Incorporating Ad Hoc Stops in Public Transport : A Study on Flex-Route Transit

Reza Shahin

Thèse

présenté pour obtenir le grade de

tbfDocteur de l'Université Gustave Eiffel

Spécialité : Informatique

présentée par

Reza Shahin

Incorporating Ad Hoc Stops in Public Transport: A Study on Flex-Route Transit

soutenue publiquement le 26/01/2024 devant le jury d'examen:

Mme Valentina Cacchiani	Professor, University of Bologna	Rapporteur
M. Olivier Péton	Professeur, Université Gustave Eiffel	Rapporteur
Mme Sourour Elloumi	Professeure, ENSTA Paris	Présidente
Mme Paola Pellegrini	Directrice de recherche, Université Gustave Eiffel	Directrice de thèse
M. Pierre-Olivier Vandanjon	Directeur de recherche, Université Gustave Eiffel	Co-directeur de thèse
M. Pierre Hosteins	Chargé de Recherche, Université Gustave Eiffel	Encadrant

préparée à l'Université Gustave Eiffel COSYS-ESTAS
20 rue Elisée Reclus, 59650 Villeneuve D'Ascq
École Doctorale MADIS 631

Acknoledgements

I would like to extend my deepest gratitude to my family, who unwaveringly stood by my side throughout the arduous journey of my doctoral studies. Their emotional sustenance and encouragement have been invaluable. I wish to acknowledge my devoted, supportive, and encouraging wife, Hediye, whose steadfast support during this academic pursuit cannot be overstated. Thank you. I owe a particular debt of gratitude to my parents, who have nurtured me with love and have consistently supported me in all of my endeavors. I dedicate this thesis to my beloved family.

Université Gustave Eiffel
January 2024

Résumé

Ces dernières années, on a assisté à une tendance croissante à l'adoption de solutions de transport flexibles telles que le transport à la demande (DRT), principalement car elles répondent mieux aux attentes des passagers. Cependant, cette flexibilité s'accompagne d'importants coûts économiques. Par conséquent, les autorités de transport explorent actuellement des méthodes pour augmenter la flexibilité des transports publics conventionnels (CPT), qui sont généralement moins innovants que la DRT. Cela amène à l'importance du Flex-Route Transit (FRT), un système innovant fusionnant les avantages du DRT et du CPT. Dans cette thèse, nous menons une enquête exhaustive sur le FRT. Plus précisément, nous examinons la littérature académique et nous mettons en évidence les lacunes existantes dans la recherche. Ensuite, nous formulons un modèle de programmation linéaire en nombres entiers mixtes (MILP) étendant l'état de l'art pour inclure des caractéristiques du problème précédemment négligées, et nous le complétons par un ensemble d'inégalités valides pour améliorer sa relaxation linéaire. Nous introduisons également un algorithme heuristique pour obtenir une solution réalisable et employons une technique de démarrage à chaud pour accélérer la résolution du problème d'optimisation mixte. Par la suite, une analyse de sensibilité est réalisée sur la base d'un plan factoriel complet. Ici, nous évaluons les niveaux de saturation du système selon divers scénarios de demande, en nous concentrant sur les temps d'attente pour les passagers. Dans le dernier chapitre, un modèle MILP stochastique pour le FRT est proposé, dans lequel certains paramètres initiaux du MILP de base deviennent des variables. Nous concluons la thèse en décrivant des pistes prospectives pour des recherches futures.

Abstract

In recent years, there has been an ascending trend towards the adoption of flexible transit solutions like Demand Responsive Transit (DRT), primarily due to the enhanced convenience they offer to passengers. However, this flexibility comes with an associated economic burden. Consequently, transport authorities are presently exploring methods to augment the flexibility of Conventional Public Transport (CPT), which is more conservative compared to DRT. This gives rise to the significance of Flex-Route Transit (FRT), an innovative system amalgamating the benefits of both DRT and CPT. In the doctoral dissertation presented here, we conduct an exhaustive investigation of FRT frameworks. Specifically, we scrutinize pertinent academic literature, highlighting extant research lacunae. Then, we formulate a Mixed Integer Linear Programming (MILP) model extending the state of the art to include previously neglected problem features, and we supplement it by a set of valid inequalities to enhance its linear relaxation. We also introduce a heuristic algorithm to procure a feasible solution and employ a warm start technique to boost the MILP solution. Subsequently, a comprehensive full factorial experimental design sensitivity analysis is carried out. Here, we evaluate the system's saturation levels under various demand scenarios, focusing on the elongated wait times for passengers. In the final contribution chapter, a stochastic MILP model for the FRT is proposed, wherein certain initial parameters of the base MILP are elevated to variable status. We conclude the dissertation by outlining prospective avenues for future research.

Contents

Chapter 1

Introduction

1.1 Context and motivation

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. Within the context of this discussion, a 'shuttle' refers to a type of vehicle used predominantly in transportation systems. These shuttles are typically medium-sized transport vehicles, designed to operate on routes, facilitating the movement of passengers.

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 as 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.

DRT systems can reduce these inconveniences, but this comes at a price. The need to effectively balance costs and flexibility motivates the introduction of 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.

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.

Even though feeder-line 646 in LA County is the main real-life example appeared in the academic literature, several examples of route-deviation services have been implemented in different places during the last 25 years. The technical report Potts et al. [2010b] 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 of between half a mile to one mile. Passengers needing an ad hoc 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].

Virtually all systems, be them actually deployed or envisaged in the literature, are meant for serving few customers in quite low density areas or at off-peak times, often at night. In these systems, the deviations requested from the planned shuttle routes are rather few, and shuttles are always big enough to accommodate all requests. Hence, the need for an optimization-based decision support tool is not felt as critical by operators. However, the potential convenience of FRT systems even for more demand-intense situations is certain. This is all the more true considering the current trend toward the progressive closure of entire zones to private cars, that is being followed in many European cities. To facilitate the generalization of the use of FRT, the academic community is focusing more and more on two research directions: advanced modeling and optimization algorithm development. This thesis contributes to both these research directions, as explained in the next section.

1.2 Research objectives and contributions

In this thesis, we contribute to the research field focusing on the design of optimization-based decision support tools for the FRT. The main objective of the research is twofold. On the one hand, we aim to deepen the understanding of FRT systems. On the other hand, we wish to enhance their scheduling solution to participate in making their practical deployment possible in situations with high demand.

These objectives are achieved through four main contributions.

In the first contribution, we propose a thorough exploration of the existing literature concerning the FRT. The chapter starts with a historical overview of FRT, commencing from its introduction in the year 2004, and discusses its evolution. Noteworthy is the discussion of the original term under which FRT was introduced — Mobility Allowance Shuttle Transit (MAST) — and its subsequent renaming to what is now commonly referred to as FRT. To identify promising possibilities to fill open research gaps, we employ comparative analyses with well-established Operations Research problems, specifically the Vehicle Routing Problem (VRP) and the DARP. By doing so, we offer a multidimensional perspective that constitutes a synthesis that could guide subsequent investigations. In the interest of maintaining terminological clarity, it is important to note that while the acronym FRT is sometimes used to describe Fixed-Route Transit in other scholarly works, in the context of this thesis, we will utilize the abbreviation CPT for Fixed-Route Transit. The term FRT will be exclusively reserved for discussions related to Flex-Route Transit. The content of this chapter makes the object of a paper that has been published in Transportation Research Part C (TRC) [Shahin et al., 2024].

In the second contribution, we focus on the extension of a state-of-the-art Mixed Integer Linear Programming (MILP) formulation for the FRT and on the improvement of its solution time. Here, the FRT is tackled as an operational problem, in which shuttle services are set and ad hoc stops must be added to the planned schedule to serve customers. With this contribution, first, we integrate to this formulation novel capacity constraints for shuttles. The consideration of these constraints is very important to study the FRT in contexts of high demand. Second, we focus on the design of a heuristic algorithm to serve

as an effective warm start to facilitate the branch and bound optimization process. This heuristic is a greedy algorithm, specifically designed to yield a feasible solution to complex FRT instances. Third, we propose various sets of novel Valid Inequalities (VI) to be exploited in the exploration of the solution space. Finally, we implement the best known VI from the literature and compare its performance with the VI we developed. The content of this chapter will be submitted to a journal in the Operations Research domain.

In the third contribution, we study the sensitivity of the FRT to the variation of its parameters. Until now, the predominant method used in FRT sensitivity analyses has been the "One Factor at a Time" (OFAT) approach. While OFAT provides valuable baseline information, its limitations become evident when one considers the interactions between multiple factors. Addressing these limitations, we employ a Full Factorial Experimental Design (FFED) to understand the intricate influence of parameters when altered simultaneously—an insight that is notably absent in OFAT-based analyses. We analyze the results of the sensitivity analysis using Analysis of Variance (ANOVA). The sensitivity analysis is performed on three input parameters. The content of this chapter is published in a conference proceeding indexed in IEEE and Scopus [Shahin et al., 2023].

In the fourth and final contribution of this thesis, we present a completely novel MILP model to tackle the FRT from the original perspective of the tactical planning. No specific demand realization is available: it is replaced by a set of possible realization scenarios with an associated probability. The FRT is hence modeled as a stochastic program, including as variables the service frequency, trips duration, fleet size and shuttles capacity. This chapter proposes a modeling of the problem but does not extend to its software implementation.

1.3 Outline of the thesis

The rest of the thesis is organized as follows.

In Chapter 2, we report the FRT definition and we detail the literature review that makes our first contribution. Moreover, we propose how some open research gaps may be filled by recurring to the literature on more classic Operations Research problems.

In Chapter 3, we formally define our modeling contributions on the operational FRT. In particular, we propose constraints to model shuttle capacity, a heuristic algorithm for the FRT, and various sets of valid inequalities. In this chapter, we assess the performance improvements brought by these modeling contributions against a MILP formulation from the literature, including some previously proposed valid inequalities, and we showcase the very positive results we can achieve.

In Chapter 4, we propose a sensitivity analysis of the performance of the FRT to the variation of some of its input parameters. We employ FEED to design the experiments and the ANOVA method to analyze the results of our sensitivity analysis.

In Chapter 5, we present a stochastic programming formulation to deal with the tactical FRT. Here, in addition to others, some of the parameters object of the previous sensitivity analysis become variables to be set to respond as well as possible to expected demand scenarios.

Finally, in Chapter 6, we summarize the main contributions of this thesis and provide future research directions.

Chapter 2

Problem description and literature review

As mentioned in Chapter 1, the FRT is a promising transport system to combine the convenience of CPT and DRT while keeping costs under control. Transport systems with similarities to the FRT exist. For example, flexible transit as proposed by Crainic et al. [2001] and 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 Figure 2.1 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.

In this chapter, 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. To point out promising research directions, we rely on the rich literature existing from some classic Operations Research problems. In particular, we focus on the Pick-up and Delivery Problem with Time Windows (PDPTW) and on the DARP. In the literature, the 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. For reasons of clarity, we only use the term FRT throughout the thesis.

This chapter is organized as follows. We begin with the system definition in Section 2.1. We review the FRT literature in Section 2.2, and we summarize our fining in Section 2.3. We discuss similarities and differences with classic problems in Section 2.4. Next, we propose a survey of selected works on relevant Operations Research problems and highlight how they may be of use to investigate the FRT further, in Section 2.5. Finally, we conclude in Section 2.6.

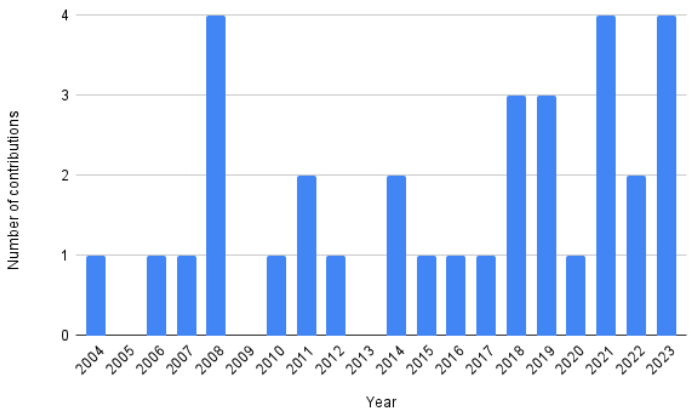


Figure 2.1: Number of contributions per year on the FRT system since 2004.

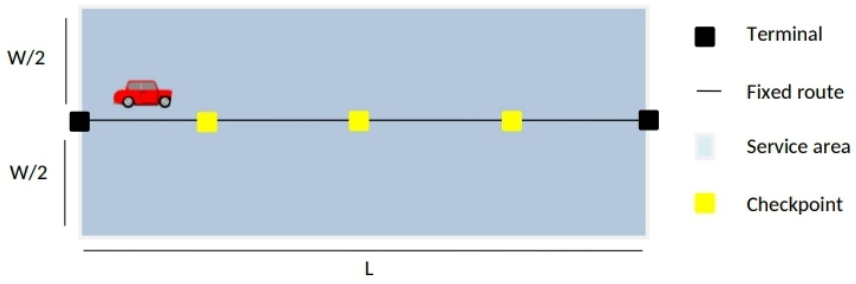


Figure 2.2: Example of the fixed route and service area of a Flex-Route Transit system.

2.1 FRT Definition

Each shuttle performs *trips* between an *origin* and a *destination terminal*. It stops at checkpoints at predetermined times and it can deviate from the direct route between checkpoints by exploiting *slack times* within the service area. The duration of a trip is named *service time* and the sum of service times is *the time horizon*. Shuttles are assumed to move at constant speed. *Longitudinal speed* is defined as the average speed held between terminals. If no deviation is performed, it is equal to the constant speed. It decreases with the distance traveled for ad hoc stops. On the one hand, customers can use checkpoints for their trips along the fixed route. On the other hand, they may request ad hoc stops within the service area. A schematic view of such system is shown in Figure 2.2. 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 the 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 the checkpoints (ad hoc stops).

When optimizing the FRT problem, the objective to minimize is typically a combination of: 1. total

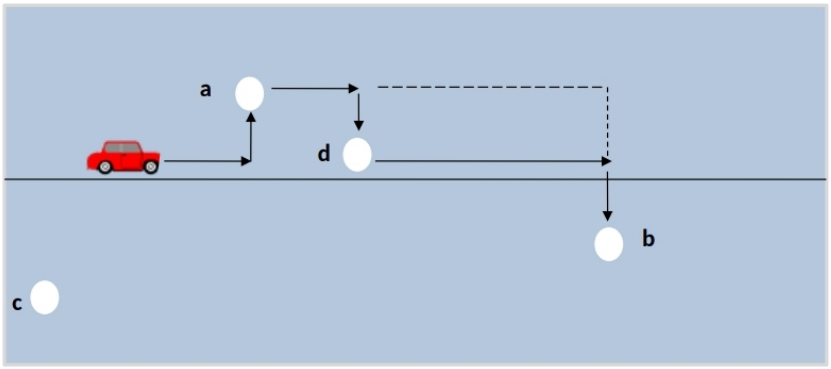


Figure 2.3: Illustration of how FRT works.

travel time of the shuttle(s); 2. total ride time of all customers; 3. total waiting time for customers to be picked-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*.

Figure 2.3 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 Figure 2.2. 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 backtrack is possible, the shuttle does not go back for request c although it just passed the corresponding location, 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.

2.2 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 the shuttle 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 survey of selected contributions on other demand-responsive systems which provide interesting approaches to tackle the hybrid public transportation systems.

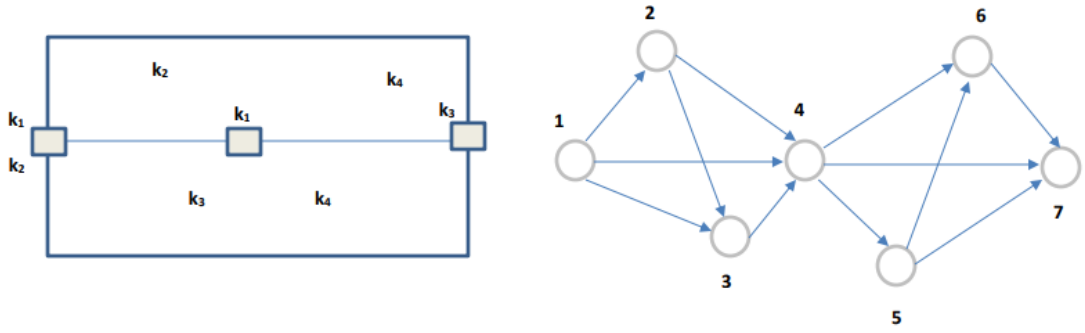


Figure 2.4: Example of a simple FRT instance: service area checkpoints and request (left) and corresponding graph (right).

2.2.1 Optimization algorithms

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.

The first paper modeling and solving the FRT [Quadrifoglio and Dessouky, 2004] does so by proposing a Mixed-Integer Linear Programming (MILP) formulation. As the model in Quadrifoglio and Dessouky [2004] serves as a basis for subsequent publications, we report it in the following. It considers a single shuttle performing multiple trips between terminals. In order to model the problem, the authors introduce a graph structure $G = (N, A)$. Nodes in $N = N_0 \cup N_n$ represent all checkpoints (N_0) and requested ad hoc (N_n) stops. If multiple stops must be performed at the same location, for example at different times, as many nodes are created. This is the case, e.g., for checkpoints and depots, which all have to be visited once per trip. Let o and d be the origin and destination terminals ($o, d \in N_0$). Set A includes a directed arc for each pair of nodes such that the shuttle is authorized to move from one to another. The distance between arc $(i, j) \in A$ is presented by $d_{i,j}$. The rectilinear travel time from node i to node $j, \forall i, j \in N$, is $\delta_{i,j} = d_{i,j}/b_i$, where b_i is the shuttle speed. The set of requests is named K. For each request k, let $o(k), d(k) \in N$ be the corresponding pick-up and drop-off nodes.

Two additional quantities are associated to nodes. Namely, b_i is the time needed for boarding and disembarking at node $i, \forall i \in N$, while θ_i is the scheduled departure time if $i \in N_0$ and the ready time of customer k is defined as τ_k. A feasible solution of the FRT is a path on G, linking origin to destination.

For example, consider the instance represented in Figure 2.4 (left). Here, we have four requests (k_1 to k_4) and three checkpoints including the origin and destination terminals. Assume that shuttle and customers travel from left to right. For each request, the first occurrence in the figure represents the pick-up location while the second is the drop-off. The corresponding graph is represented in Figure 2.4 (right). It includes seven nodes, where nodes 1, 4, 7 are checkpoints ($N_0 = \{o = 1, 4, d = 7\}$), and the remaining ones are ad hoc stops ($N_n = \{2, 3, 5, 6\}$). The pick-up and drop-off requests from k_1 to k_4 are presented as pairs of nodes, namely: (1,4), (1,2), (3,7) and (5,6), respectively.

The decision variables are chosen as follows:

- $x_{ij} = \{0, 1\}$ = binary variables indicating if an arc $(i, j), \forall (i, j) \in A$, is traversed ($x_{i,j} = 1$) or not ($x_{i,j} = 0$).
- $\bar{t}_i$ = arrival time at node $i, \forall i \in N \setminus \{o\}$.

- t_i = departure time from node $i, \forall i \in N \setminus \{d\}$.

The MILP model is as follows. The objective function (2.1) minimizes the weighted sum of three different factors, namely: 1. total travel time of the shuttle; 2. total ride time of all customers; 3. total waiting time at the pick-up stops, defined as the time interval between the customers' ready time and their corresponding pick-up time. The corresponding weights are named w_1 to w_3. This definition allows optimizing both the shuttle variable cost (first term) and the service level (the last two terms): by modifying the weights accordingly, one can emphasize one factor over the others as needed.

$$\min \quad w_1 \sum_{(i,j)\in A} d_{ij}x_{ij} + w_2 \sum_{k\in K} (\bar{t}_{d(k)} - t_{o(k)}) + w_3 \sum_{i\in N_n} (t_i - \theta_i) \tag{2.1}$$

This objective must be optimized subject to:

$$\sum_{i\in N} x_{i,j} = 1 \qquad \forall j \in N \setminus \{o\} \tag{2.2}$$

$$\sum_{j\in N} x_{i,j} = 1 \qquad \forall i \in N \setminus \{d\} \tag{2.3}$$

$$t_i = \theta_i \qquad \forall i \in N_0 \tag{2.4}$$

$$t_{o(k)} \geq \tau_k \qquad \forall k \in K, \tag{2.5}$$

$$\bar{t}_{d(k)} \geq t_{o(k)} \qquad \forall k \in K \tag{2.6}$$

$$\bar{t}_j \geq t_i + x_{i,j}\delta_{i,j} - (1 - x_{i,j})M \qquad \forall (i,j) \in A \tag{2.7}$$

$$t_i \geq \bar{t}_i + b_i \qquad \forall i \in N \setminus \{o,d\} \tag{2.8}$$

$$x_{i,j} \in \{0,1\} \qquad \forall (i,j) \in A \tag{2.9}$$

$$t_i, \bar{t}_i \geq 0 \qquad \forall i \in N \tag{2.10}$$

Network Constraints (2.2) and (2.3) allow each stop (except at node o and d) to have exactly one incoming and one outgoing arc, so that all stops will be performed. Constraints (2.4) force the departure times from the checkpoints to be fixed, while Constraints (2.5) prevent each pick-up at an ad hoc stop from having its departure time earlier than its ready time. Constraints (2.6) prevent the drop-off stop to be scheduled earlier than the pick-up stop for each request. Constraints (2.7) enforce that for each traversed arc (i, j), the arrival time at node j should be no less than the departure time from node i plus the time needed to travel between i and j. The last term ensures that for any $x_{ij} = 0$ the constraint becomes trivially satisfied. These constraints also guarantee that every feasible solution does not contain inner loops, but a single route from node o to node d, visiting all nodes in N only once. Constraints (2.8) link arrival time and departure time for each stop i in the graph.

