Transportation Research Part C 158 (2024) 104437

Contents lists available at ScienceDirect

Transportation Research Part C

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

A survey of Flex-Route Transit problem and its link with Vehicle Routing Problem

Reza Shahin [a,*], Pierre Hosteins [a,d], Paola Pellegrini [a], Pierre-Olivier Vandanjon [b], Luca Quadrifoglio [c]

[a] *COSYS-ESTAS, Univ Gustave Eiffel, IFSTTAR, Univ Lille, F-59650 Villeneuve d'Ascq, France*
[b] *AME-PLOTT, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France*
[c] *Texas A&M University, College Station, TX 77843, United States of America*
[d] *Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, 10149, Torino, Italy*

ARTICLE INFO

Keywords:
Transportation
Flex-Route Transit
Vehicle Routing problem
Demand Responsive Transit
Dial-a-Ride Problem

ABSTRACT

Flexible transport systems such as Demand Responsive Transit (DRT) are becoming more and more popular over the last few years due to their convenience for customers. However, this convenience comes at a price. Transport authorities are currently looking for ways to improve service flexibility of Conventional Public Transport (CPT), which is undoubtedly cheaper than DRT. This justifies the need for Flex-Route Transit (FRT), which combines the flexibility of DRT and the low cost of CPT. This paper surveys research developments on FRT, as a promising alternative mode of public transport. Based on this survey, we discuss current research gaps that may be filled to increase FRT applicability. Moreover, we show how literature on classic Operations Research problems is of help to do so. In particular, we study similarities and differences between FRT and Vehicle Routing Problem, and specifically with one of its variant named Dial-a-Ride Problem. The analysis illustrates promising techniques that may be of use for solving FRT.

1. Introduction

Faced with environmental problems, increasing request for mobility and congestion, transport authorities are trying to limit the use of private cars (Sims et al., 2014). They are currently looking for ways to improve service flexibility of public transport in a cost efficient way.

Today, Conventional Public Transport (CPT) operates based on planned schedules or frequency and shuttles pick-up and drop-off customers at pre-defined stops. On the contrary, Demand Responsive Transit (DRT) systems provide very flexible transport solutions for customers (Palmer et al., 2004). These customers can define *ad hoc stops*, i.e. specific pick-up and drop-off locations, as well as desired stopping times. A famous example of DRT is the Dial-a-Ride Problem (DARP) transport system (Ho et al., 2018). Undoubtedly, CPT is cheaper than DRT (Quadrifoglio et al., 2006; Palmer et al., 2004). However, customers perceive the CPT to be inconvenient for three main reasons:

1. The stop locations for pick-up and drop-off rarely coincide with their specific needs;
2. The service schedule is not flexible;
3. The total time of the trip is longer than that of private cars.

* Corresponding author.
E-mail addresses: Rezaa.shahin.1992@gmail.com (R. Shahin), Pierre.hosteins@univ-eiffel.fr (P. Hosteins), Paola.pellegrini@univ-eiffel.fr (P. Pellegrini), Pierre-olivier.vandanjon@univ-eiffel.fr (P.-O. Vandanjon), quadrifo@tamu.edu (L. Quadrifoglio).

https://doi.org/10.1016/j.trc.2023.104437
Received 4 March 2023; Received in revised form 27 September 2023; Accepted 23 November 2023
Available online 2 December 2023
0968-090X/

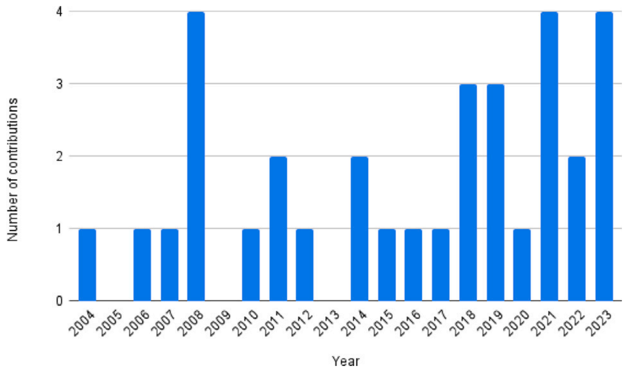

Fig. 1. Number of contributions per year on the FRT system since 2004.

DRT systems can reduce these inconveniences, but this comes at a price. The need to effectively balance costs and flexibility motivates the introduction of Demand-Adaptive Systems (DAS) (Errico et al., 2013) and among them the Flex-Route Transit (FRT). The FRT is a transport system that combines the flexibility of DRT with the low cost of CPT. In particular, shuttles follow fixed routes and schedules as in a CPT system. Nevertheless, they can deviate from these routes to perform ad hoc stops for picking-up or dropping-off customers at their desired locations within a predefined service area, by exploiting available slack times at the fixed stops. Transport systems with similarities to the FRT exist. For example, flexible transit as proposed by Crainic et al. (2001), Malucelli et al. (1999) relies on a sequence of compulsory and optional stops where customers may want to be picked-up or dropped-off. In the absence of any request for optional stops, the shuttle goes directly from one compulsory stop to another through the shortest route. Daganzo (1984a) discusses another type of transport system, called checkpoint Dial-a-Ride. Here, passengers can request ad hoc stops choosing among a finite number of possible locations. The a priori definition of optional stops or possible locations differentiates these systems from the FRT, where any location in the service area can be chosen by customers. Being able to freely choose ad hoc stops is particularly advantageous in areas where walking may be inconvenient due, for example, to security issues. In order to underline the recent increased focus on FRT in the literature, we plot in Fig. 1 the number of contributions on this specific subject since its introduction in 2004. We refer the reader to Koffman (2004) and Errico et al. (2013) for more information on other flexible services and aspects to be taken into account when designing them.

FRT systems have already been in application in practice, albeit at a reduced, simplified scale. For example, during the early 2000's, the Metropolitan Transit Authority (MTA) of Los Angeles County had decided to operate one of its bus lines as a FRT (Quadrifoglio and Dessouky, 2004). During daytime, this line operated as a fixed route bus, at a predefined frequency. During nighttime, the bus still performed its usual stops, but the passengers had the possibility to request for ad hoc stops within half a mile from the fixed route. In this application, as the number of passengers is low, the driver is able to make effective decisions concerning the shuttle routing. However, should one deal with a large number of passengers requesting deviations, an optimization based decision support system is necessary. A survey on *semi-flexible* systems has been published a decade ago (Errico et al., 2013). These systems include the FRT described as a *point deviation* system. The survey unifies the various semi-flexible systems inside a common framework and classifies the different contributions according to their level of application, namely: 1. strategic; 2. tactical; 3. operations; 4. evaluation. Besides FRT, other semi-flexible systems are described with varied characteristics, e.g.: the shuttle follows a fixed route along which customers can be picked up by issuing a flag-request, and it has to rejoin the route where it left it in case of deviation (*Route Deviation*); the shuttle can stop at additional optional stops along the fixed route (*Request Stops*). Some specific cases of FRT are sometimes labeled with specific names, such as *Demand-Responsive Connector* (a scheduled transfer point connects the service with a fixed-route network), *Flexible Route Segments* (the shuttle can switch to a demand-responsive mode only on specific portions of its route) or *Zone Routes* (no fixed checkpoint exists between the origin and destination depots of the route). An update on the literature is proposed in the recent survey of Vansteenwegen et al. (2022), which also proposes a unifying view of many on-demand public transport systems. The scope of the survey is, however, broader than the present article which concentrates on the specificities of the FRT system, and only a fraction of the works surveyed in this review are discussed by the authors of Vansteenwegen et al. (2022).

In this paper, we propose a survey of the literature on the FRT since its introduction in 2004, mainly from a combinatorial optimization point of view. We describe the methods that have been proposed to study some relevant FRT versions, including interesting recent developments. Based on this survey, we identify research gaps that have not received a lot of attention in the literature so far. These are gaps that need to be filled to provide decision support systems for applying FRT in a high demand public transport setting or to provide public transport operators with reliable tools to study the viability of switching to such systems. To point out promising research directions, we rely on the rich literature existing for some classic Operations Research problems. In particular, we focus on the Vehicle Routing Problem with Pick-up and Delivery and Time Windows (VRPPDTW) and on the Dial-a-Ride Problem (DARP). In the literature, FRT was originally introduced as Mobility Allowance Shuttle Transit (MAST) (Quadrifoglio and Dessouky, 2004; Quadrifoglio et al., 2006, 2007, 2008a; Quadrifoglio and Li, 2009; Quadrifoglio and Dessouky, 2008; Quadrifoglio and Shen, 2010; Quadrifoglio et al., 2008b; Zhao and Dessouky, 2008) and later on renamed to FRT.

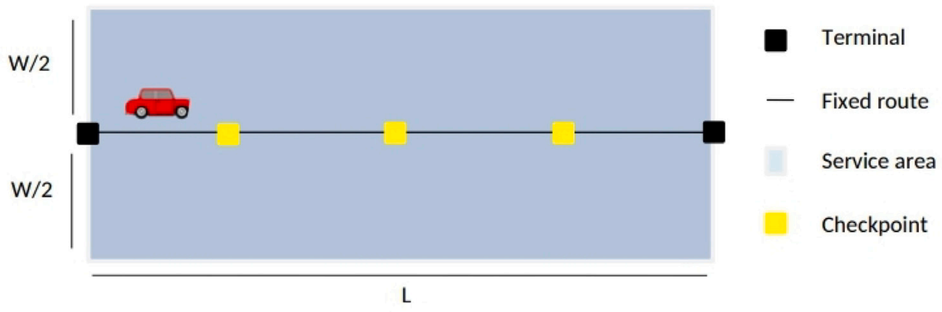

Fig. 2. Example of the fixed route and service area of a Flex-Route Transit (FRT) system.

For reasons of clarity, we only use the term FRT throughout the paper. Although *Fixed-Route Transit* is sometimes referred to as FRT in the literature on transport science, we refer to it by the abbreviation CPT and only use the FRT acronym when discussing Flex-Route Transit.

The remainder of this paper is organized as follows. We begin with the system definition in Sections Section 2. We then review the FRT literature in Section 3 as well as some real-life implementations and the characteristics of the existing case studies in Section 4 before discussing research gaps in Section 5. After concluding in Section 6, we also provide a limited survey of existing mathematical programming approaches for specific Vehicle Routing Problems of relevance to the FRT in Appendix in order to provide inspiration for tackling the FRT through more advanced exact algorithms.

2. FRT definition

An FRT system is based on a *base route* which is very similar to a traditional bus system, with an *origin* and a *destination terminal* between which the shuttles perform back-and-forth trips. Between the depots are present a set of intermediary predetermined stops called *checkpoints* and the shuttles leave these checkpoints following a predetermined timetable. The duration of a trip is named *service time interval* and the sum of service time intervals is *the time horizon*. The shuttles are assumed to move at constant speed. Nevertheless, differently from a basic bus transport system, the timetable is conceived so as to provide additional *slack time* in addition to the riding time between two consecutive checkpoints, which allows the shuttles to deviate from the base route. This additional time is used to serve some customers *on demand* at a location of their choice (ad hoc stops), as long as the available slack time is sufficient, i.e. the predetermined timetable at checkpoints cannot be violated. The *longitudinal speed* is defined as the average speed held between terminals. If no deviation is performed, it is equal to the constant speed of the shuttle. It decreases with the distance traveled for ad hoc stops. Customers can use checkpoints for their trips along the fixed route, or they may request ad hoc stops within the service area. A schematic view of such a system is shown in Fig. 2. The FRT responds to four different types of customers' requests based on their pick-up and drop-off locations (Quadrifoglio et al., 2007):

1. PD (regular): pick-up and drop-off at checkpoints.
2. PND (hybrid): pick-up at a checkpoint, drop-off not at a checkpoint (ad hoc stop).
3. NPD (hybrid): pick-up not at a checkpoint (ad hoc stop), drop-off at a checkpoint.
4. NPND (random): pick-up and drop-off not at checkpoints (ad hoc stops).

When optimizing the FRT problem, the objective to minimize is typically a combination of: 1. total travel time of shuttles; 2. total ride time of all customers; 3. total waiting time for customers' pick-up.

The FRT can be defined in a *static* or a *dynamic* environment. In the former, all requests are known before the beginning of the service. In the latter, requests can be issued at any time, and they are dynamically affecting the shuttles schedule. In this case, *backtracking* may be of use, i.e. shuttles may return to already visited locations to add an ad hoc stop. A maximum allowed backtracking distance may be set. In the following we will refer to this distance as *backtracking threshold*.

Fig. 3 exemplifies the scheduling decision process of a shuttle in the FRT in a dynamic environment where backtracking is not possible. It represents the trip portion between the origin and the first checkpoint in Fig. 2. As typically assumed in the literature, the shuttle travels only horizontally or vertically in the service area. Requests a and b are issued before the shuttle departure. While the shuttle travels to serve a, request c is issued. As no backtracking is possible, the shuttle does not go back for request c although it just passed the corresponding location along the longitudinal axis, and continues its way to a. Now, suppose request d is issued right after the shuttle serves a. Instead of moving directly to b along the dashed line, the shuttle modifies its route as shown in the solid line to pick-up request d.

