Modeling and optimizing a fare incentive strategy to manage queuing and crowding in mass transit systems

Tang, Yili; Jiang, Yu; Yang, Hai; Nielsen, Otto Anker

Published in:
Transportation Research Part B: Methodological

Link to article, DOI:
10.1016/j.trb.2020.05.006

Publication date:
2020

Document Version
Peer reviewed version

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Modeling and optimizing a fare incentive strategy to manage queuing and crowding in mass transit systems

- Bilevel formulation, solution algorithm and Copenhagen case study

Yili Tang\textsuperscript{a, b}\*, Yu Jiang\textsuperscript{c}, Hai Yang\textsuperscript{a}, Otto Anker Nielsen\textsuperscript{c}

\textsuperscript{a} The Hong Kong University of Science and Technology, Department of Civil and Environmental Engineering, Clear Water Bay, Hong Kong
\textsuperscript{b} Institute of Transportation Studies, University of California, Berkeley, CA, United States
\textsuperscript{c} Department of Technology, Management and Economics, Technical University of Denmark, Bygningstорvet116B, 2800Kgs, Lyngby, Denmark

Abstract

This paper solves the problem of optimizing a surcharge-reward scheme and analyzes equilibrium properties incorporating commuters’ departure time choice to relieve crowding and queuing congestion in mass transit systems. The surcharge-reward scheme incentivizes commuters to switch departure times from a pre-specified central period to shoulder periods. We formulate a bilevel model to design and optimize the surcharge-reward scheme. The upper-level problem minimizes the total equilibrium costs by determining the refundable surcharges, the rewards, and the corresponding central charging period. The lower-level problem determines the equilibrium of commuters’ departure times in the peak period with respect to generalized travel costs. Equilibrium properties are analyzed and a sequential iterative solution algorithm is developed. We found that the existence of an optimal solution depends on the scheme design and there exists a lower bound on the surcharge to achieve the system optimum. Numerical studies are conducted on a commuting rail line in Copenhagen. The proposed algorithm converges efficiently, and the fare incentive scheme can simultaneously reduce the individual trip costs, total crowding costs, and total queuing time costs. The performance of the scheme increases with the rewards and surcharges up to a point and beyond which it stays unchanged.

Keywords: incentive, departure time choice, queuing and crowding, mass transit system, bilevel model

1. Introduction

Peaks and troughs in demand during peak and off-peak hours have negative effects on transit operators and users. The costly capacity required to fulfill peak demand is

\* Corresponding author. Tel: +852 23587175, E-mail address: yiloftang@gmail.com
inefficiently utilized during off-peak periods. In addition, peak congestion decreases the stability and reliability of transit services, reducing user satisfaction. As demand for public transportation is growing faster than improvements in transit capacity, the effects of queuing and crowding during peak periods have received increasing attention in the literature.

1.1 Background

A number of studies have modeled transit assignment and commuters’ departure times in transit systems incorporating the effects of queuing and crowding. Sumi et al. (1990) provided a commuter response function to deduce commuters’ route choice behavior and thereby minimize their disutility. Alfa and Chen (1995) proposed an algorithm to determine the distribution of commuter demand in multiple origin-destination public transportation systems. Nielsen (2000) modeled stochastic transit assignment while distinguishing waiting time inclusive of transfer and hidden waiting times in various situations. Moreover, Parbo et al. (2014) proposed a bilevel timetable optimization to minimize transfer time and waiting time for public transit users. In addition, Szeto and Jiang (2014a) and Jiang and Szeto (2016) developed a novel approach-based formulation for the transit assignment problem. Taking into account in-vehicle crowding cost in a transit system, Huang et al. (2004) introduced a crowding cost function and modeled the equilibrium of an urban mass transit system during peak hours. In line with this work, Tian and Huang (2007) explored the equilibrium properties of a mass transit system with a rigid capacity constraint.