In Quadrifoglio and Dessouky [2004], the MILP formulation solved through a commercial solver is efficient enough to find solutions for instances with 25 customers per hour and a time horizon of 50 hours. The distribution of the customers are: 10% of PD, 10% of NPND, 40% of PND and 40% of NPD customers. The paper proposes a comparison between FRT and CPT based on the weighted sum of 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 FRT, while walking time is for 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 weights twice as much as ride time on the shuttle. The total travel time of the shuttle and the passengers' ride time are equally weighted. 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]. Here, the authors explicitly consider the shuttle trips r by adding a new binary variable $z_{k,r}$: it indicates whether the pick-up of a PND customer or the drop-off of an NPD one is scheduled at trip r. Five sets of constraints are added to the MILP as well, allowing exactly one z variable to be equal to 1 for each hybrid request and linking it to arrival and departure times of the shuttle at the corresponding ad hoc stop. 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 first group of valid inequalities links the t (departure times) and z variables of ad hoc stops for hybrid customers, as well as the z and x variables. The second group links x and z variables with θ (ready time). The third group links x and z variables by applying the results from the mentioned observations to pairs of hybrid requests. The model is tested on various sets of experiments, with up to 17 customers per hour and a time horizon of 10 hours. 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 those inequalities improve the linear relaxation greatly and some instances can be solved with a computational time reduced by up to 90%, some of the largest instances still cannot be solved in less than 10 hours of computational time.

A further extension of the model in Quadrifoglio and Dessouky [2004] is investigated in Lu, Xie, Wang and Quadrifoglio [2011], 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], except for the valid inequalities. Here, both x and z variables are indexed also on shuttles. Moreover, three sets of constraints are added to ensure that the pick-up and drop-off of each customer are performed by the same shuttle. 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: it assesses the expected value of the MILP objective function (very similar to (2.1)) considering an uniform probability distribution for ad hoc stop locations over the service area. The analytical model shows that if there are 12 customers per hour or more, 2-FRT achieves better performance. The empirical results of the solution of the MILP formulation indicate that this number is 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 following characteristics of the CPT lines are varied to design the appropriate FRT systems: number of fixed-stops (checkpoints), route length, riding time, 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.

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 alternate 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, so there is a maximum pick-up time for customers, and add a capacity constraint for the 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

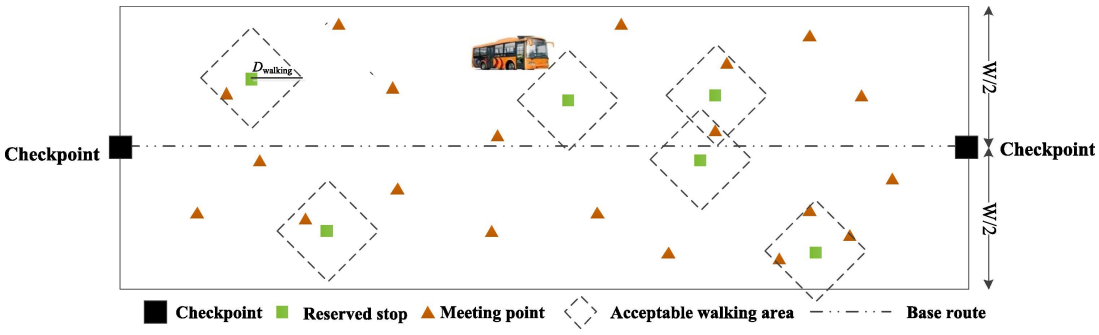


Figure 2.5: Example of the association of requests with meeting points according to the acceptable walking distance of each customer (reproduced from Zheng et al. [2019]).

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 proposed for other types of demand-responsive systems, see, e.g., Daganzo [1984a]. Figure 2.5, 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 variables $y_i = \{0, 1\}, \forall i \in N$, are introduced to indicate if node i is included in the optimal tour where the node set now includes the possible meeting points. Indeed, not all nodes need to be visited depending on whether we choose to use some meeting points or not and these variables are defined to decide if a customer is picked-up/dropped-off at their desired location or at a meeting point. A *cluster* set S_i of all the points which can substitute a non-checkpoint stop request i allows the authors to rewrite the node-visiting constraints (2.2) as:

$$\sum_{j \in N} x_{i,j} = \sum_{j \in N} x_{j,i} = y_i \qquad \forall i \in N, \tag{2.11}$$

where now $N = S_1 \cup ... \cup S_{TC}$ and variables y_i obey the relation:

$$\sum_{i \in S_h} y_i = 1 \qquad h = 1, ..., TC. \tag{2.12}$$

Moreover, constraints (2.7) must be adapted since a given demand can be served at different nodes, for example for two demand nodes $i, j \in \{1, ..., TC\}$ which do not have common meeting points, they become:

$$\bar{t}_j \geq t_i + \sum_{p \in S_i} \sum_{q \in S_j} x_{p,q} \delta_{p,q} - \left(1 - \sum_{p \in S_i} \sum_{q \in S_j} x_{p,q}\right) \theta_d \qquad \forall (i, j) \in A \tag{2.13}$$

The waiting time term in the traditional FRT objective function is replaced by:

$$w_2 \sum_{i \in N} \omega_i y_i \tag{2.14}$$

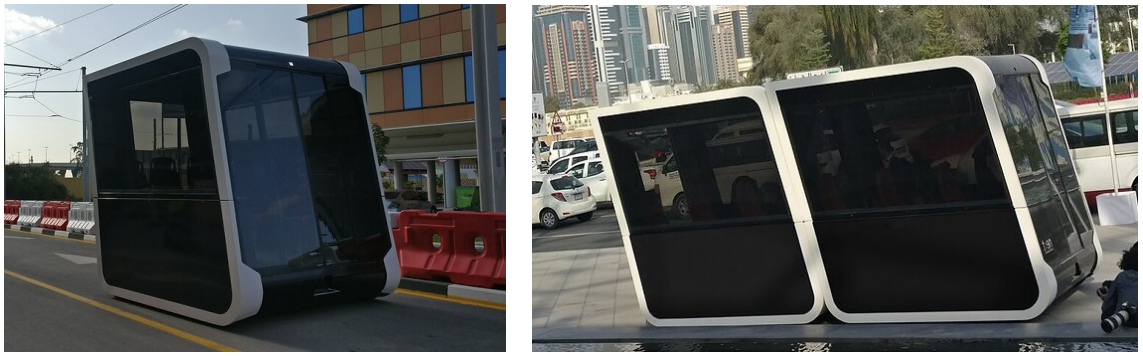


Figure 2.6: Example of a MAV (on the left) and how they can combine into a single vehicle (on the right).

where ω_i is the walking time from node i to the corresponding requested stop $\forall i \in N$. Since the MILP 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 (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 points 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 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. 2.6. 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. 2.7) 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 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 is taken into account in the objective function. The formulation opts for a *time-indexed* approach and introduces a set of departure times T. Many decision variables need to be defined to obtain a linear formulation for that problem, which are classified into three types: the *direct decision* variables include the classic FRT variables similar to the multi-shuttle model of Lu, Xie, Wang and Quadrifoglio [2011] plus a time-indexed binary variable assessing whether a MAV vehicle leaves the terminus at a certain time index and two variables tracking the number of PD pick-ups and drop-offs at each checkpoint; the *intermediate* variables track the number of PD and NPND customers in the vehicles as well as the state of the demand at each time stamp; finally, the *auxiliary* variables are introduced to linearize the model completely. The constraints are divided into *vehicle route and time constraints*, involving the direct decision variables and mainly corresponding to the constraints of existing FRT models, and *constraints associated with passenger assignment*, involving the intermediate and auxiliary variables. Overall the MILP contains a very large number of variables and constraints,

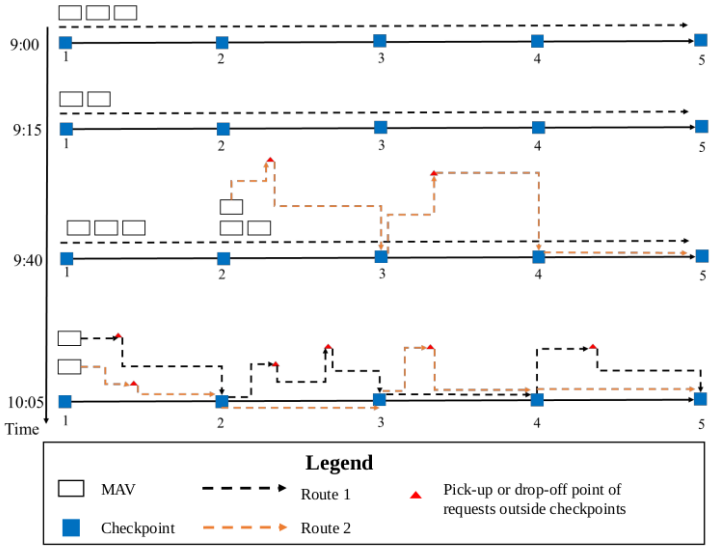


Figure 2.7: Example of four trips using MAVs (reproduced from Liu et al. [2021]).

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 of between 10 and 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.

The work of Yang et al. [2016] tackles the problem from a more strategic perspective and aims at selecting the routes of FRT lines in a specific region. 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., 2010a]), while still serving the main urban centres 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. Some routes (such as those out of our Nashville case) are most cost-effective when they are short, while others (such as those out of their Memphis instance) are most cost-effective when they are lengthy.

Dynamic FRT

In the remainder of this section, we review the works which tackle the dynamic version of the problem where not all the requests are known before the shuttle departure. These approaches mainly use heuristic algorithms.

The first work to consider a heuristic algorithm as well as a backtracking policy for the FRT is the one of Quadrifoglio et al. [2007], which studies a single shuttle performing several trips. 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 hours of time horizon, meaning two days and two hours consecutively. 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, 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 schedule customers by newly inserted ones. The results show that the heuristic schedule is not far from the static optimal one, while the system with FCFS policy is saturated very quickly.

By using the insertion heuristic of Quadrifoglio et al. [2007], Quadrifoglio and Dessouky [2008] performs a sensitivity analysis and assesses 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 hours, 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 are compared with those of a CPT. The comparison is based on the same criteria as in Quadrifoglio and Dessouky [2004] and, as in the first paper, the FRT outperforms the CPT.

The authors of Lu, Lu and Quadrifoglio [2011] use the insertion heuristic algorithm from Quadrifoglio et al. [2007], in order to understand if having a 1-FRT outperforms a 2-FRT in a dynamic environment. Three tests are performed by having 15, 20, and 25 customers per hour, over a time horizon of 50 hours. The system parameters are the same as in the analysis of Lu, Xie, Wang and Quadrifoglio [2011]. In all of the tests, the 2-FRT outperforms the 1-FRT. It is also found that having a 2-FRT reduces the waiting time for customers by half compared to a 1-FRT. 2-FRT is also shown to achieve performances comparable to those of the static MILP.

Google Maps real-time traffic information is used in Qiu, Li and An [2014] to re-schedule shuttle routes if the planned one is blocked or congested. An insertion heuristic is used. Two scenarios are examined, utilizing and omitting real-time traffic information. Little difference is detected during off-peak hours. Instead, during peak hours, the service quality decreases 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.

In Zheng and Li [2019], the authors explore the operational efficiency of a static, partially and fully dynamic environment for the FRT. A MILP formulation similar to the one of Zhang et al. [2021] addresses the initial requests that are known in advance (static). Then, an insertion heuristic is employed to insert dynamic requests. 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, and beyond the value of 75% it is stabilized, while the waiting time keeps increasing. In addition, if all of the requests are known in advance, the system performance increases by 2.6%. Later on, this study is further developed by Zheng et al. [2021] considering requests cancellation and customer no-shows. The no-show and cancellation rates are varied from 0% to 20% with a random distribution. As they increase, the riding time declines, and 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 offline and online demands 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 online demands are inserted one by one with a greedy logic. Numerical studies establish the applicability of FRT in a realistic road network and indicate that FRT 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 Lu, Xie, Wang and Quadrifoglio [2011] and the online requests are handled through an insertion heuristic very similar to the approach of Quadrifoglio et al. [2007]. However, the difference with Quadrifoglio et al. [2007] is the fact that the authors consider the slack time as a (variable) control parameter. Under the same demand circumstances, the FRT system was 5.9% to 10.8% less expensive per customer than CPT. When the demand for the five bus in Harbin's suburbs (China) is 20 to 40 customer per hour, the FRT system is more effective than CPT.

2.2.2 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 typically defines the physical characteristics of the system based on a certain demand level, including line characteristics such as depot and checkpoint stop locations, as well as the shape and size of the service area. These parameters are usually costly and challenging to modify at later stages due to their substantial reliance on physical infrastructure. However, it is important to note that the categorization of vehicle fleet decisions as purely strategic can vary. Transportation companies with extensive vehicle fleets often purchase new vehicles several times a year, allowing them to adapt their fleet to current needs in a relatively fluid manner. This flexibility suggests that fleet management can also be considered a tactical decision, reflecting a more nuanced approach to categorizing system elements based on their operational adaptability and the frequency of reassessment required. At the tactical level, service characteristics are determined, including slack time at checkpoints, trip frequency, or backtracking threshold. It's worth mentioning that the designation of the service area for on-demand stops may straddle the line between strategic and tactical planning. This is because altering the service area can be executed with relative ease at any point without necessitating changes to the physical infrastructure, suggesting a more flexible approach to system design. Other contributions in the field focus on studying the average efficiency of different transport modalities to ascertain the most suitable option under specific circumstances. This underscores the importance of a multi-level planning approach that accommodates the dynamic nature of transportation systems and the varying degrees of flexibility within strategic and tactical decisions.

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/hour 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 2.2.1, 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 2.2.1, 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 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 2.2.1, 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, Li and Zhang [2014], 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/hour). 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, Lu and Quadrifoglio [2011] 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.

2.2.3 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, Li, Zhang, Guo and Zheng [2023] and Yu, Zheng, Li, Zhang, Guo and Wu [2023]. 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.

2.2.4 Related hybrid transportation 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 2.2.2 show that analytical equations have been used to tackle 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 FRT. However, the customers cannot request a stop at a location of their choice but instead may request the use of a pre-determined optional stop. An earlier contribution on the subject of choosing the optional stops to make available to the 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 in the case their preferred optional stop is not selected. The passenger circulation is modeled through flows and the number of optional stops between two compulsory stops is usually limited to 1. Another contribution, i.e. Crainic et al. [2012], concentrates 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 at deciding 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 the master schedule as defined earlier. The mathematical model includes binary variables for selecting the compulsory stops and their schedule as well 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 specific heuristic hierarchical decomposition methods. The method 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.

2.3 Summary of the literature review

Table 2.1 provides a summary of the main aspects characterizing the different papers available on the FRT.

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

Table 2.1: Different aspects characterizing FRT studies in the literature. 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).

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

Table 2.2: Distribution of the reviewed studies by the sensitivity analyses conducted.
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).

Several studies conduct some form of partial sensitivity analysis on the system parameters. Table 2.2 provides a summary of such analysis.

2.4 Similarities and differences between the FRT and VRP

Given the similarities between the FRT and other transport problems, it is possible to get inspiration from the algorithms proposed for these problems to tackle the research gaps identified in Section 2.3. In this section, we propose some directions that may be explored to do so.

The main such problem is the VRP, which is of great importance in the sectors of transport and logistics. This motivates a large body of research on efficient algorithms for many VRP variants which need to be solved in real life situations. These algorithms can either be exact algorithms, usually fit for small to medium instances, or (meta-)heuristics. We refer to the survey of Konstantakopoulos et al. [2022]; Braekers and Kovacs [2016] for a general overview of the literature on the VRP.

One of the well-known sub-categories of VRP is the PDPTW. This specific version of the VRP has common features with the FRT. In a PDPTW, 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 PDPTW exist in the literature. The version of the PDPTW we are interested in is usually labelled as the *one-to-one* PDPTW. 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 PDPTW 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 PDPTW 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 PDPTW 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 PDPTW 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 it emerges from the reported literature review, the FRT is often defined as a single shuttle problem, however it is a multi-trip problem and can naturally be compared with the multi-vehicle PDPTW/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. have to pick-up/drop-off all of the 'customers' or 'goods'; 5. want 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 2.1, 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 transportation 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. As we will see in Section 2.5, this type of features can sometimes impact significantly the efficiency of a given algorithm.

2.5 Pick-up and Delivery Problem with Time Windows and Dial a Ride Problem

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

There are many approaches have been used to solve the VRP with exact methods. We direct readers to Ho et al. [2018]; Moghdani et al. [2021] for comprehensive details on various methodologies. In this context, we opt to focus on the Branch-and-Price and Branch-and-Cut methods as employed in literature since their inception.. The former was first proposed for the PDPTW 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. Indeed, 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 PDPTW 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. In order to exemplify the similarities with the FRT problem, we reproduce here the model proposed by Cordeau [2006], introducing first the sets and parameters:

- $P = \{1, ..., n\}$: pick-up nodes.
- $D = \{n+1, ..., 2n\}$: drop-off nodes.
- 0 = origin terminal.
- $2n+1$ = destination terminal.
- $N = P \cup D \cup \{0, 2n+1\}$: set of all nodes.
- $A = \{(i,j) : i, j \in N\}$: set of arcs between the different nodes.
- q_i = load of the request at node $i \in N$, with $q_0 = q_{2n+1} = 0$ and $q_i = -q_{n+i}$ for $i \in \{1, ..., n\}$.
- d_i = non-negativity service duration at node $i \in N$ with $d_0 = d_{2n+1} = 0$.
- e_i = earliest time, at which service may begin at node $i \in N$.
- l_i = latest time, at which service may begin at node $i \in N$.
- c_{ij} = routing cost between nodes $i, j \in N$.
- t_{ij} = travel time between nodes $i, j \in N$.
- L = the maximum allowed riding time of a user.
- K = set of shuttles. Each shuttle $k \in K$ has a capacity of Q_k and the total duration of its route cannot exceed T_k.
- $G = (N, A)$ where $N = P \cup D \cup \{0, 2n+1\}, P = \{1, ..., n\}$ and $D = \{n+1, ..., 2n\}$.
- Q_k: capacity of shuttle $k \in K$.
- T_k: maximum duration of the route of shuttle $k \in K$.

The variables adopted for the model are the following:

- $x_{ij}^k = 1$ if the shuttle $k \in K$ travels from node $i \in P$ to node $j \in P$.
- B_i^k = time at which shuttle $k \in K$ begins service at node $i \in N$.

- Q_i^k = load of shuttle $k \in K$ after visiting node $i \in N$.
- L_i^k = ride time of user $i \in P$ on shuttle $k \in K$.

The model itself is defined as follows:

$$\min \quad \sum_{k \in K} \sum_{i \in N} \sum_{j \in N} c_{ij}^k x_{ij}^k \tag{2.15}$$

Subject to:

$$\sum_{k \in K} \sum_{j \in N} x_{ij}^k = 1 \qquad \forall i \in P \tag{2.16}$$

$$\sum_{j \in N} x_{ij}^k - \sum_{j \in N} x_{n+i,j}^k = 0 \qquad \forall i \in P, k \in K \tag{2.17}$$

$$\sum_{j \in N} x_{0j}^k = 1 \qquad \forall k \in K \tag{2.18}$$

$$\sum_{j \in N} x_{ji}^k - \sum_{j \in N} x_{ij}^k = 0 \qquad \forall i \in P \cup D, k \in K \tag{2.19}$$

$$\sum_{i \in N} x_{i,2n+1}^k = 1 \qquad \forall k \in K \tag{2.20}$$

$$B_j^k \geq B_i^k + d_i + t_{ij} - M_{ij}^k (1 - x_{ij}^k) \qquad \forall i \in N, j \in N, k \in K \tag{2.21}$$

$$Q_j^k \geq Q_i^k + q_j - W_{ij}^k (1 - x_{ij}^k) \qquad \forall i \in N, j \in N, k \in K \tag{2.22}$$

$$L_i^k = B_{n+i}^k - (B_i^k + d_i) \qquad \forall i \in P, k \in K \tag{2.23}$$

$$B_{2n+1}^k - B_0^k \leq T_k \qquad \forall k \in K \tag{2.24}$$

$$e_i \leq B_i^k \leq l_i \qquad \forall i \in N, k \in K \tag{2.25}$$

$$t_{i,n+i} \leq L_i^k \leq L \qquad \forall i \in P, k \in K \tag{2.26}$$

$$\max\{0, q_i\} \leq Q_i^k \leq \min\{Q_k, Q_k + q_i\} \qquad \forall i \in N, k \in K \tag{2.27}$$

$$x_{ij}^k \in \{0, 1\} \qquad \forall i \in N, j \in N, k \in K \tag{2.28}$$

The objective function (2.15) minimizes the total routing cost, which corresponds to the first of the FRT objectives. Constraints (2.16) and (2.17) ensure that each request is served exactly once and that the pick-up and delivery nodes are visited by the same shuttle. Constraints (2.17) - (2.20) guarantee that the route of each shuttle k is connected, starts at the origin terminal and ends at the destination terminal. The consistency between the time variables and load variables is respectively ensured by Constraints (2.21) and (2.22). The ride time of each user is defined in Constraints (2.23) and bounded by Constraints (2.26). The latter constraints also act as precedence constraints because the non-negativity of the L_i^k variables ensures that node i will be visited before node $n+i$ for every pick-up request $i \in P$. Finally, Constraints (2.24) bound the duration of each route, while (2.25) and (2.27) impose time windows and capacity constraints, respectively. We can observe right away that (2.16) is similar to (2.2) and (2.3) in the FRT formulation presented in Section 2.2.1 , with the slight difference that (2.16) is only defined for the pick-up nodes P. Constraints (2.21) are akin in spirit and form to Constraints (2.7) of the FRT, even though the departure time variable of the FRT is not present in the above model. We can also partially map Constraints (2.25) with (2.4) and (2.5) with respect to the lower bound of the time windows, since the upper bound is infinite in the FRT. The precedence relation between the pick-up and delivery nodes is modeled in a somewhat different way between both models, since in the DARP it is implied by (2.23) and (2.26) through the use of the ride time variable for the users, while it has the more direct form (2.6) in the FRT. Finally, Constraints (2.17) do not have a strict equivalent in the one shuttle FRT model but have a

direct equivalent in the FRT model for multiple shuttles, see Lu, Xie, Wang and Quadrifoglio [2011]. On the contrary, since in the FRT the shuttle will go several times through the terminal, Constraints (2.18) and (2.20) have no counterpart in the FRT, as well as Constraints (2.22) given that the FRT shuttles have an infinite capacity.

In Cordeau [2006], this model was enriched with several families of complex valid inequalities, most of which are based on the precedence or the capacity constraints. 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 $U_1, ..., U_m \subset N$ are defined such that they do not contain any terminal but each U_l contains both i_l and $n+i_{l+1}$ with $i_1, ..., i_m \in P$. 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.