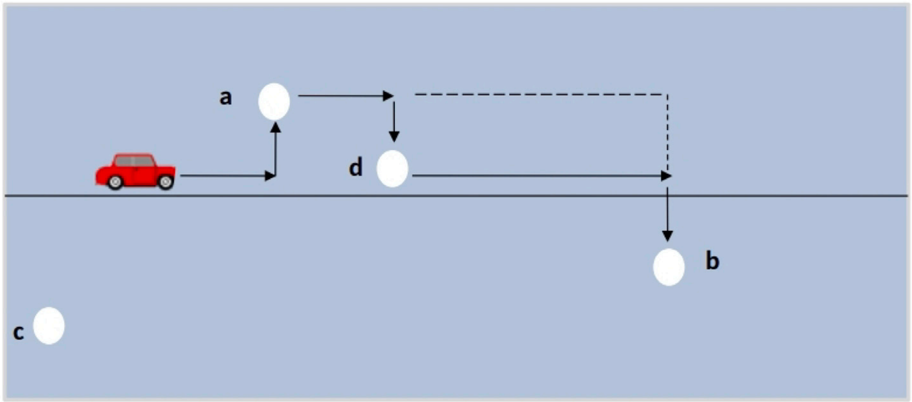

Fig. 3. Illustration of how FRT works.

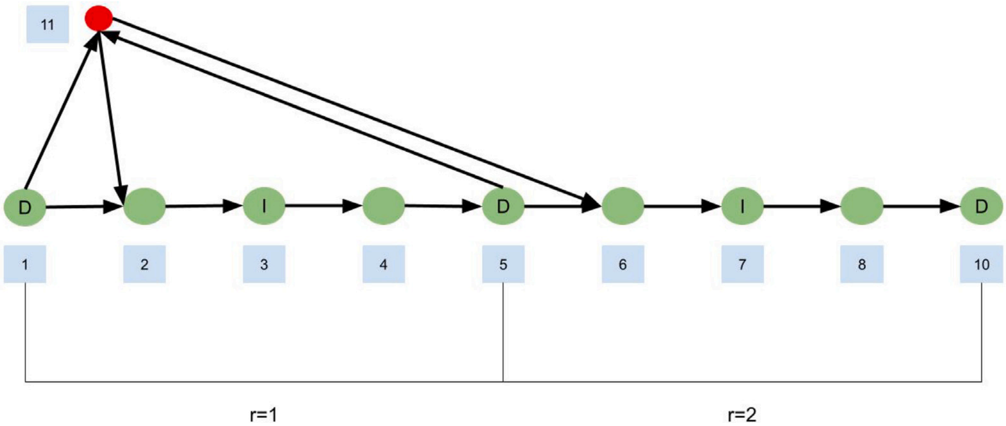

Fig. 4. Example of an FRT graph for a simple instance.

2.1. FRT graph

Even though the FRT problem may seem quite close to the (multi-trip) VRP, e.g., in the form of the mathematical programming models developed to solve them, the peculiarities of the FRT imply that the transport graph over which the problem is solved has a very specific, non-trivial structure. As some of the operational constraints of the problem are encoded in the graph structure and do not appear as such in the constraints of the mathematical programming models present in the literature (e.g., the no-backtracking constraint), we propose a brief discussion of how the transport graph is built on a very simple FRT instance.

Consider an FRT system with one shuttle performing two back-and-forth trips with a no-backtracking policy. The base route has three different physical checkpoints, including the base depot (D), the intermediate depot (I) where the shuttle turns back in the middle of its trip, and an intermediate checkpoint. Since the shuttle performs several (back-and-forth) trips, it will stop several times at the same physical checkpoint. Therefore, such physical checkpoints will be mapped to several nodes in the transport graph. These nodes are displayed in green in Fig. 4 and the two trips (index $r = 1, 2$) with a total of 10 checkpoint stops. As one can see, nodes 2, 4, 6 and 8 all correspond to the same physical checkpoint. A single non-checkpoint stop is considered and displayed in red (node number 11). Since the shuttle has to proceed forward at all times along the longitudinal axis, many arcs in the graph are forbidden. Given the characteristic of the demand associated to node 11, it can only be served between the starting depot D and the intermediate checkpoint. As a consequence, the only arcs permitted are those between two consecutive checkpoint stops (as these stops are mandatory and correspond to the fixed timetable), between nodes 1 and 11 as well as 11 and 2 (used if the non-checkpoint stop is inserted in the first trip $r = 1$), and between nodes 5 and 11 as well as 11 and 6 (used if instead the non-checkpoint stop is inserted in the second trip $r = 2$).

3. Literature review

In this section, we propose a survey of the literature on the FRT. Two categories of approaches are used to deal with the FRT in the literature. The first category addresses the problem using either exact or heuristic optimization algorithms to plan shuttles routing and scheduling for a given instance. The second category uses analytic equations to study the sensitivity of the system to different instance features. We also briefly touch upon an investigation of the possible customers' preferences and expectations. We close this section with a short discussion on problems related to the FRT. On the one hand, we propose a survey of selected contributions on other demand-responsive systems which provide interesting approaches to tackle a more strategic aspects of line planning for hybrid public transport systems. On the other hand, we highlight similarities and differences of the FRT and some classic combinatorial optimization problems, for which several solution algorithms have been presented in the literature.

3.1. Optimization algorithms

The optimization approaches to the FRT have been used both for handling a static version where all customers are known in advance, and dynamic policies where customers can request a ride after the shuttle started its service. Many of the studies surveyed below provide some measure of comparison between the results of an FRT system and another type of transport system such as the CPT.

3.1.1. Static FRT

We start this literature review on optimization algorithms for the static version of the FRT where all the demand is known in advance. The contributions mainly use exact approaches, in the form of Mixed Integer Linear Programming (MILP) models solved by a Branch-and-Cut solver, though researchers have started developing meta-heuristic algorithms to deal with the most difficult variants. We first review the contributions that establish increasingly complex MILP models for the base version of the FRT, as described in the previous section. The rest of this section is devoted to more recent extensions of FRT with additional characteristics, such as the existence of meeting points or the availability of advanced modular vehicles.

The earliest modeling contribution for FRT systems is presented in Quadrifoglio and Dessouky (2004). A MILP formulation is proposed for the restricted case with only one trip. It is solved through a commercial solver and finds solutions for instances with 25 customers per hour. The paper proposes a comparison between FRT and CPT based on the weighted sum of the following four criteria: 1. average ride time per passenger ; 2. average waiting time over pick-up at ad hoc stops ; 3. total shuttle travel time ; 4. average walking time per passenger . Waiting time is relevant only for the FRT, while walking time only for the CPT: passengers are picked-up at their desired location in the former problem, possibly after waiting some time, while they have to walk to their nearest checkpoint in the latter. The authors assume that waiting time at checkpoints weighs twice as much as riding time on the shuttle, while the total travel time of the shuttle and the passengers' ride time are equally weighted. Finally, the walking time is equivalent to the waiting time at checkpoints. In the experiments presented, the FRT outperforms the CPT according to this performance measure.

The MILP model of Quadrifoglio and Dessouky (2004) is further developed by the same research group in Quadrifoglio et al. (2008b) and Quadrifoglio et al. (2008a). The previous MILP is extended to a situation with multiple trips to perform. The peculiarity of the multi-trip problem is that non-PD customers can be handled in different trips according to the workload of the shuttle. For hybrid customers (PND or NPD), this requires a peculiar definition of the transport graph, as the checkpoint node associated to pick-up or drop-off cannot be set *a priori*. An additional family of variables needs to be added to decide in which trip to insert such customers. Three groups of valid inequalities are also identified based on the analysis of optimal solutions. All three groups are based on the following observations: 1. NPD customers disembark at the first occurrence of their drop-off checkpoint following their pick-up; 2. If the weight associated to ride time is larger than the one of waiting time in the objective function, PND customers board the shuttle at the last occurrence of their pick-up checkpoint prior to their drop-off. The model is tested on various sets of experiments, with up to 17 customers per hour and a time horizon of 10 h. The results show that the first group of inequalities is the most effective one, as it is the only one to consistently reduce the gap of the linear relaxation. Even though those inequalities improve the linear relaxation greatly and the computational time to solve some instances can be reduced by up to 90%, some of the largest instances still cannot be solved in less than 10 h of computational time.

A further extension of this model is investigated in Lu et al. (2011b), which introduces the use of two shuttles to perform the set of trips. The purpose of the paper is to understand when it is advantageous to switch from a single (1-FRT) to a double (2-FRT) shuttle configuration. The MILP formulation is based on the one of Quadrifoglio et al. (2008a). The model is tested to compare the performance of 2-FRT and 1-FRT, considering between 8 and 20 customers per hour. In particular, the paper aims to identify the critical number of customers which makes the second shuttle beneficial. To do so, it introduces an analytical model which assesses the expected value of the MILP objective function considering a uniform probability distribution for ad hoc stop locations over the service area. While the analytical model advocates for an increase in the number of vehicles starting from 12 customers per hour, the MILP formulation settles for 14 customers per hour. Finally, a sensitivity analysis is performed over the objective function weight of the total shuttle travel time. The purpose is to see how the critical number of customers is influenced by different values of this weight. It is found that if this weight increases, the number of critical customers also increases.

A case study in suburban Toronto is investigated in Alshalalfah and Shalaby (2012) to see what would be the impact of transforming three CPT lines to FRT. The study is based on the 'REFLEX' simulation software that exploits constraint programming to schedule the customers. The problem heuristically solved here cumulates the features of the above mentioned versions of the FRT (multi-trip and multiple shuttles), and includes the possibility to reject requests. The following characteristics of the CPT lines are varied to design the appropriate FRT systems: number of fixed-stops (checkpoints), route length, average checkpoint spacing, number of shuttles, slack time and number of passengers. The objective is to maximize the number of accepted requests. It is found that by increasing the slack time, the idle time will likewise increase and more requests are accepted. Also, increasing checkpoint spacing will increase accepted requests, as shuttles have more time for deviations.

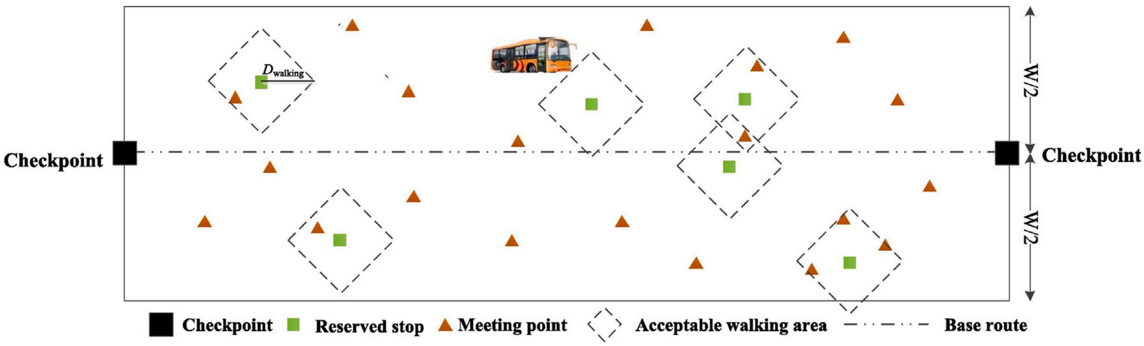

Fig. 5. Example of the association of requests with meeting points according to the acceptable walking distance of each customer.
Source: Reproduced from Zheng et al. (2019).

After a break of a few years, more refined and sophisticated versions of the base FRT have recently been investigated, demonstrating a renewed interest in studying the performances of an FRT system by exploiting new technical advances or alternative ways to handle the demand. In Zhang et al. (2021), the authors modify the problem of Quadrifoglio and Dessouky (2004) by adding time windows constraints in the form of a maximum pick-up time for customers, and a capacity constraint for shuttles. The scheduling of customers is done heuristically through a First-Come First-Served basis. The authors compare the FRT system with a CPT in which people in the service area use shared biking system to get to the closest checkpoint to be picked-up. A cost function similar to the one of Quadrifoglio and Dessouky (2004) is defined, except that bike riding time is taken into account rather than waiting time of customers. It is found that both systems have almost the same cost function value. However, the FRT is perceived to be more comfortable. A sensitivity analysis is performed to understand the influence of demand density, number of checkpoints and their spacing, and travel preference of customers among the two systems. It is found that for a low demand area and where checkpoint spacing is large, the FRT performs better.

In Zheng et al. (2019), some meeting points are available within the service area, where customers can be picked-up or dropped-off if such meeting points are within their acceptable walking distance. This strategy has already been proposed for other types of demand-responsive systems, see, e.g., Daganzo (1984a). Fig. 5, reproduced from Zheng et al. (2019), illustrates this situation where demand points are displayed as green squares and meeting points as orange triangles: the acceptable walking area around the squares is displayed as a dashed line so that one can see which meeting points are acceptable for each customer. A MILP formulation similar to the one of Quadrifoglio and Dessouky (2004) is employed but the waiting time is not considered in the objective function. It is instead replaced by a walking time from the requested non-checkpoint stops and the meeting points chosen instead. Moreover, additional binary variables are introduced to indicate if a node is included in the optimal tour (where the node set now includes the possible meeting points) since not all nodes need to be visited depending on whether we choose to use some meeting points or not. Since integer linear solvers cannot solve large-scale instances, a Memetic algorithm is designed. In this algorithm, the fitness value of feasible solutions is equal to the MILP objective function. If the solution obtained violates the scheduled departure time constraint at checkpoints, it pays an additional penalty. Various scenarios are tested for traditional FRT and FRT with meeting points. The results show that for low demand levels (i.e., 6 customers/trip), employing a meeting point strategy is not very influential. However, for higher demands (more than 12 customers/trip) the rejection rate declines by up to 20% if a meeting point strategy is employed. A further analysis over fares is conducted. Typically, non-PD customers have to pay an additional fare. However, it is assumed that if a non-PD customer is assigned to a meeting point, they pay as much as PD customers in return of added walking distance to/from meeting points. When the demand increases, more customers can be served using a meeting point strategy and transport authorities obtain higher revenues compared to traditional FRT. A sensitivity analysis is performed over the number of meeting points and acceptable walking distance of each customers. By increasing each term, the rejection rate and idle time decrease. The most recent contribution on the front of meeting point strategies is the one of Li and Tang (2023), where the authors propose a more strategic approach to decide the location of meeting points in the service area, based on different demand scenarios. They employ a simulation-based optimization method and aim to improve the performance of FRT systems under unexpectedly high-demand levels. The optimization methodology is based on a Multiple Meta-model-based Efficient Global Optimization (MMEGO) algorithm embedded in a Monte Carlo simulation (MCS) framework. The MMEGO algorithm can reduce the system cost from 3.5% to 26.6%, compared with competing algorithms. However, the walking time cost of the FRT system with optimized meeting points is higher than that of the FRT system without meeting points by up to 7.3% at slightly high demand levels.

