IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 February 2023, accepted 14 March 2023, date of publication 22 March 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260727

RESEARCH ARTICLE

Optimal Design of Integrated Semi-Flexible Transit Services in Low-Demand Conditions

SUSHREETA MISHRA AND BABAK MEHRAN
Urban Mobility and Transportation Informatics Group (UMTIG), Department of Civil Engineering, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
Corresponding author: Babak Mehran (babak.mehran@umanitoba.ca)

ABSTRACT Semi-flexible transit (SFT) is commonly discussed as a cost-effective alternative to serving public transportation users in low-demand conditions. We hypothesize that joint optimization of service headway and slack time per trip for route deviation is essential for designing a schedule for the operation of an integrated SFT that can meet both fixed-route and paratransit demand. An integrated SFT has the potential to lower the cost of transportation for regular transit users (both operators and riders) while redirecting potential paratransit riders to less expensive transit modes; thus, reducing demand for overwhelmed paratransit services operating with limited resources. The optimization problem has three competing objectives: minimizing operator costs, minimizing user costs, and maximizing service benefits. Two state-of-the-art multi-objective evolutionary algorithms NSGA-II and SMPSO are compared to obtain the most representative deterministic Pareto optimal solution set. This study has three major contributions. First, quantile regression is used to suggest multiple slack time values for a given headway that transit planners can consider when generating a static schedule for SFT operation. Second, relationships derived to analyze cost trade-offs suggest that headway governs operator cost and is negatively correlated, user cost is positively and equally influenced by both variables, and slack time governs service benefit and is positively correlated. Third, sensitivity analysis for an integrated SFT operation reveals that low-capacity minivans and standard vans offer higher vehicle occupancy and cost efficiency, mostly economical for low to medium demand (5-20 pass/hr), low permissible deviation from the fixed route is desirable during peak hours to avoid delays for passengers on-board, and extreme weather conditions dramatically and negatively influence costs. Policy recommendations for integrated SFT implementation include a recommendation for fare structure design addressing service equity through surcharges/discounts, vehicle technology and service booking technology advancements for cost reduction, and fleet mix design through estimation of passenger loading profile. The application of the study methodology is demonstrated for a low-demand bus route in Regina, Canada.

INDEX TERMS Paratransit, semi-flexible transit, service headway, slack time.

I. INTRODUCTION

Current trends in the economy and societal changes have contributed to low and dispersed travel demand, which is critical to the operation of rigid forms of transit like fixed-route bus transit (FBT). Demand responsive transit (DRT) providing on-demand curb-to-curb services to all passengers is a common alternative but is only limited to providing specialized services, like paratransit because of its relatively high cost of operation. Semi-flexible transit (SFT) which combines the rigidity of FBT and the flexibility of DRT is the most discussed alternative in the past two decades to serve low-demand travel needs [1], [2], [3]. Several transit agencies in North America suggest SFT as an interesting solution to respond to the growing demand for expensive paratransit services and the high operating cost of FBT in low-demand conditions [4]. Regina transit in Canada facing a similar situation has recently piloted a flexible transit service along a least frequently used route hoping to increase ridership and the vehicle used is accessible with spots

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

reserved for wheelchairs/mobility devices [5]. While paratransit demand continues to grow in many North American communities, some transit operators, including Regina Transit, are providing training and incentives to help shift trips from paratransit/DRT to less costly modes like FBT and SFT [6]. Regina paratransit, like most others, is currently operating at its capacity [7]. Consequently, SFT with route deviations, when provided along an under-performing low-demand bus route, is capable of substituting both FBT and DRT in the service zone and is beneficial to both shifted regular transit and paratransit users as well as the operator [4], [6]. Paratransit users will benefit from improved response times and flexible mobility, thereby improving their social inclusion. There will be more resources at the disposal of transit agencies for paratransit service in the same or other zones, in addition to cost savings. General transit users could expect a discounted fare as the total operating cost reduces. The modeling and optimization of SFT to provide an efficient service is a challenging and complex task since it requires retaining both FBT and DRT properties [1]. Due to its composite nature, SFT operates with a fixed/flexible route, stop, and schedule to accommodate a few curb-to-curb stop requests. While several design parameters govern SFT, the determination of the amount of slack time allocated in the schedule to accommodate route deviations for serving paratransit users is the most critical [8]. Other key design parameters for SFT include zone size/service area, headway, slack time distribution, and demand [1], [4].

This study hypothesizes that joint optimization of service headway (h) and slack time per trip (Δt) is necessary to design SFT that will accommodate existing FBT and paratransit demand along an under-performing low-demand bus route. There are three conflicting objectives in this optimization problem: minimization of operator cost, minimization of user cost, and maximization of service benefit. We conducted descriptive and statistical analyses of Pareto optimal solutions obtained using multi-objective evolutionary algorithms to derive meaningful relationships between the decision variables and the cost components. Policy recommendations for fares, technology, and operations are derived based on proposed models and sensitivity analyses conducted for vehicle capacity, hourly demand, permitted deviation, and weather conditions. The application of the study methodology is demonstrated for a low-demand bus route in Regina, Canada.

II. LITERATURE REVIEW

A large body of literature concurs that it is cost-efficient to operate SFT in low-demand conditions and FBT when the demand for transit is high [2], [3]. Several studies have emphasized this concept, including this study which expands on previous research by addressing how to best operate the SFT. Table 1 summarizes the studies mostly related to this study and is discussed in detail in this section.

A. KEY ELEMENTS OF FLEXIBLE TRANSIT SERVICE DESIGN

This section focuses on defining the key elements of flexible transit optimization models studied in literature including decision variables, objective(s), constraints, and solution method. The decision variables or optimized variables for SFT commonly include routing and scheduling [9], service zone size [10], [11], [12], passenger request [13], headway [11], [14], [15], velocity [16], and slack time [8], [10], [17], [18]. Most objectives can be classified into two categories: operator-related and user-related cost and service benefit. Operator costs include the minimization of fleet acquisition and operation costs [14]. User-related objectives include the minimization of travel time components such as access time, waiting time, and in-vehicle time [8], [17]. To attain system-wide savings, few studies optimize total cost considering both operator and user-related objectives [15]. The benefit associated is specific to the operators' intent like increasing revenue/fare income [9], reducing parking infrastructure investment [17], increasing mobility, reducing vehicle miles and emissions, or replacing a costly transit alternative DRT/FBT for paratransit passengers and passengers in suburban or rural areas [8]. The design of SFT includes constraints characteristic of regular bus transit design including capacity [11], vehicle arrival and departure schedule [9], travel time [8], and fleet size [19] in addition to including constraints characteristic of DRT like zoning [10] and passenger pick-up and drop-off schedule [9]. Finally, for a given set of objective functions and constraints, the optimal value of decision variables can be derived using analytical models [11], numerical approximation [20], simulation [8], and heuristics [21]. In this study, using heuristic methods, we derive optimal values of slack time and headway that minimize operator and user costs while maximizing the benefit defined as the cost of serving paratransit demand in an expensive DRT mode if not serviced by SFT, when vehicle capacity is constrained.

B. STUDIES FOCUSING ON SLACK TIME OPTIMIZATION FOR FLEXIBLE TRANSIT

Fu [8] proposed the first analytical model to determine the optimal slack time for a flex service to accommodate door-to-door paratransit requests while serving mandatory stops along the route. This model minimizes the total net cost to all stakeholders, including the operator, and regular and paratransit passengers. The fundamental relationships between system performance and design parameters revealed using an analytical model are further validated using simulation. Despite capturing some general trends, the models developed failed to capture the details of system behavior. Smith et al. [10] implemented a heuristic method to optimize two key design variables in flex-route service planning: service area and slack time distribution. The optimization problem included two objectives: maximization of feasible deviations (i.e., from the operator's perspective) and minimization of dwell time/unused slack time (i.e., from the user's perspective). Two existing fixed routes with a maximum of five major fixed stops were chosen to serve as flex routes. Assuming a deterministic scenario, this study uses the gradient method to derive Pareto-optimal solutions for transit planners to assess design trade-offs. Alshalalfah [17] implemented analytical modeling and constraint programming for static and dynamic flex-route service optimization. The system is designed to cover mandatory stops with a predetermined schedule while accommodating on-demand route deviation requests constrained by slack time in the schedule. The study derived optimal values for service area and slack time that minimizes the operator and user costs and maximizes the savings in parking costs when encouraging people to switch from using their cars to using transit when accessing a regional rail network. Zheng et al. [22] proposed a slack arrival strategy to improve the acceptance rate of the flex-route service at both expected and unexpected demand levels. Analytical and simulation models are developed to investigate the optimal slack time window based on system cost, including the operator and user costs. Studies cited above have focused on designing SFT and most studies, except these, analyze slack time differently. Alshalalfah and Shalaby [18] conducted sensitivity analyses with various slack time values (0 to 12 minutes) to study its effect on the number of accepted demand-responsive requests. Quadrifoglio et al. [16] suggested that the maximum slack time between checkpoints for mobility allowance shuttle transit (MAST) vehicles could be set by the minimum threshold longitudinal velocity value while minimizing the total distance traveled. Lai et al. [21] considered slack ratio in designing flexible transit system elements such as path, pick-up and drop-off location, and schedule that maximizes vehicle sharing and the number of accepted requests while minimizing the walking time. These studies do not, however, focus on identifying the optimal slack time value.

C. STUDIES FOCUSING ON HEADWAY OPTIMIZATION FOR FLEXIBLE TRANSIT

Kim and Schonfeld [15] implemented a probabilistic optimization model to determine optimal vehicle capacity, headway, and fleet size that minimizes the passenger transfer cost in integrated conventional and flexible feeder systems with coordinated transfers. The feeder system offers door-to-door service and follows a predetermined schedule to make timed transfers. Wang et al. [12] derived an analytical model to identify zone size and headway that minimizes both operator and user costs when designing a many-to-one DRT between a residential area and a terminal. Nourbakhsh and Ouyang [20] proposed a DRT bus service along a service area designed as a hybrid of hub-and-spoke and grid networks. The study optimizes the network layout, service area, and headway based on operator and user cost using numerical approximation. Likewise, for a many-to-one flexible door-to-door feeder system, Kim et al. [11] suggested that joint optimization of service headway and zone size is essential for minimizing the total system cost. This paper implemented Newton's method to solve for the optimal values while constraining the vehicle capacity. Estrada et al. [14] determined the optimal vehicle technology (i.e., diesel, electric, and autonomous), service pattern (i.e., SFT, FBT, and DRT), and vehicle size (i.e., mini-bus, van, bus, and car) for varying demand density. Enumeration procedure implemented to identify headway, stop spacing, and waiting time for the above scenarios that minimizes the total cost to the operator and user constrained by capacity.

As a final consideration, we note that studies optimizing the decision variables that are elements of strategic planning or tactical planning for SFT like headway and slack time are very limited and are essential to define a timetable for SFT operation [23]. Although Alshalalfah [17] and Fu [8] suggested some interaction between slack time and headway, they focused on the optimization of slack time, assuming a fixed value of headway for system design.

Joint optimization of slack time and headway for an integrated service has not yet been addressed in the literature, which mostly concentrates on optimizing them separately. To create a static schedule for integrated SFT operations that accommodates both existing fixed route transit demand and shifted paratransit demand, joint optimization is essential, and to monitor the impact of changes in headway on optimal values of slack time as well as the impact of these variations on operator and user costs and benefits. The motivation for joint optimization is derived from Kim et al. [11] who compared two optimization scenarios for flexible-bus service: 1) One decision variable, zone size considering the maximum allowable headway policy, and 2) Joint optimization of headway and zone size. According to the study, scenario 1 has a 26% greater average cost per passenger trip and an 83% larger optimal zone size than scenario 2. When compared to scenario 2, scenario 1 proposes solutions that reduce operator costs but increase in-vehicle and waiting costs.

III. PROBLEM DESCRIPTION

A. SERVICE AREA AND DEMAND

An existing underperforming fixed bus route is used as the study area, which is defined by two terminal stations and modeled as a rectangle with dimensions W (km) and L (km) (see Fig. 1). SFT in this study is designed to serve two types of passenger demand: (a) $\mathrm{Q_G}$- existing FBT demand (Type G), and (b) $\mathrm{Q_S}$- existing DRT demand referring to users that are eligible for paratransit service in the study area (Type S). It is assumed that the demand per trip is uniformly and independently distributed within the service area.