It is important to note that there are numerous other types of inequalities, although our focus was limited to only a select few. 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 PDPTW, 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 is important since the linear relaxation of the base MILP for the FRT problem is not very tight, due to 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 hours 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 PDPTW valid inequalities is a more open question, which requires careful investigation given the differences with the FRT as sketched in Section 2.4. 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 PDPTW, many such approaches exist. Unfortunately, existing neighbourhood 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 neighbourhood 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 PDPTW offers many insights and tools to tackle extensions of the FRT systems, 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])

2.6 Conclusion

The FRT is very promising for meeting the challenges of mobility. In this chapter, 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. In the next chapters, we aim at filling some of the gaps that remain open in the current literature.

Chapter 3

A formulation with new valid inequalities and warm-starting for the multi-shuttle FRT

3.1 Introduction

In this PhD, the main tool for solving optimization problems linked to FRT systems is the use of Mixed-Integer Linear Programming (MILP) formulations, solved by a commercial solver. In this chapter, we will propose a MILP formulation to solve the FRT problem with a fleet of shuttles, largely based on the contribution of [Lu, Xie, Wang and Quadrifoglio, 2011]. In order to extend both the scope and the efficiency of their formulation, we will add a capacity constraint and a set of valid inequalities to strengthen the value of the linear relaxation of the model. The seminal works in the literature introducing a MILP formulation for the canonical FRT system are:

- The MILP formulation by Quadrifoglio and Dessouky [2004] which addresses a system configuration involving a single shuttle and a single vehicle.
- The work by Quadrifoglio et al. [2008a], extending the MILP formulation to accommodate a multi-trip system.
- The formulation by Lu, Xie, Wang and Quadrifoglio [2011], which is the most comprehensive, covering both multi-trip and multi-shuttle configurations.

Furthermore, we contribute a set of valid inequalities to enhance the performance of the MILP. To benchmark the effectiveness of our proposed valid inequalities, we implement and adapt the most efficient group of valid inequalities from Quadrifoglio et al. [2008a], originally designed for single-vehicle systems, to function in multi-shuttle configurations.

For large-scale instances where the MILP may Prove it too strong, a greedy heuristic is developed to obtain a feasible solution to be used as a warm start for the linear solver.

The chapter is structured as follows. Section 3.2 provides a short review of the relevant literature. Section 3.3 details the base mathematical formulation of the problem. Section 3.4 delves into valid inequalities used to accelerate the computational resolution, focusing on existing (Section 3.4.1) and new ones (Section 3.4.2). Section 3.5 presents the logic of the heuristic algorithm used for warm starting the MILP and Section 3.6 discusses the warm start methodology. Section 3.7 presents the empirical results of several formulations, with and without valid inequalities and warm start, in order to demonstrate the

effect of the different available ingredients to strengthen the FRT formulation. Section 3.8 concludes the chapter, summarizing the key contributions and findings. By rigorously evaluating the performance of our model and comparing it with existing MILP formulations and heuristics, this chapter aims to provide a comprehensive framework for addressing capacity constraints and enhancing the efficiency of the system through valid inequalities.

3.2 Literature Review

After providing the literature review in-depth in Chapter 2, here we only summarize papers that employ optimization approaches in order to address the problems in more details.

In the MILP formulation provided by Quadrifoglio and Dessouky [2004], the problem is addressed for the first time in the literature. The MILP formulation is provided for a single trip and is compared to a CPT system. However, as the model is developed for a single trip version of the problem, it is not very challenging for CPLEX to solve the problem to optimality. A more advanced MILP formulation of the problem is addressed in Quadrifoglio et al. [2008a], where the multi-trip and a more realistic version of the problem is taken into account. Indeed, when a multi-trip version of the problem is considered, it requires a more structured graph, making it more difficult for CPLEX to find the best position to insert customers in different trips. In the MILP formulation of Quadrifoglio et al. [2008a], the main hard constraints are the scheduled departure time from each checkpoint at each trip. It means that if there are 10 requests between two consecutive checkpoints (let us call them 'a' and 'b'), the scheduled departure time of checkpoint b is set in a way that if the shuttle wants to pick-up more than five customers, the scheduled departure time of checkpoint b will be violated, making the problem infeasible. Using the graph structure corresponding to the problem instance, a linear solver can choose which customers should be assigned to a later trip in order to avoid infeasibility. However, as the FRT system is originally designed for a night shift and it is expected that during the night the demand will be low, no capacity constraint is considered when modeling the problem, making it an unrealistic problem when transport authorities want to implement such a transit system for the day shift. The model is strengthened by three groups of valid inequalities. Their results show that the first group is the most efficient one in terms of reducing the computation time. In Lu, Xie, Wang and Quadrifoglio [2011], the same research group develops further the MILP formulation of Quadrifoglio et al. [2008a] to model a multi-shuttle system while still neglecting the finite capacity of the shuttle. In Zheng et al. [2019], some meeting points are available within the service area, where customers can be grouped to be picked up or dropped off if such meeting points are within their acceptable walking distance. In Zhang et al. [2021], the authors modify the problem of Quadrifoglio and Dessouky [2004] by adding time windows constraints, so there is a maximum pick-up time for customers, and add a capacity constraint for the shuttles. The scheduling of customers is done heuristically on a First-Come First-Served basis. 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). A complex MILP formulation is developed to address the problem. Given the small size of MAVs, it is necessary to introduce 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 is taken into account in the objective function. Since some customers have pick-up and/or drop-off stops located at checkpoints and they can be assigned to different possible trips, the corresponding pick-up or drop-off node of those customers is not known in advance. As a consequence, it is not straightforward to determine the load at checkpoint stops. Liu et al. [2021] solve this problem by introducing a specific node for each relevant customer at checkpoint stops instead of using the same graph node for all customers. Later in this chapter, we will adopt a different view and keep a single graph node corresponding to a physical stop at a checkpoint.

In order to study the behavior of an FRT system with limited shuttle capacity compared to the level of demand, we chose to implement the model of Lu, Xie, Wang and Quadrifoglio [2011], which is one of the most advanced existing MILP in the literature for classic FRT systems, and we adopt and implement

References	Year	# of shuttles	Capacity	# of Customers	Methodology
Quadrifoglio and Dessouky [2004]	2004	1	×	25	MILP
Quadrifoglio et al. [2008a]	2008	1	×	17	MILP + valid ineq.
Lu, Xie, Wang and Quadrifoglio [2011]	2011	1-2	×	14	MILP
Zheng et al. [2019]	2019	1	×	25	MILP
Zhang et al. [2021]	2021	1	50	50	MILP + Heuristic
Liu et al. [2021]	2021	10-30	10	45	MILP + Heuristic
Current study	2023	1	15	25-100	MILP + warm start + valid ineq.

Table 3.1: Summary of the literature.

the first group valid inequalities of Quadrifoglio et al. [2008a] in order to reach optimality in a shorter amount of time. Moreover, we model the capacity constraints in a different way with respect to what is done in the literature. In the literature, the papers which consider finite capacity tend to fix it so that the shuttle capacity is as large as the number of served customer Zhang et al. [2021]; Liu et al. [2021]: the capacity constraints are not really saturated. In order to tackle instances with a larger number of customers in a reasonable amount of computation time, we also propose several new valid inequalities after motivating their necessity given the model structure and weaknesses.

3.3 Mathematical Formulation

Below, we introduce a large body of notation needed to model a complex transportation system such as the FRT. The FRT problem, similarly to other transportation problems like the VRP, is modeled through the introduction of a graph structure where each stop of the vehicle corresponds to one node and an arc is present between two nodes if the vehicle can reach the second node from the first. A difference with the complete graph structure often found in models for many variants of the VRP, given the policy that FRT shuttles cannot go back along the longitudinal axis of the service area, many arcs are not present in the problem graph. A specificity of the FRT is also that the shuttles perform several trips, so that they will stop several times at the same physical checkpoint. Therefore, one graph node is created for each stop at such fixed checkpoints.

Table 3.2: Notations of the MILP formulation

Notations	**Description**
Sets	
$\mathcal{K}_{\mathcal{PD}}$	Set of PD requests.
$\mathcal{K}_{\mathcal{PND}}$	Set of PND requests.
$\mathcal{K}_{\mathcal{NPD}}$	Set of NPD requests.
$\mathcal{K}_{\mathcal{NPND}}$	Set of NPND requests.
$\mathcal{K}$	Set of all requests.
$\mathcal{HYBR}$	Subset of hybrid customers with one checkpoint and one non-checkpoint stop, i.e., $HYBR = K_{NPD} \cup K_{PND}$.
N_0	Set of checkpoints stops.
N_n	Set of non-checkpoint stops.

Continued on next page

Table 3.2 – continued from previous page

Notations	Description
$\mathcal{N} = N_0 \cup N_n$	Set of all stops.
$\mathcal{A}$	Set of all arcs.
R	Last trip
A_n	Allowed arcs (i, j) with $i, j \in N_n$ and $i \neq j$.
$A_{n,0}$	Arcs $(i, j) \in A$ with $i \in N_n, j \in N_0 \setminus \{1\}$.
$A_{0,n}$	Arcs $(i, j) \in A$ with $i \in N_0 \setminus \{TC\}, j \in N_n$, where TC is the total number of stops at checkpoints in the schedule
$\mathcal{RD} = \{1, ..., R\}$	Set of trips.
$HYBR(k) \subset RD$	Feasible trips of k, $\forall k \in KHYB$.
Parameters	
R	Number of trips.
C	Number of checkpoints.
V_e	Number of vehicles.
V	Set of vehicles.
TC_0	$(C-1) \times R + 1$ = total number of stops at checkpoints in the schedule for one vehicle.
TC	$TC_0 \times V$ = total number of stops at checkpoints in the schedule.
TS	$TC + \lvert K_{PND}\rvert + \lvert K_{NPD}\rvert + 2\lvert K_{NPND}\rvert$ = total number of stops.
$o(r)$	Origin depot in trip r
$d(r)$	Destination depot in trip r
H_k	The number of customers k
θ_i	Scheduled departure time of checkpoint stop $i, \forall i \in N_0, (\theta_1 = 0)$.
τ_k	Ready time of request k, $\forall k \in K$.
$t_{o(k)}$	pick-up time of customer k
$t_{d(k)}$	drop-off time of customer k
$\delta_{i,j}$	Rectilinear travel time between i and j, $\forall (i, j) \in A$.
b_i	Service time for stop $i \in N$.
$pc(k, r)$	The checkpoint stop of customer $k \in K_{PND}$ if the customer is served at trip $r \in RD$.
$dc(k, r)$	The checkpoint stop of customer $k \in K_{NPD}$ if the customer is handled at trip $r \in RD$.
$w_1/w_2/w_3$	Objective function weights.
v_r	Vehicle handling trip $r \in RD$.
o_v, $d_v \in N_0$	Origin of first trip and destination of last trip of shuttle $v \in V$.
Variables	
$x^v_{i,j}$	Binary variables equal to 1 if arc $(i, j) \in A$ is used by vehicle $v \in V$, 0 otherwise.
$\bar{t}_i$	Arrival time at stop $i, \forall i \in N \setminus \{1\}$.
t_i	Departure time from stop $i, \forall i \in N$.
p_k	Pick-up time of request k, $\forall k \in K$.
d_k	Drop-off time of request k, $\forall k \in K$.
Q^v_i	Residual capacity of vehicles $v \in V$ when leaving node $i \in N$.
q_i	represent the net change in load at node $i \in N$ (can be positive or negative).
$z_{k,r}$	Binary variable equal to 1 if the checkpoint stop of hybrid request $k \in K_{PND} \cup K_{NPD}$ is scheduled in trip $r \in RD$ of vehicle $v \in V$

Our formulation is largely based on the one of Lu, Xie, Wang and Quadrifoglio [2011], modulo a difference on the way we model the shuttles trips. Where Lu, Xie, Wang and Quadrifoglio [2011] considers that each shuttle v performs R trips numbered from 1 to R, we consider that each trip has a specific, unique number. Hence, in our model, the FRT system performs R trips globally, distributed among the V

$(V = \{1, \ldots, |V|\})$ shuttles and each trip $r \in RD$ is assigned a specific shuttle v_r. Compared to Lu, Xie, Wang and Quadrifoglio [2011], we also introduce additional variables to handle the finite shuttle capacity: Q_v^i, the residual capacity of each shuttle at each node (similar to the way that capacity constraints are usually modeled in capacitated VRP or DARP) and q_i, the load at each node, which depends on the customers handled at the specific node and is a variable in the FRT. The formulation is detailed below:

$$\min \quad w_1 \sum_{v \in V} \sum_{(i,j) \in A} (\delta_{i,j} x_{i,j}^v) + w_2 \sum_{k \in K} (d_k - p_k) + w_3 \sum_{k \in K} (p_k - \tau_k) \tag{3.1}$$

Subject to:

$$\sum_{v \in V} \sum_{i} x_{i,j}^v = 1 \qquad \forall j \in N \setminus \cup_{v \in V} \{o_v\} \tag{3.2}$$

$$\sum_{v \in V} \sum_{j} x_{i,j}^v = 1 \qquad \forall i \in N \setminus \cup_{v \in V} \{d_v\} \tag{3.3}$$

$$\sum_{i} x_{i,j}^v = \sum_{i} x_{j,i}^v \qquad \forall j \in N \setminus (\{o_v\} \cup \{d_v\}), v \in V \tag{3.4}$$

$$t_i = \theta_i \qquad \forall i \in N_0 \tag{3.5}$$

$$p_k = t_{o(k)} \qquad \forall k \in K \setminus K_{PND} \tag{3.6}$$

$$d_k = \bar{t}_{d(k)} \qquad \forall k \in K \setminus K_{NPD} \tag{3.7}$$

$$\sum_{r \in RD} z_{k,r} = 1 \qquad \forall k \in K_{HYB} \tag{3.8}$$

$$p_k \geq t_{pc(k,r)} - M(1 - z_{k,r}) \qquad \forall k \in K_{PND}, r \in HYBR(k), \tag{3.9}$$

$$p_k \leq t_{pc(k,r)} + M(1 - z_{k,r}) \qquad \forall k \in K_{PND}, r \in HYBR(k), \tag{3.10}$$

$$d_k \geq \bar{t}_{dc(k,r)} - M(1 - z_{k,r}) \qquad \forall k \in K_{NPD}, r \in HYBR(k), \tag{3.11}$$

$$d_k \leq \bar{t}_{dc(k,r)} + M(1 - z_{k,r}) \qquad \forall k \in K_{NPD}, r \in HYBR(k), \tag{3.12}$$

$$p_k \geq \tau_k \qquad \forall k \in K \tag{3.13}$$

$$d_k \geq p_k \qquad \forall k \in K \tag{3.14}$$

$$\bar{t}_j \geq t_i + \sum_{v \in V} x_{i,j}^v \delta_{i,j} - M(1 - \sum_{v \in V} x_{i,j}^v) \qquad \forall (i,j) \in A \tag{3.15}$$

$$t_i \geq \bar{t}_i + b_i \qquad \forall i \in N \setminus \{0, TC_0, 2TC_0, \ldots (v_e - 1)TC_0\} \tag{3.16}$$

$$\sum_{j} x_{ps(k),j}^v - \sum_{j} x_{j,ds(k)}^v = 0 \qquad \forall k \in K_{PD} \cup K_{NPND}, v \in V \tag{3.17}$$

$$\sum_{r \in HYBR: v = v_r} \sum_{j} x_{pc(k,r),j}^{v_r} - \sum_{j} x_{j,ds(k)}^v = 0 \qquad \forall k \in K_{PND}, v \in V \tag{3.18}$$

$$\sum_{j} x_{ps(k),j}^v - \sum_{r \in HYBR(k): v = v_r} \sum_{j} x_{j,dc(k,r)}^{v_r} = 0 \qquad \forall k \in K_{NPD}, v \in V \tag{3.19}$$

$$d_k - p_k \leq (\theta_{d(r)} - \theta_{o(r)}) + M(1 - z_{k,r}) \qquad \forall k \in K_{HYB}, r \in HYBR(k) \tag{3.20}$$

$$Q_j^v \geq (Q_i^v + q_j) - Q(1 - x_{ij}^v) \qquad \forall (i,j) \in A, r \in RD, v \in V \tag{3.21}$$

$$q_i = \sum_{\substack{k \in K_{PD}: \\ ps(k) = i}} H_k + \sum_{r \in RD} \sum_{\substack{k \in K_{PND}: \\ pc(k,r) = i}} z_{k,r} H_k - \sum_{r \in RD} \sum_{\substack{k \in K_{NPD}: \\ dc(k,r) = i}} z_{k,r} H_k - \sum_{\substack{k \in K_{PD}: \\ ds(k) = i}} H_k \qquad \forall i \in N_0 \tag{3.22}$$

$$q_{ps(k)} = H_k \qquad \forall k \in K_{NPD} \cup K_{NPND} \tag{3.23}$$

$$q_{ds(k)} = -H_k \qquad \forall k \in K_{PND} \cup K_{NPND} \tag{3.24}$$

The objective function (3.1) has three components. The first component minimizes the distance traveled by vehicles. The second and third components minimize the riding times and waiting times of customers, respectively. Constraints (3.2) and (3.3) impose to have one incoming arc and outgoing arc from each node except for the origin and destination depots, respectively. Constraints (3.4) force the same vehicle to enter and leave a given node. Constraints (3.5) set the scheduled departure time at each checkpoint. Constraints (3.6) set the pick-up time of all customers apart from the PND type as the departure time of that node, while Constraints (3.7) set the drop-off time of all customers except for the NPD type as the arrival time at that node. The Constraints (3.6) and (3.7) are not valid for PND and NPD customers, respectively. The reasoning behind this is that the checkpoint stops of these type of customers are taken into account from Constraints (3.9) to (3.12) by $z_{k,r}$ binary variable as we do not know a priori in which trips they are assigned to. Constraints (3.8) ensures that there will be one and only one pick-up for PND customers and one and only one drop-off for NPD customers among their possible trips. Possible trips for each non-PD customer are defined as a set of trips starting from the first trip they can be assigned to, until the last performed trip by last vehicle. Constraints (3.9) and (3.10) set the pick-up time of PND customers, while Constraints (3.11) and (3.12) set the drop-off time of NPD customers. Constraints (3.13) ensure that the pick-up time of each customer is equal to or larger than the ready time (issued request time) of each customer. Constraints (3.14) ensure that the drop-off time of each customer is larger than their pick-up time. Constraints (3.15) compute the arrival time at node j based on the departure time from node i plus the time it takes to travel from i to j, should the shuttle use that arc on the transportation graph. Constraints (3.16) ensure that the departure time at each node is larger or equal to the arrival time of that node plus a minimum stop time required for pick-up or drop-off of customers. Constraints (3.17) - (3.19) ensure that if a customer is picked-up by a specific vehicle it has to be dropped-off by the same vehicle as well. The next constraints are new constraints added in order to generalize the existing model of Lu, Xie, Wang and Quadrifoglio [2011], mainly to handle the finite capacity of the shuttle. Constraints (3.20) ensure that if a customer is picked-up in a trip, the customer has to be dropped-off in the same trip. Constraints (3.21) are capacity constraints computing the remaining capacity of each vehicle v when leaving node j, taking into account the load at that node. Constraints (3.22) ensure that the load q_j at node $j \in N$ is computed correctly, depending on the customers to be picked up or dropped off at that node. q_j can take positive or negative values depending on the number of customers to pick up or drop off: for checkpoint stops, this number depends on the customers who are actually handled at the corresponding trip of the checkpoint stop (which is decided by the value of the binary variables z).

3.4 Valid inequalities

Like many formulations for routing and scheduling problems, the linear integer formulation for the FRT has a very weak linear relaxation, with a very large gap compared to the optimal solution. This feature has a very negative impact on the resolution of the linear integer problem given that it is very hard to prune efficiently in the branching tree. It is therefore not surprising that valid inequalities have been proposed to try to strengthen the formulation, in [Quadrifoglio et al., 2008a]. In the Section 3.4.1, we provide details about the most efficient group of inequalities proposed by Quadrifoglio et al. [2008a] as we will implement and test them later. Subsequently, in Section 3.4.2, we provide an analysis of why the formulation is weak and propose specific valid inequalities to counter said weaknesses.

3.4.1 Existing valid inequalities

Our primary contribution to the field of MILP for the FRT lies in the adaptation and extension of a set of valid inequalities initially formulated for single-shuttle systems. We have successfully expanded their applicability to multi-shuttle systems, thereby broadening their utility in a significant manner. These inequalities were originally designed to enhance the computational efficiency of MILP models by narrowing the feasible solution space, a property that remains intact in our adapted version. These constraints serve a dual purpose. Firstly, they act as valid inequalities, meaning they reduce the dimensions of the relaxed feasible region without excluding any integer feasible solutions. This attribute accelerates the solver journey to optimality. Secondly, a subset of these inequalities functions as "logic cuts", aiming to exclude integer feasible solutions that are provably non-optimal, based on specific logical considerations. The efficacy of these inequalities is closely tied to the weights in the objective function. For instance, their effectiveness assumes that customers would prefer to wait for pick-up rather than spend extended time on the shuttle. This is particularly relevant in multi-shuttle systems where customers are aware of the expected pick-up and drop-off times.