The contribution of Liu et al. (2021) explores the possibility of using Modular Autonomous Vehicles (MAVs) which may reduce the operational costs with respect to Traditional Vehicles (TVs). MAVs are small vehicles which can be assembled into larger unique vehicles at checkpoints and can split between any consecutive checkpoints to serve more efficiently non-PD clients. A visual example of such vehicles is provided in Fig. 6. In order to illustrate how the use of MAVs allows more sophisticated ways to handle non-PD customers, we reproduce a figure from Liu et al. (2021) (Fig. 7) which shows different trips with the routes of the different MAV clusters. The trip starting at 9:40 shows one MAV decoupling from the two others between checkpoints 2–3 and 3–4 to handle non-PD stops. A MILP formulation is developed to address the problem. Given the small size of MAVs, it is necessary to introduce

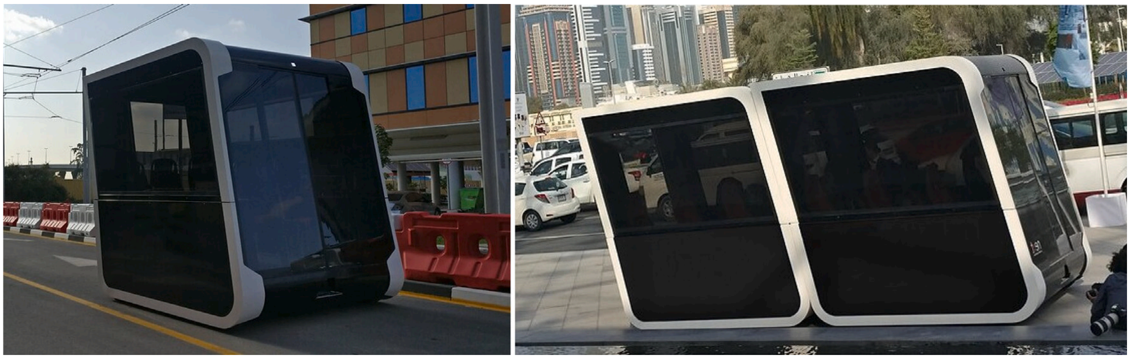

Fig. 6. Example of a MAV (on the left) and how they can combine into a single vehicle (on the right).

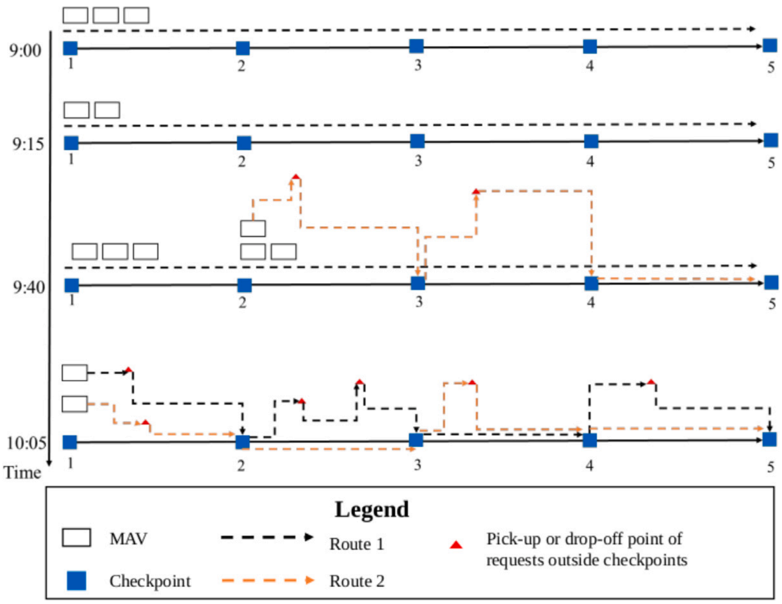

Fig. 7. Example of four trips using MAVs.
Source: Reproduced from Liu et al. (2021).

an MAV capacity with an associated capacity constraint. The possibility to discard a customer's request is also taken into account and the total rejection cost of such rejected requests makes up an additional term in the objective function. The formulation opts for a *time-indexed* approach and introduces a set of departure times T. Overall, the MILP contains a very large number of variables and constraints (many of which are big-M constraints) and is unable to solve realistic instances. A two-stage solution framework is therefore developed to decompose said MILP formulation. In the first stage, a Customized Dynamic Programming (CDP) with valid cuts is designed for route scheduling. In the second stage, a fast heuristic is employed to solve the assignment of customers to each MAV. For a system with 10 to 30 MAVs and 15 to 45 non-PD customers per trip, the results of the MILP and the heuristic are compared. Even though the CDP computation time is substantially lower, its results are very similar to the ones of the MILP (solved by Gurobi). These results suggest that using MAVs would reduce operational costs.

We close this section with the only available work on the FRT which tackles the problem from a more strategic perspective using optimization tools: Yang et al. (2016) aim at selecting the routes of FRT lines in a specific region based on an anticipated demand. The objective is to capture the largest possible demand per mile of travel, to build cost-effective routes (an objective highly valued by government agencies and service providers (Potts et al., 2010)), while still serving the main urban centers in the region. The main roads are divided in segments of equal length and the FRT base routes are built starting from the main urban centers and extending them greedily until all the necessary places are covered by the lines. The most economical path is determined and shown on a GIS map. A case study demonstrates the approach in numerous urban areas of Tennessee. Depending on the situation, routes may be more cost-effective depending on their length.

3.1.2. Dynamic FRT

The remaining optimization approaches for the FRT tackle the dynamic version of the problem, where not all the requests are known before the shuttle departure. These approaches mainly use heuristic algorithms. When rejection is considered, it is intended that both the pick-up and drop-off stops of a customer are inserted or rejected together in the shuttle schedule.

The first work considering the dynamic FRT is the one of Quadrifoglio et al. (2007). It presents a heuristic algorithm which deals with the single shuttle multi-trip problem, excludes shuttle capacity and request rejection, but introduces a backtracking policy: shuttles can move backwards to collect additional customers. In this algorithm, it is assumed that a schedule has already been devised to handle prior requests. When a new request is received, the algorithm chooses the best place to insert the customer's stops in the already planned schedule, among the feasible insertions. Feasibility is checked with respect to a backtracking threshold and the available slack time between two consecutive checkpoints. The purpose of the paper is to understand the impact of the usable slack time and backtracking threshold over the saturation level, defined as the maximum number of requests that a system can serve without becoming unstable. Different scenarios are tested to find the saturation level for the requests. Then, the increment in the objective function is computed based on the extra distance driven by the shuttle and the extra riding and waiting times of previously and newly inserted customers. A rejection of the request only occurs if the system is saturated or the request is issued too close to the end of the service (i.e. there is no feasible insertion at all). The algorithm is tested on up to 25 customers per hour and over 50 h of time horizon. Various parameter configurations are considered as for backtracking threshold and available slack time. The results of the heuristic are compared to those of the optimal solution of a MILP in a static environment, where all requests are known in advance, and with a First-Come/First-Serve (FCFS) policy. This policy assumes that no extra waiting or riding time can be imposed to already scheduled customers by newly inserted ones. The results show that the heuristic schedule is not far from the static optimal one, while the system with the FCFS policy is saturated very quickly.

The insertion heuristic of Quadrifoglio et al. (2007) has later been used by several works to investigate the behavior of a dynamic policy FRT. In particular, Quadrifoglio and Dessouky (2008) perform a sensitivity analysis and assess the system performance over different length and width of a service area with constant size (12 square miles). The paper considers instances with a time horizon of 45 h (corresponding to 54 shuttle trips), between 10 and 20 customers per hour and a single shuttle performing several trips. The results show a better performance with a slimmer service area, as the system uses lower amounts of slack and ride times for on-board customers. The impact of widening the service area on the shuttle longitudinal speed is quantified. Indeed, the wider the area, the larger the amount of time needed to serve customers at ad hoc stops. The results achieved in the narrowest service area outperform those of a CPT, based on the same criteria as in Quadrifoglio and Dessouky (2004). The authors of Lu et al. (2011a) later extend the insertion heuristic algorithm from Quadrifoglio et al. (2007), in order to understand if a 1-FRT system outperforms a 2-FRT systems in a dynamic environment, similar in spirit to the comparison of Lu et al. (2011b) for the static environment. Three tests are performed with respectively 15, 20, and 25 customers per hour, over a time horizon of 50 h. The system parameters are the same as in the analysis of Lu et al. (2011b). In all of the tests, the 2-FRT outperforms the 1-FRT. It is also found that 2-FRT reduces the waiting time for customers by half compared to 1-FRT. 2-FRT is also shown to achieve performances comparable to those of the static MILP. The same type of insertion heuristic is used together with the real-time traffic information available on Google Maps in Qiu et al. (2014a) to re-schedule shuttle routes if the planned one is blocked or congested. The results are compared with the simpler scenario omitting real-time traffic information. Little difference is detected during off-peak hours. Instead, during peak hours, the service quality improves up to 11.7% when routes are re-scheduled, and the results indicate that non-PD customers are more likely to be rejected by deviation services.

Following the meeting point philosophy of Zheng et al. (2019), the authors of Zhang et al. (2023) aim to enhance the operational service capability of the FRT by proposing a dynamic clustering meeting point strategy which sets dynamic pickup and drop-off meeting points based on real-time travel demand. The employed methodology involves analyzing the impact of checkpoint spacing on the performance of the dynamic meeting point strategy, for example to help reduce the rejection rate under different demand levels. After proposing a mathematical programming model, the authors propose a memetic algorithm to solve the problem on realistic instances. The paper concludes that the proposed dynamic clustering meeting point strategy can significantly reduce the rejection rate of customers' requests without increasing operational costs.

A specific line of research investigates the case where some requests are known in advance, while others are inserted later as the shuttle moves around. Here, Zheng and Li (2019) explore the operational efficiency of a static, partially and fully dynamic environment for the FRT. A MILP formulation addresses the initial requests that are known in advance (static). Later, an insertion heuristic is employed to insert dynamic requests on the fly. The dynamic requests are inserted if shuttle capacity constraints are not violated and enough slack time remains. The ratio of the number of dynamic requests to the total number of requests is named Degree of Dynamism (DOD). The DOD varies from DOD=0% (pure static) to DOD=100% (pure dynamic). It is found that riding time increases as the DOD increases. Beyond the value of 75%, the riding time stabilizes, while the waiting time keeps increasing. In addition, if all the requests are known in advance, the system performance increases by 2.6%. This study is later developed by Zheng et al. (2021) considering possible request cancellations and customer no-shows. The no-show and cancellation rates vary from 0% to 20% with a random distribution. As they increase, the riding time declines, and the idle time increases. However, as the in-vehicle time is the sum of riding and idle times, it remains stable. Another analysis which handles both static and dynamic demand is presented in Li et al. (2022). In this work, though, the base route of the shuttle is not fixed and the order in which the checkpoints are visited can be changed to accommodate better the non-checkpoint demands. The route choice including the demands made prior to the shuttle departure is done through a genetic algorithm, while the dynamic requests are inserted one by one with a greedy logic. Numerical studies establish the applicability of FRT in a realistic road network and indicate that it can handle demand more efficiently by saving costs by 40% compared to CPT. Finally, Sun and Liu (2022) address the same type of problem where the known requests for the following day are handled through the use of a (multi-shuttle) MILP, based on the one of Lu et al. (2011b). The dynamic requests are handled through an insertion heuristic very similar to the approach of Quadrifoglio et al. (2007), with the noticeable difference that the authors consider the slack time as a (variable) control parameter, as opposed to a fixed quantity. Under the same demand circumstances, the FRT system is 5.9% to 10.8% less expensive per customer than CPT. When the demand for the five buses in Harbin's suburbs (China) is 20 to 40 customers per hour, the FRT system is more effective than the CPT one.

3.2. Analytical equations

In this section, we review the articles that address the sensitivity of the FRT to instance characteristics using analytical equations. In order to derive closed form formulae or equations, a uniform distribution in space is considered for demand generation for all papers in this section and all but one paper consider a uniform distribution in time. Moreover, many contributions directly consider the expected values of various service characteristics, such as, e.g., customer waiting and riding times. These approaches are sometimes designed to tackle the problem at the tactical and/or strategic level. The strategic level usually determines the physical characteristics of the system given a certain level of demand, such as the line characteristics (depot and checkpoint stop locations, shape and size of the area) or the vehicle characteristics (fleet size, capacity/autonomy of the vehicles). These parameters are generally costly and difficult to modify in later stages. The tactical level instead determines the service characteristics, such as slack time at checkpoints, trip frequency or backtracking threshold.[1] Other contributions study the average efficiency of different transport modalities to determine which one is the best suited under specific circumstances.