Several empirical studies have investigated the cost of crowding and queuing time in the urban transit system. A meta-analysis of 17 British studies over 20 years found that the average crowding costs for seating and standing, in terms of time multipliers, were 1.19 and 2.32, respectively (Wardman and Whelan, 2011). Crowding not only affects individuals’ behavior and psychological health but is also the first urban externality considered by policymakers (Haywood and Koning, 2015; Tirachini et al., 2014). A survey conducted in 2010 on the Paris subway suggested that welfare loss during peak-hour transport was underestimated by 27% relative to the benchmark where crowding effects are not considered (Haywood and Koning, 2013). In Melbourne, Australia, the cost of overcrowding on public transport has been estimated to reach approximately $280 million per year (Veitch et al., 2013).

Researchers have also explored the effects of crowding and queuing congestion and management policies via the cost-benefit analysis on transit projects (de Palma et al., 2017; Parry and Small, 2009). Conducted from the perspective of transit authorities, most studies of transit queuing congestion and crowding have addressed transit capacity choice,
scheduling, and fare pricing. Theoretically, Tian et al. (2009) showed that the social optimum could be achieved by implementing time-varying fares to eliminate congestion in urban transit systems. Furthermore, de Palma et al. (2017) analyzed the transit crowding and schedule delay cost and derived the optimal pricing regimes and capacity of transit service. However, it is not always feasible to improve service supply in the peak period because transit lines with overcrowding during peak hours usually operate at the maximum capacity and expanding the capacity may incur prohibitively high costs. On the other hand, temporally differential fares are easy to implement as a way to reduce demand but politically unpopular: peak surcharges penalize peak-hour commuters and off-peak fare discounts are usually given at the expense of the government or transit operator.

Various studies have investigated more acceptable and practicable demand management strategies for road traffic in static and dynamic equilibrium models, such as the tradable credit and permit schemes to manage mobility in road traffic (Yang and Wang, 2011; Liu, W. et al., 2014; Bao et al., 2016), the revenue-refunding and Pareto-improving strategies (Daganzo and Garcia, 2000; Guo and Yang, 2010), and the incentives for managing congestion in shared mobility and truck platooning systems (Pfrommer et al., 2014; Farokhi and Johansson, 2015). The few studies in transit systems have found that a combination of fare surcharge and off-peak fare discounts has larger effects than a single strategy (Douglas et al., 2011; Whelan and Johnson, 2004). Moreover, Halvorsen et al., (2016) reviewed the effects of time-varying fares in mass transit systems on travel pattern changes. Besides, Yang and Tang (2018) and Tang et al. (2019) designed revenue-neutral fare-reward schemes in a transit bottleneck model, which indicates that a significant reduction in total time costs is possible if passengers could be incentivized to shift their travel to the shoulder period of the peak. While alternatives and incentives in road traffic are well researched, those in mass transit systems are still emerging. Moreover, researchers have devoted less attention to the effects of crowding and queuing on pricing strategies and incentives in mass transit systems than in road traffic. Yet, as demand is growing faster than transit capacity expansion in many metropolises, managing crowding and queuing in the transit systems becomes increasingly important.

1.2 Motivation and suggested fare incentive strategy

The goal of this paper is to analyze equilibrium properties and optimize the performance of a surcharge-reward scheme (SRS) for managing congestion (queuing and crowding) in a mass transit system, by using a departure time equilibrium model that incorporates user crowding cost, queuing time cost, and schedule delay cost. The SRS is implemented such that a passenger incurs fare surcharges during a central peak period which are fully refundable and accumulated in his/her personal account. Commuters can use the monetary
rewards in their account from previous surcharges if they travel in shoulder periods. We devised a bilevel model to design the optimal SRS and proposed a solution algorithm for the global optimal solution. The diagram of the overall methodology is shown in Figure 1. The upper-level model minimizes the total equilibrium costs and determines the optimal location and duration of the central period and the surcharges and rewards with the incentive constraint. The lower-level model takes into account the SRS devised by the upper-level model and considers the equilibrium of commuters’ departure time choices based on the trade-offs between schedule delay cost, crowding cost, and queuing time cost. We solved the model by analyzing the equilibrium properties for the global optimal solution and proposed the solution algorithm. In the solution algorithm, we first defined the equilibrium conditions and then proved the global optimal solution for different variables sequentially.

