et al. (2014), and Niu et al. (2015).

If the optimal solution (the optimal dispatching headways and the optimal bus dispatching order) found with low-resolution demand (see Fig. 9) is applied to high-resolution demand volumes, the average passenger waiting time increases by 15.5%, going from 3.55 to 4.10 (min/pax), due to an increase of 80% in the number of passengers left behind, going from 309 to 556 (pax). This result explicitly accentuates the advantage of having detailed demand information, especially when passenger arrival rates follow a bell-shaped pattern as time progresses. The peak inside

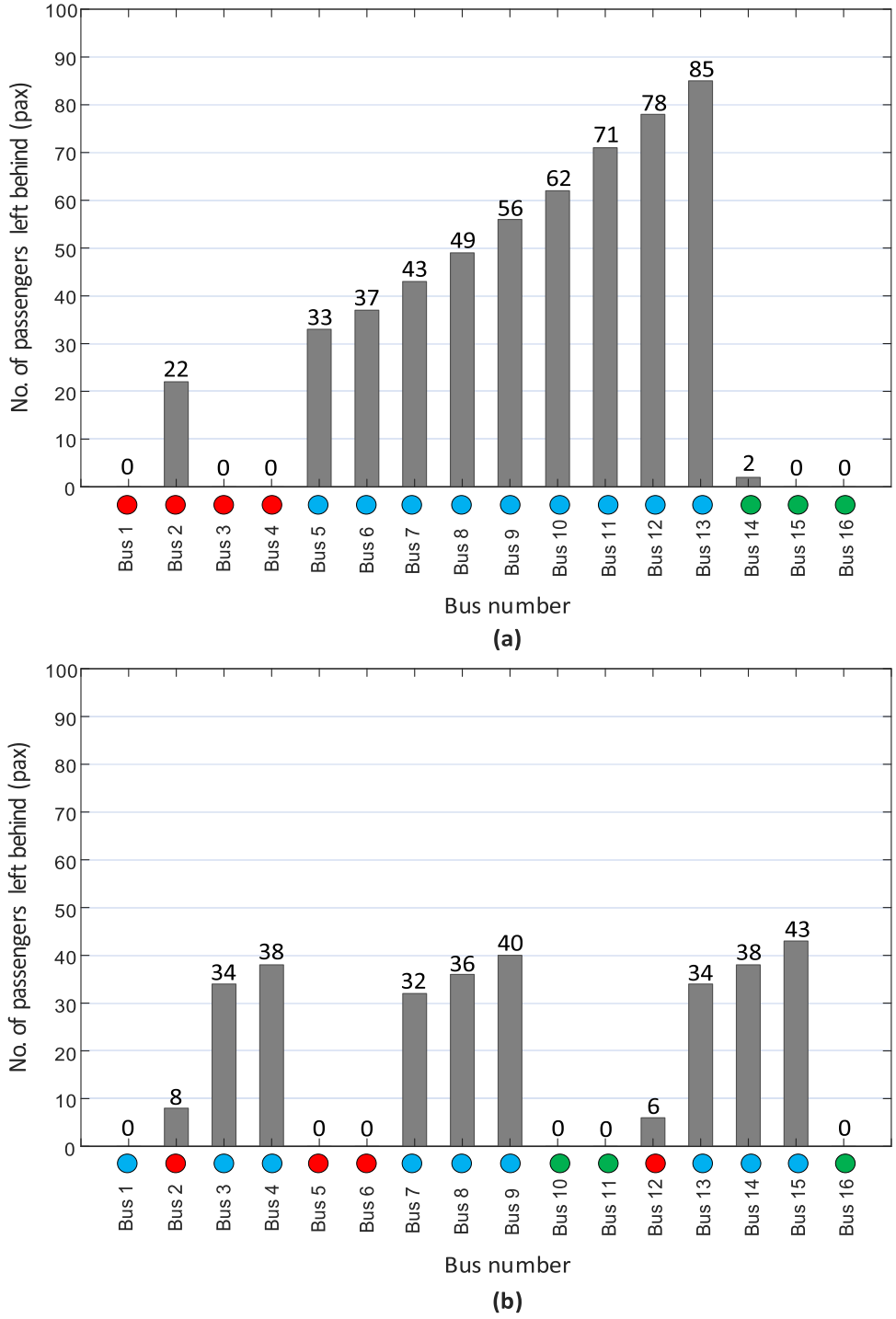

Fig. 8. Total number of passengers left behind by each bus during the entire analysis period: (a) in scenario $D_{15-12-18}$; and (b) in the optimal scenario.