An early contribution is the one of Fu (2002), in which an analytical model is proposed for an idealized operating environment, with the objective of determining the optimal slack time that should be allocated to a flex-route segment. An equation is derived for the relationship between the number of feasible deviations and various system parameters such as slack time, zone size, and dwell time. The results illustrate the impact of the different system parameters on the system output.

Focusing on a narrower scope, in Quadrifoglio et al. (2006), the authors devise sophisticated methods based on a continuous approximation to compute lower and upper bounds on the maximum longitudinal speed of the shuttle given a number of customers to be served. The longitudinal speed of the shuttle is a crucial parameter for ensuring a sufficient level of service, as it conditions the riding time of customers. Here, the base version of the FRT by Quadrifoglio and Dessouky (2004) is considered. The paper aims to identify the maximum number of requests and size of service area to avoid saturation while maintaining an average longitudinal speed above a fixed threshold. One lower bound and two upper bounds are computed for the average longitudinal speed, using a uniform request distribution over space and time. The lower bound is computed based on the average distance traveled by the shuttle with a no backtracking policy. The first upper bound is obtained by considering a subset of the customers, such that each pair of customers is separated by a minimum longitudinal distance. The distance traveled by the shuttle can then be computed exactly as for the Traveling Salesman Problem (TSP) derived by Daganzo (1984b). The second upper bound is obtained by relaxing the classic constraint of having a unique incoming arc for each node, while still requiring a single outgoing arc. Two instances are considered with a service area width of respectively 0.5 and 1 mile. According to the derived lower and upper bounds, in the first instance a shuttle can serve between 90 and 130 stops/h while in the second between 70 and 120. The bounds are somewhat close, and the authors conclude that the system capacity is not heavily affected when the service area width is doubled. A logic similar to the one of Quadrifoglio et al. (2006) is used by Zhao and Dessouky (2008) by computing approximations for the mean and variance of the shuttle travel time. This work tackles an FRT system with a dynamic customer insertion policy, as in Section 3.1.2, with no backtracking policy. Requests are uniformly distributed in space and time. Moreover, departure times from checkpoints and terminals are no more hard constraints. The aim is to deliver the service to customers as soon as possible, accepting a possible delay for the following checkpoint departure times. Therefore, when the shuttle is late by a few minutes, the requests arriving dynamically tend to accumulate, which further increases delays. Three sets of instances are tested with different values for the scheduled departure times and the service area width and length. The authors define the *level of service* as the probability of arriving on time at checkpoints. The paper studies how the actual departure times and the service level depend on the instance characteristics, and when the system tends to become unstable (accumulating delays).

As for the works of Section 3.1, some researchers compare the FRT with other transport systems, such as CPT or other on-demand systems. Alshalalfah and Shalaby (2010) propose a model to compute the appropriate amount of slack time in order to handle a certain level of demand and study the impact of the service area width and the available slack time on the percentage of on-demand requests that are accommodated by the flex-route service. It also provides approximate formulae to estimate the costs and benefits for the customers and the system operator of changing from a CPT to an FRT system, depending on system parameters such as the available slack time or the number of fixed stops. In Zheng et al. (2018a), the authors investigate two systems similar to the FRT, calling them systems A and B. The authors consider average values for different quantities, such as speed and walking time, in order to assess the impact of both systems. A single shuttle is considered, and the service area has checkpoints at the two extremities and no intermediate ones. In system A, a shuttle that deviates from the base route to serve non-PD customers must then go back and continue the trip from the point where it departed. The non-PD customers either walk to the base route and are picked-up from there (flag requests, which cannot be rejected) or ask for a door-to-door request. Door-to-door requests are asked to shift to flag requests

[1] For FRT systems, it is however debatable whether the service area for on-demand stops is part of the strategic or tactical stage since it can be changed somewhat easily at any time and does not imply any change in the physical infrastructure.

if their desired location cannot be inserted in shuttle routes due to a lack of slack time. System B is much closer to a traditional FRT system, where customers whose location cannot be inserted into shuttle routes must walk to the closest checkpoint. Several tests are performed for different demand levels. It is found that system B performs better for low demand levels (lower than 38 customers/trip), whereas system A is better for medium levels of demand.

A few contributions explore the possibility to use the FRT as a feeder transit system. A feeder transit system is a transport system which aims at transporting customers to the station of a more important transport line. The study of Qiu et al. (2015) examines the feasibility of replacing the fixed-route policy with an FRT in feeder transit systems without disrupting the coordination between the main transit and feeder services. The authors determine an upper bound on demand for applying the FRT in this feeder service, enabling planners to make better decisions when planning the FRT in response to a fluctuating customer demand. The performance metric used combines vehicle operating cost and a transit customer cost. The findings show that the designed FRT system is still expected to have a noticeable system advantage over the fixed-route service in operating situations with occasional request rejections. This suggests that a good flex-route feeder system might operate with some request rejection tolerance. In Zheng et al. (2018b), some checkpoints might act as transfer points, where customers continue their trip with another mode of transport. The authors propose a slack arrival strategy to reduce the number of rejected requests and idle time at checkpoints. This strategy relaxes the scheduled departure time constraint of each checkpoint. Customers whose desired location cannot be inserted into the schedule, either walk to the nearest checkpoint to be picked-up or choose another mode of transport. By running a large number of generated scenarios, the authors derive expressions for various parts of the objective function using expected values of customers walking, waiting and riding times. An insertion heuristic based on a FCFS policy is used to construct the vehicle schedule. For instance, if the departure time of a checkpoint is violated, the idle time of the following checkpoint is shortened accordingly. However, violating the departure time of the transfer checkpoints leads to a severe inconvenience for customers. Transfer checkpoints are placed in different positions of the service area (i.e. intermediate or terminal) in different scenarios. When the transfer points are located at intermediate checkpoints, only a small improvement is seen by implementing a slack time strategy for the system. It is also found that the system performance can improve by up to 40% if the slack time strategy is employed when transfer points are located at terminals.

As in Section 3.1, some contributions consider further refinements of the FRT systems, such as the possibility to cluster customers, similar in spirit to Zheng et al. (2019). In Qiu et al. (2014b), the authors introduce a Dynamic Station (DS) strategy where non-PD customers can get to other customers' locations (instead of walking to/from checkpoints or waiting for the next trips) if their desired pick-up location cannot be inserted in the schedule. The authors use expected values for the waiting, walking and riding times and use the same type of approximation as in Zheng et al. (2018b). Three systems are compared: CPT, FRT and FRT-DS. The FRT-DS is proven to operate better than the others in cutting down on walking time. In particular, in the scenarios tested, the walking time decreases by 50% for high demand levels (60 customers/h). Additionally, the FRT outperforms the CPT for small to medium scale instances, whereas the opposite holds for large ones. The contribution of Sun et al. (2018) uses the same type of approximation and investigates the possibility of clustering customers, which is called *optimal FRT system*. Such system is compared with the traditional FRT, based on a utility function which combines the total walking time, riding time, idle time and rejection cost of customers. In the traditional FRT, when the remaining available slack time is insufficient, passengers cannot be picked-up by the first passing shuttle and must be assigned to a later trip. Therefore, for high demand levels, the overall waiting time increases. Instead, in the optimal FRT, customers can be clustered together at one of the customers' locations to avoid large waiting times. Priority is usually given to the elderly, so they do not have to walk. This policy increases the walking time but reduces the waiting time. For low demand levels, the two systems perform similarly. For high demand levels, however, the FRT with optimized clustering performs better as the waiting time is reduced.

A couple of contributions focus on the impact of the fare policy in the single-shuttle multi-trip FRT of Quadrifoglio et al. (2008a). The authors of Shen et al. (2017) compare two different fare policies. The first one assigns a different fee to each customer depending on the quality of service provided (based on expected walking, waiting and riding times), while the second one assigns a fixed fee to each type of customer (PD and non-PD). In the second case, PD customers get lower rates for enforced delay, whilst non-PD customers pay more for personal deviation services. The authors consider average values for different quantities such as speed and walking time in order to assess the impact of both policies. They recommend a unique service-based fee structure when constructing a suitable price structure to transform a fixed-route transport to an FRT. Another contribution using a similar methodology for computing average speed and walking times is proposed by Shen et al. (2019) to assist transport authorities in determining FRT fares for various service area shapes (i.e., L/W ratio). Two scenarios where rejected customers respectively either switch to another mode of transport, or simply walk to the nearest checkpoint. The higher the service quality, the higher the fare. It is found that the resulting fares for the two systems are nearly identical for low demand levels. Instead, for high demand levels, the fare is higher when rejected customers opt out of the system.

We close this section with the more strategic approach developed in Sipetas and Gonzales (2021), where the authors aim to determine the optimal checkpoints spacing (S) and width of the service area (A). These quantities are approximated as continuous functions of the distance from the terminal on the longitudinal axis. The two-shuttle FRT variant introduced by Lu et al. (2011a) is considered here, with the addition of the use of a backtracking policy. A formula is defined for different components of the objective function. Based on these components, S and A are computed for different values of x. Three systems are compared, namely: a CPT, an FRT, and a fully flexible system. To measure the performance of each system, operating costs and user costs are measured. It is found that the CPT and the fully flexible systems have the lowest and highest operating costs, respectively. Lower demand levels and smaller service areas imply better user costs for the FRT compared to the CPT. The results show that the FRT can reduce user costs by 80% compared to the fully flexible system, and by up to 35% compared to the CPT under different scenarios. Then, a sensitivity analysis is performed over a single trip travel time, percentage of non-PD demand served and relative weight values of user and operating costs. The results show that greater headway leads to lower flexibility, shorter stop spacing, and higher user costs for non-PD customers.

3.3. Demand studies

Only a low number of contributions attempt to study the possible customers' response to a newly established FRT system, by gathering data from respondents in a specific area. The first such study, conducted by Zheng et al. (2020) in the city of Nanjing, China, aims to evaluate the potential customers' service design preferences. The authors distributed questionnaires to put in comparison a conventional fixed-route transport, private cars, and an hypothetical FRT. Walking time, waiting time, in-vehicle time, and cost are picked as alternate features that vary in each scenario. The findings of the study indicate that around 78% of the 630 respondents are eager to experience the FRT service. The target demographics for this service include women, bike-sharing users, handicapped and elderly people, and those who need to transfer to the metro. It is also found that the fare should not exceed the one applied for most bus services in the same city in order to attract many customers (42.5% of the demand). The majority of the respondents are in favor of mobile application-based travel booking and payment. The study of the impact of socio-demographic and psychological latent factors affecting FRT acceptance is conducted in the more recent studies of Yu et al. (2023a,b). The first study focuses on low-demand areas around the city of Nanjing, China, and shows that socio-demographic factors such as age, income, and education level all have a significant impact on FRT acceptance. In addition, psychological factors such as comfort, flexibility, perceived barriers, personal barriers, subjective evaluation, and use willingness are also important predictors of FRT acceptance. The second study instead focuses on the area of Beijing, China, and categorizes respondents into five ordered stages: Pre-contemplation, Contemplation, Preparation, Action, and Maintenance. In the analysis of the survey data, the influence of psychological factors is found to be more significant than the one of demographic characteristics.

3.4. Related hybrid transport problems

In order to provide relevant insights on the state of the art for FRT systems, it is useful to survey a few contributions on related problems for which optimization approaches exist for tackling the strategic and tactical level. The articles surveyed in Section 3.2 show that analytical equations have been used to study some tactical aspects (slack time at checkpoints) in relation with some strategic aspects (width of the service area). Other strategic aspects such as the definition of the checkpoints have not been addressed to our knowledge (as already stressed in Errico et al., 2013) and no optimization approach for either the tactical or strategic level has been proposed.

This is not true for any Demand Adaptive System though. The contributions we survey in this section tackle a system where compulsory stops are part of the core route of the shuttle, just as in the FRT. Here, customers cannot request a stop at the location of their choice, but may request the use of a pre-determined optional stop. An earlier contribution on the choice of the optional stops to make available to customers is the one of Pratelli and Schoen (2001). The average expected values of demand are used at each possible node, and some assumptions are made on the behavior of customers if their preferred optional stop is not selected. Passenger circulation is modeled through flows, and the number of optional stops between two compulsory ones is usually limited to 1. Later, Crainic et al. (2012) concentrate on what is called the *master schedule*, i.e. the set of time windows for the departure times at the n fixed stops. A set R of possible origin–destination trips is used, based on the fixed and optional stops, with an associated (independent) probability p_r for $r \in R$. Each fixed stop h is associated to a time window $[a_h, b_h]$ and the aim is to minimize the last upper time window value b_n, such that the probability of serving each possible request is greater than a threshold ε. This is done by creating a sampling of request sets using the probabilities p_r in order to compute probabilities of optional stops to be involved in a request and cumulative distribution functions for the arrival time at fixed stops.