We used the Copenhagen metropolitan area as a case study. The Copenhagen transport system is crowded and congested during peak periods, especially in some of the important links such as the central stations, sections of the downtown subway and some of the urban commuter rail line sections, and several downtown bus routes (Danish Ministry of Transport, 2000; Danish Ministry of Transport, 2011). Moreover, Copenhagen has already implemented simple surcharge strategies to tackle this problem, which enables analyzing the effects of the suggested improved SRS versus the existing pricing strategies.

Figure 1. Overall methodology

The SRS has several advantages from the perspective of passengers, operators as well as system development.

For passengers, all the surcharges are refunded to their personal accounts which incentivizes them to keep traveling on this system since they already have an account...
balance. The individual-based concept also improves fairness because commuters who pay the surcharges get those back in their own account directly, which is different from many anonymous pricing strategies where operators collect the surcharges during peak periods and utilize them for the overall compensation or operation (e.g. off-peak discounts). The SRS is more straightforward and can encourage commuters to shift their departure times.

For operators, the SRS has several control variables that allow them and policymakers to adjust the performance and implementation smoothly. They can start with a wide time window within the central peak period and/or a small fare difference to understand commuters’ behaviors and then adjust it gradually. This will also increase the practicality and the public’s acceptance of the proposed scheme.

In terms of system development, the SRS can be easily implemented through the current smartcard system. Additionally, in the long term, the system performance and commuter behaviors can be predicted with the collection of a large amount of data on individual commuters’ responses and travel choices that are generated every day. This set of data can provide analytics and guidance for infrastructure development and service improvement to operators and policymakers.

1.3 Outline of the paper

The rest of the paper is organized as follows. Section 2 describes the passenger generalized travel cost and formulates a general departure time equilibrium model with a rigid available capacity constraint. Section 3 devises the bilevel model and is devoted to the theoretical and algorithmic analysis of the model. We analyze the equilibrium properties of the problem to explore commuters’ departure time choice under the proposed fare incentive scheme and provide a theoretical foundation for algorithmic development. Section 4 describes the proposed sequential iterative algorithm designed to give the global optimal solution. Section 5 conducts case studies to validate the proposed solution algorithm and assesses the performance of the SRS in changing commuters’ departure times and reducing trip costs. Finally, Section 6 discusses future implementations and concludes the paper.

2. Equilibrium departure time choice with a rigid capacity constraint under a uniform fare

In this section, we introduce the generalized trip cost under a uniform fare and formulate a general equilibrium departure time choice model inclusive of queuing time costs, crowding costs and schedule delay costs with a rigid capacity constraint in a mass transit system.
2.1 The generalized trip cost

As shown in Figure 2, we consider one dedicated transit line connecting two intermediate stations (denoted by diamonds) without stops. During the peak period, there is a total of $M$ uniform service runs along the transit line. Each service run refers to a train with the full (physical) capacity $S$. Because of the intermediate stations, the actual supply of each service run between the two stations is constrained by the available capacity $s$ at the departing station. When boarding passengers occupy all the available capacity $s$, the service run is at the full capacity $S$. The schedule departure time of service run $m$ is given by $t_m$ and the headway between two consecutive runs is maintained at $h$. The in-vehicle time between the two stops is a constant $d$ for all service runs, and the travel time from home to the station as well as from the station to the workplace is ignored. Hence, commuters who take a service run at $t_m$ will arrive at work at $t_m + d$. At the demand side, a given total of $N$ homogeneous individuals commute between the two stations and have an identical starting time for work, $t^*$, which is assumed to correspond to the departure time of service run $m^*$ such that $t^* = t_m^* + d$.

![Figure 2. Transit system with uniform headway service runs](image)

A commuter who takes a service run departing at $t_m$ encounters a uniform fare cost $p_0$, a queuing time cost $q(t_m)$, a crowding cost $c(t_m)$, a schedule delay cost $\delta(t_m)$ if he/she arrives at work early or late, and an in-vehicle time cost $\phi$ which is a constant assuming that travel time between the two stations is constant. Therefore, the generalized trip cost is given by

$$TC(t_m) = \delta(t_m) + q(t_m) + c(t_m) + p_0 + \phi, m = 1, 2, \ldots, M.$$ (1)