B. OPERATING POLICY

For SFT, we adopt the route-deviation policy defined by Koffman [4] where vehicles follow a fixed route and deviate to serve curb-to-curb requests, with a maximum allowable deviation of $W/2$ on both sides (see Fig. 1). This operating policy accepts two types of stop requests: flag requests and curb-to-curb requests. Flag requests involve vehicles stopping at any location along the route, which may or may not correspond to a marked stop. Curb-to-curb requests involve vehicles deviating from their fixed route to serve pick-up and drop-off locations requested by passengers in advance

TABLE 1. Summary of studies discussed.

Year	Reference	System	Method	Objective	Constraint	Decision variable
2002	Fu [8]	SFT	AN, SM	Max. OC, Min. UC, & Max. SB	TT, SC	Slack time
2003	Smith et al. [10]	SFT	HT	Min. OC & Min UC	SC	Service area and slack time distribution
2006	Quadrifoglio et al. [16]	DRT	AN, SM	Min. OC	SC	Longitudinal velocity
2009	Alshalalfah [17]	SFT	AN, SM	Max. OC, Min. UC, & Max. SB	SC, CP	Service area and slack time distribution
2012	Alshalalfah and Shalaby [18]	SFT	AN, SM	Max. OC	SC, CP	Slack time
2012	Nourbakhsh et al. [20]	DRT	NA	Min. OC & Min UC	-	Network layout, service area, and headway
2014	Kim and Schonfeld [15]	SFT	AN, HT	Min. OC & Min UC	SC	Vehicle capacity, headway, and fleet size
2018	Wang et al. [12]	DRT	AN	Min. OC & Min UC	CP	Service area
2018	Zheng et al. [22]	SFT	AN, SM	Min. OC & Min UC	SC	Slack arrival strategy
2019	Kim et al. [11]	DRT	AN	Min. OC & Min UC	CP	Service area and headway
2019	Pei et al. [9]	SFT	HT	Max.[SB-(OC+UC)]	SC	Stop locations and routes
2021	Estrada et al. [14]	DRT &FBT	AN	Min. OC & Min UC	CP	Stop distances, headways, or waiting time
2022	Lai et al. [21]	SFT	HT	Min. OC & Min UC	CP	Path, pick-up and drop-off location, vehicle schedule
Proposed		SFT	HT	Max. OC, Min. UC, & Max. SB	CP	Service headway and slack time

Method: NA- Numerical approximation, HT- Heuristic, SM- Simulated, AN- Analytical; *Objective:* OC- Operating cost, UC- User cost, TC- Total cost, SB- Service benefit; *Constraints:* CP- Capacity, TT- travel time for vehicle/passenger, SC- vehicle/passenger schedule constraints, ZN- zoning/service area constraints, F- Fleet size

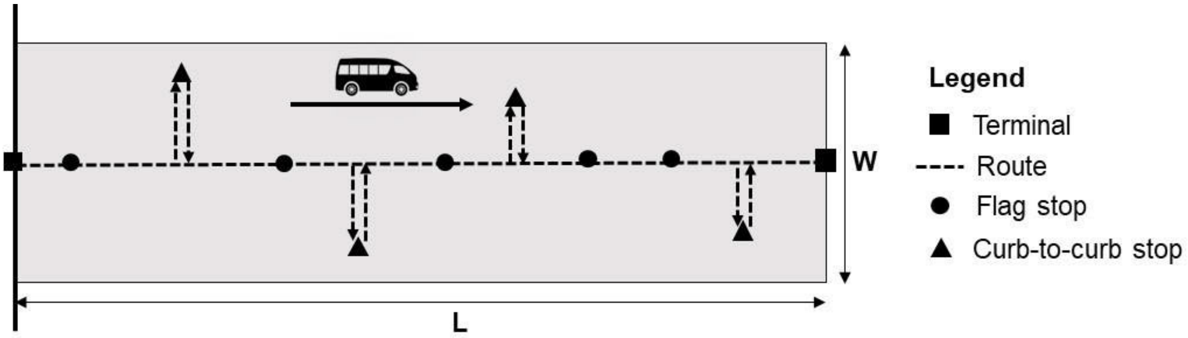


FIGURE 1. A schematic representation of route-deviation operating policy.

(usually 1 hour). Flag-stop in an integrated service can be requested by both Type G and Type S passengers; however, curb-to-curb stop requests in this study are restricted to Type S passengers. An online dispatch system is assumed to handle curb-to-curb service requests and routing, and a predetermined timetable, including slack time, is published for all or a few stops along the fixed route to assist Type G and Type S passengers in planning their arrival at the route. Within the service area, Type S passengers can request four possible types of requests with proportions of η_{R1}, η_{R2}, η_{R3}, and η_{R4} ($\eta_{R1} + \eta_{R2} + \eta_{R3} + \eta_{R4} = 1$) as follows:

R1: Both pick-up and drop-off locations are along the fixed route (Both flag stops)

R2: Drop-off location deviates from the fixed route and the pick-up location is along the fixed route (flag stop and curb-to-curb stop)

R3: Pick-up location deviates from the fixed route and the drop-off location is along the fixed route (flag stop and curb-to-curb stop)

R4: Both pick-up and drop-off locations deviate from the fixed route (Both curb-to-curb stops)

C. PROBLEM DEFINITION

In SFT, adding slack time increases the benefit derived from the service but also increases the one-way running time of vehicles, which can be compensated for by increasing the service headway to reduce fleet size requirements, this, however, will increase the waiting time and reduce the available vehicle capacity for serving Type S passengers with more Type G passengers queued up for the service. Thus, joint optimization of slack time and headway will ensure maximum vehicle utilization for a desired level of service. This is a multi-objective optimization problem where we are interested in finding a set of solutions that define the best tradeoff between competing objectives: minimizing operator costs, minimizing user costs, and maximizing service benefits.

IV. PROBLEM FORMULATION

A. DECISION VARIABLE

The study presents a multi-objective optimization model for service headway, h (hr), and slack time required to serve Type S passengers, Δt (hr) within the service area. Here, h represents the time difference between the arrival of two vehicles at any stop, and Δt represents the one-way trip time difference when the potential Type S passengers are accommodated in the trip against serving only Type G passengers.

B. OBJECTIVE FUNCTION

The three conflicting objective functions considered are f_1: minimization of operator cost (C_{OC}), f_2: minimization of user cost (C_{UC}), and f_3: maximization of service benefit (C_{SB}), as shown in Equation (1) – (3)

$$\text{Min}\, f_1 \rightarrow C_{OP}\,(h, \Delta t) \tag{1}$$

$$\text{Min}\, f_2 \rightarrow C_{UC}\,(h, \Delta t) \tag{2}$$

$$\text{Max}\, f_3 \rightarrow C_{SB}\,(h, \Delta t) \tag{3}$$

C. CONSTRAINTS

Constraint 1: Equation (4) depicts that h is constrained by a minimum and maximum service headway. Literature consensus that for a low-demand route characterized by low-frequency service, the minimum headway can be set to 10 minutes [24]. The minimum desired level of service is set through policy headway, $h_{\mathrm{p},}$ and the vehicle capacity, C governs the maximum value.

$$0.167 \leq h \leq \min\left\{\frac{C}{Q_G + Q_S}, h_p\right\} \tag{4}$$

where C is capacity in pass/veh, $Q_{G,}$ and Q_S represent hourly Type G and Type S demand in pass/hr, and h_p is the policy headway in hr.

Constraint 2: As shown in (5), slack time, Δt, should vary between 0 and a maximum value based on the SFT vehicle capacity (C) and the average time to serve one passenger/ paratransit user (δ). A minimum value corresponds to no Type S passengers served (i.e., Type G passengers only), and a maximum value corresponds to all passengers on board are Type S (i.e., No Type G passengers).

$$0 \leq \Delta t \leq C\delta \tag{5}$$

Constraint 3: Constraint 3 in (6) limits the number of Type S passengers served per one-way trip between terminals to the available vehicle capacity after serving Type G passengers. This imposes Type G flow and holds priority in the assignment of the vehicle capacity since an alternate mode DRT is available to Type S passengers if not accommodated in SFT. Also, it is assumed that the rejected boarding requests by Type S passengers may be accommodated in the next trip.

$$Q_R h + \frac{\Delta t}{\delta} \leq C \tag{6}$$

C4: Constraint 4 in (7) ensures that the number of Type S passengers served cannot exceed the demand received during the service interval (i.e., h).

$$\frac{\Delta t}{\delta} \leq Q_S h \tag{7}$$

Estimation of δ

Based on Table 2, the time required to serve a given request type, R1 to R4 is composed of (a) riding time, (b) acceleration and deceleration time, and (c) dwell time. For instance, the riding time required to serve request type R4 includes the time required to deviate an average of $W/4$ from the fixed route to serve a curb-to-curb stop and the same W/4 distance back to the fixed route for both pickup and drop-off. Equation (8) based on conditional probability theory, is used to estimate the expected time to serve one Type S passenger (δ).

$$\begin{aligned}\delta &= \eta_{R1}\delta_{R1} + \eta_{R2}\delta_{R2} + \eta_{R3}\delta_{R3} + \eta_{R4}\delta_{R4}\\ &= \frac{W}{2V_R}\left(\frac{\eta_{R2} + \eta_{R3}}{2} + \eta_{R4}\right) + 2t_{ad} + 2t_d\end{aligned} \tag{8}$$

where V_R is average riding speed (km/hr), t_{ad} is acceleration and deceleration per stopping (hr), and t_d is dwell time per stopping for boarding or alighting (hr).

D. ANALYTICAL COST MODELS FOR DETERMINISTIC ANALYSIS

1) OPERATOR COST

SFT operating cost (C_{OC}) estimated in \$/hr is defined as a function of Fleet Size, M as given in (9).

$$C_{OC} = c_1\,[M] \tag{9}$$

where c_1 (\$/veh-hr) is the unit cost of operating a transit unit including fleet acquisition cost, and distance and time-based cost [25].

Equation (10) expresses M as a ratio of the total round-trip time, T_R (hr), and headway, h (hr). T_R consists of three components: (a) time required for serving Type G passengers, T_v (hr), (b) layover time at the terminal station, T_l (hr), and (c) slack time to serve Type S passengers, Δt (hr). T_v includes total riding time, time for vehicle acceleration and deceleration, and dwell time as given in (11). For the estimation of dwell time, the number of stops is assumed to be twice the number of passengers boarded in a vehicle, which holds in low-demand situations [26].

$$M = \left[\frac{T_R}{h}\right]^+ = \left[\frac{2(T_v + T_l + \Delta t)}{h}\right]^+ = \left[\frac{2\left(T_v\left(1+\mu\right)+\Delta t\right)}{h}\right]^+ \tag{10}$$

$$T_v = \frac{L}{V_R} + \left(2t_{ad} + 2t_d\right)\left(Q_R h\right) \tag{11}$$

where $\mu = T_l \,/\, T_v$.

2) USER COST

The user cost (C_{UC}) is the sum of the costs of the three equivalent time components, access/egress time, C_A, waiting time, C_W, and in-vehicle time, C_I, estimated in \$/hr given in (12). The product of passenger value of time, c_2 (\$/pass-hr), and equivalent time components $\mathrm{T_a}$ (hr), $\mathrm{T_w}$ (hr), and $\mathrm{T_v}$ (hr) as given in Equation (12) are used to estimate $\mathrm{C_{UC}}$. Estimation of Ta and Tv are based on the microeconomic models for vehicle resource consumption derived by Mohring [27] and the model for Tw estimation is based on vehicle and passenger arrival patterns derived by Ansari Esfeh et al. [24].

$$C_{UC} = C_A + C_W + C_I = c_4(T_a + T_w + T_v) \tag{12}$$

When requesting a flag stop, Type G and Type S passengers must walk/wheel an average vertical distance of W/4 from their origin/destination to the fixed route, and curb-to-curb pick-ups/drop-offs for Type S passengers involve no walking (see Table 2). Hence, the expected walking time cost can be estimated using (13).

$$C_A = c_2\left[\frac{W}{4V_a}\left(2\eta_{R1} + \eta_{R2} + \eta_{R3}\right) \times \frac{\Delta t}{\delta h} + \frac{W}{2V_a} \times Q_R\right] \tag{13}$$

where, V_{a} (km/hr) is the passenger walking speed and $\Delta t/\delta h$ is the accepted Type S demand passenger (pass/hr) which is always less than or equal to the received demand, Q_{S}.