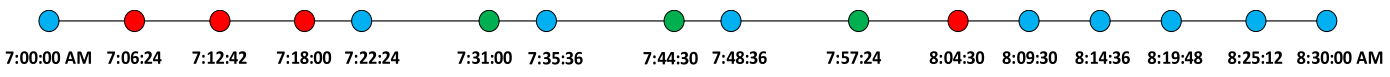

Fig. 9. Optimal dispatching pattern under low-resolution demand data.

the peak should be properly accounted for when designing dispatching schemes, which points to the relevance on investing to have detailed demand information for public transportation agencies, which includes the use of, e.g., smartcards and mobile phone data.

4.7. Uniform fleet

Next, we analyze the case in which a fleet with uniform bus sizes is available and only the dispatching time of each bus is optimized under high-resolution demand. The optimal dispatching

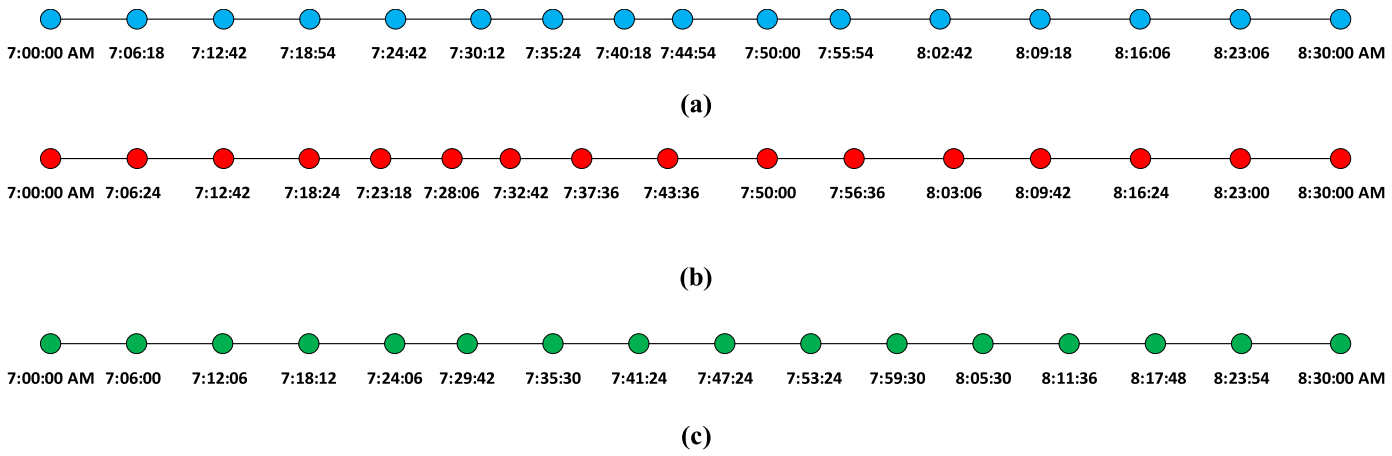

Fig. 10. Optimal dispatching headway for uniform fleets: (a) 12-meter fleet; (b) 15-meter fleet; and (c) 18-meter fleet.

solution with uniform fleets is presented in Fig. 10. In uniform fleets with 12, 15, and 18-meter long buses, the average passenger waiting time reaches the values of 5.96, 3.21, and 2.93 (min/pax) respectively, while the percentage of passengers left behind is equal to 34.2%, 5.6%, and 0% respectively in these three cases, showing that passengers are not confronted with a lack of capacity when an 18-meter fleet is used.

Regarding the optimal dispatching headways, as it is clear from Fig. 10, buses are not dispatched at quite even headways in order to deal with passenger demand fluctuation during the time operation, if capacity constraints are binding. This is clear for the case of 12-meter and 15-meter long buses, in which it is optimal to dispatch vehicles at uneven headways; in these cases the coefficient of variation of dispatching headways are 0.14 and 0.13, respectively. On the other hand, for 18-meter long buses, the dispatching headways are almost uniform, with a coefficient of variation of only 0.03. In any case, these coefficients of variation are much lower than that of the optimal solution with a mixed fleet in the base case (coefficient of variation 0.31). Therefore, we can conclude that the optimality of uneven dispatching headways stems from two elements: having a mixed fleet and having localized peaks on demand that make buses run full.