Below we reproduce the most efficient valid inequalities of [Lu, Xie, Wang and Quadrifoglio, 2011]. The logic of this group of inequalities is that when the objective weight of the riding time is at least as large as the weight associated to the waiting time, it is always beneficial to reduce the riding time of customers as much as possible. As a consequence, it is clear that since hybrid customers have one stop associated to a checkpoint, where the shuttle has to stop in any case and pick them up or drop them off at no additional cost, these customers will be picked up and dropped off during the same trip. It is important to note that trip $r+1$ is the subsequent trip following trip r, whereas trip $r-1$ is the preceding one. For PND customers, this translates into the following valid inequalities:

$$t_{ds(k)} \leq z_{k,r}\theta_{pc(k,r+1)} + M(1 - z_{k,r}) \qquad \forall k \in K_{PND}, r \in RD \setminus \{R\} \tag{3.25}$$

$$x_{ds(k),j} \leq z_{k,r} \qquad \forall pc(k,r) \leq j \leq pc(k,r+1), k \in K_{PND}, r \in RD \setminus \{R\}, (ds(k), j) \in A_{n,0} \tag{3.26}$$

$$x_{i,ds(k)} \leq z_{k,r} \qquad \forall pc(k,r) \leq i \leq pc(k,r+1), k \in K_{PND}, r \in RD \setminus \{R\}, (i, ds(k)) \in A_{0,n} \tag{3.27}$$

where Constraints (3.25) force the drop-off stop $ds(k)$ of customers $k \in K_{PND}$ to be planned before the next occurrence in the schedule of the checkpoint chosen as the pick-up. If $z_{k,r} = 1$ the k is picked up at checkpoint $pc(k,r)$ in trip r and the constraint imposes that $ds(k)$ has to be scheduled before $pc(k,r+1)$ by setting an upper bound on the departure time $t_{ds(k)}$. Constraints (3.26) ensure that if $z_{k,r} = 1$, then $ds(k)$ must be scheduled between $pc(k,r)$ and $pc(k,r+1)$ and all arcs originating from $ds(k)$ and ending at a checkpoint j cannot exist whenever j is not included in that interval. These arcs would in fact infeasibly require the vehicle to go from $ds(k)$ to a checkpoint scheduled before its pick-up $pc(k,r)$ or to skip $pc(k,r+1)$ going directly from $ds(k)$ to a checkpoint scheduled after $pc(k,r+1)$. Constraints (3.27), ensure that all arcs originating from a checkpoint i and ending at $ds(k)$ are eliminated whenever i is not included in the interval $pc(k,r), pc(k,r+1)$ identified by $z_{k,r} = 1$. A similar (symmetric) logic applies for NPD customers, which provides the following valid inequalities:

$$t_{ps(k)} \geq z_{k,r}\theta_{dc(k,r-1)} - M(1 - z_{k,r}) \qquad \forall k \in K_{NPD}, r \in RD \setminus \{1\} \tag{3.28}$$

$$x_{i,ps(k)} \leq z_{k,r} \qquad \forall dc(k,r-1) \leq i \leq dc(k,r), k \in K_{NPD}, r \in RD \setminus \{1\}, (i, ps(k)) \in A_{0,n} \tag{3.29}$$

$$x_{ps(k),j} \leq z_{k,r} \qquad \forall dc(k,r-1) \leq j \leq dc(k,r), k \in K_{NPD}, r \in RD \setminus \{1\}, (ps(k), j) \in A_{n,0} \tag{3.30}$$

3.4.2 New valid inequalities for the FRT

We can define the slack time σ_i at each checkpoint node $i \in N_0$ and observe that an NPD customer will arrive at his drop-off checkpoint node at the minimum time $\theta_{dc(k,r)} - \sigma_{dc(k,r)}$ if assigned to trip r. We can therefore propose a valid inequality for d_k :

$$d_k \geq \sum_{r \in HYBR(k)} (\theta_{dc(k,r)} - \sigma_{dc(k,r)}) z_{k,r}, \qquad \forall k \in K_{NPD}. \tag{3.31}$$

Since a strong problem of the linear relaxation is that it manages to obtain 0 customers' riding time, we can also try to force it to *at least* take into account the minimum time needed to reach the drop-off point of the customer from the pick-up point. Let us define $\delta_k^{\min}$ as the minimum time needed for the shuttle to travel between the pick-up and drop-off points of customer $k \in K$, i.e., in the optimistic case where no other customer must be inserted in between the two stops. We can write the straightforward relation:

$$d_k \geq p_k + \delta_k^{\min}, \qquad \forall k \in K. \tag{3.32}$$

In the specific case with an irregular timetable, i.e., not all trips have the same slack time for a given checkpoint node, we can differentiate the minimum time needed as $\delta_{kr}^{\min}$ for each trip $r \in RD$ and refine the inequality using the expression $\sum_r \delta_{kr}^{\min} z_{k,r}$ instead of $\delta_k^{\min}$. We can also try to compensate for the big-M term of Constraint (3.15) by working directly on the t and $\bar{t}$ variables, at least at the checkpoints. Indeed, given the slack time σ_i at checkpoint node $i \in N_0$, we know by definition that the shuttle cannot arrive more than the slack time value at i before the departure time at said node, i.e.:

$$\bar{t}_i \geq t_i - \sigma_i, \qquad i \in N_0. \tag{3.33}$$

By focusing on hybrid customers, it is possible to further refine the constraints on the pick-up and drop-off times of those customers. The idea is to consider the fixed checkpoints along the shuttle route. By considering the last checkpoint stop before the non-checkpoint pick-up or drop-off stop of hybrid customers and using the variables $z_{k,r}$ which pilot the trip in which those customers are handled, we can write stronger relations for the p_k and d_k variables. Let us introduce the following additional data structure: $lc(k,r) \in N_0$ is defined as the last checkpoint node encountered by the shuttle before reaching the drop-off node d_k (respectively the pick-up node p_k) of $k \in K_{PND}$ (respectively $k \in K_{NPD}$) in trip $r \in RD$. We also introduce $\delta_k^{\min,lc}$ as the minimum time needed to reach the drop-off stop (respectively the pick-up stop) of customer $k \in K_{PND}$ (respectively $k \in K_{NPD}$) from the last previous checkpoint stop. For PND customers, the drop-off point cannot be reached before the departure time of the last previous checkpoint plus the time needed to reach directly said drop-off point:

$$d_k \geq \sum_{r \in HYBR(k)} \theta_{lc(k,r)} z_{k,r} + \delta_k^{\min,lc}, \qquad \forall k \in K_{PND}. \tag{3.34}$$

For NPD customers, the same reasoning applies on the pick-up time:

$$p_k \geq \sum_{r \in HYBR(k)} \theta_{lc(k,r)} z_{k,r} + \delta_k^{\min,lc}, \qquad \forall k \in K_{NPD}. \tag{3.35}$$

3.5 Heuristic algorithm

Another way to help speed up the solution process of linear integer solvers is to pass a (good quality) feasible solution prior to the branch-and-cut process, in order to prune in the branching tree and guide the branching process more effectively. In this section, we introduce a heuristic algorithm that will later be used to pass the obtained feasible solution as a *warm start* to the solver.

The initial phase of the algorithm (algorithm 1) sets the groundwork for route optimization by assuming that no feasible route has yet been identified. It begins by initializing all trips within a set denoted as RD. Following this initialization, customers are assigned to initial trips based on predefined criteria, which could involve factors like proximity, priority, or earliest request time, though the specific criteria are determined by the operational context of the algorithm. Process: This stage does not involve complex logical or iterative processes; rather, it is a straightforward setup where the algorithm prepares the data (trips and customer requests) for the subsequent, more involved stages of optimization.

Algorithm 1 Initialization of Routes

```
Initialize all trips in RD
Assign customers to initial trips in RD based on some criteria
```

Then, algorithm 2 is crucial for ensuring that the trips adhere to the θ condition, which is the scheduled departure time at each checkpoint and cannot be violated. The algorithm iteratively checks each trip for any requests that violate this condition. The steps are as follows:

- Trip Iteration: The algorithm sorts and examines each trip r in RD by the earliest start time, prioritizing the processing order.
- Request Inspection and Adjustment: For each trip, requests are sorted by the largest waiting time to prioritize critical adjustments. If a request's pick-up violates the θ condition it labeled and moved to a 'candidate list,' effectively removing it from the current trip processing queue. This ensures that the trip remains viable by excluding requests that could compromise its feasibility.
- Deferred Processing: After evaluating all requests in a trip, those identified as problematic and moved to the candidate list are then assigned to the next available trip $(r + 1)$, postponing their processing in an attempt to find a feasible routing solution that complies with all constraints.

Algorithm 2 Feasibility Checking for Routes

```
while No feasible route found do
    for each trip r in RD sorted by earliest start time do
        Sort requests in candidate list by largest waiting time
        while no θ condition passed do
            Check each request of current trip sorted by largest waiting time
            if processing of this request violates the θ condition then
                Save this request in candidate list
                Remove this request from current trip request list
            end if
        end while
        for each request k in candidate list sorted by largest waiting time do
            Insert current request k into next trip (r+1) request list
        end for
    end for
end while
```

The objective of third algorithm is to ensure that the load on each vehicle does not exceed its capacity at any point during the trip. This involves a detailed assessment of each trip load and the reallocation of requests to maintain compliance with vehicle capacity limits. The steps are as follows:

- Load Computation and Maximum Load Identification: For every trip r in RD, the algorithm calculates the vehicle load at all nodes and identifies the maximum load for the trip. This step is critical for understanding the demand placed on the vehicle throughout its route.

- Capacity Evaluation and Request Reallocation:
  - The algorithm assesses whether the maximum vehicle load for a trip exceeds the vehicle capacity. If so, it initiates a corrective loop where requests contributing to this overload are identified based on their waiting time.
  - Requests causing an overload condition, particularly at their pick-up indices, are moved to the 'candidate list' for reallocation. This ensures that each trip adheres to the vehicle capacity constraints by adjusting the trip request composition.
  - Finally, requests in the candidate list are systematically added to the request list of the subsequent trip $(r + 1)$, aiming for a balanced distribution of load across all trips.
- Iterative Improvement: This process repeats, iterating through trips and adjusting request allocations to satisfy vehicle capacity requirements, thereby moving closer to identifying a feasible and optimized routing solution.

Algorithm 3 Load Balancing Among Trips

```
for each trip r in RD do
    Compute vehicle load for all nodes in current trip r
    Get maximum vehicle load of current trip r
    if current load is larger than vehicle capacity then
        Sort requests in current trip by largest waiting time
        while capacity condition not passed do
            Get requests with largest waiting time
            if pick-up index of current request have overload condition then
                Save current request in candidate list
                Remove request from current trip list
            end if
        end while
        for each request k in candidate list sorted by largest waiting time do
            Add current request k into next trip (r + 1) request list
        end for
    end if
end for
```

The heuristic algorithm aims to determine a feasible route, considering constraints related to both the (θ) condition (i.e., the scheduled departure time at each checkpoint, which is a hard constraint that cannot be violated) and the vehicle capacity. The main steps of the algorithm are as follows:

3.6 Warm start

Warm start, in the realm of optimization, refers to initiating an optimization algorithm not from a blank slate or a random point, but from a feasible (or nearly feasible) solution that has been previously computed or is known. This already-known solution provides a "warm" starting point for the algorithm, as opposed to a "cold" start where the algorithm begins with no prior knowledge about potential good solutions. In the following we provide the motivation for employing warm start:

- Speed: One of the primary reasons for using a warm start is the potential reduction in computation time. When optimization algorithms, like those in CPLEX, start from a feasible solution, they often converge to an optimal (or near-optimal) solution faster than they would from a random or default starting point.

Parameters	Values	Units
Number of checkpoints	5	-
Service time	18	seconds
Service area length	10	km
Service area width	2	km
Vehicle speed	30	km/h
Capacity	15	
Slack time	10	minutes
PD	10	%
PND	40	%
NPD	40	%
NPND	10	%
w_1, w_2, w_3	1	-

Table 3.3: Default parameter setting

- Improved Solution Quality: A warm start might help the algorithm escape local optima if the starting points are different enough from one start to another. By beginning from a known feasible solution, the optimization can navigate the solution space more effectively.
- Resource Efficiency: Warm starting can also lead to more efficient use of computational resources. By reducing the number of iterations or nodes explored (in tree-based methods), it can reduce the memory and CPU requirements.

Our heuristic algorithm aims to find a feasible solution by considering specific constraints and conditions. By its very design, a heuristic might not always guarantee an optimal solution, but it can often find good solutions in a relatively short amount of time. The heuristic provides a solution that already respects the problem constraints and conditions, giving the linear solver a good foundation to start its search. Moreover, in MILP formulations, the linear solver attempts to close the gap between the lower and upper bounds of the solution. By providing a feasible solution as a warm start, one can often reduce the initial gap, enabling the linear solver to converge faster.

Warm starting is a powerful technique that leverages prior knowledge or quickly obtained solutions to enhance the efficiency and effectiveness of optimization algorithms. By using heuristic results for warm start, one can harness the strengths of both heuristic methods and branch-and-cut solvers to solve complex problems more efficiently. For the above reasons, we use the solution of our heuristic as warm start, making it possible to close the gap faster and shorten the computation time.

3.7 Results

In this section, we evaluate the effectiveness of the groups of inequalities defined in Section 3.4 by solving different instances of the problem. In Table 3.3, we summarize the parameters common to all cases. In order to determine how the different formulations scale with the level of demand, we run experiments for five different demand levels, namely, 25, 30, 35 and 40 customers, to check the efficiency of the cuts and compare the different computational times. In each set, we solve the problem with four different formulations: without adding any group of inequalities ("none"), adding only the (most efficient) inequalities of Quadrifoglio et al. [2008a] ("#1"), adding only some of the new inequalities described in this chapter ("#2", regrouping valid inequalities (3.32) to (3.35)[1]), or adding both groups of inequalities together ("all"). For each demand level, we generated 5 different instances with random uniform spatial

[1]We decided not to use the valid inequalities proposed before (3.32) since they tended to increase the needed computational time to close the different instances.

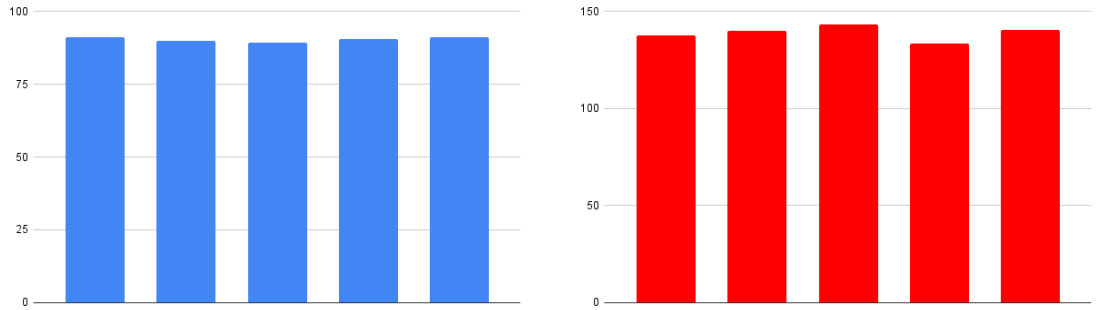
and time distribution for the customers, and the results are averaged over those 5 instances. In the following, we first assess the impact of adding only valid inequalities to the base formulation, using the CPLEX solver, where the MILP has been implemented using the CPLEX C++ Application Programming Interface (API) and solved using the version 20.1 of the CPLEX optimization solver. Computational experiments are executed on a high-performance server equipped with 24 central processing units (CPUs) and 125 GB of Random Access Memory (RAM). The server's architecture is based on an Intel(R) Xeon(R) CPU E5-2643 v4, operating at a clock speed of 3.40 GHz. The C++ code was compiled using version 9.3.0 of the GNU Compiler Collection (GCC), specifically the g++ compiler. Later, we will also consider the impact of warm starting on the computational efficiency, using larger instances with 50, 75 or 100 customers.

The computational experiments reveal that the different sets of valid inequalities have a profound impact on the solution of the MILP formulation, as can be seen in Table 3.4. Across the different demand levels, ranging from 25 to 40 customers, the addition of valid inequalities consistently resulted in a marked reduction in computational time. For demand levels up to 35 customers, all configurations with valid inequalities significantly outperform the baseline in terms of computational time, though all formulations manage to solve the instances to optimality within the 18,000-second time limit. For instance, at the 25-customer demand level, the new set of valid inequalities (#2) exhibits an efficiency gain of approximately 85.71% compared to the baseline, in terms of computational time. This trend is consistent across the different demand levels. Comparatively, the second set of valid inequalities (#2) consistently outperforms the first set (#1) with respect to computational efficiency. Specifically, at the 30-customer level, the time required for #2 is approximately 48.99% lower than that for #1. This trend continues to be evident at the 35-customer level, where the computational time for #2 is roughly 25.9% less than #1, further emphasizing the computational superiority of the second set of valid inequalities. For the 35-customer level, it is noteworthy that the 'none' configuration could achieve the optimal solution; however, the lower bound remains 4.2% below even after 5 hours. The formulation based on our new set of valid inequalities (#2) is the only one managing to actually solve to optimality the 40-customer instance. All other formulations do not even manage to obtain a feasible solution, hence the obvious necessity to study the impact of providing the solver with a warm start solution, as we will investigate below. An interesting trend in our results is the fact that combining both sets of inequalities #1 and #2 results in poorer results compared to the sole use of inequalities #2, suggesting a negative interaction of the two sets. Though it would be interesting to provide insights as to why this phenomenon consistently appears for all demand levels, we have as of now no convincing explanation for this empirical result.

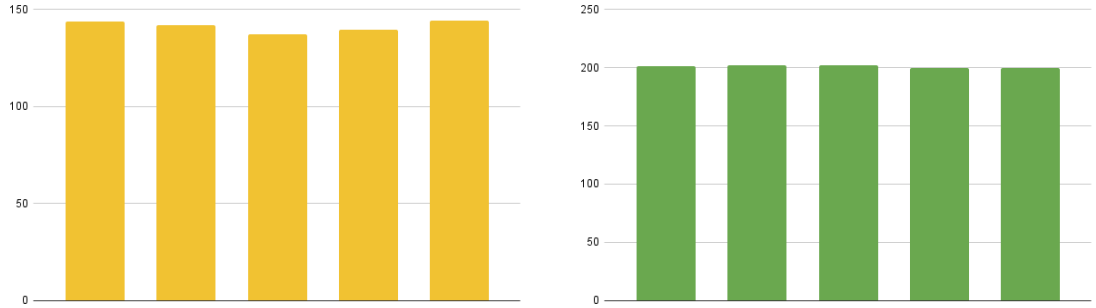
Table 3.4: Average results for the different formulations and different levels of demand.

Cuts	# of Customers	Time (s)	Objective	Gaps (%)
None	25	98	90	0
#1	25	55	90	0
#2	25	14	90	0
All	25	25	90	0
None	30	5,000	138	0
#1	30	643	138	0
#2	30	328	138	0
All	30	365	138	0
None	35	11,141	141	4.2%
#1	35	1,315	141	0%
#2	35	974	141	0%
All	35	1,273	141	0%
None	40	18,000	-	-
#1	40	18,000	-	-

Continued on next page



(a) Objective function values for number of demand of 25 (b) Objective function values for number of demand of 30



(c) Objective function values for number of demand of 35 (d) Objective function values for number of demand of 40

Figure 3.1: Distribution of objective functions values for different demand levels

Table 3.4 – *Continued from previous page*

Cuts	# of Customers	Time (s)	Objective	Gaps (%)
#2	40	6,863	201	0%
All	40	18,000	-	-

In order to provide an idea of the variability between the different instances generated for the different levels of demand, we plot in Figure 3.1 the values of the objective function for the five instances of each customer demand level. We can see that the value of the objective function exhibits minimal variation within each demand level, thereby underscoring the uniformity of the objective function outcomes across disparate benchmark scenarios.

An indication of the value of valid inequalities can be partially evaluated by looking at the value of the linear relaxation of the model. In Table 3.5 we provide the average value of the objective function for the relaxations of the different models investigated previously. An interesting observation is the fact that the inequalities of set #1 do not increase the value of the objective of the linear relaxation, neither used alone nor together with set #2. On the contrary, set #2 provides a mild increment with respect to the value of the optimal integer solution. We therefore hypothesize that set #2 also increases the effect of branching in subsequent nodes of the branching tree, so that the number of nodes pruned by the bounding procedure may be much higher.

Table 3.5: Data representing a linear relaxation for different demand level

Cuts	# of Customers	Linear Relaxation
None	25	11.64
#1	25	11.64
#2	25	12.75
All	25	12.75
None	30	16.22
#1	30	16.22
#2	30	17.95
All	30	17.95
None	35	17.49
#1	35	17.49
#2	35	19.5
All	35	19.5
None	40	27.73
#1	40	27.73
#2	40	29.94
All	40	29.94

We move next to the study of the impact of supplying a warm start solution to the solver, as summed up in Table 3.6. The table provides the same results in Table 3.4 for the formulations using valid inequalities #1, #2 or both, all supplemented by the same warm start solution, across four different demand levels: 40, 50, 75, and 100 customers. We begin by examining the 40-customer level to demonstrate the significance of implementing the greedy heuristic, which notably reduces computation time and achieves solutions of optimal quality. The results once again are in favor of set #2 of valid inequalities. This is best seen for scenarios with 50 customers, where the difference in computational time between both sets of inequalities is more than two orders of magnitude. Let us observe that for formulation #1, some instances have reached the 18,000s time limit. As the demand level increases to 75 customers, the computational times systematically reach the upper limit of 18,000 seconds for all configurations and the gap percentages begin to diverge. For 75 customers or more, formulation #2 consistently provides both a better optimality gap and a better feasible solution compared to formulation #1 or the combination of both #1 and #2, confirming that both sets of valid inequalities should not be used together. The optimality for the formulation using our new set of valid inequalities #2 and a warm start solution provide solutions with a gap on average less than 12%, which allows us to provide solutions of sufficiently good quality.

Table 3.6: Average results for the formulations using a heuristic solution as warm start, for different demand levels.

Cuts	# of Customers	Time (s)	Objective	Gaps (%)
Heu + #1	40	7,453	201	0%
Heu + #2	40	53	201	0%
Heu + all	40	124	201	0%
Heu + #1	50	15,599	239	2.29%
Heu + #2	50	113	239	0%
Heu + all	50	176	239	0%
Heu + #1	75	18,000	335	9.7%
Heu + #2	75	18,000	331	3.1%
Heu + all	75	18,000	332	3.81%
Heu + #1	100	18,000	521	21.58%
Heu + #2	100	18,000	518	11.72%

Continued on next page

Table 3.6 – *Continued from previous page*

Cuts	# of Customers	Time (s)	Objective	Gaps (%)
Heu + all	100	18,000	519	12.93%

3.8 Conclusion

In this chapter, we provided a MILP formulation for the FRT problem, starting from an existing formulation for multi-shuttle and multi-trip instances. We recalled existing valid inequalities from the literature and after discussing some of the weaknesses of the linear relaxation of the model, we proposed new valid inequalities for the problem. We also provided a simple heuristic algorithm to compute a feasible solution to be passed to the linear integer solver to help it converging faster, as we observed that, with a growing number of customers, it becomes difficult for CPLEX to find even one feasible solution.