Most studies assume that passengers arrive at bus stops at random; therefore, the mean waiting time equals half of the service headway. Low-demand routes usually have a higher headway (i.e., $h > 10$ minutes) and published timetable; thus, passengers may or may not exhibit random arrival, and instead may adjust their arrival time at the departure stop to minimize the waiting time; thus, the mean waiting time is less than half the headway [24]. The mean waiting time for passengers requesting a flag stop pick-up (i.e., R1 and R2) can be calculated using (14), proposed by Ansari Esfeh et al. [24] for low-demand routes. Passengers requesting curb-to-curb pickup (i.e., R3 and R4) does not incur any waiting time since the pick-up time is scheduled and passengers spend their time at origins home/work location instead of waiting at the transit stop. The expected value of waiting time can therefore be estimated using (15) derived from conditional probability theory.

$$E\left(W\right) = \left[\frac{1}{2} - \frac{\alpha\left(1-\beta\right)}{2}\right]h \tag{14}$$

$$C_W = c_2\left[\left[\frac{1}{2} - \frac{\alpha\left(1-\beta\right)}{2}\right]h\left(\eta_{R1} + \eta_{R2}\right) \times \frac{\Delta t}{\delta h} + \left[\frac{1}{2} - \frac{\alpha\left(1-\beta\right)}{2}\right]h \times Q_R\right] \tag{15}$$

where α and β are the proportion of planning passengers and the proportion of planning passengers with fixed arrival times, respectively.

Similarly, passengers can be dropped off/picked up uniformly anytime in the trip between two terminals. Thus, the average in-vehicle time for Type G and Type S passengers is half of the total travel time between the two terminals with the expected value given in (16).

$$C_I = c_2\left[\frac{T_v + \Delta t}{2}\left(\eta_{R1} + \eta_{R2} + \eta_{R3} + \eta_{R4}\right) \times \frac{\Delta t}{\delta h} + \frac{T_v + \Delta t}{2} \times Q_R\right] = c_2\left[\frac{T_v + \Delta t}{2}\left(\frac{\Delta t}{\delta h} + Q_R\right)\right] \tag{16}$$

3) SERVICE BENEFIT

SFT service benefits are specific to the operators' intent [8]. This analysis defines the service benefit in (17) as the cost incurred to serve paratransit passengers (Type S) using a dedicated DRT service if not served by SFT.

$$C_{SB} = c_3 \times \frac{\Delta t}{\delta h} \tag{17}$$

where $\mathrm{c_3}$ (\$/pass) is the average operating cost of providing paratransit service (assumed).

V. SOLUTION METHOD

Based on the general formulation of the optimization problem presented in Equations (1)-(3), we define the objective functions as functions of the decision variables by expanding them using the cost component equations in (8) - (17) before reducing them to a simpler form as given in equations (18) – (20) after initialization. Essentially, the constants θ_0 to θ_8 are estimated based on initialized values of cost coefficients (i.e., c_1, c_2, and c_3); and other parameters for the case study (i.e., L, W, t_{ad}, t_d, etc.). Equation (18) suggests that f_1 is inversely related to h and includes an interaction term ($\Delta t/h$) which indicates that the effect of one decision variable (h or Δt) on f_1 is based on the level or magnitude of another decision variable. f_2 in equation (19) is linearly related to h and Δt and includes interaction terms with linear and quadratic relationships and f_3 in equation (20) is only a function of interaction term $\Delta t/h$. Essentially, the optimization problem is non-linear since the

TABLE 2. Values of δ and user time components are classified by demand and request type.

Flexibility increases ↓	Demand type	Request type	Demand proportion	δ	Access time	Waiting time	In-vehicle time
	Type G	-	-	-	$\frac{W}{2V_a}$	$\left[\frac{1}{2}-\frac{\alpha(1-\beta)}{2}\right]h$	$\frac{T_v+\Delta t}{2}$
	Type S	R1	η_{R1}	$2t_{ad}+2t_d$	$\frac{W}{2V_a}$	$\left[\frac{1}{2}-\frac{\alpha(1-\beta)}{2}\right]h$	$\frac{T_v+\Delta t}{2}$
		R2	η_{R2}	$\frac{W}{2V_R}+2t_{ad}+2t_d$	$\frac{W}{4V_a}$	$\left[\frac{1}{2}-\frac{\alpha(1-\beta)}{2}\right]h$	$\frac{T_v+\Delta t}{2}$
		R3	η_{R3}	$\frac{W}{2V_R}+2t_{ad}+2t_d$	$\frac{W}{4V_a}$	0	$\frac{T_v+\Delta t}{2}$
		R4	η_{R4}	$\frac{W}{V_R}+2t_{ad}+2t_d$	0	0	$\frac{T_v+\Delta t}{2}$

effect of both decision variables on $f_1 f_2$, and f_3 takes both linear and non-linear effects.

$$f_1=\theta_0+\theta_1\left(\frac{1}{h}\right)+\theta_2\left(\frac{\Delta t}{h}\right) \tag{18}$$

$$f_2=\theta_3+\theta_4\,(h)+\theta_5\,(\Delta t)+\theta_6\left(\frac{\Delta t}{h}\right)+\theta_7\left(\frac{\Delta t^2}{h}\right) \tag{19}$$

$$f_3=\theta_8\left(\frac{\Delta t}{h}\right) \tag{20}$$

Multi-objective evolutionary algorithms (MOEAs) are generally considered mainstream methods for solving these problems. The optimization problem is handled using the MOEAs based on the classical method of Pareto optimality which uses the concept of domination to obtain a set of solutions that are not dominated by any member of the feasible solution set with respect to all objective values and are strictly better in at least one objective. Our study compares the quality of Pareto solutions obtained using two state-of-the-art MOEAs: Non-dominated Sorting Genetic Algorithm-II (NSGA-II) proposed by Deb et al. [28] and Speed-constrained Multi-objective PSO Algorithm (SMPSO) proposed by Nebro et al. [29]. Pseudocodes to implement both algorithms are provided in Fig. 10 of the Appendix.

To obtain the most representative deterministic Pareto optimal solution set, the procedure described below is followed.

Step 1: Obtaining Pareto optimal solution sets for each MOEA based on different parameter settings.

- NSGA-II and SMPSO implementation require the initialization of the following parameters: chromosome population size (N_P) or swarm size (N_S), number of generations (G), crossover probability (p_c), mutation probability (p_m), crossover distribution index (τ_c), and mutation distribution index (τ_m).
- For 'X' combinations of N_S or N_P and G we perform 'Y' runs of each combination obtaining XY Pareto sets for each algorithm.

Step 2: Performance evaluation of Pareto optimal sets.

- For performance evaluation, we use five indicators: hypervolume (HV), generation distance (GD), inverted generational distance (IGD), epsilon (ε), and computation time per run (CT).
- HV, GD, and IGD measure the diversity and/or convergence of solutions compared to a reference Pareto front RF or reference point RP which can be obtained by selecting non-dominated solutions from all Pareto solutions obtained in Step 1 (discussion in detail in Ishibuchi et al. [30]).
- Finally, the average indicator values across 'Y' runs are reported for 'X' combinations in each algorithm.

Step 3: Ranking of Pareto optimal sets.

- TOPSIS ranking method proposed by Tzeng and Huang [31] is implemented.
- In TOPSIS, we rank Pareto optimal sets by minimizing GD, IGD, ε, and CT, and maximizing HV while assigning equal weightage to indicators. We then select a solution set at random from all 'Y' runs corresponding to the top-ranked parameter setting of the best-performing algorithm (i.e., the lowest sum of the ranks).

VI. RESULTS AND DISCUSSIONS

A. STUDY AREA DESCRIPTION

As a case study, the optimization problem is applied to a low-demand bus route served by Regina Transit, Canada. Analysis conducted by CUTA [32] indicates that a total annual ridership of 6,434,022 is served by Regina Transit with revenue to cost ratio of 26.27%; thus, the non-passengers (i.e., the city and taxpayers) subsidize the remaining 73.73%. The analysis of ridership and operating data obtained from Regina Transit for 2015 shows that routes 6, 14, 15, and 16 exhibit underperformances based on average R/C [2]. Route 6 selected for the case study highlighted in Fig. 2 is attributed to low ridership (Q_R) of 9 passengers/hour, R/C of

15.3%, the high operating cost of \$10.84/pass, and declining ridership of 1.04% from 2014.

B. MULTI-OBJECTIVE OPTIMIZATION RESULTS

1) IDENTIFICATION OF THE BEST SET OF PARETO OPTIMAL SOLUTIONS

Initialization of parameters in analytical cost models and parameter settings for NSGA-II and SMPSO are outlined in Table 6 in Appendix B. Constraint 1 in the optimization problem limits h from 10 minutes to 1 hour, Constraint 2 limits Δt from 0 to 24 minutes, and the average time required to serve one Type S passenger/paratransit user (δ) is estimated as 1.6 minutes. NSGA-II and SMPSO algorithms are implemented in Python to obtain Pareto solution sets for 100 runs of 18 different parameter settings in each algorithm, subsequently used to estimate the reference front and reference point. According to TOPSIS analysis, NSGA-II yields a lower sum of ranks than SMPSO; therefore, it outperforms SMPSO. Table 3 shows the top-ranked parameter settings for both algorithms. SMPSO requires twice the number of function evaluations and hence, much higher computation time than NSGA-II to obtain Pareto fronts that do not differ greatly in terms of convergence and diversity. From 100 runs of the NSGA-II algorithm with the parameter settings described in Table 3, a Pareto set is selected at random for further analysis.

2) DESCRIPTIVE ANALYSIS OF PARETO SOLUTIONS

The hypervolume indicator value for NSGA-II-based Pareto solutions shown in Fig. 3(a) is 0.79, which indicates a good convergence of solutions since 79% of the volume represents the dominated space. For NSGA-II, the knee-point solution with the highest hypervolume contribution marked in Fig. 3(a) is usually a preferred trade-off solution. The knee-point solution has the following properties: $h = 41$ minutes; $\Delta t = 6.4$ minutes; operator cost, $f_1 = \$122$/hr; user cost, $f_2 = \$393$/hr; and service benefit, $f_3 = \$253$/hr. With 10 passengers per trip, 6 Type G and 4 Type S, and 60% capacity utilization, the knee-point solution is feasible from an operator cost and service benefit perspective, but it isn't from a user cost perspective. According to Fig. 3(b), operator cost (f_1) and user cost (f_2) are inversely related; for a given value of f_1, an increase in f_2 increases f_3; for a given value of f_2, an increase in f_1 increases f_3; and the lowest values of f_1 and f_2 provide almost no service benefit (f_3). Thus, the transit operator must carefully compromise between f_1 and f_2 to maximize f_3. Fig. 3(c) indicates that adopting solutions with h and Δt above 16.3 minutes and 2.5 minutes, respectively, would result in lower costs for serving the same number of Type S passengers via integrated SFT service compared to the existing DRT service. Thus, operators will gain monetary benefits from an integrated SFT service when high values of h and Δt are used. Additionally, h has a U-shaped distribution with a mean of 31 minutes and Δt has a right-skewed distribution with a mean of 2.6 minutes. Mean values of f_1, f_2, and f_3 are \$187.1/hr, \$281.1/hr, and \$129.1/hr respectively.