5. Concluding remarks

Bus dispatching strategy has a profound impact on passenger waiting times. In some cities, bus agencies have to combine vehicles of different sizes (e.g., rigid and articulated buses) in the form of a heterogeneous fleet to serve demand along a route. Such a situation can take place due to resource limitations together with historical reasons, e.g., when different sizes of buses are purchased at different times through different contracts. A heterogeneous fleet of buses considers services with different in-vehicle capacities during operations. In such a condition, the fundamental question that needs to be addressed is how to optimally deploy a given mixed fleet with buses of different sizes (capacities) to provide services that minimize passenger waiting time. In the present research, we formulated a novel heterogeneous fleet dispatching problem as a Mixed-Integer Nonlinear Programming (MINLP) model to optimize vehicle dispatching schemes (in terms of dispatching order and dispatching time) in the case of time-dependent demand volumes. We consider stochastic travel times between stops to reflect the actual operating conditions in urban bus systems. Moreover, for a more proper representation of operating conditions in mixed-fleet operations, which can provide services with different capacities during operations, in-vehicle capacity constraints are explicitly modeled in our mathematical formulation. The main objective of the model is to minimize the average passenger waiting time in order to improve the level of service provided to users. Due to the existence of a discrete set of feasible dispatching sequences, which can be prescribed for vehicles in our mixed-fleet operation, the proposed dispatching problem is a complex permutation-based combinatorial optimization problem. Given the complex nature of the problem formulated as an MINLP model, a Simulated Annealing (SA) algorithm with state-of-the-art features is employed to solve large-sized dispatching instances. Moreover, a Monte Carlo simulation framework is implanted as a subroutine into the SA to handle travel time uncertainty in the presence of stochastic travel times between stops. We also offer a full integer space enumeration method, whereby the master MINLP problem is decomposed into a certain number of NLP subproblems, to provide a direction towards the global optimal solutions for small and medium-sized instances.

To evaluate the effectiveness of the proposed model and the solution algorithm, a series of numerical experiments were conducted based on data from a real bus corridor under high-resolution demand volumes (15-minute-dependent demand volumes). The results showed that, in addition to bus dispatching headway, bus dispatching sequence can strongly affect passenger waiting time in the mixed-fleet operation. For example, in the optimal dispatching plan found by the SA, buses of the same size were not necessarily dispatched sequentially, leading to a better utilization of the existing fleet (i.e., a better utilization of vehicles' capacity) under time-dependent passenger demand. Indeed, this is explained by the fact that with an optimal dispatching sequence, a more precise adjustment of supply to demand is possible under time-varying demand volumes and the total savings in waiting time are mainly caused by a marked reduction in the number of passengers left behind (i.e., by preventing denied boarding problems from further exacerbation). Moreover, to highlight the importance of the bus dispatching order, we also tested six different dispatching scenarios, in which buses of one size were always dispatched consecutively one after the other. By comparing the optimal dispatching order with the worst-case scenario, we saw that the percentage of passengers left behind declined markedly from a peak of 23.6% to 9.9%, and consequently the average passenger waiting time went down by 25.7% to 3.55 (min/pax). We found that the desirability of programming uneven bus headways

depends on two factors: the existence of a fleet of vehicles of different sizes and of binding capacity constraints.

To highlight the value of having fine-grained demand information (every 15 minutes instead of every 60 minutes) when designing a dispatching scheme, the experiments were also tested with low-resolution demand volumes (one-hour-dependent demand volumes). The findings revealed that if the optimal dispatching scheme is designed based on low-resolution demand data and then this scheme is prescribed for real-world operations, the number of passengers left behind, and thus passenger waiting time can climb. Hence, in heavy-demand bus corridors in which passenger arrival rates follow a bell-shaped pattern, not taking the detailed demand information into account can lead to an unrealistic estimation of the passenger waiting time.

The findings of this study have a number of important implications for future practice. Our results with a focus on the operational planning level can also illustrate the need for the integration of both tactical and operational planning levels when bus operators aim to purchase (acquire) a heterogeneous fleet of buses to serve demand on their routes. Indeed, even at the tactical planning stage of fleet procurement (i.e., when determining the number and type of vehicles to be purchased), bus agencies should prospectively assess different dispatching cases (e.g., dispatching order) which later become visible (as an indisputable fact) during the operational planning of a mixed fleet. Importantly, to model tactical planning decisions and balance the benefits from the social welfare perspective in future studies, an integrated total cost (social cost) minimization model is needed to establish the trade-off between both user costs and operator costs (as conflicting objectives), as the solutions preferred by passengers are typically different from those preferred by public transport providers (e.g., see Proboste, Muñoz & Gschwender, 2020; Tirachini & Antoniou, 2020; Tirachini et al., 2014, which are some studies focusing on the tactical planning level). For example, although the deployment of a larger fleet size with bigger buses can reduce user costs (e.g., waiting time costs and in-vehicle discomfort costs) due to an increase in capacity, this situation can significantly increase capital (fleet acquisition) costs, and operating costs (e.g., energy/fuel consumption, tires, maintenance) for bus operators (Jara-Díaz & Gschwender, 2003; Tirachini & Antoniou, 2020; Tirachini et al., 2014). Hence, the needs and interests of both users (on the demand side) and operators (on the supply side) should be addressed when adjusting service supply items at the tactical public transit planning level (for an in-depth review of users' and operators' cost formulations, see Hörcher & Tirachini, 2021).