We then tested different formulations, with or without inequalities, and demonstrated the superiority of our new inequalities over the existing ones in all cases considered. We empirically assessed that our new inequalities do not interact favorably with other existing inequalities, leading in any case to larger resolution times or poorer feasible solutions. We also demonstrated that solving instances with 40 or more customers requires the use of a warm start solution and that warm starting is extremely beneficial, especially when combined with our new valid inequalities.

Although instances with 100 customers are still very challenging, even with our proposed improvements, we managed to find solutions with a reasonable optimality gap, which motivates us to use this formulation for an in-depth sensitivity analysis in our next chapter. As the field advances, further refining mathematical programming and heuristic techniques and exploring other innovative strategies will be paramount to efficiently tackle even more extensive and complex scenarios.

Chapter 4

Sensitivity analysis

4.1 Introduction

Sensitivity analysis in the context of optimization is the study of how the uncertainty in the output of a mathematical model can be apportioned, qualitatively or quantitatively, to different sources of uncertainty in its inputs. Essentially, it determines how sensitive the results of an optimization are to changes in the parameters of the model. This understanding is critical in modeling scenarios where inputs can vary, and provides insights into the robustness of the solution. Sensitivity analysis is commonly divided into two categories: local and global sensitivity analysis. Local sensitivity analysis studies the sensitivity of the output with respect to small changes in the input parameters near a specific point in the parameter space, typically using derivatives. It is often concerned with understanding the behavior of the model in a particular region. Global sensitivity analysis, on the other hand, considers the entire possible input space and examines the influence of one or more input variables over the entire range of parameter values. It provides a more comprehensive understanding of the relationships between input and output variables, allowing for the identification of the most critical variables across all possible values.

In operations research, sensitivity analysis plays a crucial role in understanding how changes in the parameters of a mathematical model affect the optimal solution. It helps in evaluating the robustness and reliability of the solutions provided by optimization models, especially in linear programming. By analyzing how a small change in the coefficients, constraints, or objective function could alter the optimal solution, decision-makers are able to anticipate how different scenarios or changes in conditions might affect outcomes. This insight aids in strategic planning and risk management, allowing for the adjustment of the model or the formulation of contingency plans. In operations research, conducting sensitivity analysis is crucial not merely for achieving an optimal solution set under current parameters, but also for enhancing the system resilience. By rigorously examining the stability of these solutions under a spectrum of potential parameter variations, sensitivity analysis enables us to fine-tune the parameters to reach an optimal configuration. This, in turn, fortifies the decision-making process against uncertainties, ensuring that the system remains robust and capable of adapting to changing conditions.

Performing sensitivity analysis is of paramount importance to ensure a resilient and adaptable transit system. Flex-route transit systems are typically complex, with many variables such as demand patterns, route flexibility, vehicle capacities, and timing constraints that can have a profound impact on efficiency and service quality. Sensitivity analysis allows transit planners and operators to understand how changes in these variables could affect the overall system performance. By identifying the key parameters that have the most significant impact on outcomes, planners can prioritize efforts to collect more accurate data or create contingency plans for those areas. Furthermore, it aids in uncovering potential vulnerabilities in the system, such as how sudden changes in demand or disruptions in routes might affect service

levels. Overall, sensitivity analysis in the context of the flex-route transit problem enables more robust decision-making, facilitating the development of transit services that are more responsive to the actual and potential future needs of the community.

As we have already discussed in the literature review chapter, several Mixed Integer Linear Programming (MILP) formulations have been developed to tackle the vehicle scheduling in FRT systems [Quadrifoglio and Dessouky, 2004; Quadrifoglio et al., 2008a; Lu, Xie, Wang and Quadrifoglio, 2011; Zheng et al., 2019; Liu et al., 2021]. However, only limited investigations on the sensitivity of the system to its parameters have been carried out. Specifically, "One-Factor-at-a-Time" (OFAT) is the sole approach used to perform sensitivity analysis. In this approach, one parameter is allowed to vary within a certain range while the others are held constant. The parameter is then reset to its default value and another parameter is varied. In this chapter, we apply a more advanced statistical method to achieve a clearer understanding of the sensitivity of the FRT objective function. Therefore, we perform a full factorial experimental design in which we not only understand the impact of changing one parameter at a time but also can investigate the impact of the interaction among parameters.

In the remainder of the chapter, we first review the existing studies on sensitivity analysis in Section 4.2. After which, we discuss the methodology in Section 4.3. The case study and the results are presented in the two following sections 4.4 and 4.5. Results are discussed in Section 4.6 conducting to a focus on the capacity factor. We conclude in Section 4.7.

4.2 Sensitivity analysis in the literature

Table 4.1 displays all the studies in the literature that perform sensitivity analysis. The first column of the table presents the references and the other columns are the parameters on which the sensitivity analysis is performed over.

In the article delineated by Lu, Xie, Wang and Quadrifoglio [2011], the authors undertake a comprehensive sensitivity analysis with respect to the weight attributed to the cumulative distance traversed by shuttles in the objective function. The empirical outcomes incontrovertibly demonstrate that an augmentation in the weight accorded to the total distance traversed engenders a concomitant elevation in the critical demand prerequisite for transiting from a single model vehicular system to a two model counterpart. This phenomenon is logically foreseeable, as the transition from a singular to a dual vehicular system precipitates a marked diminution in the latter two constituents of the objective function, which are indicative of service quality, while concomitantly effecting a near-doubling of the inaugural term.

In the publication by Zhang et al. [2021], the authors execute an exhaustive sensitivity analysis concentrating on the variability in demand levels and the quantity of checkpoints. The empirical evidence posits that the incorporation of FRT for the facilitation of pick-up and drop-off activities for shared bicycle users yields salutary effects on both the modality of passenger travel and the governance of urban traffic management. Furthermore, the authors elucidate that, as the density of passenger demand escalates, the temporal expenditure associated with the shared bicycle system experiences a corresponding increment, whilst the time cost attributed to the FRT system maintains relative stability. In addition, it is revealed that the temporal cost incurred by the FRT system remains largely impervious to alterations in the number of predetermined stations when deviating from its foundational route to accommodate passengers. Contrarily, the time cost of the shared bicycle system is found to be inversely proportional to the spacing between fixed stations, with the exception that this cost achieves a state of relative stability when said spacing exceeds 2.5 miles.

In the work presented by Zheng et al. [2019], the authors introduce the possibility of gathering customers at a set of predetermined meeting points. They conduct an extensive sensitivity analysis focusing on two distinct parameters: the acceptable walking distance and the quantity of meeting points within the service area. With respect to the latter parameter, the research reveals that the number of designated

References	RTW	DL	CS	MP	AWD	ST	SC
Lu, Xie, Wang and Quadrifoglio [2011]	✓						
Zhang et al. [2021]		✓	✓				
Zheng et al. [2019]				✓	✓		
Current study		✓				✓	✓

Table 4.1: Summary of sensitivity analysis in the literature. RTW (Riding Time Weight); DL (Demand Level); CS (Checkpoints Spacing); Meeting Points (MP); AWD (Acceptable Walking Distance); ST (Slack Time); SC (Shuttle Capacity).

meeting points exerts a considerable influence on the overall performance metrics of the system. More specifically, an augmentation in the number of meeting points leads to a decrease in the 'rejection rate' of ride requests, this term refers to the proportion of customer requests that cannot be accommodated within the service framework, albeit at the expense of increasing the pedestrian transit time for passengers. The authors propose that viable locations for these meeting points could encompass public parking zones, intersections on side roads, and vehicular turning areas. Moreover, they contend that amenities such as lighting, shelters, and seating arrangements can contribute to enhancing the overall passenger satisfaction. In regard to the parameter of acceptable walking distance, the authors ascertain that an increase in this variable similarly leads to a decrease in the rejection rate of ride requests, but with a concomitant increase in the pedestrian transit time. Specifically, the study delineates that when the acceptable walking distance is set at 0.3 miles, the rejection rate stands at 7.6%, the riding time averages 15.13 minutes, idle time is 0.87 minutes, and the walking time is 0.35 minutes. When the acceptable walking distance is extended to 0.5 miles, the rejection rate plummets to 1.2%, with riding and walking times of 15.04 and 1.04 minutes, respectively. Finally, when the acceptable walking distance is further increased to 0.8 miles, the rejection rate diminishes to a negligible 0%, while the riding and walking times are 15.07 and 1.09 minutes, respectively.

The outcomes of the sensitivity analyses presented in the existing literature could hold significant implications for decision-makers within transportation authorities. Such analyses provide invaluable insights into the impact of specific parameters on system performance, thereby facilitating a more nuanced understanding for these decision-makers. However, there are two principal lacunae in the existing sensitivity analyses:

- The capacity of the shuttles has been conspicuously overlooked, despite its relevance as a key variable that decision-makers must consider.
- Existing studies predominantly employ the One-Factor-At-a-Time (OFAT) method for conducting sensitivity analyses on a pair of parameters. This approach is inherently limited in that it precludes the examination of interactive effects between parameters. Specifically, the interplay between demand levels and shuttle capacity represents a critical area of interest for decision-makers.

To circumvent these shortcomings, we propose the execution of a more comprehensive sensitivity analysis that encompasses three parameters: Shuttle Capacity, Slack Time and demand Level. The position of our study in comparison with the literature is presented in Table 4.1.

Methodologically, we advocate the utilization of Analysis of Variance (ANOVA) techniques to scrutinize the interactive effects among the parameters under consideration.

4.3 Analysis of Variance (ANOVA) method

In this chapter, we aim at understanding the impact of the values of multiple parameters, both in isolation and jointly. To do so, we exploit a full factorial experimental design. Here, we consider various values for

each parameter (respectively called *levels* (values) and *factor* (parameter) in the ANOVA terminology), and we run the MILP of Chapter 3 with all their possible combinations (which are called *effects*). For instance, with three parameters and two values for each, this leads to perform $2^3 = 8$ runs.

ANOVA is used to analyze the variance between groups and within groups to determine if the differences in the means of the groups are statistically significant or if they can be attributed to chance. ANOVA is based on the F-test, which compares the ratio of the variance between groups to the variance within groups. If the ratio is large enough, then it indicates that there is a significant difference between the means of the groups. ANOVA is commonly used in experimental and research studies to determine if there is a significant difference between the means of the groups being studied. It can be used to test hypotheses and to analyze the effect of different factors on the outcome being measured.

ANOVA can be used in the field of sensitivity parameters to identify which parameters have significant impact on the output of a model. This is done by comparing the variance in the output of the model for different levels of each parameter. The steps for using ANOVA in the field of sensitivity parameters are as follows:

- Define the model: Define the model to analyze and identify the parameters that are sensitive to changes in the output.
- Define the parameter ranges: Define the ranges of values for each parameter to analyze.
- Generate the input combinations: Generate a set of input combinations by varying the values of the parameters over their ranges. The number of input combinations should be large enough to cover the parameter space adequately.
- Run the model: Run the model for each input combination and record the output.
- Perform ANOVA: Perform ANOVA on the output data to identify which parameters have a significant impact on the output. This is done by comparing the variance in the output for different levels of each parameter.
- Interpret the results: Interpret the results of ANOVA to identify the most significant parameters and their impact on the output.

In the literature on the FRT, the sole method employed to perform sensitivity analysis is OFAT. The OFAT method allows understanding the main effects but if the interactions among parameters have to be studied, it is recommended to employ ANOVA. It considers the effects of qualitative variables (in our case, the problem parameters), on one output. The main reason why ANOVA is useful is that it can provide a comprehensive understanding of the interaction impact between parameters on the output. This helps us to understand if the impact of one parameter changes depending on the value of the other parameters. Therefore, we employ ANOVA to trace this behavior.

The ANOVA method is considered superior to the OFAT one, especially when examining complex systems or processes with multiple variables. One of the main reasons for this is that ANOVA allows for the simultaneous examination of multiple factors and their interactions. The interaction effect, which denotes the combined influence of two or more factors on the outcome, can be critically important. In some cases, the interaction effect can be even more significant than the main effects of individual factors. This information is readily available in the results of ANOVA, as seen in factorial designs. Instead, in the OFAT method, only one factor is varied while keeping the others constant. This method inherently ignores interaction effects, making it impossible to determine how multiple factors might jointly influence the output. Thus, OFAT can lead to incorrect or incomplete conclusions, especially when interactions play a pivotal role. Another reason is that ANOVA, by considering all factors and their interactions, concurrently provides a more robust analysis against random variations in the system. The isolation of each factor in OFAT can sometimes amplify random variations, leading to false conclusions about the importance or impact of a particular factor. While the OFAT method can provide quick insights into the

effect of individual factors, it is limited in its scope and can lead to an oversimplified understanding of the system. ANOVA, on the other hand, offers a detailed and comprehensive analysis, capturing both main effects and interactions, ensuring that critical nuances and relationships are not missed.

ANOVA relies on a statistical model which compares differences in means. In order to determine if the means between different groups of data are significantly different (which points to a significant effect on the output), it uses the variance inside the different groups [Saltelli et al., 2010]. A group corresponds to a number of independent measures of the output of the system under study, which corresponds for us to the solutions of the MILP presented above for different FRT instances. To each combination of parameters therefore corresponds a number of solutions called *repetitions*.

The usual way to display the output of an ANOVA analysis is through what we can call the ANOVA table. This table lists a certain number of properties of the variation of the system output (the objective function for our optimization problem), depending on the combination of input parameters considered. It has one row for each such combination, listed in the column labeled *Effect*. The Degree of Freedom in the Denominator (DFd) represents the total number of observations minus the number of groups being compared. The 'F' stands for Fisher-Statistic or F-Statistic, which is the ratio of two variances: between-group variance and the within-group variance. The F-Statistic is used to test the null hypothesis that the means of all groups are equal. The interpretation of the F-statistic, presented in the ANOVA table, is crucial for understanding the variance within the data. A higher F-statistic value indicates that the between-group variance is significantly larger than the within-group variance, suggesting that the groups are not all the same. In contrast, a lower F-statistic implies that the variance within each group is similar to the variance across the groups, pointing to a lack of significant differences among group means. Therefore, the F-statistic is a critical component in assessing whether the observed differences in means across groups are statistically significant. The 'P' stands for the p-value and is a measure of the probability of observing a test statistic as extreme as the one computed from the sample data, assuming the null hypothesis is true [Helton et al., 2006; Oakley and O'Hagan, 2004]. If the p-value is less than the significance level (less than 0.05), then we reject the null hypothesis and conclude that there is a significant difference between at least two of the group means. The magnitude of the impact of the corresponding 'Effect' is provided by the 'ges' (generalized eta square) column. The 'ges' is a measure of effect size in ANOVA that quantifies the proportion of variance accounted for by a predictor variable (i.e., independent variable) in a linear model. It ranges from 0 to 1, with larger values indicating a stronger effect. If the value of 'ges' is less than 0.02 the impact is small, if it is between 0.02 and 0.26, the impact is moderate, whereas a value above 0.26 implies a large impact [Wilcox, 2011]. The most important parameters for assessing the impact of the different effects are the $p-$value (under the column labeled P) and the 'ges' (under the column ges)

4.4 Case study

In this study, we implemented the MILP formulation of Lu, Xie, Wang and Quadrifoglio [2011] extended as explained in Chapter 3, by adding capacity constraints, valid inequalities, and a heuristic to be used for warm starting.

To perform the sensitivity analysis, we have the following parameter setting, also shown in Table 4.2. The entire spatial domain is segmented into 5 checkpoints. The expected service time for each stop (both checkpoint and non-checkpoint) is 18 seconds. The service area has a rectangular shape, with a length of 10 km and a width of 2 km. The vehicles operating within this service area are assumed to move at a constant speed of 30 km/h. The distribution of different types of customers in the service area are 10% of PD, 10% of NPND, 40% of PND and 40% of PND. The demand is uniform in space and time. The problem also involves a set of weights for the different components of the objective function, which are represented by w_1, w_2 and w_3. All these weighted factors are assigned an equal value of 1, signifying that they have an equal impact on the problem's objective function or the quality of solutions. Table 4.3

Parameters	Values	Units
Number of checkpoints	5	-
Service time	18	seconds
Service area length	10	km
Service area width	2	km
Vehicle speed	30	km/h
PD	10	%
PND	40	%
NPD	40	%
NPND	10	%
w_1, w_2, w_3	1	-

Table 4.2: Nominal value of each parameter

References	Year	# of shuttles	Capacity	# of Customers
Quadrifoglio and Dessouky [2004]	2004	1	×	25/hour
Quadrifoglio et al. [2008a]	2008	1	×	17/hour
Lu, Xie, Wang and Quadrifoglio [2011]	2011	1-2	×	14/hour
Zheng et al. [2019]	2019	1	×	25/trip
Zhang et al. [2021]	2021	1	50	50/trip
Liu et al. [2021]	2021	10-30	10	45/trip
Current study	2023	1	15-20-25	20-25-30/hour

Table 4.3: Case studies of the literature in comparison with our case study.

displays the positions of the nominal value in comparison with the case studies in the literature.

We conduct a sensitivity analysis on three parameters: demand level (uniformly distributed in time and space over a 5-hour time horizon), slack time and shuttle capacity. Each of these parameters has three possible values. They are shown in Table 4.4. We have considered these specific ranges in order to reach the saturation level in the system and see how the model reacts. The saturation level corresponds to a level of demand too large to handle each customer at the first possible trip after they become available for pick-up, so that, due to the lack of capacity and/or slack time, an increasing number of customers has to be assigned to later trips. Therefore, we have chosen these configurations to see the model reaction. A full factorial analysis with these configurations requires 27 parameter combinations. To perform a robust sensitivity analysis, we solve 5 repetitions for each parameter combination in which the random demand realization varies.

The computation time for the above mentioned configuration requires 135 runs and for each one we let the heuristic find a feasible solution and pass it to CPLEX using the warm start method, in order to see if the results can be improved in a time limit of one hour. This leads to a maximum of 135 hours of computation time.

Parameters	Values	Unit
Slack time	5-10-15	minutes
Level of demand	20-25-30	per hour
Shuttle capacity	15-20-25	-

Table 4.4: Sensitivity analysis parameters

Effect	DFn	DFd	F	P	P < 0.05	ges
Slack time	2	108	513	7.07e-56	Yes	0.905
Demand	2	108	3819	6 e-101	Yes	0.986
Capacity	2	108	1118	6 e-73	Yes	0.954
Slack:Demand	4	108	144	2 e-42	Yes	0.843
Slack:Capacity	4	108	216	1 e-50	Yes	0.889
Demand:Capacity	4	108	57	7 e-26	Yes	0.680
Slack:Demand:Capacity	8	108	6	2 e-06	Yes	0.307

Table 4.5: The results of ANOVA table for uniform demand distribution

4.5 Results

Table 4.5 summarizes the results of ANOVA applied to the sensitivity analysis using a full factorial experimental design according to the explanation of the methodology presented in Section 4.3.

The factors being considered are Slack time, Demand, and Capacity, as well as their interactions. All the main factors and interactions are significant. However, they do not all have the same influence on the variance of the results. In the following, we describe each of the factors and their interactions in descending order of influence.

1. Slack time with an F-value of 513 and a p-value of 7.07×10^{-56}. The 'ges' value of 0.905 suggests a very large effect size, meaning that changes in slack time have a substantial impact on the response variable. Slack time also plays an important role in having a high quality level of service for customers.

2. As expected, for the main effects, 'ges' value of 0.986 for Demand indicates a huge effect size, suggesting that Demand is likely the most influential factor among the three.

3. The statistical significance of shuttle capacity is evident from its F-value of 1118 and p-value of 6.64×10^{-73}. A 'ges' of 0.95 further solidifies its substantial impact on the response. Indeed, this high 'ges' value shows the importance of investigating capacity as one of the contributing factor and proves the importance of our choice to study this parameter.

4. With respect to the interaction effects, the interaction between slack time and demand is significant with an F-value of 144 and p-value of 1.75×10^{-42}. A 'ges' value of 0.843 suggests that the joint effect of these two factors is also considerably large.

5. Considering the interaction of Slack time and Capacity (Slack:Capacity) as evidenced by the F-value of 216 and p-value of 1.19×10^{-50} this interaction has a high impact. The 'ges' value of 0.89 shows that their combined impact is profound.

6. Demand:Capacity: the interaction effect between Demand and Capacity has an F-value of 57 and p-value of 6.93×10^{-26}. Although this interaction is weaker compared to the first two, it is still statistically significant with a moderately large effect size of 0.680.

7. Slack:Demand:Capacity: this three-way interaction, although statistically significant with a p-value of 2.46×10^{-6}, has the smallest effect size among all listed interactions, at 0.31. This means while the combined effect of all three factors does influence the outcome, it is relatively less impactful than the individual factors and two-way interactions.

It is notable to mention that the extremely small p-values listed in the table, such as 7.07e-56 for Slack time or 6e-101 for Demand, indicate a very high statistical significance. These values are far below the conventional significance level of 0.05, suggesting that the differences in group means are not due to random chance but are indeed statistically significant. In practical terms, such minuscule p-values

strongly reject the null hypothesis, affirming that the observed differences between the group means are meaningful and reliable. This is further corroborated by the 'Yes' entries under the column '$P < 0.05$', which explicitly denote that the p-values fall below the threshold of statistical significance. Therefore, the results in columns P, despite being very small, unequivocally point towards significant differences among the groups for each of the effects studied in the ANOVA.

4.6 Discussion

4.6.1 Remarks on the results

From the given ANOVA results, we can understand that all the factors - Slack time, Demand, and Capacity - as well as their interactions are statistically significant, suggesting they play a pivotal role in influencing the outcome of the problem. Among the main effects, Demand has the strongest influence followed by Capacity and then Slack time. This revelation underscores the pivotal role that demand levels assume in guiding transportation authorities during the conceptualization and design of such a system. Indeed, the 'ges' value of 0.986 reveals that the demand variable alone accounts for roughly 98.6% of the variance, making it an extremely influential factor in the study.

In terms of interaction effects, the combined impact of Slack and Capacity appears to be the most influential, closely followed by the Slack and Demand interaction. The empirical evidence substantiates the notion that both the isolated effects of individual parameters and their interactive consequences are instrumental in influencing the system's output. Furthermore, the parameter of capacity emerges as a crucial variable, warranting meticulous consideration by transportation authorities, particularly when proposing the implementation of such a system to prospective clients. However, the three-way interaction is least influential, implying that the synergy of all three factors together does not drastically change the outcome beyond their individual or two-way interaction effects. This observation illuminates the imperative that the individual or interactive effects of other parameters warrant a level of scrutiny that surpasses the collective impact of the three aforementioned factors.