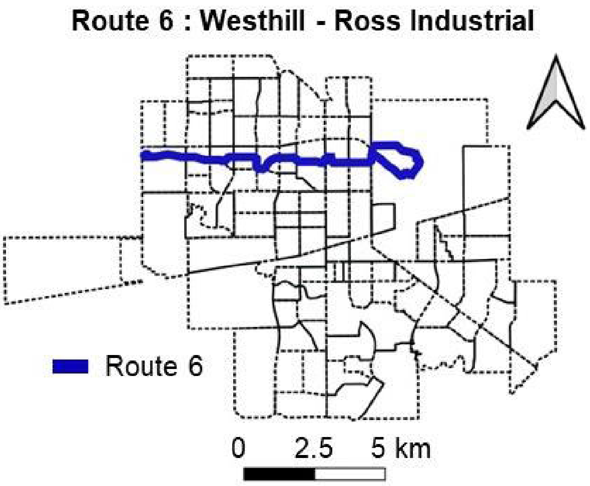


FIGURE 2. Route 6: Westhill to Ross industrial.

3) STATISTICAL ANALYSIS OF PARETO SOLUTIONS

a: RELATIONSHIP BETWEEN H AND ΔT

The relationship between h and Δt cannot be modeled using linear regression approaches, as indicated by the violation of homogeneity in Fig. 4(a); thus, quantile regression is applied in this case. For quantile levels ranging from 0.1 to 0.9, Fig. 4(a) shows the fitted quantile regression line, including the median regression line (i.e., the 50th quantile regression). For a significance level of 95%, the coefficients for all quantile levels are significant. The quantile regression models the conditional quantiles of Δt given the input variable, h. For example, when the quantile level of regression is 0.9, we obtain an intercept value of -0.23 and a coefficient of 0.15; therefore, for $h = 30$ minutes, the 90th percentile of Δt is expected to be 4.37 minutes. Table 4 illustrates that as the quantile level increases, the slope coefficient increases, but the rate of increase in slope decreases. The lower and upper quartiles differ significantly from the least squares estimate. Thus, with an increase in h, we are more likely to obtain higher Δt values, but we are less likely to obtain Pareto solutions in general. The intercepts for most quantile levels are smaller than those for least squares; thus, quantile regression is more predictive than linear regression. Based on the quantile regression model parameters shown in Table 4, the value of Δt is estimated for values of h from 10 to 60 minutes and illustrated in Fig. 4(b). Considering that it takes approximately 1.6 minutes to serve one Type S passenger (δ), Fig. 4(b) suggests that the probability P (Number of Type S passengers served per trip $\geq$ 1) increases with h, while P (Number of Type S passengers served per trip = 0) decreases with h. Also, Δt increases by 0.2 minutes for each 0.1 increment in quantile level for h between 10 and 20 minutes, and by 0.4, 0.6, 0.7, and 0.9 minutes for h between 20 to 30, 30 to 40, 40 to 50, and 50 to 60 minutes. If an agency wishes to operate its fleet at $h = 25$ minutes, the optimal slack time, Δt, is between 0.4 and 3.6 minutes which means the optimal solutions are 10% likely to be ≤ 0.4 minutes and 90% likely to

TABLE 3. TOPSIS results.

Algorithm		NSGA-II	SMPSO
Sum of ranks		306	360
Rank		1	6
N_P or N_S		500	1000
G		50	50
Runs		50	50
Values averaged across runs	CT (sec)	49.82	271.44
	GD	0.0007	0.0019
	IGD	0.0198	0.0185
	ϵ	0.1113	0.1012
	HV	0.782	0.801

be $\leq$ 3.6 minutes. The average value of Δt as a percentage of one-way operating time without route deviation (T_v) is 8.9%. Furthermore, 50% of the ratio Δt /$\mathrm{T_v}$ falls between 2.86% and 13.48%, with a median of 6.49% and a maximum and minimum of 28.5% and 0%, respectively. In contrast to our study, most studies recommend a fixed slack time for a given set of conditions. Fu [8] suggested that Δt of 6 minutes is optimal to accommodate two deviated stops requested per analysis period of $T_v + \Delta\mathrm{t}$ with a maximum allowable deviation ratio of $\Delta\mathrm{t}$ / $T_v + \Delta t$ of 40% and vehicle capacity of C = 9 seats. To accommodate route deviation requests, Potomac and Rappahannock Transportation Commission (PRTC) has included approximately 20% slack time in their basic schedules for medium-duty, 28-passenger buses [4]. Careful consideration should be given to the amount of slack time to be built into each route (in each direction, i.e., inbound/outbound). Δt should be sufficient to process the desired number of Type S requests without causing excessive idle time downstream if no Type S requests are received.

b: RELATIONSHIP BETWEEN H AND ΔT AND F$_1$, F$_2$, AND F$_3$

Fig. 5(a) and 5(b) confirms the relationship in equation (18) and suggest that f_1 and M, the operator cost and fleet size, are negatively correlated with h and positively correlated with Δt. By increasing h, the fleet size requirements will be reduced, thereby reducing f_1. An increase in Δt will result in more Type S users being served per trip, increasing round-trip time; hence, larger fleet size is needed to maintain the same service frequency, which ultimately increases f_1. As shown in Fig. 5(a), the range of Δt values observed is highest for h between 50-60 minutes with the standard deviation in f_1 values observed within this range being 6.3. When Δt is between 0-2 minutes, h values range from 10-60 minutes with the standard deviation in f_1 values observed within this Δt range being 81.8. This standard deviation in f_1 values decreases with an increase in h and Δt ranges. Thus, f_1 is primarily influenced by h with the influence of decision variables reducing for high h and Δt values. Alshalalfah [17] suggested that fleet size would increase by a fraction equal to the ratio of slack time to original headway, similar to (10). Fig. 5(c) suggests that user cost, f_2 is positively correlated to h and Δt. The standard deviation in f_2 values for h between 50-60 minutes is 59.4 while the standard deviation is 40.9 for Δt between 0-2 minutes. The standard deviation increases with h and reduces with Δt with low values occurring for h and Δt between 40-50 and 2-4 minutes, respectively. Thus, f_2 is influenced by both variables with Δt having slightly greater influence than h and the variation in f_2 values is lowest for the medium range of values of decision variables followed by high range values and low range values. $\mathrm{C_A}$ in Fig. 5(d) is primarily a function of Δt since the number of Type S passengers increases with Δt but decreases with h since the number of Type G passengers increases with h (as per Constraint 3). As shown in Fig. 5(e), $\mathrm{C_W}$ is a function of both h and Δt as the wait time per passenger varies directly with h, and the accepted Type S passenger demand, which increases with Δt but decreases with h, with h having a greater impact on the former than the latter. $\mathrm{C_I}$ in Fig. 5(f) increases with Δt as it increases the one-way travel time (i.e., $\mathrm{C_I}$ exhibits positive quadratic growth as it is proportional to the square of Δt). h increases Type G passengers in the system, thus increasing T_V, but also limits Type S passengers, with a greater impact on the latter. From Fig. 5(g) it is evident that f_3 is directly proportional to Δt and inversely proportional to h. Additionally, the standard deviation in f_3 values does not vary significantly with h ranges and is averaged at 68.4, while the standard deviation values decrease from 67.1 to 11.3 with the increase in Δt value ranges. Thus, Δt primarily impacts f_3 irrespective of h, and as Δt increases variation in f_3 reduces. Fig. 5(h) graphically confirms that total cost, TC defined as the sum of f_1 and f_2 minus f_3, is a convex function with the minimum value of TC = \$242.2/hr. The Pareto solution corresponding lowest TC is $h = 21$ minutes and $\Delta t = 3$ minutes as shown in Fig. 5(i). In comparison to the "knee-point solution", this solution with $f_1 = \$197.8$/hr, $f_2 = \$285.3$/hr, and $f_3 = \$240.9$/hr would drive f_2 down by 27.4% and f_3 down by 4.8%, and f_1 up by 62%. Additionally, Fig. 5(i) suggests that extreme values of h and Δt yield higher TC values and that opting for mid-range values would minimize TC. For pruning the Pareto optimal set, the analysis described here can be useful.

C. SENSITIVITY ANALYSIS

1) VEHICLE CAPACITY

Vehicle capacity (C) is an important design parameter for flexible transit [11], [14]. In this study, constraints limit the range of decision variables and ensure that no proposed solution exceeds C. Accordingly, vehicle occupancy, passenger composition, and system costs also vary with C. We consider three common vehicle types in SFT with varying capacities: Mini-van (7-passenger vehicle excluding the driver), standard van (15-passenger vehicle excluding the driver), and mini-bus (25-passenger vehicle excluding the driver). As C increases, vehicle occupancy decreases due to low passenger demand along the route, as shown

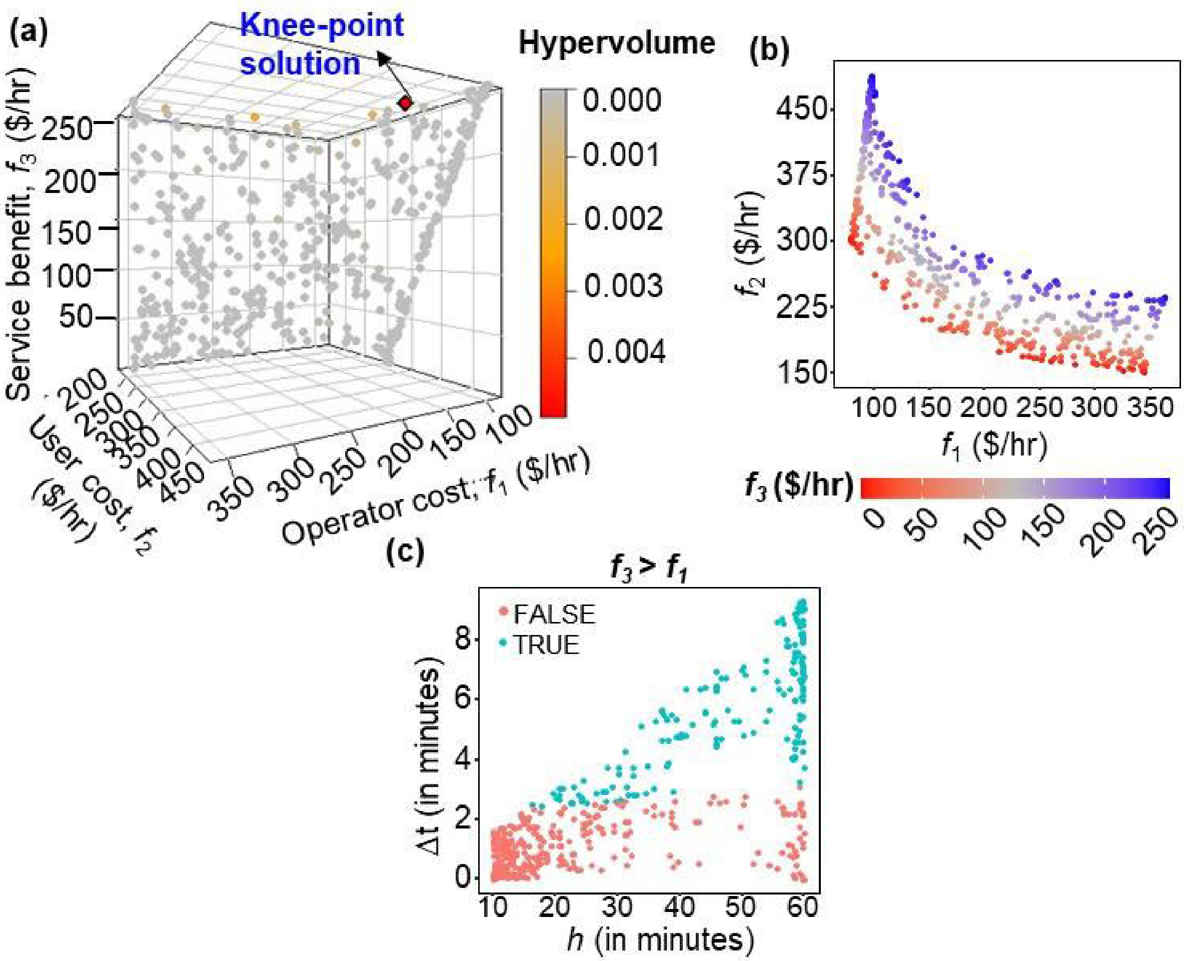


FIGURE 3. (a) 3D-Pareto front and knee-point solution; (b) variation of f_2 with f_1 and f_3; and (c) Operator cost –Benefit.