The schedule delay cost is the disutility of an individual from not arriving at work on time. From much of the analysis, the schedule delay cost is assumed to be in a piecewise linear form. It is zero if one arrives at the official starting time for work, $t^* = t_m^* + d$, and is strictly
decreasing with the departure time if $t_m < t^*$ while strictly increasing with the departure time if $t_m > t^*$. $\delta(t_m)$ is thus given by

\[
\begin{align*}
\delta(t_m) &= \beta(t^* - t_m) \text{ and } \delta'(t_m) < 0, \text{ if } t_m < t^* \\
\delta(t_m) &= 0, \text{ if } t_m = t^*, m = 1, 2, \ldots, M \\
\delta(t_m) &= \gamma(t_m - t^*), \text{ and } \delta'(t_m) > 0, \text{ if } t_m > t^*
\end{align*}
\] (2)

where the unit costs $\beta$ and $\gamma$ are the unit costs of arriving early and arriving late respectively.

The crowding costs in public transport have been examined by several studies and formulated as a function of the number of commuters and the in-vehicle travel time (Hamdouch et al. 2014; de Palma et al., 2015; de Palma et al., 2017; Tian et al., 2007). In this paper, we assume a general form of the crowding cost incurred by commuters, which is given by

\[
c(t_m) = g(n_m) d, m = 1, 2, \ldots, M
\] (3)

where $n_m$ is the total number of commuters taking service run $m$, and $g(n_m)$ is the unit cost of crowding, which is assumed to be monotonically increasing with $n_m$ and twice differentiable convex that $g'(\cdot) > 0$, $g''(\cdot) > 0$, and $g(0) = 0$.

The queuing time cost $q(t_m)$ is induced by the capacity constraint. In line with the literature (Tian and Huang, 2007; Tian et al., 2007; Yang and Huang, 2005), the queuing time cost is defined to be no less than zero if a service run is at its full capacity and is zero otherwise. The equilibrium condition of queuing time costs can be mathematically expressed as

\[
\begin{align*}
q(t_m) &= 0, \text{ if } n_m < s \\
q(t_m) &\geq 0, \text{ if } n_m = s, m = 1, 2, \ldots, M
\end{align*}
\] (4)

Hence commuters can board the train if it is not full upon its arrival or they would have to wait for the next one and incur additional queuing time costs if the train is full. Such
queuing time costs can be regarded as the expected queuing time costs that commuters will incur in each full service run\(^1\).

Since both the uniform fare and in-vehicle time costs are constant, in equilibrium, commuters face the trade-offs between schedule delay cost, queuing time costs and crowding costs. The equilibrium generalized trip cost can be simplified as

\[
v(t_m) = \max \left\{ \beta (t_m - t_m), \gamma (t_m - t_m) \right\} + q(t_m) + c(t_m), \quad m = 1, 2, \ldots, M.
\]  

(5)

2.2 Equilibrium condition

At equilibrium, no commuter can reduce his/her generalized trip cost by unilaterally changing departure times or service runs. Hence all commuters should have the same generalized trip cost. The equilibrium condition can be mathematically expressed as

\[
\begin{cases}
    v(t_m) = v, & \text{if } n_m > 0 \\
    v(t_m) \geq v, & \text{if } n_m = 0, \quad m = 1, 2, \ldots, M
\end{cases}
\]  

(6)

where \( v \) is the equilibrium generalized trip cost, and \( v(t_m) \) is the generalized trip cost of taking service run \( m \). The condition (6) states that if a service run is utilized by any commuter, then the generalized trip cost associated with the service run is the same as the equilibrium generalized trip cost, and if a service run is empty, then the generalized trip cost associated with this run is no less than the equilibrium generalized trip cost.

With a fixed number of commuters, a uniform fare has no effect on the fare-free user equilibrium. Moreover, since commuters travel the same distance between the two stations, the in-vehicle time cost is identical for all commuters and has no effect on their departure time distribution equilibrium. Hence this latter equilibrium can be derived by solving the following minimization problem:

\[
\min L(n) = \sum_{m=1}^{M} G(n_m) \cdot d + \sum_{m=1}^{M} n_m \cdot \delta(t_m)
\]  

(7)