4.6.2 Methodological remarks

The specificity of our results is the strong interaction of any type of parameter combination, which implies that it is not straightforward to generalize a conclusion drawn from an OFAT analysis.

In the realm of research surrounding sensitivity analysis in FRT literature, this study marks the inaugural instance where shuttle capacity has been considered as a critical parameter. Drawing from the data presented in the previously shared table, the influence of shuttle capacity emerges as both significant and profound. The F-values and the associated minuscule p-values underscore its pivotal role in the outcome of the optimization problem. Moreover, the 'ges' value further attests to the magnitude of the impact of capacity on the overall system. These findings not only highlight the relevance of incorporating shuttle capacity into sensitivity analyses but also emphasize the need for stakeholders and decision-makers to accord it considerable attention in future undertakings and optimizations. In order to have a better understanding of the phenomenon, more analysis are carried out to study the combination of capacity with the other parameters in the following section.

In future research, we suggest to investigate the impact and interactions of a larger number of parameters, such as the frequency departure at the depot when the system can use a fleet of shuttles instead of a single shuttle. Moreover, as the computation time for performing a Full Factorial Experimental Design is expensive one can investigate Design of Experiments (DOE), which is a powerful statistical tool used to systematically explore and optimize complex systems. The use of DOE can offer reduced testing costs, efficient parameter estimation, improved understanding of the system, and robustness analysis.

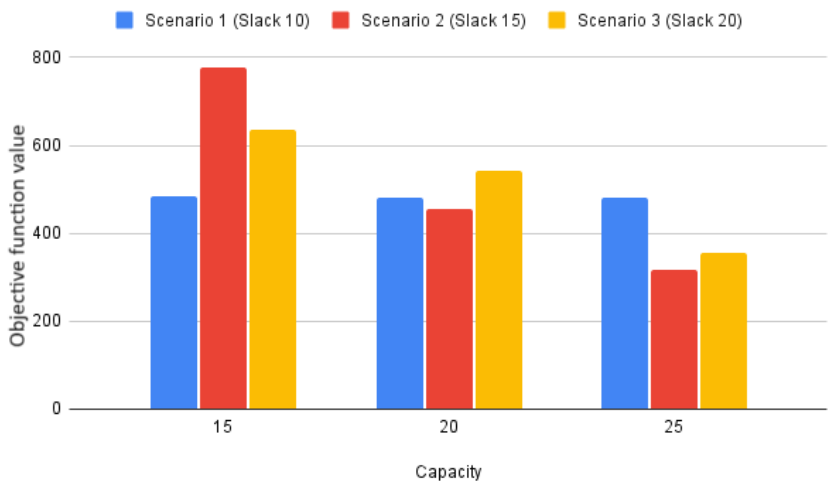


Figure 4.1: Total objective function value (hours) for different values of capacity and slack time (for 100 customers).

4.6.3 Focus on the interaction with capacity factor

We place our investigation in a situation already encountered in the FRT literature where the system is said to be *saturated* when the demand reaches a critical level. With this level, the shuttle cannot satisfy all the non-PD customer requests as soon as possible. The result is that some requests must be handled at later trips, which increases their waiting time. The fact that non-PD customers tend to accumulate and their waiting time tends to increase signals that the FRT system is not well calibrated with respect to the expected level of demand.

The objective is to clarify the impact of capacity on a saturated FRT system. We first consider the combination of slack time and capacity. We consider a fixed level of demand of 100 customers and define three sets of scenarios as outlined in Table 4.6. Each scenario set considers a different value of the slack time at checkpoints. It gathers the values of the objective function for all three values of the shuttle capacity. The average values of the objective function for the resulting nine different combinations of slack time and capacity are displayed in Fig. 4.1. As expected, the general trend is that the objective function decreases as capacity increases. However, the specific behavior depends on the specific slack time scenario considered. If we compare scenarios 1 and 2, we can measure a difference of 293 hours of objective function for capacity 15, which is approximately 60%. When considering a capacity to 25, this difference shrinks to approximately 23%. This illustrates how differently the system will react to a change in capacity depending on the values of the other system parameters, for example in terms of operating costs.

Moreover, the ratio of objective functions between scenarios 2 and 1 changes from a value of 1.6 for capacity 15 to 0.65 for capacity 25. As a consequence, it is very difficult to evaluate the system behavior by changing only one parameter at a time, for example starting from a base configuration with a slack time of 10 minutes and a capacity of 15. In such a case, one can first investigate the effect of a capacity of 25, and then of a slack time of 10 minutes, however, neither the differences nor the ratios are conserved when both parameters evolve. This demonstrates the importance of performing sensitivity analyses considering multiple parameters concurrently. The figure also illustrates that the effect of slack time is not monotonic for a given level of demand and that the variation in the system behavior will change depending on the chosen shuttle capacity.

No.	Slack time	Demand	Shuttle capacity
Scenario 1	10	100	15-20-25
Scenario 2	15	100	15-20-25
Scenario 3	20	100	15-20-25

Table 4.6: Scenarios for assessing the objective function value based on increasing slack time and capacity.

No.	Shuttle capacity	Slack time	Demand
Scenario 4	15	15	100-125-150
Scenario 5	20	15	100-125-150
Scenario 6	25	15	100-125-150

Table 4.7: Scenarios for assessing the shuttle capacity threshold

We also investigate the combination of demand and capacity, along similar lines, since it can also be considered as having a strong impact on the FRT behavior. As before, we consider only nine configurations divided in three sets of scenarios, each associated to a different capacity value (15, 20 and 25), as detailed in Table 4.7. The slack time value is fixed at 15 minutes in order to make sure that the saturation phenomenon comes from a lack of capacity. We will adopt a customer-oriented view and display the average waiting time per customer, which is only one part of the full objective function of the MILP. The fact that we consider an average waiting time *per customer*, as opposed to the *total waiting time*, rules out the natural expectation that the cost function would scale proportionally to the demand level. The results are plotted in Figure 4.2 in a similar fashion to Figure 4.1.

While the waiting time per customer generally increases with the level of demand, we can see that the increment value depends on the exact shuttle capacity: while the difference between demand of 100 and demand of 150 is around 182 minutes for a capacity of 15, it decreases at around 108 minutes for a capacity of 25. These results are very interesting for practitioners. By increasing the shuttle's capacity by just 5 seats, the system became less saturated. The drop between the blue and red bars, which corresponds to an increase in capacity of 5 seats, is remarkable.

Having said that, while the observed reduction in waiting time per customer with increasing shuttle capacity is notable, particularly the marked decrease when capacity is expanded from 20 to 25 seats, it is important to consider the potential 'asymptotic nature' of this relationship. As the capacity continues to increase beyond the range studied here, it is plausible that we may encounter a point where additional seats result in progressively smaller reductions in waiting time. This asymptotic behavior implies that there could be an optimal capacity level beyond which further increments are less effective in decreasing waiting times.

Interestingly, if we plot instead the riding time per customer for the same configurations of parameters, see Figure 4.3, we see a slight increase with respect to the shuttle capacity (as expected since the shuttle can perform more detours to handle more customers within a given trip) but no particular variation when the demand increases. This is a hint that the FRT should also be studied with respect to each of its objectives as an overall variation of the total objective function can hide very different variations of its components.

These results are interesting from a strategic perspective to anticipate the impact of the choice of shuttle depending on the different levels of demand, in order to weigh the investment costs compared to the level of service provided for customers.

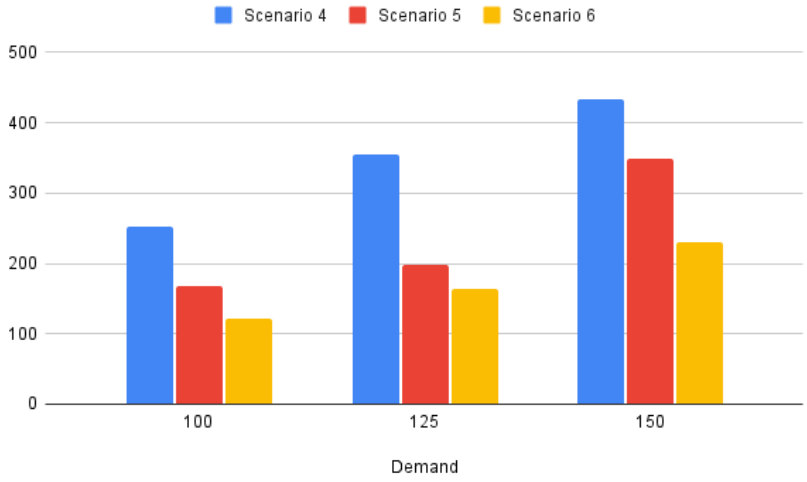


Figure 4.2: Average waiting time per customer (minutes) according to capacity and level of demand

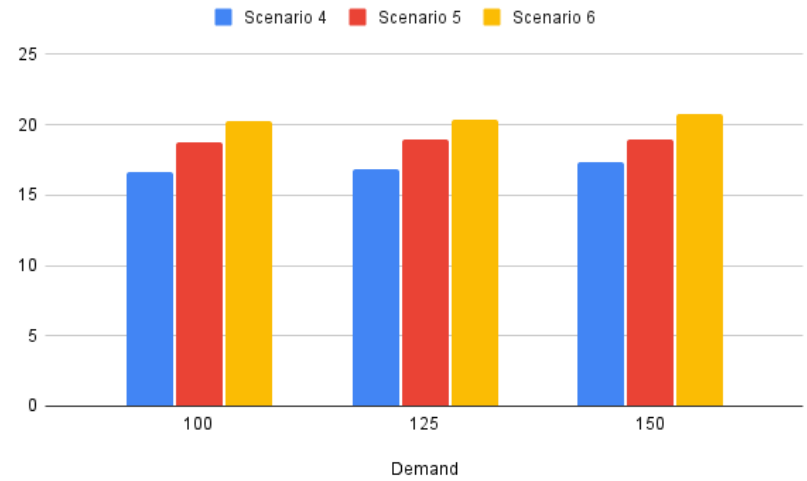


Figure 4.3: Average riding time per customer (minutes) according to capacity and level of demand

4.7 Conclusion

We performed a sensitivity analysis with respect to the shuttle capacity, the level of demand and the slack time at checkpoints, using a full factorial experimental design. We analyzed it using ANOVA in order to assess parameters interdependence on the FRT system behavior. The results emphasize that capacity is a key factor for the FRT system, although this factor is not generally treated in the literature.

This sensitivity analysis also highlights the importance of the interaction between capacity and level of demand and between capacity and slack time. It is the reason why specific experiments were designed and implemented to study specifically these interactions when the FRT system is saturated. These specific analyses quantify the interaction effect between capacity and demand level on average user waiting times. This information is crucial for decision-makers, helping them choose the appropriate shuttle capacity

Based on these results, it would be tempting to extend this sensitivity analysis to other FRT parameters as well (e.g., non-uniform demand distributions, relative weights in the objective function, backtracking threshold, etc...). Unfortunately, the full factorial experimental design falls victim to the curse of dimensionality and requires an excessive computation time for more than three parameters. Other techniques based on design of experiments should be considered. For example, methods such as Fractional Factorial Design, Latin Hypercube Sampling, and Taguchi Methods offer a more efficient exploration of the parameter space. Fractional Factorial Design enables the study of multiple factors simultaneously while reducing the total number of experimental runs [Pigeon, 2006]. Moreover, the Taguchi Method is renowned for its robustness in identifying optimal settings for multiple parameters through orthogonal arrays [Tsui, 1992].

This chapter provides a more detailed analysis of an article in which I served as the first author, published in 8^{th} International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) which was held in Nice, France between 14^{th} and 16^{th} of June 2023. This conference was indexed in **IEEE** and **Scopus** [Shahin et al., 2023].

Chapter 5

A stochastic optimization approach for the Tactical FRT

5.1 Introduction

The two previous chapters have been devoted to devise an efficient formulation for the operational aspects of the FRT, i.e., devising the partition of the customers in the different available trips of the shuttle fleet, and using this formulation to perform an advanced sensitivity analysis of the problem. In this chapter, we study the FRT from a more tactical perspective. While the operational perspective is interested in optimizing the schedule and route of the shuttle over a given time horizon, tactical and strategic aspects are more focused on finding out the best system parameters to handle a given level of demand. These parameters at the tactical level will involve, for example, the schedule of the shuttles in each trip (and consequently the slack time at checkpoints) or the width of the service area. Instead, the strategic level is often interested in the design of the base shuttle route and the location of depots and checkpoints.

As underlined in the earlier literature survey about the FRT problem and its variants, the main existing approaches to tackle tactical or strategic aspects of the problem are based on analytical approximations of specific indicators (such as the average longitudinal speed of the shuttle). Though such approaches are interesting in order to obtain quickly some information about the behavior of the system in certain situations, they have a few drawbacks, one of them being the fact that they almost systematically use a uniform spatial and time probability distribution for the customers. Such assumptions on the probability distributions is necessary in order to obtain closed form expressions and tractable dependencies on system variables. These analytical approaches have been used, e.g., to understand the impact of changing the value of the slack time at checkpoints, and hence of the shuttle schedule, on the level of service. The only optimization-based approach to tackle the FRT at the strategic level is the one of Yang et al. [2016], which tries to determine the base routes for several FRT lines at a large scale in order to capture as much demand as possible. No optimization approach exists in order to determine the shuttle fleet composition or schedule, or even the width of the service area (a parameter which can easily be changed by the transporter at virtually no cost) depending on the type of demand distribution that has to be served on a regular basis.

In this chapter, we propose a MILP formulation to handle the tactical version of the FRT where a transporting authority needs to adjust the parameters of an FRT line given some (known from historical data or anticipated) recurring demand scenarios. The approach is stochastic in nature, as it tries to determine system parameters in advance to be able to serve as best as possible an anticipated future level of demand. We suppose that the transporter possesses a limited number of demand scenarios, each

sufficiently representative of a specific situation, for example one scenario for each day of the week, plus a few scenarios representative of unusual situations like national holidays or days with exceptional peaks of demand (e.g., sports events). This approach allows us to work with a limited set of scenarios instead of having to determine an analytical form for the customers' probability distribution, which is in general more in line with the type of data available to real transporting authorities. Our approach is therefore more suited to cases with potentially complicated and non-uniform demand probability distributions. In our setting, several system parameters are promoted to variables submitted to optimization within a certain range. The impact of changing these parameters is evaluated through the resolution of the shuttle fleet scheduling for the different demand scenarios. The natural form of our model is therefore the one of a *stochastic model* where, in the traditional stochastic programming language, the system parameters allowed to vary are modeled as "here-and-now" variables and the shuttle fleet scheduling variables (the base variables of the operational MILP described in the last chapters) are modeled as "wait-and-see" variables. The here-and-now variables need to be decided before the stochastic parameters (in our case, the specific demand requirements) are known, while the wait-and-see variables are left to decide after the stochastic parameters have assumed a specific value. In a typical stochastic model, a copy of the wait-and-see variables is introduced for each scenario the model has to consider.

In the following sections, we will first introduce a necessary new body of notations and then propose a generalization of the operational model for the FRT described in precedence to the case of the tactical FRT. Our new tactical model considers as new here-and-now variables the schedule of the shuttles at checkpoints, the composition of the fleet (number of vehicles and capacity of each vehicle) as well as the width of the service area. We instead suppose that strategic parameters such as the base route and the checkpoints location are already decided and fixed. The wait-and-see variables contain the traditional variables of the operational FRT model along with new customer rejection variables, to take into account the customers that will be left aside by our tactical choices.

5.2 Stochastic programming

Stochastic programming manifests as a specialized subset of mathematical programming, explicitly crafted to address optimization issues replete with uncertainties. Contrary to deterministic programming—which operates on the presumption that all parameters are unequivocally defined at the time the decisions are made—stochastic programming purposefully integrates uncertainties into its mathematical models. This proves to be pivotal in devising solutions that are more congruent with the complexities inherent to some classes of real-world problems. The framework has found extensive utility in diverse fields such as finance, healthcare, energy systems, and supply chain management [Birge and Louveaux, 2011]. In their most general form, these problems rely on estimated probability distribution functions for their uncertain parameters. When these distributions are continuous, the stochastic problems involve (implicit) computations over an infinity of scenarios, which makes them very hard to solve. The Sample Average Approximation (SAA) consists in estimating the necessary indicators through a limited sample over the scenario space. The stochastic programs can then be expressed in a standard MILP form. This metamorphosis enables the deployment of standard optimization algorithms, thus facilitating a more efficient solution methodology. However, employing SAA necessitates caution, especially regarding the convergence of the approximate solution to the genuine optimal outcome of the original stochastic problem. Fortunately, under certain well-defined conditions—such as the bounded nature of the objective function and constraints, as well as Lipschitz continuity—SAA has been demonstrated to converge with high probability to the optimal solution [Kleywegt et al., 2002].

The reliability of SAA is substantiated by a robust mathematical foundation, confirming its proficiency as an approximation method. Explicitly, when the sample size escalates and the prescribed conditions are satisfied, the SAA solution gravitates toward the actual optimal solution, thereby solidifying its role as a dependable instrument in stochastic programming [Shapiro, 2014]. In the following, we will suppose that a limited sample of scenarios has been extracted that provides a forecast for customers' requests in

the future. Therefore, we will implicitly formulate our stochastic model in the form of an SAA.

5.3 Notations and conventions

Though some of our notations will be directly inspired by the ones introduced in Chapter 3, some specific changes need to be made to take into account the stochastic nature of our problem. Moreover, new notations need to be introduced to model the new tactical aspects with respect to the simpler operational perspective.

In most cases in stochastic programming, the value of some problem parameters is by nature stochastic, following a specific probability distribution. The value of those parameters will assume a different value in different possible scenarios, hence the wait-and-see variables, with a value to be decided after the exact value of such parameters is known, are naturally indexed over the set of possible scenarios. The tactical FRT problem, however, is based on demand scenarios, with different sets of customers and different graphs of stopping points for each such scenario. We will therefore choose a slightly different approach from most stochastic programs, indexing most of the sets of the operational FRT model instead of its variables.

Let us first introduce the change in notation for the operational scheduling problem to be solved in each scenario. The sets to be used are displayed in Table 5.1, the main difference being the indexation on the set of scenarios.

Table 5.1: Sets used for the MILP tactical formulation

Notations	Description
$RD = \{1, ..., R\}$	Set of trips with R the maximum possible number of trips.
$N^s_{0,r}$	Set of checkpoint stops of trip $r \in RD$ for scenario $s \in \mathcal{S}$, with $N^s_0 = \cup_{r \in RD} N^s_{0,r}$.
N^s_n	Set of non checkpoint stops for scenario $s \in \mathcal{S}$.
A^s	Set of arcs between the nodes of $N^s_{0,r}$.
$G^s = (N^s, A^s)$	Transportation graph with $N^s = N^s_0 \cup N^s_n$.
$V = \{1, ..., V_e\}$	Set of possible vehicles for composing the fleet, with a maximum of V_e vehicles.

Most of the parameters introduced in Chapter 3 are used in the following with the same meaning, e.g., the ready times τ for customers, travel times δ between graph nodes or service times b at nodes. However, to model the new tactical aspects, we need to introduce the new sets and parameters listed in Table 5.2.

Table 5.2: New parameters for the MILP tactical formulation

Notations	Description
ρ_k	Cost of rejecting customer $k \in K^s \setminus K^s_{PD}$ in scenario $s \in \mathcal{S}$.
k_i	Demand associated to a non-checkpoint node $i \in N^s_n$ for each scenario $s \in \mathcal{S}$.
(x_i, y_i)	Coordinates of non-checkpoint node $i \in N^s_n$ in scenario $s \in \mathcal{S}$.
o^s_r and d^s_r	Respectively origin and destination nodes of trip $r \in RD$ (starting and ending depot) for scenario $s \in \mathcal{S}$ ($o^s_r, d^s_r \in N^s_{0,r}$).
T	Time horizon of the service (the exact number of trips used by the shuttle fleet will depend on the value of T and on the number of shuttles).
ϕ_s	Probability associated with scenario $s \in \mathcal{S}$.
W_{min}	Minimum acceptable width of the service area
ω	Minimum stopping time of a shuttle between two consecutive trips.
Γ	Set of available shuttle capacities, with $Q_{max} = \max\{c \in \Gamma\}$.
γ_c	Cost of buying a shuttle with capacity $c \in \Gamma$.

Continued on next page

Table 5.2 – continued from previous page

Notations	Description
y_{max}	Maximum distance of any customer to the central route: $y_{max} = \max_{i \in \cup_{s \in \mathcal{S}} N_n^s} \{\lvert y_i \rvert\}$.
u	Customer walking speed.
Σ	Shuttle speed.
N_0	Set of checkpoint stops, independently of any specific scenario (all scenarios will have the same schedule and therefore the same physical checkpoint stops).
χ_i	Checkpoint in N_0 corresponding to a given checkpoint $i \in N_0^s$ used in scenario $s \in \mathcal{S}$.
ν_i	Next checkpoint/depot after $i \in N_0$ in the schedule.

In our model, we define γ_c and ρ_k as the total fleet costs and the costs associated with rejecting customers, respectively. For calculating the total fleet cost, we draw inspiration from [Liu et al., 2021], who provide a detailed rationale behind the computation of fleet costs. The evaluation of fleet performance hinges on the setting of realistic monetary value parameters, encompassing both operational and capital expenses. The hourly energy cost, for example, is determined by the vehicle energy consumption rate, pegged at US$ 0.19 per kilometer for pure electric vehicles. Additionally, the hourly purchase cost is derived by dividing the vehicle purchase price, typically US$ 450,000 for a traditional vehicle with an operational lifespan of 400,000 km, by its total life cycle in hours. Given that the operational speed of such vehicles ranges between 20 km/h and 30 km/h, the combined hourly operational and purchase costs amount to approximately US$ 36 per hour for a 40-seat vehicle. This cost metric can be adjusted proportionally based on different vehicle capacities, allowing for a nuanced cost evaluation tailored to specific fleet compositions. Regarding the cost of rejecting customers (ρ_k), this factor is vital for understanding the broader economic implications of service denial. To estimate this cost, one must consider several potential repercussions, including but not limited to the cost of alternative transportation modes that the customer might resort to, such as a private taxi or an additional shuttle service. The cost of rejecting a customer can be modeled as a function of these alternative costs, reflecting the immediate financial impact on the customer and the indirect cost to the service provider in terms of lost goodwill and potential future business. A reasonable assumption might involve benchmarking the rejection cost against the average market rate for a private taxi fare within the operational area, adjusted for distance and time. For example, if the average taxi fare for a similar distance and duration as the rejected service is US$ 20, this figure could serve as a baseline rejection cost. However, to account for the intangible losses associated with customer dissatisfaction and potential damage to the service provider reputation, a multiplier (e.g., 1.5x) might be applied, resulting in a calculated rejection cost of US$ 30 per incident.