in Fig. 6(a), 6(b), and 6(c). Average occupancy drops from 52.5% to 42.4% when C increases from 7 to 15 seats/vehicle, and even further to 25.1% when C increases from 15 to 25 seats/vehicle. From Fig. 6(d) and 6(e), as C increases from 7 to 15 seats/vehicle, the average h across the Pareto set ($\bar{h}$) increases from 18 to 31 minutes, while $\overline{\Delta t}$ increases from 1.6 to 2.6 minutes. When C increases from 15 to 25 seats, little or no difference is observed in $\bar{h}$ and $\overline{\Delta t}$. As expected, the average number of Type G and Type S passengers served per trip increases from 3 to 5 and 1 to 2, respectively, when C increases from 7 to 15 seats/vehicle. A similar finding was reported by Kim et al. [11] where optimal zone size and optimal headway for operating flexible buses increased rapidly when C increased from 5 to 15 seats; however, when vehicles had sufficient capacity (i.e., greater than 15 seats/bus), the increase was less rapid. Owing to an increase in $\bar{h}$ and $\overline{\Delta t}$ when C increases from 7 to 15 seats/vehicle, f_1 decrease significantly, while user costs f_2 increase. When $C > 15$ seats/vehicle, the difference becomes less evident as little, or no difference is observed in $\bar{h}$ and $\overline{\Delta t}$ while f_3 does not improve much across vehicle sizes. Precisely, with an increase of C from 5 to 15 and 15 to 25 seats/vehicle, $\bar{f_1}$ decreases by 21.5% and 2%, respectively, while $\bar{f_2}$ increases by 22.7% and 1%. Accordingly, a larger fleet with smaller capacity vehicles results in shorter passenger travel times, but from the operator's perspective, a few larger capacity vehicles are more cost-effective. We can conclude that minivans are more appropriate for SFT services in terms of vehicle occupancy and user costs, and that standard vans are the most cost-effective from the standpoint of operator costs, while minibuses offer no benefit from an operator, user, or service standpoint. If more paratransit users are accommodated, the standard van and minibus occupancy may increase, but the service may become less competitive as the overall travel time increases. Estrada et al. [14] also reported that flexible services with cars ($C = 4$ pass/veh) are most economical in terms of operator cost than minibuses ($C = 22$ pass/veh), and standard buses ($C = 70$ pass/veh).

2) HOURLY DEMAND (Q_G AND Q_S)

Fig. 7 illustrates the sensitivity of Pareto solutions to average hourly Type G (Q_G) and Type S (Q_S) passenger demand, with each varying from 5 to 30 passengers/hour. According to Fig. 7(a), when the demand for the service increases,

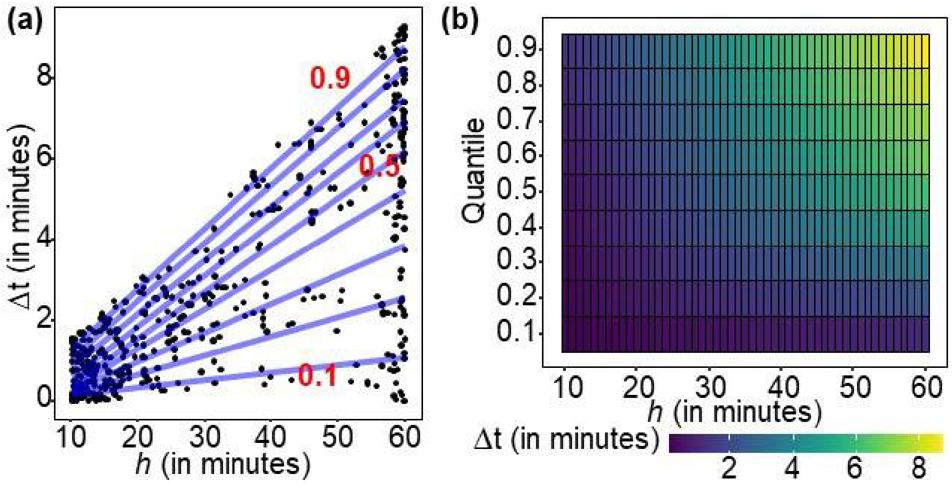


FIGURE 4. (a) Quantile regression of h and Δt for quantile levels and (b) Variation of Δt with h and quantile level.

TABLE 4. Quantile regression coefficient estimates.

Quantile level	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Least square method
Intercept	-0.07	-0.29	-0.49	-0.66*	-0.76*	-0.69*	-0.50*	-0.39*	-0.23*	-0.39
Slope	0.02*	0.05*	0.07*	0.10*	0.12*	0.13*	0.13*	0.14*	0.15*	0.10

Note: Significance levels: $^*p < .05$

a high-frequency service is required to achieve system optimality. With Q_S at 5 pass/hr, the average decrease in $\bar{h}$ with Q_G is 10.89%, reaching 12.29% at $Q_S = 10$ pass/hr, and continuously declining to 6.61% at $Q_S = 30$ pass/hr. Fig. 7(b) demonstrates an expected phenomenon, $\overline{\Delta t}$ increases with Q_S since more Type S passengers are available, and decreases with Q_G as the capacity available to serve Type S passengers reduces with Q_G. The decrease in $\overline{\Delta t}$ with Q_G reducing as Q_S increases, varying from 12.39% to 7.34%. Fig. 7(c) shows that optimization problem constraints always ensure that the number of Type S passengers served per SFT trip is always less or equal to the total observed Type S demand. An increase in Q_G suggests high service frequency (i.e., $\bar{h}$ reduces) and reduced available capacity for Type S passengers (i.e., $\overline{\Delta t}$ reduces) while the increase in Q_S suggests higher slack time to accommodate Type S passengers (i.e., $\overline{\Delta t}$ increases) and increased service frequency as demand increases (i.e., $\bar{h}$ reduces). Thus, as demand increases, average operating cost ($\bar{f}_1$) and average user cost ($\bar{f}_2$) increases but the rate of increase with Q_G reduces as Q_S increases, ranging from 6.3% to 3.5% for $\bar{f}_1$ and 30.5% to15.5% for $\bar{f}_2$. $\bar{f}_3$ increases with Q_S while decreasing with Q_G, with the rate being significantly higher in the former case. Simply put, operator cost and user cost are directly proportional to Q_G and Q_S whereas service benefit is directly proportional to Q_S but inversely proportional to Q_G. Hence, we can say that a reasonable trade-off in cost and benefit is possible when demand is low to medium (5-20 passes/hr), whereas high demand dramatically increases costs. In their initial feasibility analysis along this route, Mishra et al. [3] recommended regular bus transit (FBT) over SFT when Type G demand exceeds 27 passengers/hour. A transit planner can utilize this analysis to develop an integrated service schedule based on the observed temporal distribution of passenger demand along the route, which is typically bimodal with two distinct peak periods for Type G, and peaks during noon off-peak periods for Type S. For example, it may be recommended to adopt relatively lower Δt and h values during peak hours than in off-peak hours.

3) PERMITTED DEVIATION (D_P)

When transit operators agree to serve curb-to-curb requests outside of the designated service area shown in Fig. 1, sensitivity analysis with permitted deviations (D_P) will help us understand its impact on Pareto solutions. D_P is 0.5 km in the base case, which is equivalent to half the width of the service area, 1 km (see Appendix Table 6). Typically, in real-world situations, D_P range between 0.5km and 2.5km from the fixed route [4]. Fig. 8 depicts the results of the sensitivity analysis

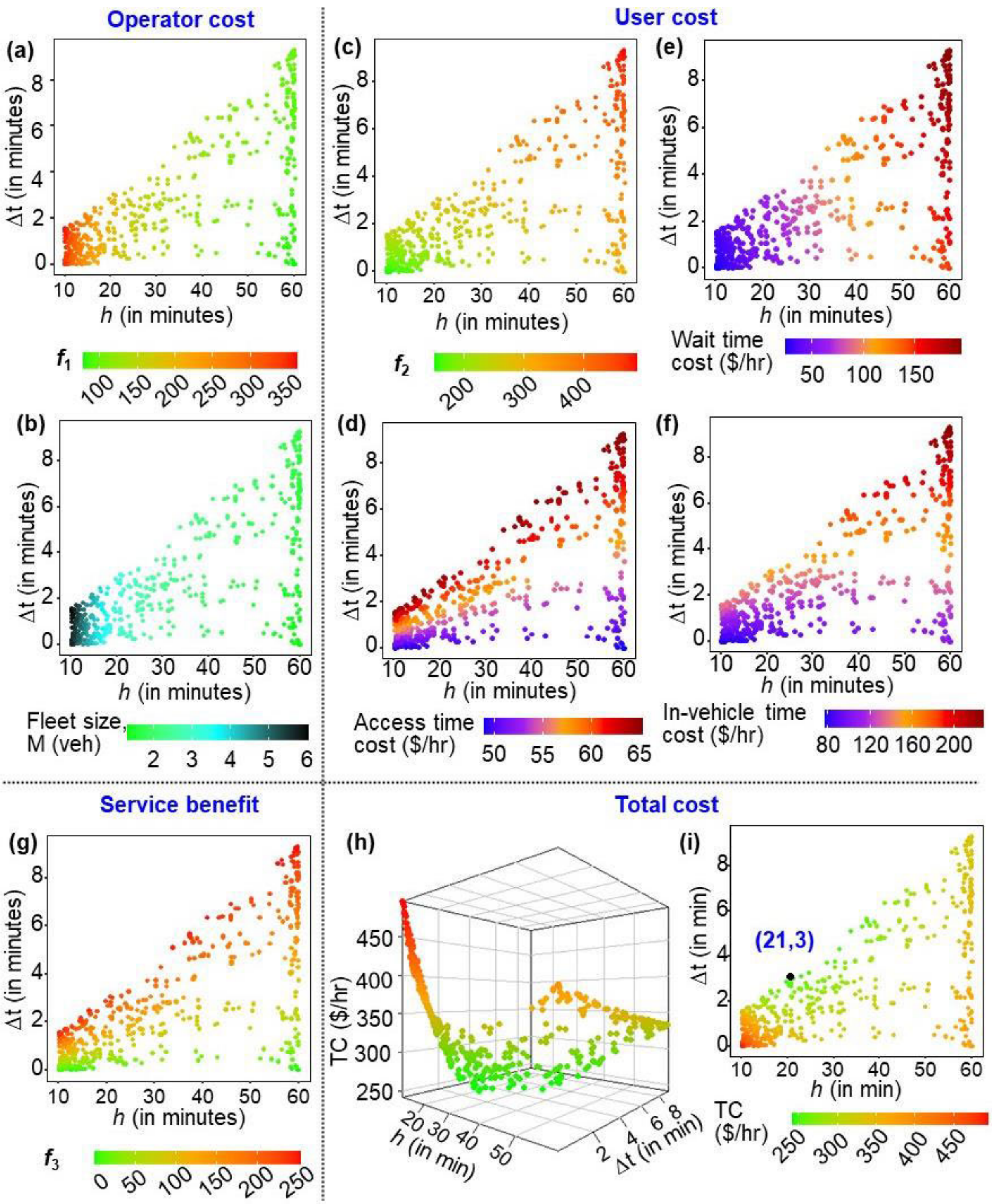


FIGURE 5. For Pareto optimal solutions F*, heatmap illustrating the variation of (a) f_1, (b) fleet size, (c) f_2, (d) access time cost, (e) wait time cost, (f) in-vehicle time cost, (g) f_3, and total cost (h) 3D-plot and (i) 2D-plot.

using (21) in this situation to estimate δ.

$$\delta = \frac{D_P}{V_R}\left(\frac{\eta_{R2}+\eta_{R3}}{2}+\eta_{R4}\right)+2t_{\text{ad}}+2t_d \qquad (21)$$

where, D_P represents the permitted deviation from a fixed route, in km.