Another possible extension is considering dwell times at bus stops to be stochastic. Future research can expand the model into a network of bus routes operated with mixed fleets rather than a bi-directional bus route merely. Moreover, the proposed dispatching problem can be solved by means of further metaheuristics or a hybrid of them to compare their performance with the simulated annealing algorithm proposed in this study.

Acknowledgments

This research has been supported by Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering-IGSSE (MO3 project), by CONICYT Chile (Grant PIA/BASAL AFB180003), and by the Humboldt Foundation, Germany. The authors would like to sincerely thank the editor (Prof. José Fernando Oliveira) and three anonymous reviewers, whose constructive comments allowed us to greatly improve the content and presentation of this paper.

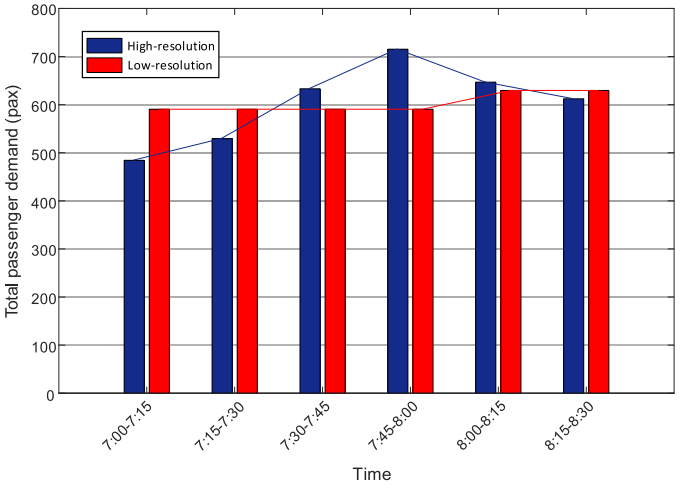

Fig. A.11. The total number of passengers arriving in the bus corridor during each 15-minute time interval based on Tables A.6 and A.7.

Table A.4
Parameter values for model application.

Parameter	Unit	Value
Acceleration time	s	6
Deceleration time	s	6
Time for opening and closing bus doors	s	6
Average alighting time per passenger	s/pax	1.5
Average boarding time per passenger	s/pax	2.5

Appendix A. Parameter values

Parameter values, which are used for both test instances and real-life instances (our case study of a bus corridor in Sydney, Australia), are listed in Table A.4. Note that size-dependent items (e.g., vehicle capacity), which are dependent on the size of vehicles, are presented in Table A.5.

For small and medium-sized test problems presented in Table 2 (Section 4.1), travel times between stops are assumed to be deterministic and set as 0.5 minutes between each two consecutive stops. For real-world applications, in which the model is applied to a real bus corridor in Sydney Australia (Section 4.2), travel times between stops are stochastic and the relevant travel time distribution parameters are given here. As introduced in Section 2, we assume a lognormal distribution of bus travel times between stops $j-1$ and j with mean and standard deviation of r_j and σ_j respectively.

$r_2 = 1.36$ (min); $r_3 = 1.35$; $r_4 = 1.37$; $r_5 = 0.95$; $r_6 = 1.25$; $r_7 = 1.59$; $r_8 = 0.79$; $r_9 = 0.77$; $r_{10} = 0.91$; $r_{11} = 1.09$; $r_{12} = 1.36$; $r_{14} = 1.49$; $r_{15} = 1.50$; $r_{16} = 1.48$; $r_{17} = 1.08$; $r_{18} = 1.38$; $r_{19} = 1.74$; $r_{20} = 0.92$; $r_{21} = 0.90$; $r_{22} = 1.03$; $r_{23} = 1.21$; $r_{24} = 1.49$.

$\sigma_2 = 0.11$ (min); $\sigma_3 = 0.11$; $\sigma_4 = 0.12$; $\sigma_5 = 0.06$; $\sigma_6 = 0.09$; $\sigma_7 = 0.15$; $\sigma_8 = 0.05$; $\sigma_9 = 0.04$; $\sigma_{10} = 0.06$; $\sigma_{11} = 0.08$; $\sigma_{12} = 0.11$; $\sigma_{14} = 0.14$; $\sigma_{15} = 0.15$; $\sigma_{16} = 0.14$; $\sigma_{17} = 0.08$; $\sigma_{18} = 0.13$; $\sigma_{19} = 0.19$; $\sigma_{20} = 0.06$; $\sigma_{21} = 0.06$; $\sigma_{22} = 0.08$; $\sigma_{23} = 0.10$; $\sigma_{24} = 0.14$.

For the bus corridor in Sydney, Tables A.6 and A.7 list the time-dependent passenger arrival rate ($\lambda_j[t]$) at each stop during the considered time horizon for every 15 minutes (high-resolution demand) vs. every 60 minutes (low-resolution demand). Indeed, in the high-resolution demand case, the passenger arrival rates are assumed to be constant during each 15-minute time interval;

Table A.5
Size-dependent parameters (Source: Tirachini et al., 2014).