\(^1\) This paper focuses on the overall queuing time cost which is not directly modeled but is derived from the equilibrium conditions by the capacity constraint. Some researchers discussed the queuing costs under different queuing principles (Yoshida, 2008; Kraus and Yoshida, 2002). In this paper, we assume passengers follow the first-in-first-out (FIFO) queuing principle. In practice, it might happen that some passengers access the service run randomly (not following FIFO) and bypass the queue, however, we believe this is rare. And at the aggregate level, this would yield the same average queuing time since the benefit of a bypass would equal to the loss of another access. We therefore keep the FIFO principle. At the equilibrium, the queuing time costs are the difference between equilibrium trip costs and the sum of schedule delay costs and crowding costs in each service run.
\[
\sum_{m=1}^{M} n_m = N \quad (8)
\]

\[
n_m \leq s, \ m = 1,2,\ldots, M \quad (9)
\]

\[
n_m \geq 0, \ m = 1,2,\ldots, M \quad (10)
\]

The first term of the objective function, \( G(n_m) = \int_0^{n_m} g(n_m) \), is the integral of the unit crowding cost function which has no economic interpretation, similar to Beckmann’s mathematical model (Beckmann et al., 1956). The second term is the aggregate schedule delay cost, where the schedule delay cost is as defined in (2). Constraint (8) requires that all demand be served and constraint (9) represents the rigid available capacity constraint.

Since both the objective function and constraints are convex, the commuter departure time equilibrium can be obtained by deriving the first-order condition of the minimization problem,

\[
\begin{align*}
&n_m \left( \delta(t_m) + q(t_m) + c(t_m) - v \right) = 0, \ m = 1,2,\ldots, M \\
&\delta(t_m) + q(t_m) + c(t_m) - v \geq 0, \ m = 1,2,\ldots, M \\
&q(t_m)(s - n_m) = 0, \ m = 1,2,\ldots, M \\
&q(t_m) \geq 0, \ m = 1,2,\ldots, M \\
&n_m \leq s, \ m = 1,2,\ldots, M \\
&\sum_{m=1}^{M} n_m = N \\
&n_m \geq 0, \ m = 1,2,\ldots, M
\end{align*}
\quad (11)
\]

where the first two equations are the equilibrium conditions of generalized trip costs as indicated in condition (6) and the third to fifth equations are the equilibrium conditions of queuing time costs as indicated in condition (4). The last two equations are demand and variable constraints.

The first-order optimality conditions indicate the equilibrium of commuters’ departure time choices is based on their generalized trip cost which is incorporated to investigate the effects of the SRS on commuters’ departure times. In the following sections, we model and examine the proposed fare incentive strategies with departure time choices and assess their efficiency and practicality.
In Section 3, we introduce the design of the SRS and establish a bilevel model to optimize the proposed scheme. The equilibrium properties will be analyzed and proved to interpret the performance of the scheme and find the global optimal solution. After that in Section 4, we propose the solution algorithm to obtain the solutions based on the properties and theorems.

3. Optimization of the surcharge-reward scheme

The unpriced departure time equilibrium reveals commuter departure time choice which is not socially optimal as queuing time and crowding lead to efficiency loss. It has been well acknowledged in the literature of road traffic congestion management that the aggregate equilibrium time costs can be reduced by implementing pricing strategies, incentives, mobility credits, etc. In this section, we depict a revenue-neutral SRS to manage transit congestion and develop a bilevel model for the optimal configuration of the scheme.

3.1 The surcharge-reward scheme in a mass transit system

As discussed above, a few studies have investigated reward schemes and incentives to reduce congestion and queuing during the peak period. For instance, Daganzo and Garcia (2000) proposed a pricing strategy that offers a proportion of users exemption from tolls. Rouwendal et al. (2012) rewarded commuters at the two shoulders of the peak period. Yang and Tang (2018) proposed a fare-reward scheme where one assigned free trip is rewarded to commuters during the shoulder period after a certain number of paid journeys within the peak period and they further investigated a hybrid fare scheme with heterogeneous commuters (Tang et al., 2018; Tang et al., 2019). While previous studies have focused on queuing congestion and many of them derived an optimal coarse toll scheme for the road bottleneck model, this study considers crowding effects and derives equilibrium travel patterns with a discrete set of alternatives rather than a continuous departure time choice set.