Observe that we keep a "basic" set of checkpoint stops N_0, independent of the graph nodes of each scenario. This structure proves useful to link the departure times at the different checkpoints in the different scenarios (using the data structure χ, which is a map between the checkpoint nodes in the graphs G^s for all scenarios $s \in \mathcal{S}$ and the nodes of N_0). We also observe that we cannot abandon the use of the different sets N_0^s of checkpoint stops for each scenario $s \in \mathcal{S}$ for the use of the common set of checkpoints N_0, since some related quantities like the arrival time $\bar{t}$ at checkpoints can differ in each scenario. Observe that some of these variables will be needed only in certain cases, depending on the way that the customers' behavior is modeled (we will elaborate on that front later on). Note that in our approach, the wait-and-see variables are only indirectly linked to the scenarios through their belonging to scenario-indexed sets, instead of being directly indexed on the demand scenarios.

5.4 MILP formulation

5.4.1 Variables

We first introduce the wait-and-see variables related to solving the operational aspect of a given demand scenario, which generalize the variable list proposed in Chapter 3 to model the operational FRT model, in Table 5.3.

Table 5.3: Wait-and-see variables for the MILP tactical formulation

Notations	Description
$x_{i,j}^v$	Binary variable equal to 1 if arc $(i,j) \in A^s$ is used by shuttle $v \in V$ in scenario $s \in \mathcal{S}$, 0 otherwise.
$z_{k,r}$	Binary variable equal to 1 if demand $k \in K^s$ is assigned to trip $r \in RD$, 0 otherwise.
p_k, d_k	Respectively, pick-up and drop-off times for customer $k \in K^s$ of scenario $s \in \mathcal{S}$.
$\bar{t}_i, t_i$	Respectively, arrival and departure times of a shuttle at node $i \in N^s$ of scenario $s \in \mathcal{S}$.
Q_i^v	Load of vehicle $v \in V$ leaving node $i \in N^s$ in scenario $s \in \mathcal{S}$.
q_i	Actual load at node $i \in N^s$ in scenario $s \in \mathcal{S}$.
e_k	Binary variable equal to 1 if customer $k \in K^s \setminus K_{PD}^s$ from scenario $s \in \mathcal{S}$ is rejected, 0 otherwise (we never reject PD customers by convention).
Δ_i^s	Distance walked by customer $k_i \in K^s \setminus K_{PD}^s$ from/to node $i \in N_n^s$ if they are outside the service area but choose to walk to/from its border.
f_i	Binary variable equal to 1 if node $i \in N_n^s$ of scenario $s \in \mathcal{S}$ is out of the service area.
$\delta_{i,j}$	Travel time between two nodes $i, j \in N^s$ for a scenario $s \in \mathcal{S}$ can either be a fixed value or a variable depending on the constraints chosen to handle customers whose request stop ends up being located out of the service area.
$\delta_{ij}'^v$	Linearization variables representing the product $x_{i,j}^v \delta_{ij}$ in the policy where people can walk to or from the border of the service area (δ_{ij} is promoted to a variable since the coordinates of the stops can vary in that context).
$\delta_{i,j}^y$	Time to travel along the y axis between two customers, which will be necessary to define $\delta_{ij}'^v$ in the policy where customers can walk to/from the border of the service area.

Observe that some of these variables will be needed only in certain cases, depending on the way that the customers' behavior is modeled (we will elaborate on that front later on). Note that in our approach, the wait-and-see variables are only indirectly linked to the scenarios through their belonging to scenario-indexed sets, instead of being directly indexed on the demand scenarios.

There again, in order to model the new tactical decisions, we introduce here-and-now variables that will be the same for all scenarios and will mathematically provide the link between such scenarios in the constraint matrix of our linear integer model, in Table 5.4.

Table 5.4: Here-and-now for the MILP tactical formulation

Notations	Description
W	**Half** width of the service area.
θ_i	Departure time of shuttles at checkpoint nodes in N_0 (used to align the t variables on sets N_0^s for all scenarios $s \in \mathcal{S}$).
κ_{cv}	Binary variable equal to 1 if shuttle $v \in V$ is assigned capacity $c \in Q$, 0 otherwise.
α_{vr}	Binary variable equal to 1 if vehicle $v \in V$ is assigned trip $r \in RD$, 0 otherwise.

If the solution of the model proves that some of the V_e possible shuttles are not strictly necessary, these shuttles will be assigned a capacity $0 \in \Gamma$ with a related buying cost of $\gamma_0 = 0$. We now need to introduce a variable α which assigns the different shuttles to the different trips, as we do not know in advance both the number of trips and the shuttles that will be used.

5.4.2 Candidate quantities as objective functions

In a tactical setting, there are several different indicators which can be of interest to decision makers. Instead of choosing an arbitrary combination of some of these indicators, we choose a more modular

approach and list below the sensible choices to consider. Obviously, in a practical approach, a decision maker will need to decide which of these indicators will serve as true objective functions to optimize and which ones will simply be bounded by a fixed threshold. This list is composed of the following possible quantities:

1. The average total riding time of shuttles over scenarios:
$$\sum_{s\in\mathcal{S}} \pi_s \sum_{v\in V} \sum_{(i,j)\in A^s} (\delta_{i,j} x^v_{i,j})$$

2. The average customers' riding time:
$$\sum_{s\in\mathcal{S}} \frac{\pi_s}{K^s} \sum_{k\in K^s} (d_k - p_k)$$

3. The average customers' waiting time:
$$\sum_{s\in\mathcal{S}} \frac{\pi_s}{K^s} \sum_{k\in K^s} (p_k - \tau_k)$$

4. The total fleet cost:
$$\sum_{v\in V} \sum_{c\in\Gamma} \gamma_c \kappa_{vc}$$

5. The average total customer rejection cost:
$$\sum_{s\in\mathcal{S}} \pi_s \sum_{k\in K^s \setminus K^s_{PD}} \rho_k e_k$$

6. The average total time walked by rejected customers:
$$\sum_{s\in\mathcal{S}} \pi_s \sum_{i\in N^s_n} \frac{\Delta^s_i}{u}$$

The last item of this list corresponds to a situation where we suppose that the customers with a stop outside the service area will choose to walk to or from the border of the area to use the FRT system anyway. Not all the above indicators may be used as part of a global objective, however the threshold imposed on some indicators and the relative weight of the ones included in an objective function, should be chosen with great care to provide balanced and insightful solutions.

5.4.3 Constraints for a model based on a pure rejection policy

In order to write a model for the tactical FRT, we must first decide how to handle the customers who might be inconvenienced by the choice of the width of the service area. This is the case when the width is chosen in a way that either the pick-up or the drop-off (or both) of the customer ends up being located out of the service area. In such a case, the customer could be considered as rejected since we cannot serve the demand as intended in the first place. This will be the policy advocated in this subsection. We remind that customer rejection has already been handled in the literature in MILP models in Liu et al. [2021]. Later in this chapter we will consider that the customers may still decide to walk to or from the border of the service area to still partially benefit from the FRT system. In first instance, we therefore consider that a customer who asks to be picked up or dropped off outside of the area is purely and simply rejected and consequently not served. We will provide a description of the model's constraints along the

same lines as for the parameters or variables. This means that we will first provide the constraints of the operational part, i.e. the scheduling of the shuttles, for the different scenarios. Most of these constraints can be organized in the constraint matrix as an independent block involving only the operational variables of a given scenario, except for the few constraints that involve scenario-independent variables (as some of the constraints corresponding to the operational MILP now involve tactical variables). We will later list the new constraints involving the tactical variables, which link the different scenario-related blocks in the constraint matrix. The operational constraints are a modification of the operational FRT model presented in Chapter 3 and are presented below:

$$\sum_{i \in N^s} x^v_{i,j} = \alpha_{rv} \qquad \forall s \in \mathcal{S}, r \in RD, v \in V, j \in N^s_{0,r} \setminus \{o^s_r\} \qquad (5.1)$$

$$\sum_{j \in N^s} x^v_{i,j} = \alpha_{rv} \qquad \forall s \in \mathcal{S}, r \in RD, v \in V, i \in N^s_{0,r} \setminus \{d^s_r\} \qquad (5.2)$$

$$\sum_{v \in V} \sum_{i \in N^s} x^v_{i,j} = 1 - e_{k_j} \qquad \forall s \in \mathcal{S}, j \in N^s_n \qquad (5.3)$$

$$\sum_{v \in V} \sum_{j \in N^s} x^v_{i,j} = 1 - e_{k_i} \qquad \forall s \in \mathcal{S}, i \in N^s_n \qquad (5.4)$$

$$\sum_{i \in N^s:(i,j) \in A^s} x^v_{i,j} = \sum_{i \in N^s:(j,i) \in A^s} x^v_{j,i} \qquad \forall s \in \mathcal{S}, j \in N^s_n, v \in V \qquad (5.5)$$

$$p_k = t_{ps(k)} \qquad \forall s \in \mathcal{S}, k \in K^s \setminus (K^s_{PND} \cup K^s_{PD}) \qquad (5.6)$$

$$d_k = \bar{t}_{ds(k)} \qquad \forall s \in \mathcal{S}, k \in K^s \setminus (K^s_{NPD} \cup K^s_{PD}) \qquad (5.7)$$

$$\sum_{r \in RD} z_{k,r} = 1 - e_k \qquad \forall s \in \mathcal{S}, k \in K^s \qquad (5.8)$$

$$p_k \geq t_{pc(k,r)} - M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, k \in (K^s_{PND} \cup K^s_{PD}), r \in RD \qquad (5.9)$$

$$p_k \leq t_{pc(k,r)} + M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, k \in (K^s_{PND} \cup K^s_{PD}), r \in HYBR(k) \qquad (5.10)$$

$$d_k \geq \bar{t}_{dc(k,r)} - M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, k \in (K^s_{NPD} \cup K^s_{PD}), r \in HYBR(k) \qquad (5.11)$$

$$d_k \leq \bar{t}_{dc(k,r)} + M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, k \in (K^s_{NPD} \cup K^s_{PD}), r \in RD \qquad (5.12)$$

$$p_k \geq \tau_k \qquad \forall s \in \mathcal{S}, k \in K^s \qquad (5.13)$$

$$d_k \geq p_k \qquad \forall s \in \mathcal{S}, k \in K^s \qquad (5.14)$$

$$t_i = \theta_{\chi_i} \qquad \forall s \in \mathcal{S}, i \in N^s_0 \qquad (5.15)$$

$$\bar{t}_j \geq t_i + \sum_{v \in V} x^v_{i,j} \delta_{i,j} - M(1 - \sum_{v \in V} x^v_{i,j}) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s \qquad (5.16)$$

$$t_i \geq \bar{t}_i + b_i \qquad \forall s \in \mathcal{S}, i \in N^s \setminus N^s_o \qquad (5.17)$$

$$\sum_j x^v_{ps(k),j} - \sum_j x^v_{j,ds(k)} = 0, \qquad \forall v \in V, s \in \mathcal{S}, k \in K^s_{NPND} \qquad (5.18)$$

$$\sum_{r \in RD} \sum_j x^v_{pc(k,r),j} - \sum_j x^v_{j,ds(k)} = 0 \qquad \forall v \in V, s \in \mathcal{S}, k \in K^s_{PND} \qquad (5.19)$$

$$\sum_j x^v_{ps(k),j} - \sum_{r \in RD} \sum_j x^v_{j,dc(k,r)} = 0 \qquad \forall v \in V, s \in \mathcal{S}, k \in K^s_{NPD} \qquad (5.20)$$

$$q_i = \sum_{r \in RD} \sum_{k \in (K^s_{PD} \cup K^s_{PND}): pc(k,r) = i} z_{k,r} H_k - \sum_{r \in RD} \sum_{k \in (K^s_{PD} \cup K^s_{NPD}): dc(k,r) = i} z_{k,r} H_k \qquad \forall s \in \mathcal{S}, \forall i \in N^s_0 \qquad (5.21)$$

$$q_{ps(k)} = H_k \qquad \forall s \in \mathcal{S}, \forall k \in K^s_{NPD} \cup K^s_{NPND} \qquad (5.22)$$

$$q_{ds(k)} = -H_k \qquad \forall s \in \mathcal{S}, \forall k \in K^s_{PND} \cup K^s_{NPND} \qquad (5.23)$$

$$Q^v_j \geq (Q^v_i + q_j) - Q_{max}(1 - x^v_{ij}) \qquad \forall s \in \mathcal{S}, \forall (i,j) \in A^s, r \in RD, v \in V \qquad (5.24)$$

$$p_k \geq t_{o^s_r} - M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, \forall k \in K^s, r \in RD \qquad (5.25)$$

$$d_k \leq \bar{t}_{d^s_r} + M(1 - z_{k,r}) \qquad \forall s \in \mathcal{S}, \forall k \in K^s, r \in RD \qquad (5.26)$$

$$t_{o^s_r} \geq t_{o^s_{r'}} \qquad \forall s \in \mathcal{S}, r, r' \in RD : r' < r \qquad (5.27)$$

$$f_i \leq e_{k_i} \qquad \forall s \in \mathcal{S}, i \in N^s_n \qquad (5.28)$$

The above constraints are very similar to those presented in Chapter 3, with a few exceptions. For example, in order to accommodate the flexibility of the tactical point of view, we had to split the first two constraints of the base model to distinguish the cases of checkpoint stops and non-checkpoint stops. In constraints (5.1) and (5.2) the right-hand side can now be equal to 0 if the shuttle is not assigned to the trip, as we do not know in advance anymore which shuttle handles each trip (or if the trip itself is performed by a shuttle). Instead, the corresponding constraints for the non-checkpoint stops, (5.3) and (5.4), are more similar to the base model, with the difference that the right-hand-side can now be equal to zero if the customer is rejected ($e_k = 1$). The departure times of the shuttles at checkpoints in the different trips are aligned between the different scenarios thanks to constraints (5.15) through the use of the map χ and tactical variables θ. We also need to impose that the customers are handled (both pick-up and drop-off) in the same trip, which is now done in constraints (5.25) and (5.26). Finally, we impose that the different trips are chronologically ordered in constraints (5.27) (which act as a sort of symmetry breaking constraints) and we introduce the new variables f in constraints (5.28) which enforces that a customers whose non-checkpoint stop is out of the service area will be rejected.

Below, we provide the new constraints involving the tactical variables which will link the different scenarios:

$$Q^v_i \leq \sum_{c \in \Gamma} c\kappa_{cv} \qquad \forall s \in \mathcal{S}, i \in N^s, v \in V \qquad (5.29)$$

$$y_{max} f_i \geq |y_i| - W \qquad \forall s \in \mathcal{S}, i \in N^s_n \qquad (5.30)$$

$$\sum_{c \in Q} \kappa_{cv} = 1 \qquad \forall v \in V \qquad (5.31)$$

$$\sum_{v \in V} \alpha_{rv} \leq 1 \qquad \forall r \in RD \qquad (5.32)$$

$$t_{o^s_r} \geq \bar{t}_{d^s_{r'}} + \omega - M(2 - \alpha_{rv} - \alpha_{r'v}) \qquad \forall v \in V, s \in \mathcal{S}, r, r' \in RD : r' < r \qquad (5.33)$$

$$z_{k,r} \leq \sum_{v \in V} \alpha_{rv} \qquad \forall s \in \mathcal{S}, k \in K^s, r \in RD \qquad (5.34)$$

We set the upper bound on the Q variables using the fleet decision variables κ in constraints (5.29). We link the value of the variables f, i.e., the fact that a customer has a request out of the service area, to the half width of the service area W in (5.30). Constraints (5.31) force to attribute a capacity to each shuttle (Q contains capacity 0 if we do not need to use V_{max} shuttles). Constraints (5.32) assign a maximum of one shuttle per trip. Constraints (5.33) force a shuttle starting trip r to start after the end of any previous trip that may have been assigned to the same shuttle (trips are ordered chronologically). Finally, (5.34) forces the z variables to be 0 if the α of the corresponding trip is 0, and vice-versa, should we assign a request to a couple shuttle-trip, then that shuttle must serve that same trip.

5.4.4 Constraints for a model based on a walking policy

Instead of simply considering that the customers whose request is outside the service area will be automatically rejected, we can suppose that at least some of them will still want to use the service but will

simply walk to or from the border of the service area from or to their original request location. This means that they will transform their request for the closest location on the border to their original request. We must therefore, in this case, take into account the additional walking time of those customers (which can be done through the last indicator as discussed in Section 5.4.2) and transfer the stop of the customers correctly into our model so that we will only take into account the real shuttle movement and customer's riding time. Many constraints will be similar to the model we provided for the pure rejection policy. An important difference is the fact that now, the travel time on the graph arcs will be considered as a variable, meaning that we do not know for sure the coordinates of a node associated with a non-checkpoint request as it can be moved to the border of the service area. We estimate that we can still reject customers but now we need to introduce the constraints linked to the walking distance variables Δ_k^s. The constraints are therefore similar to the previous case except for constraints (5.16) and (5.28). We now need to introduce the new constraint:

$$\Delta_i^s \geq |y_i| - W - e_{k_i}(y_{max} - W_{min}) \qquad \forall s \in \mathcal{S}, i \in N_n^s \tag{5.35}$$

which pilot the value of Δ_k^s, in the case where demand k_i, associated to non-checkpoint node i, is not rejected. Given that the pick-up and drop-off stops of some customers can now be moved to the closest point on the border of the service area, parameter $\delta_{i,j}$ actually becomes a variable and the expressions $\sum_{v \in V} x_{i,j}^v \delta_{i,j}$ need to be linearized. This can be done using the newly defined variables $\delta_{i,j}^{\prime v}$ and $\delta_{i,j}^y$, which will depend on whether or not each node is out of the service area (i.e., on the value of variables f), through the following constraints:

$$\delta_{i,j}^{\prime v} \geq \frac{1}{\Sigma}|x_i - x_j| + \delta_{i,j}^y - M(1 - x_{i,j}^v) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s, v \in V \tag{5.36}$$

$$\delta_{i,j}^y \geq \frac{1}{\Sigma}|y_i - y_j| - M(f_i + f_j) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s \tag{5.37}$$

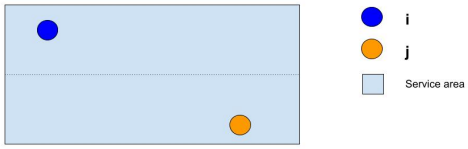
$$\delta_{i,j}^y \geq \frac{1}{\Sigma}(W - \text{sign}(y_i)y_j) - M(1 - f_i + f_j) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s \tag{5.38}$$

$$\delta_{i,j}^y \geq \frac{1}{\Sigma}(W - \text{sign}(y_j)y_i) - M(1 - f_j + f_i) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s \tag{5.39}$$

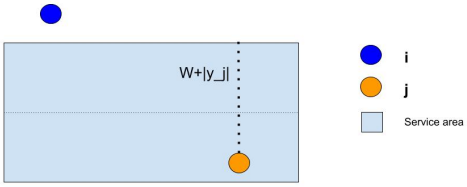
$$\delta_{i,j}^y \geq \frac{1}{\Sigma}W(1 - \text{sign}(y_i)\text{sign}(y_j)) - M(2 - f_j - f_i) \qquad \forall s \in \mathcal{S}, (i,j) \in A^s \tag{5.40}$$

$\delta_{i,j}^{\prime v}$ is supposed to replace the expression $x_{i,j}^v \delta_{i,j}$ and $\delta_{i,j}^y$ is the orthogonal part of the Manhattan distance between the two nodes (we imagine, to simplify the notation, that f_i is defined for all nodes, even though there is no sense in defining it for checkpoint nodes: we suppose it is equal to zero for such nodes). Constraints (5.36) define the product $x_{i,j}^v \delta_{i,j}$, through the use of the new variable $\delta_{i,j}^y$ representing the travel time along the y axis. Constraints (5.37) define the value of $\delta_{i,j}^y$ in the case both nodes i, j are inside the service area ($f_i = f_j = 0$), in which case the value is the same as in the classical case, i.e., the distance between the two stops along the y-axis divided by the shuttle speed. The computation is slightly more complicated when at least one of the stops is out of the service area. When, e.g., stop i is out of the service area while stop j is inside, the customer who required stop i will have their stop at the border, i.e. at y-coordinate W or $-W$ depending on the sign of the original coordinate y_i. The distance between the actual stops associated to i and j is given by the formula $W - \text{sign}(y_i)y_j$, as can be seen in constraints (5.38). Constraints (5.39) handle the symmetric case where j is out of the area while i is not. The case where both nodes are out of the service area is handled by (5.40). In this case, either both requests are on the same side of the service area and there is no travel along the y-axis, or they are on opposite sides and we need to travel the whole width of the service area, meaning $2W$.