Parameter	Unit	Standard bus (12-m long)	Rigid bus (15-m long)	Articulated bus (18-m long)
Vehicle capacity	pax/veh	70	90	120
PPA*=PPB#	%	60	43	30

* PPA stands for the proportion of passengers alighting through the busiest door.
# PPB stands for the proportion of passengers boarding through the busiest door.

Table A.6
High-resolution passenger arrival rates (unit: pax/min).

Stop	7:00–7:15	7:15–7:30	7:30–7:45	7:45–8:00	8:00–8:15	8:15–8:30
1	2.56	3.22	3.94	4.42	4.21	3.81
2	0.85	0.89	1.01	1.27	1.13	1.06
3	0.67	0.77	0.82	0.99	0.97	0.89
4	1.08	0.96	1.31	1.38	1.35	1.15
5	0.88	1.05	1.21	1.47	1.37	1.33
6	1.66	2.32	2.32	2.88	2.29	2.72
7	0.72	0.79	0.93	1.15	1.02	0.96
8	2.87	2.72	3.02	3.56	3.64	3.11
9	1.55	1.95	2.15	2.59	2.49	2.39
10	2.14	2.52	2.66	3.17	3.15	2.98
11	0.79	0.83	1.06	1.17	1.11	1.04
12	0	0	0	0	0	0
13	2.04	3.54	3.96	4.34	4.04	3.62
14	1.11	0.71	1.31	1.07	1.41	0.79
15	0.47	0.51	0.82	0.99	0.87	1.16
16	1.40	1.06	1.51	1.73	1.01	1.09
17	1.05	1.36	1.15	1.69	1.37	1.26
18	1.49	2.08	2.90	2.45	2.40	2.99
19	0.83	0.91	0.93	0.75	0.76	0.81
20	3.01	1.91	2.11	3.38	2.91	1.87
21	1.86	1.76	2.58	3.37	2.24	2.98
22	2.35	2.77	3.19	3.17	2.05	1.93
23	0.91	0.66	1.32	0.70	1.33	0.88
24	0	0	0	0	0	0

Table A.7
Low-resolution passenger arrival rates[a] (unit: pax/min).

Stop	7:00–8:00	8:00–8:30
1	3.54	4.01
2	1.01	1.09
3	0.81	0.93
4	1.18	1.25
5	1.15	1.35
6	2.29	2.51
7	0.89	0.99
8	3.04	3.37
9	2.06	2.44
10	2.62	3.06
11	0.96	1.07
12	0	0
13	3.47	3.83
14	1.05	1.10
15	0.69	1.02
16	1.43	1.05
17	1.31	1.32
18	2.23	2.69
19	0.86	0.78
20	2.60	2.39
21	2.39	2.61
22	2.87	1.99
23	0.89	1.11
24	0	0

[a] The values in Table A.7 are obtained through the average of passenger arrival rates during the relevant 15-minute time intervals presented in Table A.6.

however, in the low-resolution demand case, the passenger arrival rates remain constant during each one-hour period. The unit of arrival rate is passengers/min, i.e., the number of passengers arriving at a stop per minute.

References

Abdolmaleki, M., Masoud, N., & Yin, Y. (2020). Transit timetable synchronization for transfer time minimization. *Transportation Research Part B: Methodological, 131*, 143–159.

Achterberg, T., & Wunderling, R. (2013). Mixed integer programming: Analyzing 12 years of progress. *Facets of combinatorial optimization* (pp. 449–481). Springer.

Aguiléra, V., Allio, S., Benezech, V., Combes, F., & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies, 43*, 198–211.

Alkaya, A. F., & Duman, E. (2015). Combining and solving sequence dependent traveling salesman and quadratic assignment problems in PCB assembly. *Discrete Applied Mathematics, 192*, 2–16.

Altazin, E., Dauzère-Pérès, S., Ramond, F., & Trefond, S. (2020). A multi-objective optimization-simulation approach for real time rescheduling in dense railway systems. *European Journal of Operational Research, 286*, 662–672.

Askarzadeh, A., dos Santos Coelho, L., Klein, C. E., & Mariani, V. C. (2016). A population-based simulated annealing algorithm for global optimization. In *Proceedings of the IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 4626–4633). IEEE.

Barrena, E., Canca, D., Coelho, L. C., & Laporte, G. (2014a). Single-line rail rapid transit timetabling under dynamic passenger demand. *Transportation Research Part B: Methodological, 70*, 134–150.

Barrena, E., Canca, D., Coelho, L. C., & Laporte, G. (2014b). Exact formulations and algorithm for the train timetabling problem with dynamic demand. *Computers & Operations Research, 44*, 66–74.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2013). *Nonlinear programming: Theory and algorithms*. John Wiley & Sons.