A greater step is taken in the direction of designing optimization methods for tactical and strategic aspects of DAS in Errico et al. (2021). The problem is the same, but now the authors also aim to decide where to locate the compulsory stops. A set of stops is supposed to have been pre-selected and only some of them can be chosen as compulsory. A specific time slice of the operational horizon is considered, where the probability for the demand between each pair of stops is uniform in time. The problem then consists in selecting the compulsory stops and their sequence, assigning the optional stops to a *segment*, i.e. a pair of consecutive compulsory stops, as well as the master schedule as defined earlier. The mathematical model includes binary variables for selecting the compulsory stops and their schedule as well as the assignment of optional stops to segments, which obey classic assignment constraints in the routing literature. Moreover, it includes time variables to fix the time windows of the master schedule and to estimate the travel time between stops. The related constraints include non-linear expressions with complicated convolution operators and quantile functions. The problem form makes it practically intractable, and the authors have to resort to a specific hierarchical decomposition heuristic. It first determines the set of compulsory stops, then their sequence and finally the master schedule. While the last phase can be solved using the approach of Crainic et al. (2012), the second phase is akin to a traveling salesman problem with generalized latency. This problem has been tackled by the same authors in Errico et al. (2017) through a branch-and-cut approach based on a Benders reformulation with valid inequalities.

We close this section by emphasizing that even though the above tactical approaches are quite sophisticated and provide a good starting point to design optimization approaches for tactical aspects of the FRT, the absence of a limited set of optional stops requires non-trivial adaptations and specific developments or approximations.

3.5. Related classic transport problems

Most FRT variants tackled in the literature show evident similarities with other well studied logistics and transport problems. The main such problem is the VRP. For this problem, and many of its variants, a large body of research exists, and efficient algorithms have been proposed. We refer to the surveys of Mor and Speranza (2022), Kumar and Panneerselvam (2012) and Laporte (1992) for a general overview of the literature on the VRP.

One of the well-known sub-categories of VRP is the Vehicle Routing Problem with Pick-up and Delivery and Time Windows (VRPPDTW). This specific version of the VRP has common features with the FRT. In a VRPPDTW, a customer is asking for a service which includes the pick-up and delivery of some merchandise within a specific time frame. Several variants of the VRPPDTW exist in the literature. The version of the VRPPDTW we are interested in is usually labeled as the one-to-one VRPPDTW. This variant has a single origin terminal and a single destination terminal. The shuttles must run between these two terminals for transporting goods. Each request consists of transporting a load from one pickup vertex to one destination vertex in a graph (Battarra et al. 2014). The VRPPDTW naturally integrates situations of reverse logistics, as companies become interested in gaining more control over the whole life-cycle of their products. Examples of reverse logistics are, e.g. the soft drink industry, where empty bottles must be returned, or the delivery to grocery stores, where reusable pallets/containers are used for the transport of merchandise. The VRPPDTW also has applications in various other contexts such as the management of returned goods, urban courier services or less-than-truckload transport. An additional important application of VRPPDTW is the door to door transport service for the elderly and the disabled. In this case, narrow time windows are often considered and ride time constraints are imposed to control the time spent by a passenger in the shuttle. This specific variant of VRPPDTW is called DARP.

In the DARP, users formulate requests from a specific origin to a specific destination. Transport is carried out by shuttles that provide a shared service in the sense that several users may be in a shuttle at the same time (ride sharing). The aim is to design a minimum cost set of shuttle routes accommodating all requests under a number of side constraints. The most frequent objective of the DARP is to minimize the operating costs and user inconvenience. Operating costs are mostly related to the fleet size and distance traveled, while user inconvenience is often measured in terms of deviations from desired pick-up/drop-off times and in terms of excess riding time.

As discussed in Section 3.1, the FRT in the literature is often defined as a single shuttle problem, however it is a multi-trip problem and can naturally be compared with the multi-vehicle VRPPDTW/DARP. These systems share strong similarities, i.e. both systems: 1. are routing problems between terminals trying to find the best routes for the shuttles; 2. have to pick-up and deliver commodities or people; 3. need time constraints on both pick-up and delivery operations; 4. aim to minimize the total distance driven by shuttles.

In spite of these similarities, we can however identify several crucial differences which make existing methods for VRP problems not trivial to apply to the FRT. For example, in the VRP: 1. each customer has to be served by one shuttle only, while in the FRT each checkpoint must be served at each trip. 2. there is usually no imposed time for leaving the origin terminal. Instead, in the FRT, shuttles need to leave the terminal at specific times and no two shuttles will leave the terminal at the same time. 3. usually two optimization objectives are considered: i. minimizing the number of shuttles; ii. reducing the transport distance (Shi et al., 2020). However, in the FRT as described in Section 2, the objectives to minimize are: i. the total distance traveled by the shuttle (the same as the VRP); ii. the total waiting time at the pick-up stops; iii. the total ride time of all customers. The second objective is akin to transforming a strict time window constraint into a soft constraint, which changes the structure of feasible solutions. The last objective is implemented as a hard constraint in the DARP, which again will affect the structure of feasible solutions and their respective ranking. 4. logistics companies or their logistics sectors are dealing with transport of goods. However, the FRT is designed to transport people from one point to another. At the same time, in Quadrifoglio and Dessouky (2008), the authors mention that slim service areas in the FRT are appropriate for transporting people, while wider service areas are a good choice for transporting goods. Instances based on goods or people transport can have different features, like the average number of requests at a given node.

4. Summary of case studies and real-life implementations of FRT

As underlined in the previous section, several works in the literature consider case studies based on real bus lines and their associated demand levels and physical structures, or even based on existing Flex-Route Transit systems (with some of them possibly discontinued). The most used case study of an FRT line is the one of feeder-line 646 of the Metropolitan Transit Authority (MTA) of Los Angeles County, USA, active in the early 2000's (see, e.g., Table 1, where 17 contributions base their numerical experiments on this specific line). This real-life FRT system is described as a rectangular service area of length $L = 12$ miles and width $W = 0.5$ mile. The fixed route contains three physical checkpoints (including the two terminals and an intermediate checkpoint) and the service was active 4.5 h each night. In one night, a single shuttle performed 9 trips, handling an average of 4 to 5 customers per hour. Many works in the literature based on the LA case modify some of the real-case features to perform a sensitivity analysis and explore the effect of changing these features on the system performances. For example, the length of the service area can vary between 6 and 16 miles and the width from 0.5 to 2 miles, while the number of trips of the shuttle can vary from 1 to 6. A few works also explore the possibility of a second shuttle. The distribution between the different types of customers is always of 40% of NPD and PND customers and 10% of both PD and NPND customers. This FRT line, however, is now closed in Los Angeles County.

Most other works use real data from standard bus lines to compare CPT FRT systems. Many such contributions still consider a rectangular service area, similar to the LA case study, though the geometry of this area can be quite different depending on the specific case study. For example, Sun and Liu (2022) consider a quite long and straight service area of 27 km length and 1 km width. Interestingly, some works consider case studies with a planned route based on existing bus lines which cannot be approximated as a straight line. In such cases, the service area is always chosen to be a simple rectangle containing the planned route, as in Alshalalfah and Shalaby (2012), Liu et al. (2021) or Li et al. (2022). The number of checkpoints varies, sometimes greatly, depending on the case study: this number can be minimal (3 as in Qiu et al., 2014a), intermediate (7 in Liu et al., 2021, 10 in Sun and Liu, 2022) or larger (up to 21 in Zhang et al., 2021 who study the impact of different numbers of checkpoints). The studies are based on traditional bus lines systematically select a subset of the stops of the CPT line to act as checkpoints. Indeed, an FRT system cannot serve the same number of fixed stops as a CPT one while maintaining a viable amount of slack time between stops. The distribution between the different types of customers is very often the same as in the LA case, with two exceptions: in Sun and Liu (2022) the proportion of PD customers is increases to 20% while the proportion of both NPD and PND customers is decreased to 35%; in Liu et al. (2021), out of 819 customers generated for the case study in Wangjing, China, 801 are generated as PD customers and the rest as non-PD, which incidentally may explain why the MAV-based FRT system is able to serve a much larger temporal density of customers than other contributions on the problem.

Even though feeder-line 646 in LA County is the main real-life example used in the literature, several examples of route-deviation services have been implemented in different places during the last 25 years. The technical report Potts et al. (2010) lists several such systems in the USA, e.g.: South Central Adult Services Council (Nevada); Mason County Transportation Authority (Washington); Jacksonville Transportation Authority (Florida); Potomac and Rappahannock Transportation Commission (Virginia); City of St. Joseph (Montana); and Pierce Transit (Washington). These services operate in areas with very different population density, ranging from 5/sq mile to 1800/sq mile. The number of route-deviation lines for each operator ranges from 1 to 8, with a possible deviation (i.e., half-width in FRT language) of between half a mile to one mile. The passengers needing a non-checkpoint stop always need to book at least one or two hours before their ride in all such systems. A few years back, such systems started using smartphone app-based booking systems to allow customers to book services more easily, such as the Late night bus of Belleville, USA, which allows users to book a trip between any two bus stations of the city area (Zhang et al., 2022).

5. Summary of the literature review and research gaps

Based on the above analysis of the literature, we now discuss research gaps that shall be addressed in order for the FRT to answer a wider spectrum of public transport needs. Indeed, the FRT was originally designed for a night shift (Quadrifoglio and Dessouky, 2004). During such a shift, the number of requests is not expected to be very high. Hence, many constraints and problem features do not really need to be modeled, while they become critical when many customers need to be served.

Table 1 provides a summary of the main aspects characterizing the different papers available on the FRT. Given this summary, we think that FRT extensions shall be studied for addressing the main aspects discussed in the following, for which some approaches may be more or less suitable.

5.1. Finite shuttle capacity

In the literature, only three of the existing papers take into account finite capacity for shuttles, while the majority of papers consider infinite capacity. By having infinite capacity, the PD customers (the ones that have pick-up and drop-off at checkpoints) can be picked-up and dropped-off from the shuttles at any checkpoint they want. They do not actually need to be taken into account when deciding the shuttle routes.

With finite capacity constraints, the way PD customers are treated has a large impact on the system. In the existing literature, they are asked to book their trip like any non-PD customer. However, this may be rather impractical during daytime as those customers may want to use the FRT as a classic public transport system without having to use a booking system in advance. In this case, the routes may be defined considering expected numbers of PD customers and reserving some capacity for the on-demand users. By doing so, we may preserve the nature of the FRT while guaranteeing a minimum level of service to all customers. A further option is designing a stochastic or robust algorithm to deal with uncertain request distributions.

So far, in the literature, each request represents one customer only. In most instances, the number of customers has no impact on the system as any number can fit a shuttle when capacity is not taken into account, except for the few contributions when customer rejection is a possibility. However, in general it is possible to have requests with more than one customer, so that more realistic data instances may be generated for future case studies. This may have a sizeable impact when taking into account finite capacity.

5.2. Number and type of shuttles

In the literature, the number of shuttles in the FRT is mostly either one or two. Only three articles consider more than two shuttles (Zheng and Li, 2019; Alshalalfah and Shalaby, 2012; Liu et al., 2021). Considering a fleet of shuttles may be important to provide a high level of service during a day shift with many requests: with several shuttles, it may be possible to plan for a high frequency of departures and therefore improve the level of service for customers, albeit at a larger cost. A crucial point to clarify on this front is the estimation of the cost of extending the fleet compared to the added value of a better customer service. In particular, it is interesting to study the effect of increasing the capacity of the shuttles versus increasing the frequency of the service (and therefore the size of the fleet).

Table 1
Different aspects characterizing FRT studies in the literature.

References	# of shuttles	Environment	Case study	Approach	Capacity	# of clients	Rejection policy	Backtracking
Quadrifoglio and Dessouky (2004)	1	Static	LA, USA	MILP	×	25/h	×	×
Quadrifoglio et al. (2006)	1	Static	LA, USA	AE	×	40–200 stops/h	×	×
Quadrifoglio et al. (2007)	1	Dynamic	LA, USA	Insertion heuristic	×	15–30/h	×	✓
Quadrifoglio et al. (2008a,b)	1	Static	LA, USA	MILP	×	3–17	×	×
Quadrifoglio and Dessouky (2008)	1	Dynamic	LA, USA	Insertion heuristic	×	10–20/h	×	✓
Zhao and Dessouky (2008)	1	Dynamic	LA, USA	AE	×	–	×	×
Alshalalfah and Shalaby (2010)	1	Static	×	AE	×	2–5/h	✓	×
Lu et al. (2011b)	1–2	Static	LA, USA	MILP	×	8–20	×	×
Lu et al. (2011a)	1–2	Dynamic	LA, USA	Insertion Heuristic	×	15–25/h	×	✓
Alshalalfah and Shalaby (2012)	1	Static	Toronto, Canada	CP	×	12–14/trip	✓	×
Qiu et al. (2014a)	1	Dynamic	Nanjing, China	Insertion Heuristic	×	10–30/h	✓	×
Qiu et al. (2014b)	1	Dynamic	LA, USA	AE	×	20–60/h	✓	×
Qiu et al. (2015)	1	Static	×	AE	×	48/h	✓	×
Yang et al. (2016)	1	Static	Tennessee, USA	Heuristic	×	–	×	×
Shen et al. (2017)	1	Static	LA, USA	AE	×	20/h	×	×
Sun et al. (2018)	1	Static	LA, USA	AE + Simulation	×	5–60/h	✓	×
Zheng et al. (2018a)	1–2	Static	Zhengzhou, China	AE	×	26–50/h	✓	×
Zheng et al. (2018b)	1	Static	LA, USA	AE	×	8–28/h	✓	×
Shen et al. (2019)	1	Static	LA, USA	AE	×	7–25/h	✓	×
Zheng and Li (2019)	fleet (unknown)	Dynamic	LA, USA	MILP & Insertion heuristic	✓	12–16/trip	✓	×
Zheng et al. (2019)	1	Static	LA, USA	MILP & Memetic Alg.	×	5–25/trip	✓	×
Liu et al. (2021)	10–30	Static	Wangjing, China	MILP & DP+Fast heuristic	✓	546/h	✓	×
Sipetas and Gonzales (2021)	2	Static	Massachusetts, USA	AE	×	2.5–7.5 pax/mi^2/h	✓	×
Zhang et al. (2021)	1	Static	Nanjing, China	MILP	✓	10–50/trip	×	×
Zheng et al. (2021)	1	Dynamic	LA, USA	MILP & Insertion heuristic	✓	12/trip	✓	×
Sun and Liu (2022)	fleet (unknown)	Static	Harbin, China	Heuristic	×	20–40/h	×	×
Li et al. (2022)	fleet (unknown)	Dynamic	Guangzhou, China	MILP & Genetic Alg.+Greedy Alg.	×	100/h	✓	×
Zhang et al. (2023)	1	Dynamic	LA, USA	MILP & Memetic Alg.	×	10–50/trip	✓	×
Li and Tang (2023)	1	Static	Shenzhen, China	MMEGO + MCS	×	50–100/h	×	×

Legend: MILP (Mixed-Integer Linear Programming); AE (Analytical Equations); CP (Constraint Programming); DP (Dynamic Programming); MMEGO (Multiple Meta-model-based Efficient Global Optimization); MCS (Monte Carlo Simulation).