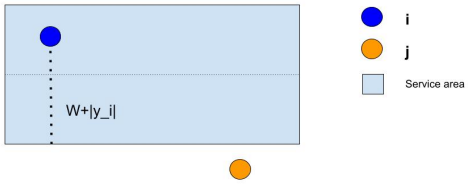
We acknowledge the necessity to further refine our model assumptions regarding customer behavior following service rejection. The initial premise, that customers whose requests are unmet within the service area might either abandon the service or walk to its border, oversimplifies a complex decision-making process and fails to consider the diverse capabilities and needs of our customer base. This



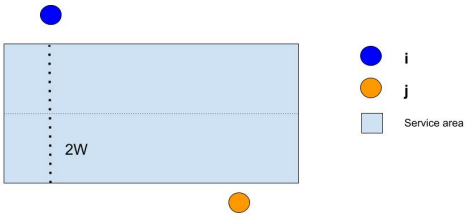
(a) Figure representing Constraints (5.37)



(b) Figure representing Constraints (5.38)



(c) Figure representing Constraints (5.39)



(d) Figure representing Constraints (5.40)

Figure 5.1: Figures representing walking policy constraints

generalization is especially inadequate for representing elderly individuals and those with disabilities, for whom walking to the service area border may not be a viable or safe option. Moreover, the feasibility and desirability of waiting at the service area border, a location that may be unsuitable or inaccessible for shuttle stopovers, were not adequately considered. In response to these insights, One could propose a model enhancement that more accurately captures the nuanced decisions customers might make when faced with service rejection. This revised model could introduce sophisticated criteria, including the option for customers to walk to the nearest checkpoint or the closest customer within the service area, thus enabling a more realistic reflection of potential customer actions. Specifically:

- Customers opting to walk to/from the closest checkpoint will see their status change from NP status to P or from ND to D, acknowledging a modified engagement with the service that accommodates for their willingness and ability to reach an agreed-upon, accessible location.
- Alternatively, customers may choose to walk to the closest customer within the service area, facilitating a collective pick-up point that supports community-centric service adjustments and promotes operational flexibility.

Incorporating these options allows the model to account for individual mobility, the availability and safety of walking paths, and the urgency or necessity of the trip, thus providing a more inclusive and realistic framework for understanding customer responses to service rejections. This approach not only ensures that our model is more reflective of diverse customer needs and preferences but also enhances the practical applicability and social responsiveness of our service design.

5.5 Solving the tactical FRT

Although we do not have numerical results to present yet, the formulations proposed in this chapter form the basis of an optimization approach to provide decisions on tactical aspects of setting up an FRT system. They should therefore be tested and refined accordingly. One specific problem of stochastic integer models is that the difficulty of solving them exactly increases sharply with the number of scenarios. This is because the scenario-related subproblem (in our case the operational scheduling of the shuttles) is often an NP-hard problem by itself. However, in our case, instead of following the classic practice in stochastic programming to generate a sample of scenarios using a mathematically well-defined probability distribution for the stochastic parameters, it may be more realistic to suppose that a decision maker will have at their disposal a small number of typical demand scenarios (corresponding, e.g., to a few week days, a weekend day or a holiday) with a limited variance. In such a case, it may be possible to obtain relevant results using our proposed formulation, depending of course on the level of demand and the time horizon considered.

An interesting use of our model would be to try to confirm or infirm results obtained using approximate analytical formulae, concerning, e.g., the level of demand at which it is necessary to switch from a one shuttle to a two shuttle roster [Lu, Xie, Wang and Quadrifoglio, 2011]. These results, though, were obtained using a uniform probability distribution of customers in space and time. This type of distribution usually has a large variance, which calls for a rather large sample of scenarios. Given the difficulty to solve stochastic MILPs when the number of scenarios grows, it might be necessary to introduce more sophisticated methods than simply solving the above MILP with a large number of scenarios. A simple option will be to create several smaller samples of scenarios and perform a statistical analysis on the results from the different samples, estimating carefully confidence intervals for the estimated optimal value [Bayraksan and Morton, 2009].

As the model becomes more difficult to solve (for, e.g., a growing level of demand or number of scenarios), solving it will require either more sophisticated mathematical programming methods (such as the introduction of valid inequalities through Branch-and-Cut or decomposition methods), or the design of an efficient (meta)heuristic algorithm.

5.6 Conclusion

In this chapter, we started from the observation that no optimization approach exists yet to tackle tactical aspects of the FRT, such as the establishment of the shuttles schedule, the width of the service area or the fleet composition. We proceeded to provide a stochastic linear integer model to solve the tactical version of the FRT where these aspects are output of our new model, based on different forecasted demand scenarios to be faced by the FRT system in the future. We listed several relevant indicators that might be used to form a proper objective function, depending on the interests of the decision makers, and provided a description of the model constraints. The model builds on the operational FRT model described and used in the previous chapters as for each demand scenario, we must solve the shuttles scheduling problem. We considered two possible cases, where the customers' requests ending up out of the service area are either rejected, or possibly accepted with the customer filling the missing part of their trip by foot, and discussed how to possibly combine and refine these policies.

In the future, this model should be tested through the use of a linear solver, in order to establish its limits and applicability. Realistically, more advanced mathematical programming or heuristic methods will be needed when the forecasted level of demand increases. This approach may be of great help to confirm or infirm previous studies on the limit of the FRT system performed through the use of analytical approximate formulae for uniform customers' distributions.

Chapter 6

Conclusion

This thesis proposes contributions to the literature on the FRT. Specifically, we aim to make a step toward the use of this flexible type of public transport system in high-demand contexts. In this section, we summarize our main contributions and we analyze some topics that we think still deserve research efforts.

6.1 Summary and conclusions

We organized this manuscript is six chapters, including this concluding one.

In Chapter 1, we introduced our research work. In particular, we described its context and motivation, and we set its objectives. Specifically, we aim to deepen the understanding of FRT systems and to enhance their scheduling solution.

In Chapter 2, we proposed a thorough literature review, in which we formally described some important MILP models for different variants of the FRT. We also underlined some similarities and differences existing with respect to the VRP and the DARP, which may be exploited to speed up the evolution of FRT solution approaches.

In Chapter 3, we presented the first modeling contributions of this thesis. These contributions concretize in additional constraint formulations, a heuristic definition to be used as a warm start for linear solvers, and some new valid inequalities to strengthen the linear relaxation of the MILP. In an experimental analysis, we studied the impact these contributions may have on the resolution of difficult instances. We observed that the combination of our new valid inequalities and warm start strongly improves the solver performance in instances with different characteristics compared to the best set of valid inequalities in the literature.

In Chapter 4, we analyzed the sensitivity of the FRT solution to the variation of some instance parameters. In particular, we study the impact of slack time, demand, and capacity on the results, as well as their interactions. It appears that demand has the strongest influence, followed by capacity and slack time, and that these parameters strongly interact. This observation underlines the central role that demand levels must assume when authorities design such a transport system. On the one hand, our results highlight the relevance of incorporating shuttle capacity into sensitivity analyses. On the other hand, they emphasize the need for stakeholders and decision-makers to accord it considerable attention in future undertakings.

In Chapter 5, we detailed a novel model for the FRT. This model allows solving the problem emerging in the tactical phase, in which most parameters have not been set yet. The model is stochastic and aims at setting these parameters to answer as well as possible to the expected level of demand. We have

considered different ways to handle customers whose requests end up being outside the service area (since its width is now a variable), which can have a non-trivial impact on the problem modeling.

6.2 Future research

In this section, we discuss the main research topics that we think deserve further investigation, including those for which we proposed a first step contribution in this thesis.

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.

Fleet 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), which is a possible extension of the sensitivity analysis conducted in this PhD thesis.

Type of shuttles
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 travelled 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. The flexibility brought by the availability of a heterogeneous fleet will necessarily have to be taken into account when designing optimization algorithms. This aspect is taken into account at modeling level in the model developed in Chapter 5, where the model chooses the shuttle capacity in accordance with various factors e.g. level of demand, etc.

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 in a dynamic environment. 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. 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, which is discussed in the literature in one paper only [Zheng et al., 2019].

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. 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.

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 and number 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. In Chapter 5, we make a first step in the direction of filling this gap by providing a model for a tactical optimization approach, however much work remains to be done to build efficient algorithms. 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 to depart 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.

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, 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.

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 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. 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, along the lines of what was proposed in Chapter 5. The model proposed in this chapter still needs to be tested to understand its computational limitations and potential.

Sensitivity analysis

Despite the fact that several studies conduct some form of partial sensitivity analysis on the system parameters, the sole approach utilized is *one factor at a time*, but for our work in Chapter 4. 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. We have proposed such an analysis in Chapter 4, focusing on the shuttle capacity. However, many other factors can be added to study the system reaction to a change in parameters, such as the probability distribution of customers, in time or in space. Additionally, advanced methodologies, such as Design of Experiments, may be employed to mitigate the computational cost associated with extensive calculations.

Bibliography

Alshalalfah, B. and 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. and Shalaby, A. [2012], 'Feasibility of flex-route as a feeder transit service to rail stations in the suburbs: case study in Toronto', Journal of Urban Planning and Development **138**(1), 90–100.

Balas, E., Fischetti, M. and Pulleyblank, W. R. [1995], 'The precedence-constrained asymmetric traveling salesman problem', Mathematical Programming **68**, 241–265.

Battarra, M., Cordeau, J.-F. and Iori, M. [2014], Chapter 6: pickup-and-delivery problems for goods transportation, in 'Vehicle Routing: Problems, Methods, and Applications, Second Edition', SIAM, pp. 161–191.

Bayraksan, G. and Morton, D. [2009], 'Assessing solution quality in stochastic programs via sampling'.

Bektaş, T. and Laporte, G. [2011], 'The pollution-routing problem', Transportation Research Part B: Methodological **45**(8), 1232–1250.

Birge, J. R. and Louveaux, F. [2011], Introduction to stochastic programming, Springer Science & Business Media.

Braekers, K. and Kovacs, A. A. [2016], 'A multi-period dial-a-ride problem with driver consistency', Transportation Research Part B: Methodological **94**, 355–377.

Cordeau, J.-F. [2006], 'A branch-and-cut algorithm for the dial-a-ride problem', Operations Research **54**(3), 573–586.

Crainic, T. G., Errico, F., Malucelli, F. and Nonato, M. [2012], 'Designing the master schedule for demand-adaptive transit systems', Annals of Operations Research **194**, 151–166.

Crainic, T., Malucelli, F. and 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', Transportation Research Part B: Methodological **18**(4-5), 315–327.

Daganzo, C. F. [1984b], 'The length of tours in zones of different shapes', Transportation Research Part B: Methodological **18**(2), 135–145.

Demir, E., Bektaş, T. and Laporte, G. [2011], 'A comparative analysis of several vehicle emission models for road freight transportation', Transportation Research Part D: Transport and Environment **16**(5), 347–357.

Demir, E., Bektaş, T. and Laporte, G. [2014], 'A review of recent research on green road freight transportation', European Journal of Operational Research **237**(3), 775–793.

Desrosiers, J., Dumas, Y. and Soumis, F. [1986], 'A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows', American Journal of Mathematical and Management Sciences **6**(3-4), 301–325.

Drexl, M. [2021], 'On the one-to-one pickup-and-delivery problem with time windows and trailers', Central European Journal of Operations Research **29**(3), 1115–1162.

Drexl, M. and Schneider, M. [2015], 'A survey of variants and extensions of the location-routing problem', European Journal of Operational Research **241**(2), 283–308.

Dumas, Y., Desrosiers, J. and Soumis, F. [1991], 'The pickup and delivery problem with time windows', European journal of Operational Research **54**(1), 7–22.

Errico, F., Crainic, T. G., Malucelli, F. and Nonato, M. [2013], 'A survey on planning semi-flexible transit systems: Methodological issues and a unifying framework', Transportation Research Part C: Emerging Technologies **36**, 324–338.

Errico, F., Crainic, T. G.and Malucelli, F. and Nonato, M. [2017], 'A benders decomposition approach for the symmetric tsp with generalized latency arising in the design of semiflexible transit systems', Transportation Science **52**, 706–722.

Errico, F., Crainic, T. G.and Malucelli, F. and Nonato, M. [2021], 'The single-line design problem for demand-adaptive transit systems: A modeling framework and decomposition approach for the stationary-demand case', Transportation Science **55**, 1300–1321.

Fu, L. [2002], 'Planning and design of flex-route transit services', Transportation Research Record **1791**(1), 59–66.

Gschwind, T., Irnich, S., Rothenbächer, A.-K. and Tilk, C. [2018], 'Bidirectional labelling in column-generation algorithms for pickup-and-delivery problems', European Journal of Operational Research **266**, 521–530.

Helton, J. C., Johnson, J. D., Sallaberry, C. J. and Storlie, C. B. [2006], 'Survey of sampling-based methods for uncertainty and sensitivity analysis', Reliability Engineering & System Safety **91**(10-11), 1175–1209.

Ho, S. C., Szeto, W. Y., Kuo, Y.-H., Leung, J. M., Petering, M. and Tou, T. W. [2018], 'A survey of dial-a-ride problems: Literature review and recent developments', Transportation Research Part B: Methodological **111**, 395–421.

Kleywegt, A. J., Shapiro, A. and Homem-de Mello, T. [2002], 'The sample average approximation method for stochastic discrete optimization', SIAM Journal on optimization **12**(2), 479–502.

Koffman, D. [2004], Operational experiences with flexible transit services, Vol. 53, Transportation Research Board.

Konstantakopoulos, G. D., Gayialis, S. P. and Kechagias, E. P. [2022], 'Vehicle routing problem and related algorithms for logistics distribution: A literature review and classification', Operational research **22**(3), 2033–2062.

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

Li, M. and 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. and Ma, X. [2021], 'Improving flex-route transit services with modular autonomous vehicles', Transportation Research Part E: Logistics and Transportation Review **149**, 102331.

Lo, S.-C. and 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. and Quadrifoglio, L. [2011], 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. and Quadrifoglio, L. [2011], 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 Journal of Operational Research **175**, 210–223.

Malucelli, F., Nonato, M. and Pallottino, S. [1999], Demand Adaptive Systems: Some Proposals on Flexible Transit, Palgrave Macmillan UK, London, pp. 157–182.

Masmoudi, M. A., Hosny, M., Demir, E., Genikomsakis, K. N. and Cheikhrouhou, N. [2018], 'The dial-a-ride problem with electric vehicles and battery swapping stations', Transportation research part E: Logistics and Transportation Review **118**, 392–420.

Moghdani, R., Salimifard, K., Demir, E. and Benyettou, A. [2021], 'The green vehicle routing problem: A systematic literature review', Journal of Cleaner Production **279**, 123691.

Oakley, J. E. and O'Hagan, A. [2004], 'Probabilistic sensitivity analysis of complex models: a bayesian approach', Journal of the Royal Statistical Society Series B: Statistical Methodology **66**(3), 751–769.

Palmer, K., Dessouky, M. and Abdelmaguid, T. [2004], 'Impacts of management practices and advanced technologies on demand responsive transit systems', Transportation Research Part A: Policy and Practice **38**(7), 495–509.

Pigeon, J. G. [2006], 'Statistics for experimenters: Design, innovation and discovery'.

Potts, J. F., Marshall, M. A., Crockett, E. C. and Washington, J. [2010a], 'A guide for planning and operating flexible public transportation services', TCRP report **140**.

Potts, J., Marshall, M., Crockett, E. and Washington, J. [2010b], Transit Cooperative Research Programme Report 140: A guide for planning and operating flexible public transportation services, Technical report, Transportation Research Board.

Pratelli, A. and Schoen, F. [2001], 'A mathematical programming model for the bus deviation route problem', Journal of the Operational Research Society **52**, 494–502.

Qiu, F., Li, W. and An, C. [2014], 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. and Haghani, A. [2015], 'An exploration of the demand limit for flex-route as feeder transit services: a case study in salt lake city', Public Transport **7**, 259–276.

Qiu, F., Li, W. and Zhang, J. [2014], 'A dynamic station strategy to improve the performance of flex-route transit services', Transportation Research Part C: Emerging Technologies **48**, 229–240.

Quadrifoglio, L. and 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. Calgary: Acta Press, 6pp'.

Quadrifoglio, L. and Dessouky, M. M. [2008], Sensitivity analyses over the service area for mobility allowance shuttle transit (mast) services, in M. Hickman, P. Mirchandani and S. Voß, eds, 'Computer-aided Systems in Public Transport', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 419–432.

Quadrifoglio, L., Dessouky, M. M. and Ordóñez, F. [2008a], 'Mobility allowance shuttle transit (MAST) services: MIP formulation and strengthening with logic constraints', European Journal of Operational Research **185**(2), 481–494.

Quadrifoglio, L., Dessouky, M. M. and Ordóñez, F. [2008b], 'Mobility allowance shuttle transit (MAST) services: MIP formulation and strengthening with logic constraints', 5th International Conference, CPAIOR 2008 Paris, France **5015**, 398–404.

Quadrifoglio, L., Dessouky, M. M. and Palmer, K. [2007], 'An insertion heuristic for scheduling mobility allowance shuttle transit (mast) services', Journal of Scheduling **10**(1), 25–40.

Quadrifoglio, L., Hall, R. W. and Dessouky, M. M. [2006], 'Performance and design of mobility allowance shuttle transit services: bounds on the maximum longitudinal velocity', Transportation Science **40**(3), 351–363.

Quadrifoglio, L. and Li, X. [2009], 'A methodology to derive the critical demand density for designing and operating feeder transit services', Transportation Research Part B: Methodological **43**(10), 922–935.

Quadrifoglio, L. and Shen, C.-W. [2010], Performance analysis of the "zoning" strategies for ada para-transit services, Southwest Region University Transportation Center, Texas Transportation.
URL: *https://static.tti.tamu.edu/swutc.tamu.edu/publications/technicalreports/169114-1.pdf*

Rahmani, N., Detienne, B., Sadykov, R. and 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. and Salani, M. [2006], 'Symmetry helps: Bounded bi-directional dynamic programming techniques for the elementary shortest path problem with resource constraints', Discrete Optimization **3**(3), 255–273.

Ropke, S. and Cordeau, J.-F. [2009], 'Branch and cut and price for the pickup and delivery problem with time windows', Transportation Science **43**(3), 267–283.

Ropke, S., Cordeau, J.-F. and Laporte, G. [2007], 'Models and branch-and-cut algorithms for pickup and delivery problems with time windows', Networks **49**(4), 258–272.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S. [2010], 'Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index', Computer physics communications **181**(2), 259–270.

Savelsbergh, M. and Sol, M. [1998], 'Drive: Dynamic routing of independent vehicles', Operations Research **46**(4), 474–490.

Shahin, R., Hosteins, P., Pellegrini, P. and Vandanjon, P.-O. [2023], A full factorial sensitivity analysis for a capacitated flex-route transit system, in '2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)', IEEE, pp. 1–6.

Shahin, R., Hosteins, P., Pellegrini, P., Vandanjon, P.-O. and Quadrifoglio, L. [2024], 'A survey of flex-route transit problem and its link with vehicle routing problem', Transportation Research Part C: Emerging Technologies **158**, 104437.

Shapiro, A. [2014], Chapter 5: Statistical inference, in 'Lectures on Stochastic Programming: Modeling and Theory, Second Edition', SIAM, pp. 163–269.

Shen, J., Qiu, F., Zheng, C. and 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. and 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. and Grunder, O. [2020], 'A lexicographic-based two-stage algorithm for vehicle routing problem with simultaneous pickup–delivery and time window', Engineering Applications of Artificial Intelligence **95**, 103901.

Sims, R., Schaeffer, R., Creutzig, F., Cruz-Núñez, X., D'Agosto, M., Dimitriu, D., Meza, M. F., Fulton, L., S. Kobayashi, O. L., McKinnon, A., Newman, P., Ouyang, M., Schauer, J., Sperling, D. and 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, chapter 8, pp. 599–670.

Sipetas, C. and Gonzales, E. J. [2021], 'Continuous approximation model for hybrid flexible transit systems with low demand density', Transportation Research Record **2675**(8), 198–214.

Sun, C., Li, W., Shao, Z. and 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. and Liu, S. [2022], 'Research on route deviation transit operation scheduling—a case study in suburb no. 5 road of harbin', Sustainability **14**(2), 633.

Tsui, K.-L. [1992], 'An overview of taguchi method and newly developed statistical methods for robust design', Iie Transactions **24**(5), 44–57.

Wilcox, R. R. [2011], Introduction to robust estimation and hypothesis testing, Academic press.

Yang, H., Cherry, C. R., Zaretzki, R., Ryerson, M. S., Liu, X. and Fu, Z. [2016], 'A gis-based method to identify cost-effective routes for rural deviated fixed route transit', Journal of Advanced Transportation **50**(8), 1770–1784.

Yu, J., Li, W., Zhang, J., Guo, R. and Zheng, Y. [2023], 'Understanding the effect of socio-demographic and psychological latent characteristics on flex-route transit acceptance', Plos one **18**(2), ahead of print.

Yu, J., Zheng, Y., Li, W., Zhang, J., Guo, R. and Wu, L. [2023], 'Understanding flex-route transit adoption from a stage of change perspective', Transportation Research Record p. ahead of print.

Yu, Y., Wang, S., Wang, J. and Huang, M. [2019], 'A branch-and-price algorithm for the heterogeneous fleet green vehicle routing problem with time windows', Transportation Research Part B: Methodological **122**, 511–527.

Zhang, J., Li, W., Wang, G. and 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. and Guo, R. [2023], 'Dynamic clustering meeting points strategy to improve operational service capability of flex-route transit', Journal of Transportation Engineering, Part A: Systems **149**(6).

Zhang, Y., Farber, S. and Young, M. [2022], 'Eliminating barriers to nighttime activity participation: the case of on-demand transit in belleville, canada', Transportation **49**, 1385–1408.

Zhao, J. and Dessouky, M. [2008], 'Service capacity design problems for mobility allowance shuttle transit systems', Transportation Research Part B: Methodological **42**(2), 135–146.

Zheng, Y., Gao, L. and Li, W. [2021], 'Vehicle routing and scheduling of flex-route transit under a dynamic operating environment', Discrete Dynamics in Nature and Society **2021**, 1–10.

Zheng, Y. and Li, W. [2019], Flex-route transit service with different degree of dynamism, in 'CICTP 2019', ASCE library, pp. 4369–4380.

Zheng, Y., Li, W. and Qiu, F. [2018a], 'A methodology for choosing between route deviation and point deviation policies for flexible transit services', Journal of Advanced Transportation **49**(3), 496–509.

Zheng, Y., Li, W. and Qiu, F. [2018b], 'A slack arrival strategy to promote flex-route transit services', Transportation Research Part C: Emerging Technologies **92**, 442–455.

Zheng, Y., Li, W., Qiu, F. and Wei, H. [2019], 'The benefits of introducing meeting points into flex-route transit services', Transportation Research Part C: Emerging Technologies **106**, 98–112.

Zheng, Y., Li, W., Qiu, F. and Wei, H. [2020], 'Travelers' potential demand toward flex-route transit: Nanjing, China, case study', Journal of Urban Planning and Development **146**(1), 05019018–10.