Berrebi, S. J., Watkins, K. E., & Laval, J. A. (2015). A real-time bus dispatching policy to minimize passenger wait on a high frequency route. *Transportation Research Part B: Methodological, 81*, 377–389.

Bianco, L., Rinaldi, G., & Sassano, A. (1987). A combinatorial optimization approach to aircraft sequencing problem. *Flow control of congested networks* (pp. 323–339). Springer.

Biegler, L. T. (2010). *Nonlinear programming: Concepts, algorithms, and applications to chemical processes*. SIAM.

Bonami, P., Kilinç, M., & Linderoth, J. (2012). Algorithms and software for convex mixed integer nonlinear programs. *Mixed integer nonlinear programming* (pp. 1–39). Springer.

Burer, S., & Letchford, A. N. (2012). Non-convex mixed-integer nonlinear program-

ming: A survey. *Surveys in Operations Research and Management Science, 17*, 97–106.
Canca, D., Barrena, E., Algaba, E., & Zarzo, A. (2014). Design and analysis of demand-adapted railway timetables. *Journal of Advanced Transportation, 48*, 119–137.
Cats, O., Larijani, A. N., Koutsopoulos, H. N., & Burghout, W. (2011). Impacts of holding control strategies on transit performance: Bus simulation model analysis. *Transportation Research Record, 2216*, 51–58. https://doi.org/10.3141/2216-06.
Cats, O., West, J., & Eliasson, J. (2016). A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. *Transportation Research Part B: Methodological, 89*, 43–57. https://doi.org/10.1016/j.trb.2016.04.001.
Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research Part A: General, 18*, 439–453.
Ceder, A., & Marguier, M. H. J. (1985). Passenger waiting time at transit stops. *Traffic Engineering and Control, 26*.
Ceder, A. A., Hassold, S., & Dano, B. (2013). Approaching even-load and even-headway transit timetables using different bus sizes. *Public Transport, 5*, 193–217.
Chen, J., Liu, Z., Zhu, S., & Wang, W. (2015). Design of limited-stop bus service with capacity constraint and stochastic travel time. *Transportation Research Part E, 83*, 1–15. https://doi.org/10.1016/j.tre.2015.08.007.
Dai, Z., Liu, X. C., Chen, X., & Ma, X. (2020). Joint optimization of scheduling and capacity for mixed traffic with autonomous and human-driven buses: A dynamic programming approach. *Transportation Research Part C: Emerging Technologies, 114*, 598–619.
Delgado, F., Munoz, J. C., & Giesen, R. (2012). How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological, 46*, 1202–1217. https://doi.org/10.1016/j.trb.2012.04.005.
dell'Olio, L., Ibeas, A., & Ruisánchez, F. (2012). Optimizing bus-size and headway in transit networks. *Transportation (Amst), 39*, 449–464.
Desaulniers, G., & Hickman, M. D. (2007). Public transit. *Handbooks in Operations Research and Management Science, 14*, 69–127.
Drabicki, A., Kucharski, R., Cats, O., & Szarata, A. (2021). Modelling the effects of real-time crowding information in urban public transport systems. *Transportmetrica A: Transport Science, 17*, 675–713.
Duran-Micco, J., Vermeir, E., & Vansteenwegen, P. (2020). Considering emissions in the transit network design and frequency setting problem with a heterogeneous fleet. *European Journal of Operational Research, 282*, 580–592.
Eglese, R. W. (1990). Simulated annealing: A tool for operational research. *European Journal of Operational Research, 46*, 271–281.
Farahani, R. Z., Miandoabchi, E., Szeto, W. Y., & Rashidi, H. (2013). A review of urban transportation network design problems. *European Journal of Operational Research, 229*, 281–302.
Furth, P. G., & Wilson, N. H. M. (1981). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record, 818*, 1–7.
Gao, Y., Kroon, L., Schmidt, M., & Yang, L. (2016). Rescheduling a metro line in an over-crowded situation after disruptions. *Transportation Research Part B: Methodological, 93*, 425–449. https://doi.org/10.1016/j.trb.2016.08.011.
Gkiotsalitis, K., & Cats, O. (2018). Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. *Transportation Research Part C: Emerging Technologies, 88*, 187–207.
Gkiotsalitis, K. (2020). A model for the periodic optimization of bus dispatching times. *Applied Mathematical Modelling, 82*, 785–801.
Gkiotsalitis, K., & Alesiani, F. (2019). Robust timetable optimization for bus lines subject to resource and regulatory constraints. *Transportation Research Part E: Logistics and Transportation Review, 128*, 30–51.
Gkiotsalitis, K., & Cats, O. (2018). Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. *Transportation Research Part C: Emerging Technologies, 88*, 187–207. https://doi.org/10.1016/j.trc.2018.01.026.
Gkiotsalitis, K., & Van Berkum, E. C. (2020a). An exact method for the bus dispatching problem in rolling horizons. *Transportation Research Part C: Emerging Technologies, 110*, 143–165.
Gkiotsalitis, K., & Van Berkum, E. C. (2020b). An analytic solution for real-time bus holding subject to vehicle capacity limits. *Transportation Research Part C: Emerging Technologies, 121*, Article 102815.
Gomes, A. M., & Oliveira, J. F. (2006). Solving irregular strip packing problems by hybridising simulated annealing and linear programming. *European Journal of Operational Research, 171*, 811–829.
Hadas, Y., & Shnaiderman, M. (2012). Public-transit frequency setting using minimum-cost approach with stochastic demand and travel time. *Transportation Research Part B: Methodological, 46*, 1068–1084.
Hadas, Y., Shnaiderman, M., & Cedar, A. (2010). Public transit frequency setting using minimum cost approach with stochastic demand, in: Proceedings of the 45th annual conference of the ORSNZ, Auckland, New Zealand. pp. 353–362.
Hassannayebi, E., & Zegordi, S. H. (2017). Variable and adaptive neighbourhood search algorithms for rail rapid transit timetabling problem. *Computers & Operations Research, 78*, 439–453.