In the literature, the FRT is always solved neglecting the consideration of shuttle range. Indeed, this is a not particularly restrictive choice when using fuel powered shuttles. The typically short tank refilling time and the wide availability of petrol stations, makes it is reasonable to imagine that shuttle ranges will be sufficient to complete the decided routes. No additional time needs to be considered when modeling the problem as a quick refill will be possible at some point without impacting the schedule.

A research gap to be filled consists in considering different types of shuttles. Among them, electric shuttles are an obvious option, as they may diminish the pollution emissions in cities. However, electric shuttles require high charging times and charging stations are often few. The problem formulation must take into account the time spent at stations either for charging or for waiting for an available charging point, and the additional distance traveled to reach them. Optimization algorithms must then be able to deal with the additional decisions on charging schedules.

It may also be interesting to consider heterogeneous shuttles with different capacity and operation costs. This would allow, for instance, to use larger shuttles during the rush hours and smaller ones for the rest of the time. The flexibility brought by the availability of a heterogeneous fleet will necessarily have to be taken into account when designing optimization algorithms.

An interesting recent contribution on this topic is the use of modular autonomous vehicles in Liu et al. (2021) which allows the vehicle to separate into smaller modules between checkpoints and potentially serve many more on-demand customers. This option, however, complicates the scheduling problem from a combinatorial perspective and requires more advanced algorithmic methods (we will elaborate on that more precisely in Appendix).

5.3. Backtracking and dynamic environment

In the literature, backtracking and dynamic environment are only considered in a few articles. When they are, a shuttle can go back to pick-up a customer who just made a request after the shuttle itself has passed the ad hoc stop location. This may shorten total waiting times and travel distances with respect to the case in which the customer must be assigned to a later trip, because either the request is not taken into account in real time (static environment) or the shuttle can only travel toward the destination terminal. Nevertheless, if backtracking is allowed, the travel time of the on-board customers may increase. Hence, a trade-off must be solved. When taking into account backtracking, the search space to be explored by an algorithm is much bigger than in the case of only forward travel allowed: if we consider the instance representation described in Section 2, the number of arcs in the associated graph strongly increases. Dealing with the larger instances may require the use of specific optimization approaches. For example, advanced mathematical programming approaches such as Column Generation and Branch-and-Price might be beneficial. When the instances are too large however, stepping to (meta-)heuristic algorithms may be necessary, as introduced in some of the most recent contributions.

In addition to make backtracking pertinent, the consideration of a dynamic environment opens a further research direction consisting in the design of algorithms to specifically fit the policies chosen for handling customers. Indeed, the policy chosen to accept and schedule customers may have an impact on the way the algorithm is designed to decide shuttle schedules and routes.

For instance, one possible policy is to schedule a customer's pick-up request in a shuttle route as soon as it arises. Here, the algorithm must be capable of handling a single new request at a time. In another possible policy, the customers may be asked to formulate requests in advance, e.g. one hour, and the operator may have some time to handle them, e.g. 30 minutes. Here, the algorithm may have to deal with several new requests simultaneously, i.e. all those received in the available handling time. The most efficient approaches to fit the two policies may hence be different.

Additional research gaps emerge when a dynamic environment is considered, as here a proper algorithm performance assessment requires the integration of the optimization in a closed-loop framework with a simulator. In such a framework, the simulator replaces reality. It constantly shares information on traffic and request status, and executes the algorithm decisions on shuttle schedules and routes. Indeed, implementing such a framework requires focusing on a large number of technical aspects, such as the communication standards, the software synchronization and the consistency between the optimization and simulation models. How to deal with these aspects is a specific research direction in itself. Moreover, the inclusion of an algorithm in a closed-loop framework requires paying attention to its design under various perspectives, that may be somehow neglected otherwise. In particular, the computational time available for the solution of instances must be in line with the closed-loop setting: typically this time will need to be rather short. The coherence between this short computational time and the size of the instances to be dealt with must then be taken into account. For example, instance decomposition methods may become appropriate, as well as (meta-)heuristic approaches. In addition, in a closed-loop framework, the algorithm will simultaneously have to deal with some previously made decisions and new ones to make. Previously made decisions may be unalterable, or modifiable up to a certain time, with a given flexibility or at a certain cost. The optimization algorithm must be designed so as to make the best of the closed-loop setting.

5.4. Tactical and strategic problems

In the literature, little effort has been dedicated to the study of strategic planning problems associated to the FRT, or to other semi-flexible transport systems (Errico et al., 2013). Examples of these problems are the location of terminals and checkpoints, as well as the design of the service area or the fleet size, given a certain demand. Considerably more effort has been dedicated to tactical planning problems, such as the relation between service area and slack time availability at checkpoints. The works available on these problems mainly use analytical equations to study the impact of parameter values. Unfortunately, these approaches typically need to rely on very strong assumptions and approximations, e.g. on demand distribution both in space and time. The use of optimization algorithms in which system parameters are promoted to decision variables may allow better FRT performance without increasing the complexity of operational management. Such parameters may include: the slack time at checkpoints and depots, the capacity of the shuttle(s), the number of shuttles or the width of the service area. The latter will impact the customers according to whether their origin or destination ends up being located outside the service area and specific policies should be proposed and studied to handle this matter. The use of such optimization algorithms would allow public transport operators to design their service characteristics using a typical day of demand or a few recurrent demand scenarios, which could represent any type of demand distribution. We believe this is a crucial step to motivate operators to study the possibility of switching to hybrid modes of public transportation. Moreover, tackling the strategic aspects will allow departing from simple rectangular service areas and fit better real demand data which has sometimes very specific and non-uniform spatial distributions. The design of algorithms for strategic and tactical problems may take inspiration, e.g. from the survey of Drexl and Schneider (2015) that reviews the literature on the location routing problem and deals with the planning of facilities, including plants, depots, warehouses and hubs, or from the articles discussed in Section 3.4 about Demand Adaptive Systems. Despite the similarities between FRT and DAS in the operational context, when moving to a tactical or strategic perspective, non-trivial adaptation of the methods designed for DAS would be necessary to deal with the fact that the non-checkpoint stops can vary and change at each trip.

5.5. Objective function

In the literature, the objective functions considered when optimizing the FRT are all rather similar. However, many extensions of this problem can be investigated by refining or diversifying the objective function. For example, the rejection of customers has only been dealt with at the modeling level in Liu et al. (2021). Possibly, request acceptances can be integrated into the models, introducing in the objective function a rejection penalty or cost. This type of modeling extension allows the algorithms to provide a feasible solution to the operator at all times, although it requires to carefully estimate the cost of a rejection.

So far, moreover, no paper has considered environmentally friendly objective functions for FRT systems. Lowering greenhouse gas emission is of great importance nowadays (Lo and Shih, 2021), and the transport sector is one of the major source of carbon dioxide emissions (Demir et al., 2011, 2014). In Bektaş and Laporte (2011), authors consider minimizing greenhouse gas emissions as the objective functions of a VRP variant called *Pollution-Routing Problem*. For this purpose, they consider elements such as vehicle weights, route slope, vehicle speed, etc. Indeed, similar elements can be considered for the FRT, to make it a greener transport mode.

5.6. Stochasticity

All existing approaches for the FRT consider deterministic problems and assume the absence of stochasticity on the problems inputs except for Zheng et al. (2021) which considers no-shows and cancellations of on-demand stops. It is clear that many

Table 2
Distribution of the reviewed studies by the sensitivity analyses conducted.

References	SAS	WTW1	RTW	ST	DL	MP	AWD	NV	IT	NC	PS
Quadrifoglio and Dessouky (2008)	✓										
Alshalalfah and Shalaby (2010)	✓			✓							
Lu et al. (2011b)			✓								
Qiu et al. (2014b)		✓									
Zheng et al. (2018a)	✓	✓						✓			
Zheng et al. (2018b)	✓							✓	✓	✓	✓
Shen et al. (2019)		✓									
Zheng et al. (2019)						✓	✓				
Zhang et al. (2021)					✓					✓	

Legend: SAS (Service Area Shape); WTW1 (Walking Time Weight); WTW2 (Waiting Time Weight); RTW (Riding Time Weight); ST (Slack Time); DL (Demand Level); CS (Checkpoint Spacing); Meeting Points (MP); AWD (Acceptable Walking Distance); NV (Number of Vehicles); IT (Idle Time); NC (Number of Checkpoints); PS (Proportion of customers).

additional events can perturb the pre-computed schedule of the shuttles, such as travel or boarding time variations, shuttle temporary unavailability, etc... Moreover, even though Zheng et al. (2021) consider stochastic events, they do not provide a stochastic approach to handle them a priori but instead react in real time to those events.

Providing Stochastic or Robust Programming approaches to the FRT would be very beneficial to design more resilient solutions. This is particularly true for strategic and tactical approaches which aim at (re)designing system features before the shuttles have to be scheduled, see e.g. the contributions discussed in 3.4 on strategic approaches for Demand Adaptive Systems. In such a case, the strategic and tactical aspects can be modeled through here-and-now variables and the shuttle scheduling decision by wait-and-see variables.

5.7. Sensitivity analysis

Despite the fact that several studies conduct some form of partial sensitivity analysis on the system parameters (see Table 2 for a summary of such analysis), the sole approach utilized is *one factor at a time*. This technique involves evaluating a single parameter at a time, as opposed to testing numerous parameters concurrently. This approach has several drawbacks, including the inability to quantify the interaction between parameters. For instance, if two parameters are selected for a sensitivity analysis, this approach may be used to analyze the influence of each parameter individually. However, the influence of the simultaneous interaction of two parameters cannot be quantified. Therefore, this technique cannot lead to a comprehensive investigation of the objective function sensitivity. To overcome this limitation, one may conduct sensitivity analysis via factorial experiments to not only analyze the influence of each parameter individually, but also the impact of their interactions at various levels.

In addition, more advanced methods, such as the *Sobol* technique (Sobol, 1993), may be used to analyze the sensitivity of the objective function. The Sobol technique is a global sensitivity analysis that operates within a probabilistic framework; it decomposes the output variance of a model or system into fractions that may be assigned to inputs or input sets. Given a model with two inputs and one output, for instance, one may discover that 60% of the output variation is related to the variance in the first input, 30% to the variance in the second input, and 10% to interactions between the two. These percentages may be instantly interpreted as assessments of sensitivity. Such methods are desirable because they quantify the sensitivity throughout the whole input space.

5.8. Demand studies

Only three, very recent contributions are surveyed in this review concerning the behavior and acceptance of customers towards FRT systems. In order for such systems to become more attractive for public transport systems deciders, more research needs to be done in this direction to assess the conditions under which potential customers will use a local FRT systems. Moreover, the research surveyed above only concerns cities in China and there is no guarantee that the results can be generalized to other cities in other countries around the world. Finally, an assessment of the perception of the customers of FRT systems already implemented in the past would provide interesting insights on the perceived advantages of this specific mode of transportation.

6. Conclusion

The FRT is very promising for meeting the challenges of mobility. In this paper, we provided a summary of the research on the FRT. We also provided a summary of the problem variants studied, the methodologies developed, and the different features of the problem that have been addressed in the literature. Based on this summary, we pointed out research gaps to be filled to increase the applicability of the FRT. In particular, various aspects such as the fleet characteristics shall be deeply investigated. Moreover, further strategical and tactical analysis may help transport authorities to understand how to shape the characteristics of a FRT system to serve a certain neighborhood. On the methodological side, sophisticated exact algorithms such as Branch-and-Price can be developed, using the advanced literature existing for such methodologies applied to variants of the VRP. In order to guide further research in this direction, we provided a short and focused survey on advanced mathematical programming methods published on the VRPPDTW and DARP in Appendix.

Appendix. Vehicle routing problem with pick-up and delivery and time windows and dial a ride problem

In the following, we propose a focused survey of VRPPDTW and DARP, which can offer insights on available efficient methods that could be adapted to the FRT problem, given the similarities between these problems, from a mathematical point of view. We discuss the limits of these approaches for the FRT and focus mainly on exact algorithms based on advanced mathematical programming methods, which allow us to understand better the similarities between the FRT and the problems presented here.