Based on Fig. 8, D_P has a minimal effect on h, but increases δ, thereby affecting Δt in the Pareto set significantly. When D_P increases by 0.25km, $\bar{h}$ and $\overline{\Delta t}$ increase by 0.15% and 10%, respectively; therefore, $\bar{f_1}, \bar{f_2}$, and $\bar{f_3}$ increase by 0.95%, 0.83%, and 0.5%, respectively; and the rate of increase decreases with increasing D_P. Percentage change in $\bar{f_3}$ with

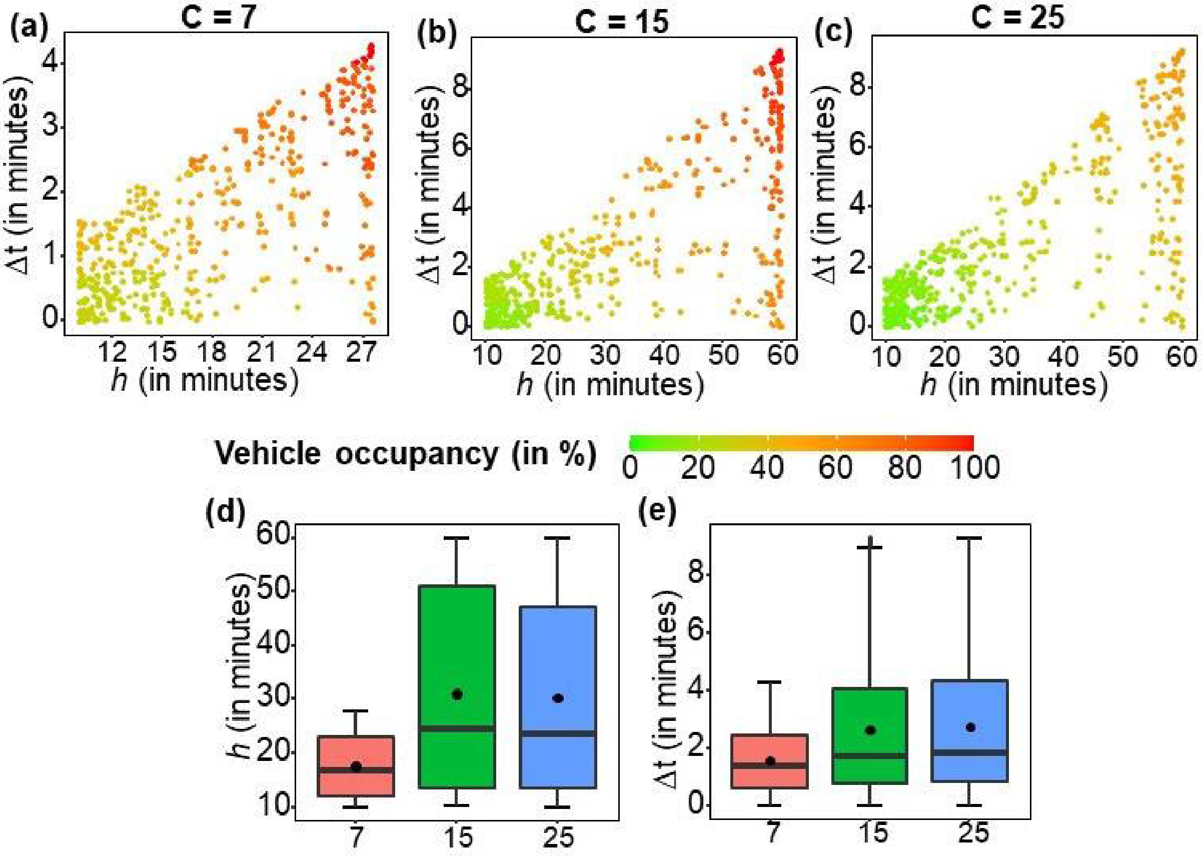


FIGURE 6. Sensitivity analysis results for vehicle capacity.

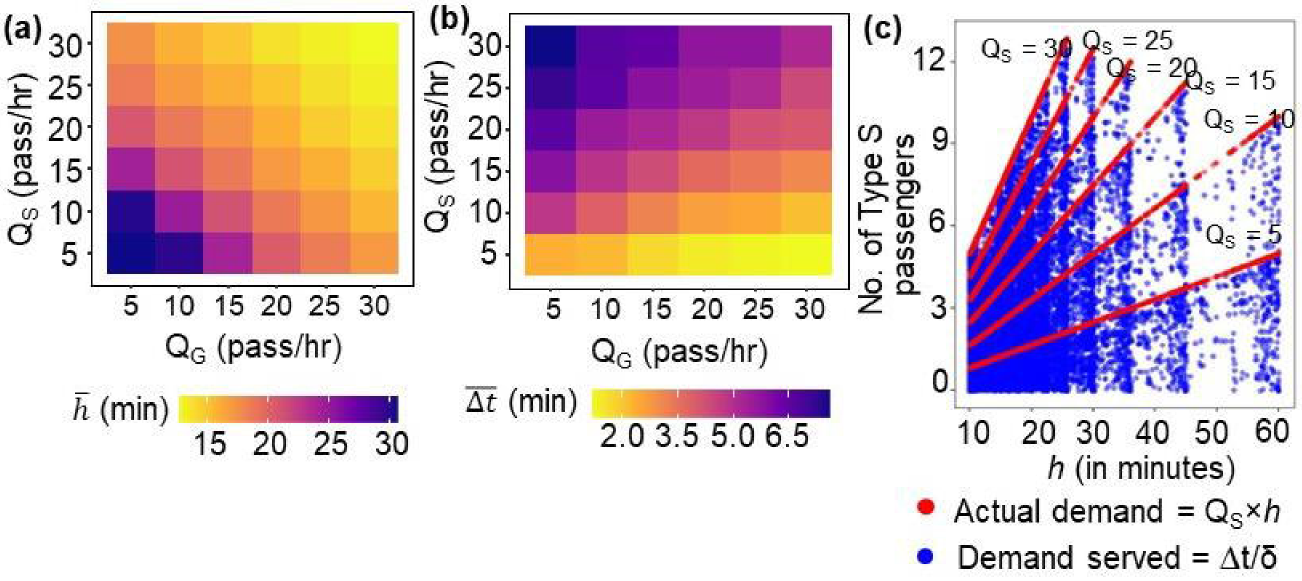


FIGURE 7. Sensitivity analysis results for hourly Type S and Type G demand.

D_P is the least, since the increase in Δt does not necessarily help serve more type S passengers as the time required to serve a paratransit request δ also increases with D_P. Alshalalfah [17] also demonstrated that as the width of the service area increases, the percentage of feasible deviations decreases systematically. The benefits derived from adopting

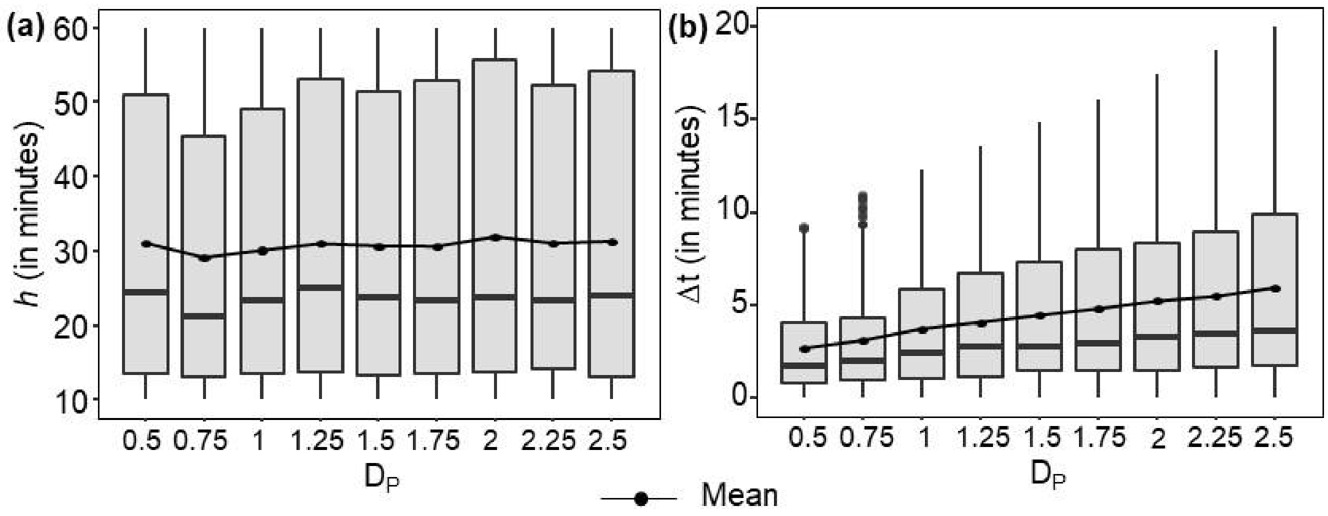


FIGURE 8. Sensitivity analysis results for permitted deviation.

TABLE 5. System characteristics based on weather conditions.

Weather condition	Q_G	Q_S	C	V_R	V_a	η_{R1}	η_{R2}	η_{R3}	η_{R4}	δ
	pass/hr	pass/hr	seats	km/hr	km/hr	-	-	-	-	hr/pass
Case 1: Adverse	9	6	15	15	0.5	0	0	0	1	0.053
Case 2: Normal	9	6	15	35	4	0.25	0.25	0.25	0.25	0.026

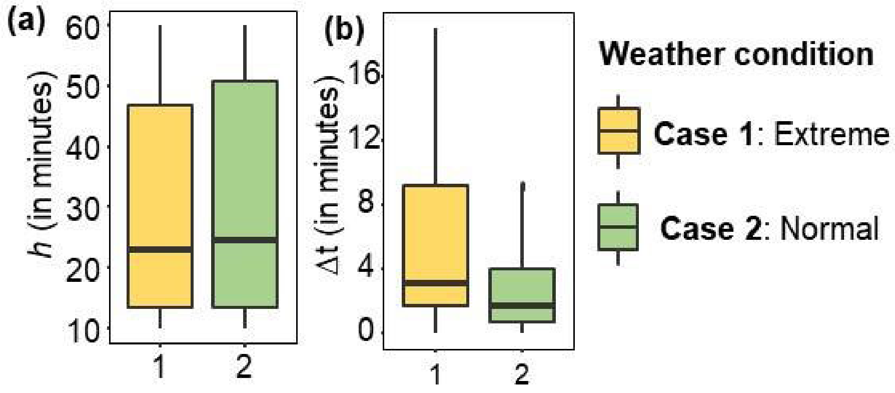


FIGURE 9. Sensitivity analysis results for weather conditions.

high Δt values suggested for systems with higher D_P are less than the increased travel time and the delay imposed on regular passengers by deviation services. By providing higher D_P at off-peak hours, when fewer passengers are on board, the user cost can be reduced while operator costs can be reduced by using larger headways.

4) WEATHER CONDITIONS

In extreme weather conditions, such as snow, ice, or sleet, commonly observed in parts of Canada during Winter, using a wheelchair or walking to the bus stop often causes significant discomfort or poses a safety risk to transit users. We replicated this scenario by reducing walking/wheeling speed, vehicle riding speed, and increasing the proportion of Type S passengers requesting curb-to-curb service for both pickup and drop-off, as shown in Table 5. For a route-deviated pickup and drop-off in extreme weather conditions (Case 1), it takes 3.16 minutes (δ), almost twice as long as for normal weather conditions (Case 2). According to Fig. 9(a), the probability of observing lower values of h is relatively higher in Case 1 than in Case 2, suggesting more frequent service is required to achieve system optimality during extreme weather conditions; the $\bar{h}$ for Case 1 is 26.9 minutes, which is 5% less than Case 2. As δ increases in Case 2, the upper limit of Δt in Constraint 2 also increases; therefore, $\overline{\Delta t}$ for Case 1 is 1.5 times higher than Case 2, as shown in Fig. 9(b). Case 1 and Case 2 serve the same number of Type S passengers per trip, but Case 1 serves two passengers on average, while Case 2 serves one passenger. $\bar{f}_1$ and $\bar{f}_2$ are higher for Case 1 compared to Case 2 by 2.2 and 2.7 times, respectively, whereas the increase in $\bar{f}_3$ for Case 1 compared to Case 2 is negligible by 6.1%. This is expected since in Case 1 walking time increases dramatically, and Δt values in Case 1 are higher, meaning that one-way travel times are longer, resulting in a need to increase fleet size requirements. Thus, normal weather conditions are more conducive to the operation of an integrated SFT than extreme weather conditions, and in extreme weather conditions, slack time should be approximately twice as long as normal weather conditions, and headway should be approximately the same as normal weather conditions or slightly higher. Under adverse weather conditions, Nourbakhsh and Ouyang [20] compared two systems, FBT and SFT, by lowering walking speed to 0.1 km/hr and found that FBT bears a significant increase in total costs (i.e., operator and user costs) as compared to SFT, which can handle a broader range of demand densities.