In the spirit of incentive and revenue neutrality, this paper proposes a SRS with the integration of surcharge and reward. In the SRS, there are two reward fare intervals (RFIs) and one surcharge fare interval (SFI). The SFI is the central peak period from $t_i$ to $t_f$ and including the work starting time $t^* = t_{m}$. The two RFIs refer to the two shoulder intervals before and after the SFI.

Implementation of the SRS introduces a surcharge $\Delta s$ to the original uniform fare $p_0$ during the predetermined central peak period, i.e., SFI. The surcharges are fully reimbursed and accumulated in a commuter’s personal account and can be used during the shoulder
periods for a valid period. Therefore, the SRS does not require commuters to change their departure time every day but rather occasionally on those days that are convenient to them. During the shoulder periods, i.e., RFIs, for each trip, commuters can use the monetary reward $\Delta r$ from their account balance accumulated by previous surcharges. Commuters are not allowed to use the reward if the account balance is less than the value of the reward. Since the SRS aims to incentivize commuters to shift their travel to the reward period without incurring more trip costs, we have the following design criterion:

**Design criterion:** The SRS is designed such that the generalized trip costs for commuters traveling in the RFIs are no greater than those for commuters traveling in the SFI. Mathematically, it is stated as

$$v^\text{RFI} - \Delta r \leq v^\text{SFI} + \Delta x$$

where $v^\text{RFI}$ and $v^\text{SFI}$ denote the equilibrium generalized trip costs for the commuters traveling in the reward period and the surcharge period, respectively. Under the proposed SRS with the design criterion, commuters with rewards will only travel in the RFIs while commuters without reward eligibility will travel in the SFI. Hereinafter we define commuters using rewards traveling in the RFIs as RFI commuters and commuters paying surcharges and traveling in the SFI as SFI commuters. Note that a commuter who travels in the central period as an SFI commuter on one day can be an RFI commuter traveling in a shoulder period on another day. Commuters’ departure times are then divided into three intervals as shown in Figure 3.

![Figure 3. Commuters’ travel choice under the SRS](image)

From the equation (12), it is straightforward to see that $\Delta s/\Delta r$ represents the number of shoulder period trips that a commuter can take with one surcharge. If $\Delta s/\Delta r > 1$, the surcharge is greater than the reward, and hence a commuter can use a reward after a trip involving a surcharge and needs to take more trips during the shoulder periods than during the central period to get fully refunded. On the contrary, if $\Delta s/\Delta r < 1$, a surcharge is less than a reward, and hence a commuter can use a reward after several trips involving a
surcharge and in the meantime, he/she takes fewer trips during the shoulder periods than during the central period to get fully refunded.

Based on the interpretation of $\Delta s/\Delta r$, we further define the reward ratio as

$$
\lambda = \frac{\Delta s/\Delta r}{1 + \Delta s/\Delta r}
$$

which can be regarded as the ratio of the number of trips with a reward to the total number of trips made by one commuter. Besides, assuming the number of commuters is sufficiently large, the reward ratio also represents the proportion of commuters with rewards to the total number of commuters on an average day. Hereinafter, we use the reward ratio to represent the reward value $\Delta r$, where $\Delta r = (1-\lambda)\Delta s/\lambda$, and a higher reward ratio indicates a smaller reward for a given surcharge. The equation (12) can thus be reformulated as

$$
\nu^{\text{RFI}} \leq \nu^{\text{SFI}} + \frac{\Delta s}{\lambda}.
$$

With such a reward scheme, every day a proportion of commuters will be encouraged to shift their travel to the reward period because of the lower trip cost with rewards.

### 3.2 The bilevel model of surcharge-reward scheme optimization

The implementation and design of the SRS require determining not only the amount of surcharge and reward but also how to split the peak period into the SFI and RFIs. Thus, we devise four decision variables for the SRS design problem, namely, 1) surcharge $\Delta s$, 2) reward ratio $\lambda$, 3) SFI location $e$, which is the index of the starting service run of the SFI, and 4) SFI duration $l$, which refers to the total number of service runs in the SFI, i.e., the length of the SFI. While the surcharge and reward ratio are continuous variables, the SFI location and duration are integer variables. As shown in Figure 3, the SFI and RFIs can be inferred from the SFI location and duration. Once the SFI interval is determined, the two RFIs are defined correspondingly. The bilevel optimization problem can be formulated as follows:

Upper level:

$$
\min_{\lambda, e, l, \Delta s, n^{\text{SFI}}, n^{\text{RFI}}} \ TEC = \nu(n^{\text{SFI}})(1-\lambda)N + \nu(n^{\text{RFI}}) \cdot \lambda N
$$

subject to:

$$
\nu(n^{\text{RFI}}) \leq \nu(n^{\text{SFI}}) + \frac{\Delta s}{\lambda}
$$
\[ e \leq M, e \in \mathbb{Z}^{+} \] (17)
\[ l \leq M, l \in \mathbb{Z}^{+} \] (18)
\[ \Delta s \geq 0 \] (19)

Lower level:

\[
\min_{n_{sfi}^{m}, n_{rﬁ}^{m}} \sum_{m=1}^{M} G(n_{sfi}^{m}) \cdot d + \sum_{m=1}^{M} n_{sfi}^{m} \cdot \delta_{sfi}^{m}(t_{m}) \\
+ \sum_{m=1}^{M} G(n_{rﬁ}^{m}) \cdot d + \sum_{m=1}^{M} n_{rﬁ}^{m} \cdot \delta_{rﬁ}^{m}(t_{m})
\] (20)

subject to:

\[ \sum_{m=1}^{M} n_{sfi}^{m} = (1-\lambda)N \] (21)
\[ n_{sfi}^{m} \leq s, m = 1, \ldots, M \] (22)
\[ n_{sfi}^{m} \geq 0, m = 1, \ldots, M \] (23)
\[ \sum_{m=1}^{M} n_{rﬁ}^{m} = \lambda N \] (24)
\[ n_{rﬁ}^{m} \leq s, m = 1, \ldots, M \] (25)
\[ n_{rﬁ}^{m} \geq 0, m = 1, \ldots, M \] (26)

where \( \delta_{sfi}^{m}(t_{m}) \) and \( \delta_{rﬁ}^{m}(t_{m}) \) can be expressed as

\[
\delta_{sfi}^{m}(t_{m}) = \begin{cases} 
\bar{\delta}(t_{m}) & \text{if } e \leq m \leq e+l-1 \\
\inf & \text{otherwise}
\end{cases}
\] (27)
\[
\delta_{rﬁ}^{m}(t_{m}) = \begin{cases} 
\inf & \text{if } e \leq m \leq e+l-1 \\
\delta(t_{m}) & \text{otherwise}
\end{cases}
\] (28)

In the upper-level problem, equation (15) minimizes the total equilibrium costs of all the commuters, where \( v(n_{sfi}) \) and \( v(n_{rﬁ}) \) are the equilibrium costs of SFI and RFI commuters respectively, which are obtained by solving the lower-level problem. The equation (16) is the incentive design criterion constraint requiring that commuters using rewards have no greater trip costs than those paying surcharges. Equations (17) - (19) are
the definitional constraints. The upper-level problem is a mixed integer nonlinear programming problem due to the nonlinearity in the objective function and constraint (16) and the integrality of $e$ and $l$. The lower-level problem extends the general equilibrium departure time choice model introduced in Section 2 by dividing the commuters into SFI and RFI commuters. The first two terms in the objective function and constraints (21)-(23) together represent the departure time choice equilibrium of SFI commuters, while the last two terms in the objective function and constraints (24)-(26) together represent that of RFI commuters. Therefore, for a given upper-level SRS design represented by $(e, l, \lambda)$, the solutions to the lower-level problem can be obtained by solving the convex optimization problem in the lower-level model. For mathematical analysis, we treat $n^{SFI}$ and $n^{RFI}$ as continuous variables hereinafter.

3.3 Properties of the optimization problem

Various algorithms have been proposed to solve bilevel models (i.e., Yang and Bell, 2001; Yang and Huang, 2005; Szeto and Jiang, 2012; Jiang and Szeto 2015; Szeto and Jiang, 2014b) and several studies have developed methods to find the global optimal solution for the network design problem (Wang and Lo, 2010; Li et al., 2012; Wang et al., 2013; Liu and Wang, 2015; Liu et al., 2019) via either branch and bound or linearization techniques. In this study, we develop a solution algorithm that attains the global optimal solution by analyzing and utilizing the properties of the proposed bi-level optimization model.