Hörcher, D., & Tirachini, A. (2021). A review of public transport economics. *Economics of Transportation, 25*, Article 100196.
Jara-Díaz, S., & Gschwender, A. (2003). Towards a general microeconomic model for the operation of public transport. *Transport Reviews, 23*, 453–469.
Karimi-Mamaghan, M., Mohammadi, M., Meyer, P., Karimi-Mamaghan, A. M., & Talbi, E. G. (2021). Machine learning at the service of Meta-heuristics for solving combinatorial optimization problems: A state-of-the-art. *European Journal of Operational Research*.
Kronqvist, J., Bernal, D. E., Lundell, A., & Grossmann, I. E. (2019). A review and comparison of solvers for convex MINLP. *Optimization and Engineering, 20*, 397–455.
Li, Y., Xu, W., & He, S. (2013). Expected value model for optimizing the multiple bus headways. *Applied Mathematics and Computation, 219*, 5849–5861.
Lin, S. W., Cheng, C. Y., Pourhejazy, P., & Ying, K.-. C. (2021). Multi-temperature simulated annealing for optimizing mixed-blocking permutation flowshop scheduling problems. *Expert Systems with Applications, 165*, Article 113837.
Liu, Z., Yan, Y., Qu, X., & Zhang, Y. (2013). Bus stop-skipping scheme with random travel time. *Transportation Research Part C: Emerging Technologies, 35*, 46–56.
Luo, X., Liu, Y., Yu, Y., Tang, J., & Li, W. (2019). Dynamic bus dispatching using multiple types of real-time information. *Transportmetrica B: Transport Dynamics, 7*, 519–545.
Marseguerra, M., Zio, E., & Podofillini, L. (2002). Condition-based maintenance optimization by means of genetic algorithms and Monte Carlo simulation. *Reliability Engineering & System Safety, 77*, 151–165.
Martínez, H., Mauttone, A., & Urquhart, M. E. (2014). Frequency optimization in public transportation systems: Formulation and metaheuristic approach. *European Journal of Operational Research, 236*, 27–36.
Meiri, R., & Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research, 171*, 842–858.
Meng, L., Luan, X., & Zhou, X. (2016). A train dispatching model under a stochastic environment: Stable train routing constraints and reformulation. *Networks and Spatial Economics, 16*, 791–820.
Meng, L., & Zhou, X. (2019). An integrated train service plan optimization model with variable demand: A team-based scheduling approach with dual cost information in a layered network. *Transportation Research Part B: Methodological, 125*, 1–28.
Mou, Z., Zhang, H., & Liang, S. (2020). Reliability optimization model of stop-skipping bus operation with capacity constraints. *Journal of Advanced Transportation*, 2020.
Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies, 24*, 9–18.
Nachtigall, K., & Voget, S. (1997). Minimizing waiting times in integrated fixed interval timetables by upgrading railway tracks. *European Journal of Operational Research, 103*, 610–627.
Newell, G. F. (1971). Dispatching policies for a transportation route. *Transportation Science, 5*, 91–105.
Niu, H., & Zhou, X. (2013). Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. *Transportation Research Part C: Emerging Technologies, 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 Research Part B: Methodological, 76*, 117–135.
Osman, I. H., & Potts, C. N. (1989). Simulated annealing for permutation flow-shop scheduling. *Omega, 17*, 551–557.
Osuna, E. E., & Newell, G. F. (1972). Control strategies for an idealized public transportation system. *Transportation Science, 6*, 52–72.
Pishvaee, M. S., Kianfar, K., & Karimi, B. (2010). Reverse logistics network design using simulated annealing. *International Journal of Advanced Manufacturing Technology, 47*, 269–281.
Proboste, F., Muñoz, J. C., & Gschwender, A. (2020). Comparing social costs of public transport networks structured around an Open and Closed BRT corridor in medium sized cities. *Transportation Research Part A: Policy and Practice, 138*, 187–212.
Robenek, T., Azadeh, S. S., Maknoon, Y., de Lapparent, M., & Bierlaire, M. (2018). Train timetable design under elastic passenger demand. *Transportation Research Part B: Methodological, 111*, 19–38.
Ruiz, R., & Stützle, T. (2007). A simple and effective iterated greedy algorithm for the permutation flowshop scheduling problem. *European Journal of Operational Research, 177*, 2033–2049.
Sánchez-Martínez, G. E., Koutsopoulos, H. N., & Wilson, N. H. M. (2016). Real-time holding control for high-frequency transit with dynamics. *Transportation Research Part B: Methodological, 83*, 1–19. https://doi.org/10.1016/j.trb.2015.11.013.
Shaabani, H., & Kamalabadi, I. N. (2016). An efficient population-based simulated annealing algorithm for the multi-product multi-retailer perishable inventory routing problem. *Computers & Industrial Engineering, 99*, 189–201.
Sun, Y., & Xu, R. (2012). Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. *Transportation Research Record, 2275*, 58–67.
Szeto, W. Y., & Wu, Y. (2011). A simultaneous bus route design and frequency setting problem for Tin Shui Wai, Hong Kong. *European Journal of Operational Research, 209*, 141–155.
Talbi, E.-. G. (2009). *Metaheuristics: From design to implementation*. John Wiley & Sons.
Tirachini, A. (2014). The economics and engineering of bus stops: Spacing, design and congestion. *Transportation Research Part A: Policy and Practice, 59*, 37–57. https://doi.org/10.1016/j.tra.2013.10.010.
Tirachini, A., & Antoniou, C. (2020). The economics of automated public transport: Effects on operator cost, travel time, fare and subsidy. *Economics of Transportation, 21*, Article 100151.
Tirachini, A., Hensher, D. A., & Rose, J. M. (2014). Multimodal pricing and optimal design of urban public transport : The interplay between traffic congestion and