VII. POLICY RECOMMENDATIONS

A. FARE POLICY

Several fare policies can be adopted during the pilot phase. First, existing bus transit users (Type G) and paratransit users (Type S) may pay the same fare, and deviations outside the service area ($D_P > W$) are not subject to a surcharge. Secondly, transit operators may impose a deviation surcharge on Type S users requesting deviations outside the designated service area or permitted deviation set by them. Transit agencies may offer Type G passengers a discounted fare for their degraded service quality since their travel time increases with deviations [33]. To encourage Type S passengers to switch from an overburdened paratransit service to the SFT, a reduced fare may be offered. When optimizing the transit system under different fare policies, the fare surcharge/discount can be incorporated into the service benefit objective.

B. TECHNOLOGY

Developments in automotive technology and technology for deviated service booking are fundamental decisions regarding technology in integrated SFT. When requesting curb-to-curb service, Type S passengers may reserve a ride via phone or mobile app about an hour in advance, providing trip details only for the deviated portion of the trip. Type S passengers may then be provided with real-time information, such as available pick-up or drop-off times based on the request type R2, R3, or R4, and slack time available Unit operator costs accounting for fleet purchases, fuel purchases, and driver wages (nearly 40-80%) can be reduced technology advancement. Through electrification, a reduction in energy costs, and automation, which eliminates the need for drivers, operating costs can be reduced.

C. OPERATIONS

Paratransit demand combined with existing regular transit demand in an integrated SFT service will result in a passenger loading profile along the route different from the loading profile observed for the existing bus transit service. Using headway, slack-time, and available demand data, transit planners can simulate passenger loading profiles for each trip along a route, and then use the observed occupancy information to optimize the vehicle size mix. To ensure maximum resource utilization for an integrated SFT operation, decision-makers can determine whether purchasing new vehicles, utilizing in-house paratransit vehicles or taxis, or contracting with private operators that already operate in the zone is the most economical option.

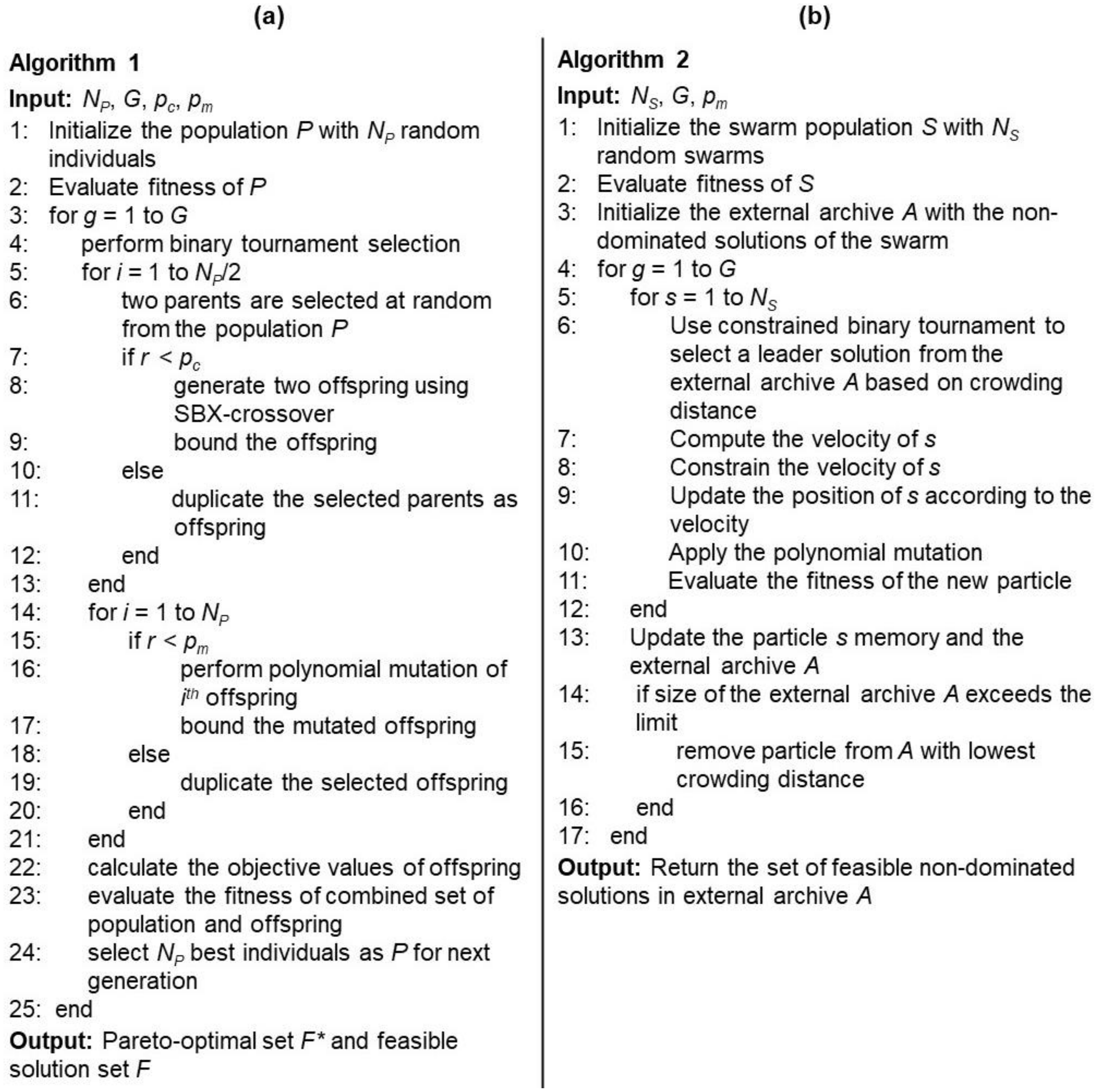
VIII. CONCLUSION

We performed joint optimization of service headway (h) and slack time per trip (Δt) utilizing operator cost (f_1), user cost (f_2), and service benefit (f_3) to design a semi-flexible transit (SFT) system along an existing low demand bus route that serves both general and special-need passengers. A detailed case study is conducted to demonstrate the methodology application for a low-demand bus route, Route 6 in Regina, Canada. These are the main contributions.

1) A relationship between optimal h and Δt is modeled using quantile regression, where conditional quantiles of Δt can be suggested for a given value or a range of h. This analysis helps transit planners evaluate different levels of flexibility that can be introduced into the timetable through slack time for a given service frequency to generate a static schedule for SFT operation that maximize cost efficiency.

2) We established a relationship between decision variables and objective functions to analyze the trade-offs in costs between alternative Pareto solutions essential for decision-makers in pruning the Pareto optimal sets. f_1 is negatively correlated and primarily impacted by h, f_2 is positively correlated to both variables, but Δt has a greater impact than h, and Δt is positively correlated and primarily influences f_3. f_1 and f_3 and f_2 exhibit relatively low sensitivity to decision variables for high and medium value ranges, respectively.

3) Sensitivity analysis reveals that low-capacity vehicles are more cost competitive with 7-seater minivans offering higher vehicle occupancy and lower user costs and 15-seater standard vans offering lower operator costs. A reasonable trade-off can be achieved between cost and benefit under low to medium demand (5-20 passes/hr), but high demand significantly increases costs. To minimize the possibility of causing passenger delays on board and to reduce user costs, a low permissible deviation from the fixed route is desirable during peak hours. When extreme weather conditions prevail, vehicle and passenger walking speeds are reduced and door-to-door services are demanded more frequently, resulting in



(a)

Algorithm 1

```
Input: $N_P$, $G$, $p_c$, $p_m$
Initialize the population $P$ with $N_P$ random individuals
Evaluate fitness of $P$
for $g$ = 1 to $G$
    perform binary tournament selection
    for $i$ = 1 to $N_P/2$
        two parents are selected at random from the population $P$
        if $r < p_c$
            generate two offspring using SBX-crossover
            bound the offspring
        else
            duplicate the selected parents as offspring
        end
    end
    for $i$ = 1 to $N_P$
        if $r < p_m$
            perform polynomial mutation of $i^{th}$ offspring
            bound the mutated offspring
        else
            duplicate the selected offspring
        end
    end
    calculate the objective values of offspring
    evaluate the fitness of combined set of population and offspring
    select $N_P$ best individuals as $P$ for next generation
end
Output: Pareto-optimal set $F^*$ and feasible solution set $F$
```

(b)

Algorithm 2

```
Input: $N_S$, $G$, $p_m$
Initialize the swarm population $S$ with $N_S$ random swarms
Evaluate fitness of $S$
Initialize the external archive $A$ with the non-dominated solutions of the swarm
for $g$ = 1 to $G$
    for $s$ = 1 to $N_S$
        Use constrained binary tournament to select a leader solution from the external archive $A$ based on crowding distance
        Compute the velocity of $s$
        Constrain the velocity of $s$
        Update the position of $s$ according to the velocity
        Apply the polynomial mutation
        Evaluate the fitness of the new particle
    end
    Update the particle $s$ memory and the external archive $A$
    if size of the external archive $A$ exceeds the limit
        remove particle from $A$ with lowest crowding distance
    end
end
Output: Return the set of feasible non-dominated solutions in external archive $A$
```

FIGURE 10. Pseudocode for NSGA-II and SMPSO.

higher operator and user costs. The purpose of sensitivity analysis for transit operators is to gain a greater understanding of the cost-effectiveness of the system under varying environmental conditions as well as determine if changes should be made to the schedule design based on variation in optimal slack time and headway.

Policy recommendations for integrated SFT implementation include a recommendation for fare structure design addressing service equity through surcharges/discounts, vehicle technology and service booking technology advancements for cost reduction, and fleet mix design through estimation of passenger loading profile.

Certain aspects limiting the implementation of this study will be investigated in future extensions. The simplified environment in terms of the service area and demand for defining analytical cost models could be enhanced to reflect a more realistic environment, including accounting for stochasticity in vehicle arrival and demand.