Two main approaches have been used to solve the VRP with exact methods, i.e. Branch-and-Price and Branch-and-Cut. The former was first proposed for the VRPPDTW in Dumas et al. (1991). The authors consider a set partitioning formulation of the problem in which each column corresponds to a feasible shuttle route and each constraint is associated to a request that must be satisfied exactly once. The resulting pricing sub-problem is a shortest route problem with time window, capacity, pairing, and precedence constraints. This problem is solvable by dynamic programming, and the authors use an algorithm similar to the one developed in Desrosiers et al. (1986) for the single-vehicle pick-up and delivery problem with time windows. As discussed in Section 5, even though the FRT is often presented as a single shuttle problem, it is inherently a multi trip problem. Therefore, the approach of Dumas et al. (1991) could be adopted by defining a specific column for each shuttle trip. Nevertheless, a clear limit of this type of approach applied to a problem like the FRT is the fact that according to the authors of Dumas et al. (1991), it works best with a large request at each customer node, i.e. tight capacity constraints, and a small number of requests per shuttle route. Instead, in the FRT, the transport demand at each node is usually small. Another Branch-and-Price approach for the VRPPDTW was later described by Savelsbergh and Sol (1998). It improves over Dumas et al. (1991) by using: improved primal heuristics and a specific column management mechanism to limit the size of the master problem; construction-improvement heuristics for the pricing problem; and a higher level branching scheme. The applicability of the obtained approach is better suited to handle the FRT problem as it can handle a larger demand and longer routes, and therefore instances with a less tight capacity constraint.

The second family of exact approaches for the VRP is Branch-and-Cut. In Branch-and-Cut, valid inequalities are added to the formulation to strengthen the relaxations. Cordeau (2006) develops a Branch-and-Cut algorithm for the DARP based on a three-index formulation of the problem, with variables and constraints very similar to the base FRT formulations.

The model was enriched with several families of complex valid inequalities, most of which are based on the precedence or the capacity constraints. In order to discuss those inequalities, we introduce the set $P = \{1, \dots, n\}$ of pick-up nodes, the set $D = \{n+1, \dots, 2n\}$ of drop-off nodes (with nodes i and $n+i$ representing respectively the pick-up and drop-off nodes of a request) as well as 0 and $2n+1$ the origin and destination terminals. Unfortunately, some of these inequalities cannot always be generalized to the FRT straightforwardly, at least in their present form. For example, the *Bounds on Load Variables* or *Capacity Inequalities* (which estimate the minimum number of shuttles needed to visit all nodes inside a set $S \subseteq P \cup D$) use the capacity constraint, while the FRT shuttles are often considered with infinite capacity in the literature. *Infeasible Route Inequalities*, which forbid routes with infeasible ride-time, also have no application to the FRT. Finally, The *Order-Matching Inequalities* are complex inequalities, an easier version of which introduces subsets $H \subset N$ which contain pick-up nodes $i, j \in P$ but neither $n+i$, $n+j$ nor the terminals. Therefore, it is not possible to select $|H|-1$ arcs inside the subgraph generated by H as well as arcs in both subsets $\{i, n+i\}$ and $\{j, n+j\}$ at the same time. However, they happen to be redundant when the graph of an instance is directed and many arcs are clearly not present in the FRT due to the fact that the shuttle is supposed to go forward at all times, unless a backtracking policy is implemented. From the perspective of a possible generalization to the FRT, the most relevant inequalities are listed below:

1. **Bounds on Time Variables**: the values of the time variables B can be tightened using the identity of the predecessor and successor nodes through variables x. These inequalities are applied on an alternative formulation where B variables are aggregated on their shuttle index $k \in K$. The same reasoning can be applied on the load variables Q when the shuttles are assigned a finite capacity.
2. **Subtour Elimination Inequalities**: the traditional subtour elimination constraints for the TSP can here be lifted similar to what has been proposed for the precedence-constrained asymmetric TSP (Balas et al., 1995).
3. **Precedence Constraint Inequalities**: the inequalities simply exploit the fact that a pick-up node $i \in P$ must be visited before its corresponding drop-off node $n+i$, defining a node set S containing nodes 0 and $n+i$ but neither i nor $2n+1$.
4. **Generalized Order Inequalities**: a set of mutually disjoint subsets $U_1, \dots, U_m \subset N$ are defined such that they do not contain any terminal but each U_l contains both i_l and $n+i_{l+1}$ with $i_1, \dots, i_m \in P$. One can use such sets together with precedence constraint inequalities to adapt the *precedence cycle breaking inequalities* from Balas et al. (1995) to form valid inequalities using the subsets U_l.

Many of the above inequalities involve subsets which come in an exponential number with respect to the instance size. Therefore, they are systematically implemented as user constraints and separation heuristics are used to generate new constraints at a subset of the branching nodes (but not any such node).

The above DARP model is solved by a commercial solver in Cordeau (2006). The advocated valid inequalities allow increasing the objective function of the linear relaxation of the model by more than 5% on average on the adopted benchmark instances. Interestingly, pre-processing techniques such as time-window tightening and arc elimination tend to increase the value of the linear relaxation by more than 15% on average on the same instances. The number of branching nodes necessary to solve an instance without the added inequalities can be three orders of magnitude larger than it is when incorporating said inequalities and the average time needed to solve those instances is more than ten times larger. This model is able to tackle instances with up to four shuttles and 32 requests. It is later improved by Ropke et al. (2007), who propose two different formulations for the VRPPDTW, which can be also used to solve the DARP. These formulations improve on Cordeau (2006) since they provide tighter linear relaxations, thanks to an exponential number of constraints and a smaller number of variables. In addition to the Subtour Elimination Inequalities and Generalized Order Inequalities imported from Cordeau (2006), several new families of valid inequalities are introduced. The inequalities which provide the largest numerical impact on the strength of the linear relaxation are:

1. **Fork Inequalities**: fork inequalities consider groups of infeasible routes which share some common arcs. A simple example is a feasible route R for which each route (i, R, j) for every $i \in S$ and $j \in T$ belonging to subsets $S, T \subset N$ is infeasible.
2. **Reachability Inequalities**: these inequalities are derived for the VPRTW in Lysgaard (2006). They are based on node sets $T \subset N$ where each node must be visited by a different shuttle. The nodes in a set are defined as *conflicting*.

It is clear that a finite capacity is an important part of the possible infeasibility of certain sub-routes, so that these inequalities will have a larger impact in versions of the FRT where a finite shuttle capacity is considered, provided that capacity is tight enough that the shuttle can be saturated in certain trips. Moreover, inequalities that consider different shuttles (and therefore different tours), must be applied to different trips (possibly of the same shuttle) in the FRT. Other inequalities considered in Ropke et al. (2007) are: *Strengthened Capacity Inequalities*, which are the traditional capacity constraints from Cordeau (2006) strengthened by considering node pairs $(k, n+k)$, with $k \in P$ visited before entering a given node subset S and $n+k$ visited after S; and *Strengthened Infeasible Route Inequalities*, which identify feasible sub-routes R that cannot be part of any feasible solution. Ropke et al. (2007) solves DARP instances with up to eight shuttles and 96 requests.

Branch-and-Price and Branch-and-Cut approaches can be mixed inside a Branch-and-Cut-and-Price framework. In Ropke and Cordeau (2009), the authors show that, by using a pricing sub-problem which takes the form of an *elementary* resource constrained shortest path problem (unlike Dumas et al., 1991), the set covering exponential formulation already satisfies the fork and reachability inequalities, as well as a certain type of (strengthened) precedence constraint inequality. They also show that the addition of other valid inequalities ruins crucial properties of the pricing sub-problem. However, they introduce a perturbation of the costs matrix which is sufficient to recover the necessary properties. The Branch-and-Cut-and-Price improves on the Branch-and-Cut of Ropke et al. (2007) and can even solve a few instances with up to 500 requests, if the time windows are tight enough. However, the valid inequalities added to the Branch-and-Price formulations are not a big help and it is observed that the improvement mainly comes from the speed-up on the solution of the pricing sub-problems. Further improvements on that front are presented in Gschwind et al. (2018) where the method of *bidirectional labeling* introduced by Righini and Salani (2006) is generalized to pick-up and delivery problems to respect *pick-up and delivery triangle inequalities* at the same time.

The above analysis seems to indicate that the Branch-and-Price technologies for the VRP with pick-up and delivery is mature enough to be applied to the FRT. Nevertheless, one should be very careful when applying existing VRP results pertaining to Column Generation approaches. An example of the difficulties of applying such an approach to FRT-like systems are highlighted in, e.g., Rahmani et al. (2016) which tackles a DARP variant through a Branch-and-Price algorithm. The objective is the same as in the base FRT problem and it is observed that the price of a column depends on the time of visit of the different nodes involved in the vehicle tour. This is why the authors make the simplifying assumption that the waiting and riding costs of the customers are equal to 0. Fortunately, such drastic simplifications are not strictly necessary from an FRT perspective: it can be assumed that it is always beneficial to reach a node and board a customer as soon as possible if the objective weight associated to the waiting time is larger than the one associated to the riding time, an otherwise reasonable assumption. It is then possible to use classic DP approaches to solve the pricing problem with state dominance rules. However, since the objective in the pricing sub-problem depends on the pick-up and drop-off times of the customers, it is *a priori* not possible to apply a backward labeling approach, and *a fortiori* impossible to use the advanced bi-directional labeling techniques of Gschwind et al. (2018). Even this may seem like a serious complication for solving FRT problems, we can observe that the structure of the route is more structured than for VRP or DARP due to the constraint of stopping at checkpoint nodes and leaving such nodes at a specific time. This means that partial solutions of the pricing sub-problem can be easily compared at such checkpoints, allowing for an intensified use of dominance rules and the potential elimination of a large subset of partial solutions. The resulting reduction of the proliferation of DP states should speed up considerably the resolution of the pricing sub-problem, a key ingredient for the efficiency of a Branch-and-Price approach, as noted earlier. The potential gain of using a set-covering reformulation is important since the linear relaxation of the base MILP for the FRT problem is not very tight, due to several big-M constraints. For example, even the MILP of Quadrifoglio et al. (2008a) with valid inequalities does not manage to close some instances with up to 30 stops in 10 h of computational time, exhibiting a final gap of more than 15%. Moreover, in Liu et al. (2021), the shuttles are composed of several modules grouped together which are separated between consecutive checkpoints and a route must be designed for each module. The use of Branch-and-Price could be beneficial as the MILP is very hard to solve in general, given the large number of big-M constraints. The question of the adaptability and numerical impact of VRPPDTW and DARP valid inequalities is a more open question, which requires careful investigation given the differences with the FRT as sketched in Section 3.5. An important question, though, is whether it would be possible to include the FRT valid inequalities of Quadrifoglio et al. (2008a) into a Branch-and-Price formulation similar to the work of Ropke and Cordeau (2009) by modifying correctly the costs matrix for the pricing sub-problems, or whether it is possible to prove that such inequalities are already implied by the set covering formulation.

Even though we do not delve here into the literature focused on heuristic algorithms for the VRPPDTW, many such approaches exist. Unfortunately, existing neighborhood structures and search algorithms designed for related problems, e.g. the Vehicle Routing Problems with Time Windows (VRPTW), may not be well suited for the FRT problem. Indeed, as per the structure of the public transport system, shuttles have to depart from the origin terminal at regular intervals. This is in sharp contrast with the VRPTW and implies that assigning a customer to an alternate trip departing much later will change considerably the cost of the solution. Moreover, barring the possibility of backtracking, once the customers have been assigned to a given trip, which often happens as a first step in VRPTW heuristics, there is very little choice in the order in which they must be visited. It is clear that the structure of solutions will be very specific and that smart neighborhood search algorithms will be crucial to the efficiency of a meta-heuristic algorithm for the FRT (Drexl, 2021).

We close this section by observing that a quick survey of the literature on the VRPPDTW offers many insights and tools to tackle extensions of the FRT systems discussed in Section 5, particularly with respect to the inclusion of a finite capacity (see, e.g. Ropke and Cordeau, 2009). Even though we did not focus on electric vehicles in this section, much knowledge of methods for the Green VRP can also be transposed to a green version of the FRT (see, e.g. Yu et al., 2019; Masmoudi et al., 2018)

References

Alshalalfah, B., Shalaby, A., 2010. Development of important relationships for the planning of flex-route transit services. In: Proc. 89th Annual Transportation Research Board Meeting (CD-ROM), Washington, DC.. Transportation Research Board, Washington DC, United States.

Alshalalfah, B., Shalaby, A., 2012. Feasibility of flex-route as a feeder transit service to rail stations in the suburbs: case study in Toronto. J. Urban Plann. Dev. 138 (1), 90–100.

Balas, E., Fischetti, M., Pulleyblank, W.R., 1995. The precedence-constrained asymmetric traveling salesman problem. Math. Program. 68, 241–265.

Bektaş, T., Laporte, G., 2011. The pollution-routing problem. Transp. Res. B 45 (8), 1232–1250.

Cordeau, J.-F., 2006. A branch-and-cut algorithm for the dial-a-ride problem. Oper. Res. 54 (3), 573–586.

Crainic, T.G., Errico, F., Malucelli, F., Nonato, M., 2012. Designing the master schedule for demand-adaptive transit systems. Ann. Oper. Res. 194, 151–166.