bus crowding. *Transportation Research Part B, 61*, 33–54. https://doi.org/10.1016/j.trb.2014.01.003.

Wang, Y., Tang, T., Ning, B., van den Boom, T. J. J., & De Schutter, B. (2015). Passenger-demands-oriented train scheduling for an urban rail transit network. *Transportation Research Part C: Emerging Technologies, 60*, 1–23. https://doi.org/10.1016/j.trc.2015.07.012.

Wang, Z., & Haghani, A. (2020). Column generation-based stochastic school bell time and bus scheduling optimization. *European Journal of Operational Research, 286*, 1087–1102.

Wardman, M. (2004). Public transport values of time. *Transport Policy, 11*, 363–377. https://doi.org/10.1016/j.tranpol.2004.05.001.

Wu, W., Liu, R., & Jin, W. (2017). Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour. *Transportation Research Part B: Methodological, 104*, 175–197.

Xumei, C., Qiaoxian, L. I. U., & Guang, D. U. (2011). Estimation of travel time values for urban public transport passengers based on SP survey. *Journal of Transportation Systems Engineering and Information Technology, 11*, 77–84.

Yap, M., Cats, O., & van Arem, B. (2020). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science, 16*, 23–42.

Zhang, F., & Liu, W. (2019). Responsive bus dispatching strategy in a multi-modal and multi-directional transportation system: A doubly dynamical approach. *Transportation Research Procedia, 38*, 119–138.

Zhang, L., Huang, J., Liu, Z., & Vu, H. L. (2020). An agent-based model for real-time bus stop-skipping and holding schemes. *Transportmetrica A: Transport Science*, 1–33.

Zhang, T., Li, D., & Qiao, Y. (2018). Comprehensive optimization of urban rail transit timetable by minimizing total travel times under time-dependent passenger demand and congested conditions. *Applied Mathematical Modelling, 58*, 421–446.

Zhang, Y., Qi, M., Lin, W.-. H., & Miao, L. (2015). A metaheuristic approach to the reliable location routing problem under disruptions. *Transportation Research Part E: Logistics and Transportation Review, 83*, 90–110.

Zhao, J., Bukkapatnam, S., & Dessouky, M. M. (2003). Distributed architecture for real-time coordination of bus holding in transit networks. *IEEE Transactions on Intelligent Transportation Systems, 4*, 43–51. https://doi.org/10.1109/TITS.2003.809769.

Zhu, Y., Koutsopoulos, H. N., & Wilson, N. H. M. (2017). Inferring left behind passengers in congested metro systems from automated data. *Transportation Research Procedia, 23*, 362–379.