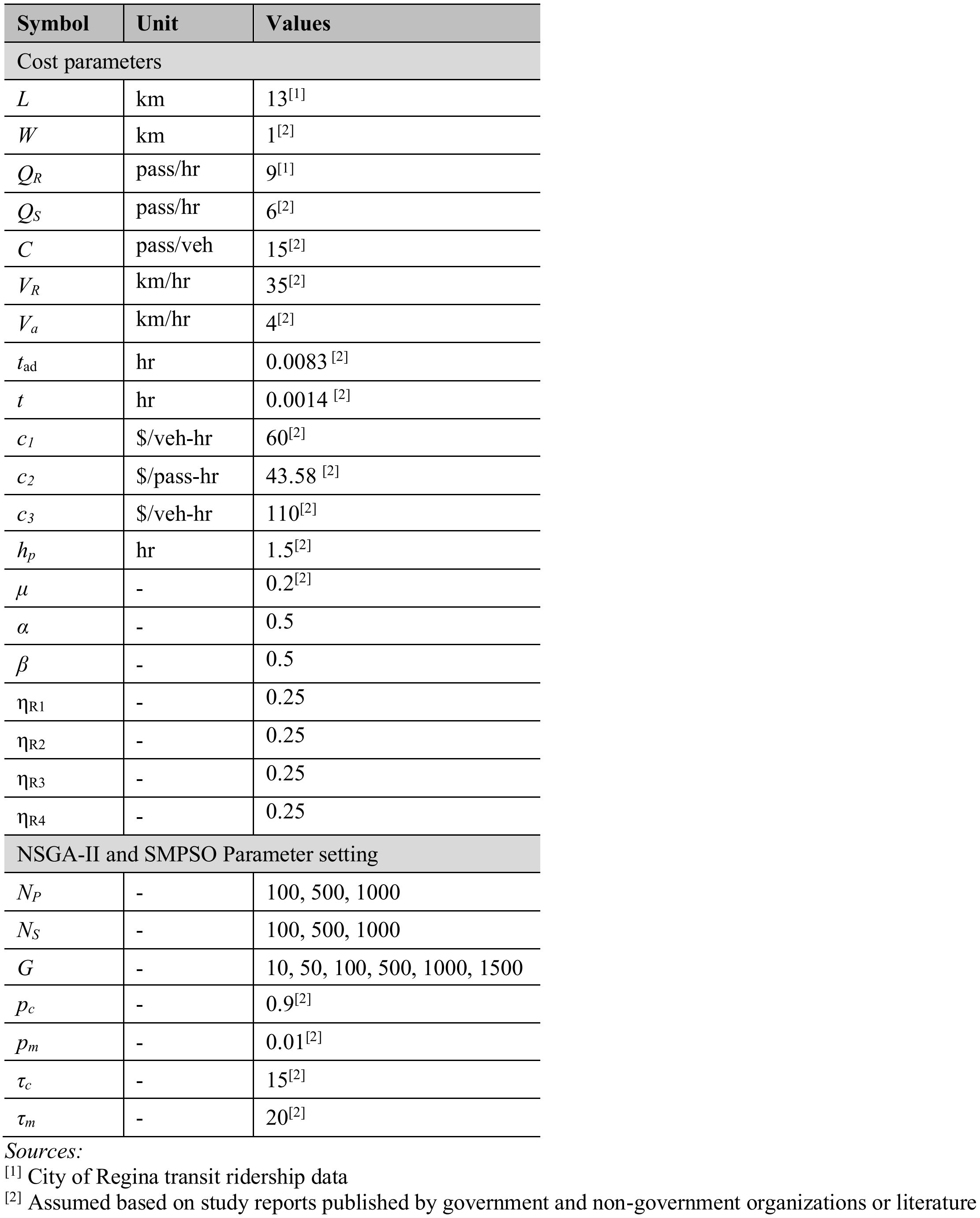
APPENDIX A

See Figure 10.

APPENDIX B

See Table 6.

TABLE 6. Case study parameter setting.



Symbol	Unit	Values
Cost parameters		
L	km	13[1]
W	km	1[2]
Q_R	pass/hr	9[1]
Q_S	pass/hr	6[2]
C	pass/veh	15[2]
V_R	km/hr	35[2]
V_a	km/hr	4[2]
t_{ad}	hr	0.0083 [2]
t	hr	0.0014 [2]
c_1	\$/veh-hr	60[2]
c_2	\$/pass-hr	43.58 [2]
c_3	\$/veh-hr	110[2]
h_p	hr	1.5[2]
μ	-	0.2[2]
α	-	0.5
β	-	0.5
η_{R1}	-	0.25
η_{R2}	-	0.25
η_{R3}	-	0.25
η_{R4}	-	0.25
NSGA-II and SMPSO Parameter setting		
N_P	-	100, 500, 1000
N_S	-	100, 500, 1000
G	-	10, 50, 100, 500, 1000, 1500
p_c	-	0.9[2]
p_m	-	0.01[2]
τ_c	-	15[2]
τ_m	-	20[2]

Sources:

[1] City of Regina transit ridership data

[2] Assumed based on study reports published by government and non-government organizations or literature

REFERENCES

[1] F. Errico, T. G. Crainic, F. Malucelli, and M. Nonato, "A survey on planning semi-flexible transit systems: Methodological issues and a unifying framework," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 324–338, Nov. 2013, doi: 10.1016/J.TRC.2013.08.010.

[2] B. Mehran, Y. Yang, and S. Mishra, "Analytical models for comparing operational costs of regular bus and semi-flexible transit services," *Public Transp.*, vol. 12, pp. 147–169, Jan. 2020, doi: 10.1007/s12469-019-00222-z.

[3] S. Mishra, B. Mehran, and P. K. Sahu, "Assessment of delivery models for semi-flexible transit operation in low-demand conditions," *Transp. Policy*, vol. 99, pp. 275–287, Dec. 2020, doi: 10.1016/j.tranpol.2020.09.004.

[4] D. Koffman, *Operational Experiences With Flexible Transit Services*. Washington, DC, USA: TCRP, 2004. [Online]. Available: http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_syn_53.pdf

[5] W. Klumpenhouwer, A. Shalaby, and L. Weissling, "The state of demand-responsive transit in Canada," Univ. Toronto, Toronto, ON, Canada, 2020, doi: 10.13140/RG.2.2.12310.47684.

[6] R. Weiner, *Integration of Paratransit and Fixed-Route Transit Services—TCRP Synthesis 76*. Washington, DC, USA: Transportation Research Board, 2008, doi: 10.17226/13993.

[7] *Paratransit Newsletter*, City of Regina, Regina, SK, Canada, 2021. [Online]. Available: https://www.regina.ca/export/sites/Regina.ca/transportation-roads-parking/transit/.galleries/Transit-Route-PDF/Paratransit-Newsletter.pdf

[8] L. Fu, "Planning and design of flex-route transit services," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1791, no. 1, pp. 59–66, 2002, doi: 10.3141/1791-09.

[9] M. Pei, P. Lin, J. Du, and X. Li, "Operational design for a real-time flexible transit system considering passenger demand and willingness to pay," *IEEE Access*, vol. 7, pp. 180305–180315, 2019, doi: 10.1109/ACCESS.2019.2949246.

[10] B. Smith, M. Demetsky, and P. Durvasula, "A multiobjective optimization model for flexroute transit service design," *J. Public Transp.*, vol. 6, no. 1, pp. 81–100, Mar. 2003, doi: 10.5038/2375-0901.6.1.5.

[11] M. E. Kim, J. Levy, and P. Schonfeld, "Optimal zone sizes and headways for flexible-route bus services," *Transp. Res. B, Methodol.*, vol. 130, pp. 67–81, Dec. 2019, doi: 10.1016/J.TRB.2019.10.006.

[12] L. Wang, S. C. Wirasinghe, L. Kattan, and S. Saidi, "Optimization of demand-responsive transit systems using zonal strategy," *Int. J. Urban Sci.*, vol. 22, no. 3, pp. 366–381, Jul. 2018, doi: 10.1080/12265934.2018.1431144.

[13] Y. Zheng, W. Li, F. Qiu, and H. Wei, "The benefits of introducing meeting points into flex-route transit services," *Transp. Res. C, Emerg. Technol.*, vol. 106, pp. 98–112, Sep. 2019, doi: 10.1016/j.trc.2019.07.012.

[14] M. Estrada, J. M. Salanova, M. Medina-Tapia, and F. Robusté, "Operational cost and user performance analysis of on-demand bus and taxi systems," *Transp. Lett.*, vol. 13, no. 3, pp. 229–242, Mar. 2021, doi: 10.1080/19427867.2020.1861507.

[15] M. E. Kim and P. Schonfeld, "Integration of conventional and flexible bus services with timed transfers," *Transp. Res. B, Methodol.*, vol. 68, pp. 76–97, Oct. 2014, doi: 10.1016/J.TRB.2014.05.017.

[16] L. Quadrifoglio, R. W. Hall, and M. M. Dessouky, "Performance and design of mobility allowance shuttle transit services: Bounds on the maximum longitudinal velocity," *Transp. Sci.*, vol. 40, no. 3, pp. 351–363, Aug. 2006, doi: 10.1287/trsc.1050.0137.

[17] B. W. Alshalalfah, *Planning, Design and Scheduling of Flex-Route Transit Service*. Toronto, ON, Canada: Univ. of Toronto, 2009.

[18] B. Alshalalfah and A. Shalaby, "Feasibility of flex-route as a feeder transit service to rail stations in the suburbs: Case study in Toronto," *J. Urban Planning Develop.*, vol. 138, no. 1, pp. 90–100, Mar. 2012, doi: 10.1061/(ASCE)UP.1943-5444.0000096.

[19] J. Zhao, S. Sun, and O. Cats, "Joint optimisation of regular and demand-responsive transit services," *Transportmetrica A, Transp. Sci.*, pp. 1–24, Oct. 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9739653, doi: 10.1080/23249935.2021.1987580.

[20] S. M. Nourbakhsh and Y. Ouyang, "A structured flexible transit system for low demand areas," *Transp. Res. B, Methodol.*, vol. 46, no. 1, pp. 204–216, Jan. 2012, doi: 10.1016/J.TRB.2011.07.014.

[21] Y. Lai, F. Yang, G. Meng, and W. Lu, "Data-driven flexible vehicle scheduling and route optimization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23099–23113, Dec. 2022, doi: 10.1109/TITS.2022.3204644.

[22] Y. Zheng, W. Li, and F. Qiu, "A slack arrival strategy to promote flex-route transit services," *Transp. Res. C, Emerg. Technol.*, vol. 92, pp. 442–455, Jul. 2018, doi: 10.1016/j.trc.2018.05.015.

[23] P. Vansteenwegen, L. Melis, D. Aktaş, B. D. G. Montenegro, F. S. Vieira, and K. Sörensen, "A survey on demand-responsive public bus systems," *Transp. Res. C, Emerg. Technol.*, vol. 137, Apr. 2022, Art. no. 103573, doi: 10.1016/j.trc.2022.103573.

[24] M. A. Esfeh, S. C. Wirasinghe, S. Saidi, and L. Kattan, "Waiting time and headway modelling for urban transit systems—A critical review and proposed approach," *Transp. Rev.*, vol. 41, no. 2, pp. 141–163, Mar. 2021, doi: 10.1080/01441647.2020.1806942.

[25] C. F. Daganzo, "An approximate analytic model of many-to-many demand responsive transportation systems," *Transp. Res.*, vol. 12, no. 5, pp. 325–333, Oct. 1978, doi: 10.1016/0041-1647(78)90007-2.

[26] S. Kikuchi and V. R. Vuchic, "Transit vehicle stopping regimes and spacings," *Transp. Sci.*, vol. 16, no. 3, pp. 311–331, Aug. 1982. [Online]. Available: http://www.jstor.org/stable/25768057

[27] H. Mohring, "Optimization and scale economies in urban bus transportation," *Amer. Econ. Rev.*, vol. 62, no. 4, pp. 591–604, 1972. [Online]. Available: http://www.jstor.org/stable/1806101

[28] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002, doi: 10.1109/4235.996017.

[29] A. J. Nebro, J. J. Durillo, J. Garcia-Nieto, C. A. C. Coello, F. Luna, and E. Alba, "SMPSO: A new PSO-based metaheuristic for multi-objective optimization," in *Proc. IEEE Symp. Comput. Intell. Milti-Criteria Decis.-Making*, Mar. 2009, pp. 66–73, doi: 10.1109/MCDM.2009.4938830.

[30] H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima, "Difficulties in specifying reference points to calculate the inverted generational distance for many-objective optimization problems," in *Proc. IEEE Symp. Comput. Intell. Multi-Criteria Decis.-Making (MCDM)*, Dec. 2014, pp. 170–177, doi: 10.1109/MCDM.2014.7007204.

[31] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. New York, NY, USA: Springer, 1981.

[32] *Canadian Transit Fact Book 2015 Operating Data*, CUTA, Toronto, ON, Canada, 2015.

[33] J. Shen, F. Qiu, C. Zheng, and C. Ma, "Fare strategy for flex-route transit services: Case study in Los Angeles," *IEEE Access*, vol. 7, pp. 82038–82051, 2019, doi: 10.1109/ACCESS.2019.2924320.

SUSHREETA MISHRA received the bachelor's degree in civil engineering from BPUT, India, and the master's degree in transportation engineering from BITS Pilani, India. She is currently pursuing the Ph.D. degree with the University of Manitoba, Winnipeg, Canada. Her Ph.D. thesis focuses on optimizing the operation of semi-flexible transit for low-demand conditions. Her primary research interests include transit operations, planning, and electrification.

BABAK MEHRAN is currently an Associate Professor of civil engineering with the University of Manitoba, Canada. His research interests include applications of big data analytics, spatial data science, and artificial intelligence in transportation operations planning and analysis, modeling and design of autonomous shared mobility services, optimization of traffic operations and public transportation systems, and transportation network resilience and reliability analysis.