Crainic, T., Malucelli, F., Nonato, M., 2001. Flexible many-to-few+ few-to-many=an almost personalized transit system. In: TRISTAN IV, Vol. 2. pp. 435–440.

Daganzo, C.F., 1984a. Checkpoint dial-a-ride systems. Transp. Res. Part B: Methodol. 18 (4-5), 315–327.

Daganzo, C.F., 1984b. The length of tours in zones of different shapes. Transp. Res. Part B: Methodol. 18 (2), 135–145.

Demir, E., Bektaş, T., Laporte, G., 2011. A comparative analysis of several vehicle emission models for road freight transportation. Transp. Res. Part D: Transp. Environ. 16 (5), 347–357.

Demir, E., Bektaş, T., Laporte, G., 2014. A review of recent research on green road freight transportation. European J. Oper. Res. 237 (3), 775–793.

Desrosiers, J., Dumas, Y., Soumis, F., 1986. A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. Amer. J. Math. Management Sci. 6 (3–4), 301–325.

Drexl, M., 2021. On the one-to-one pickup-and-delivery problem with time windows and trailers. CEJOR Cent. Eur. J. Oper. Res. 29 (3), 1115–1162.

Drexl, M., Schneider, M., 2015. A survey of variants and extensions of the location-routing problem. European J. Oper. Res. 241 (2), 283–308.

Dumas, Y., Desrosiers, J., Soumis, F., 1991. The pickup and delivery problem with time windows. Eur. J. Oper. Res. 54 (1), 7–22.

Errico, F., Crainic, T.G., Malucelli, F., Nonato, M., 2013. A survey on planning semi-flexible transit systems: Methodological issues and a unifying framework. Transp. Res. C 36, 324–338.

Errico, F., Crainic, F., Nonato, M., 2017. A benders decomposition approach for the symmetric TSP with generalized latency arising in the design of semiflexible transit systems. Transp. Sci. 52, 706–722.

Errico, F., Crainic, F., Nonato, M., 2021. The single-line design problem for demand-adaptive transit systems: A modeling framework and decomposition approach for the stationary-demand case. Transp. Sci. 55, 1300–1321.

Fu, L., 2002. Planning and design of flex-route transit services. Transp. Res. Rec. 1791 (1), 59–66.

Gschwind, T., Irnich, S., Rothenbächer, A.-K., Tilk, C., 2018. Bidirectional labelling in column-generation algorithms for pickup-and-delivery problems. European J. Oper. Res. 266, 521–530.

Ho, S.C., Szeto, W.Y., Kuo, Y.-H., Leung, J.M., Petering, M., Tou, T.W., 2018. A survey of dial-a-ride problems: Literature review and recent developments. Transp. Res. B 111, 395–421.

Koffman, D., 2004. Operational Experiences with Flexible Transit Services, Vol. 53. Transportation Research Board.

Kumar, S.N., Panneerselvam, R., 2012. A survey on the vehicle routing problem and its variants. Sci. Res..

Laporte, G., 1992. The vehicle routing problem: An overview of exact and approximate algorithms. Eur. J. Oper. Res. 59 (3), 345–358.

Li, J., He, Z., Zhong, J., 2022. The multi-type demands oriented framework for flex-route transit design. Sustainability 14 (15), 9727.

Li, M., Tang, J., 2023. Simulation-based optimization considering energy consumption for assisted station locations to enhance flex-route transit. Energy 277, 127715.

Liu, X., Qu, X., Ma, X., 2021. Improving flex-route transit services with modular autonomous vehicles. Transp. Res. Part E: Logist. Transp. Rev. 149, 102331.

Lo, S.-C., Shih, Y.-C., 2021. A genetic algorithm with quantum random number generator for solving the pollution-routing problem in sustainable logistics management. Sustainability 13 (15), 8381.

Lu, W., Lu, L., Quadrifoglio, L., 2011a. Scheduling multiple vehicle mobility allowance shuttle transit (m-MAST) services. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems. (ITSC), IEEE, pp. 125–132.

Lu, W., Xie, Y., Wang, W., Quadrifoglio, L., 2011b. An analytical model to select the fleet size for MAST systems. In: 2011 IEEE Forum on Integrated and Sustainable Transportation Systems. IEEE, pp. 152–158.

Lysgaard, J., 2006. Reachability cuts for the vehicle routing problem with time windows. European J. Oper. Res. 175, 210–223.

Malucelli, F., Nonato, M., Pallottino, S., 1999. Demand adaptive systems: Some proposals on flexible transit. In: Ciriani, T.A., Gliozzi, S., Johnson, E.L., Tadei, R. (Eds.), Operational Research in Industry. Palgrave Macmillan UK, London, pp. 157–182.

Masmoudi, M.A., Hosny, M., Demir, E., Genikomsakis, K.N., Cheikhrouhou, N., 2018. The dial-a-ride problem with electric vehicles and battery swapping stations. Transp. Res. Part E: Logist. Transp. Rev. 118, 392–420.

Mor, A., Speranza, M.G., 2022. Vehicle routing problems over time: a survey. Ann. Oper. Res. 314 (1), 255–275.

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

Potts, J., Marshall, M., Crockett, E., Washington, J., 2010. Transit Cooperative Research Programme Report 140: A Guide for Planning and Operating Flexible Public Transportation Services. Technical report, Transportation Research Board.

Pratelli, A., Schoen, F., 2001. A mathematical programming model for the bus deviation route problem. J. Oper. Res. Soc. 52, 494–502.

Qiu, F., Li, W., An, C., 2014a. A google maps-based flex-route transit scheduling system. In: CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems. ASCE, pp. 247–257.

Qiu, F., Li, W., Haghani, A., 2015. An exploration of the demand limit for flex-route as feeder transit services: A case study in salt lake city. Public Transp. 7, 259–276.

Qiu, F., Li, W., Zhang, J., 2014b. A dynamic station strategy to improve the performance of flex-route transit services. Transp. Res. C 48, 229–240.

Quadrifoglio, L., Dessouky, M.M., 2004. Mobility allowance shuttle transit (MAST) services: formulation and simulation comparison with conventional fixed route bus services. In: Modelling, Simulation, and Optimization – Proceedings of the 4th IASTED International Conference. Kauai, HI, USA, 17–19 August. Acta Press, Calgary, p. 6.

Quadrifoglio, L., Dessouky, M.M., 2008. Sensitivity analyses over the service area for mobility allowance shuttle transit (MAST) services. In: Hickman, M., Mirchandani, P., Voß, S. (Eds.), Computer-Aided Systems in Public Transport. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 419–432.
Quadrifoglio, L., Dessouky, M.M., Ordóñez, F., 2008a. Mobility allowance shuttle transit (MAST) services: MIP formulation and strengthening with logic constraints. Eur. J. Oper. Res. 185 (2), 481–494.
Quadrifoglio, L., Dessouky, M.M., Ordóñez, F., 2008b. Mobility allowance shuttle transit (MAST) services: MIP formulation and strengthening with logic constraints. In: 5th International Conference, CPAIOR 2008 Paris, France 5015. pp. 398–404.
Quadrifoglio, L., Dessouky, M.M., Palmer, K., 2007. An insertion heuristic for scheduling mobility allowance shuttle transit (MAST) services. J. Sched. 10 (1), 25–40.
Quadrifoglio, L., Hall, R.W., Dessouky, M.M., 2006. Performance and design of mobility allowance shuttle transit services: bounds on the maximum longitudinal velocity. Transp. Sci. 40 (3), 351–363.
Quadrifoglio, L., Li, X., 2009. A methodology to derive the critical demand density for designing and operating feeder transit services. Transp. Res. B 43 (10), 922–935.
Quadrifoglio, L., Shen, C.-W., 2010. Performance analysis of the "Zoning" strategies for ADA paratransit services. URL https://static.tti.tamu.edu/swutc.tamu.edu/publications/technicalreports/169114-1.pdf.
Rahmani, N., Detienne, B., Sadykov, R., Vanderbeck, F., 2016. A column generation based heuristic for the dial-a-ride problem. In: 6th International Conference on Information Systems, Logistics and Supply Chain, June 1–4, Bordeaux, France.
Righini, G., Salani, M., 2006. Symmetry helps: Bounded bi-directional dynamic programming techniques for the elementary shortest path problem with resource constraints. Discrete Optim. 3 (3), 255–273.
Ropke, S., Cordeau, J.-F., 2009. Branch and cut and price for the pickup and delivery problem with time windows. Transp. Sci. 43 (3), 267–283.
Ropke, S., Cordeau, J.-F., Laporte, G., 2007. Models and branch-and-cut algorithms for pickup and delivery problems with time windows. Networks 49 (4), 258–272.
Savelsbergh, M., Sol, M., 1998. Drive: Dynamic routing of independent vehicles. Oper. Res. 46 (4), 474–490.
Shen, J., Qiu, F., Zheng, C., Ma, C., 2019. Fare strategy for flex-route transit services: Case study in Los Angeles. IEEE Access 7, 82038–82051.
Shen, J.-x., Zhou, Y.-h., Liu, Y.-n., Qiu, F., 2017. A service-based fare policy for flex-route transit services. In: International Conference on Green Intelligent Transportation System and Safety. Springer, pp. 87–95.
Shi, Y., Zhou, Y., Boudouh, T., Grunder, O., 2020. A lexicographic-based two-stage algorithm for vehicle routing problem with simultaneous pickup–delivery and time window. Eng. Appl. Artif. Intell. 95, 103901.
Sims, R., Schaeffer, R., Creutzig, F., nez, X.C.-N., D'Agosto, M., Dimitriu, D., Meza, M.F., Fulton, L., S. Kobayashi, O.L., McKinnon, A., Newman, P., Ouyang, M., Schauer, J., Sperling, D., Tiwari, G., 2014. Transport. In: Climate Change 2014 Mitigation of Climate Change Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, pp. 599–670, chapter 8.
Sipetas, C., Gonzales, E.J., 2021. Continuous approximation model for hybrid flexible transit systems with low demand density. Transp. Res. Rec. 2675 (8), 198–214.
Sobol, I., 1993. Sensitivity analysis for non-linear mathematical models. Math. Model. Comput. Exp. 1, 407–414.
Sun, C., Li, W., Shao, Z., Zheng, Y., 2018. A new strategy to optimize the flex-route transit service. In: CICTP 2017: Transportation Reform and Change—Equity, Inclusiveness, Sharing, and Innovation. American Society of Civil Engineers Reston, VA, pp. 3524–3533.
Sun, X., Liu, S., 2022. Research on route deviation transit operation scheduling—A case study in suburb no. 5 road of Harbin. Sustainability 14 (2), 633.
Vansteenwegen, P., Melis, L., Aktaş, D., David, B., Montenegro, G., Sartori Vieira, F., Sörensen, K., 2022. A survey on demand-responsive public bus systems. Transp. Res. C 137, 103573.
Yang, H., Cherry, C.R., Zaretzki, R., Ryerson, M.S., Liu, X., Fu, Z., 2016. A GIS-based method to identify cost-effective routes for rural deviated fixed route transit. J. Adv. Transp. 50 (8), 1770–1784.
Yu, J., Li, W., Zhang, J., Guo, R., Zheng, Y., 2023a. Understanding the effect of socio-demographic and psychological latent characteristics on flex-route transit acceptance. Plos One 18 (2), ahead of print.
Yu, Y., Wang, S., Wang, J., Huang, M., 2019. A branch-and-price algorithm for the heterogeneous fleet green vehicle routing problem with time windows. Transp. Res. B 122, 511–527.
Yu, J., Zheng, Y., Li, W., Zhang, J., Guo, R., Wu, L., 2023b. Understanding flex-route transit adoption from a stage of change perspective. Transp. Res. Rec. ahead of print.
Zhang, Y., Farber, S., Young, M., 2022. Eliminating barriers to nighttime activity participation: the case of on–demand transit in belleville, Canada. Transportation 49, 1385–1408.
Zhang, J., Li, W., Wang, G., Yu, J., 2021. Feasibility study of transferring shared bicycle users with commuting demand to flex-route transit—A case study of nanjing city, China. Sustainability 13 (11), 60–67.
Zhang, J., Li, W., Zheng, Y., Guo, R., 2023. Dynamic clustering meeting points strategy to improve operational service capability of flex-route transit. J. Transp. Eng., Part A: Syst. 149 (6).
Zhao, J., Dessouky, M., 2008. Service capacity design problems for mobility allowance shuttle transit systems. Transp. Res. B 42 (2), 135–146.
Zheng, Y., Gao, L., Li, W., 2021. Vehicle routing and scheduling of flex-route transit under a dynamic operating environment. Discrete Dyn. Nat. Soc. 2021, 1–10.
Zheng, Y., Li, W., 2019. Flex-route transit service with different degree of dynamism. In: CICTP 2019. ASCE library, pp. 4369–4380.
Zheng, Y., Li, W., Qiu, F., 2018a. A methodology for choosing between route deviation and point deviation policies for flexible transit services. J. Adv. Transp. 49 (3), 496–509.
Zheng, Y., Li, W., Qiu, F., 2018b. A slack arrival strategy to promote flex-route transit services. Transp. Res. Part C: Emerg. Technol. 92, 442–455.
Zheng, Y., Li, W., Qiu, F., Wei, H., 2019. The benefits of introducing meeting points into flex-route transit services. Transp. Res. C 106, 98–112.
Zheng, Y., Li, W., Qiu, F., Wei, H., 2020. Travelers' potential demand toward flex-route transit: Nanjing, China, case study. J. Urban Plann. Dev. 146 (1), 05019018–10.