The rest of this section examines the linkage among the decision variables. As shown in Figure 4, we first introduce definitions of a free-flow starting service and duration and the set of efficient reward ratios regarding the departure time distribution equilibrium. After that, we investigate the properties of the optimal solution with SFI location, reward ratio and surcharge sequentially for a given SFI duration.
3.3.1 Properties of departure time distribution equilibrium

In the equilibrium departure time choice model stated in (7) - (10), a commuter’s equilibrium departure time essentially depends on the trade-offs between the crowding cost, schedule delay and queuing time cost. The queuing time cost exists only if the arriving service run is at its full capacity. For non-full trains, a commuter’s equilibrium generalized trip costs consist of only schedule delay costs and crowding cost. For a given total number of commuters and the timetable, the number of non-empty service runs and the index of non-empty starting service runs is endogenously given by equilibrium conditions. Accordingly, we introduce the following definition to facilitate our analysis.

**Definition 1:** Free-Flow Starting Service and Duration: For a given total number of \( N \) commuters and timetable, under the equilibrium departure time choice conditions:

i) The free-flow duration is defined as the number of non-empty service runs and is denoted by \( L_N \).

ii) The free-flow starting service run is defined as the first non-empty service run and is denoted by \( E_N \).
Since free-flow duration and service runs are endogenous, the value of $L_N$ depends on the mass transit demand, cost functions, and capacity constraints. The second definition is then introduced.

**Definition 2:** Efficient Reward Ratio Set: For a given SFI location and duration $(e, l)$ and the total number of $N$ commuters, an efficient reward ratio set is defined such that all service runs are non-empty within the SFI, namely, $\Omega_l = \left\{ \lambda_l \left| L_{(1-l\cdot s/N)} = l \right. \right\}$, where $L_{(1-l\cdot s/N)}$ is the free-flow duration for total demand $(1-l\cdot s/N)$.

Definition 2 demonstrates the range of efficient reward ratios. The minimum efficient reward ratio is denoted by $\lambda_{\text{min}} = \max \{(1-l\cdot s/N), 0\}$. $\lambda_{\text{min}} = (1-l\cdot s/N)$ represents the full capacity utilization in the case that the SFI duration is short, whereas $\lambda_{\text{min}} = 0$ represents a very long SFI duration $l$ such that $(1-l\cdot s/N) \leq 0$. Hence the reward ratio is set to 0 and all commuters travel within the SFI and the scheme is not effective in this case. The maximum efficient reward ratio is denoted by $\lambda_{\text{max}}$, which implies that the free-flow duration with the reward ratio $\lambda_{\text{max}}$ is exactly equivalent to the SFI duration and an increase in reward ratio will result in empty service runs.

Next, we can derive the following property to refine the solution space of the optimal reward ratio.

**Proposition 1:** An optimal reward ratio must be an efficient reward ratio.

**Proof:** Proof of Proposition 1 is provided in Appendix 1.

Proposition 1 indicates that the optimal solution belongs to the efficient reward ratio set such that all service runs in the SFI are non-empty. We can then derive Theorem 1 below.

**Theorem 1:** Optimal SFI location: For a given SFI duration $l$ and the corresponding efficient reward set $\Omega_l$, the optimal SFI location $e^*_l$ (first service run of the SFI) is set such that the schedule delay cost of the SFI is minimized, namely

$$e^*_l = \arg \min_{m=1,2,3,...,M} \left\{ \max \{\delta(t_m), \delta(t_{m+l-1})\} \right\}.$$

**Proof:** Proof of Theorem 1 is provided in Appendix 2.

Proposition 1 and Theorem 1 indicate that the optimal SFI location can be determined with a given SFI duration. Meanwhile, Proposition 1 implies that given an SFI duration solution,
the space of the optimal reward ratio can be refined to an efficient reward ratio set. It is therefore important to explore the performance of the SRS with efficient reward ratios and the properties of the optimal solution.

### 3.3.2 Properties of the surcharge-reward scheme with efficient reward ratios

To derive the properties of the SRS with efficient reward ratios, we first introduce the following assumption.

**Assumption 1:** The crowding cost at full capacity is greater than the maximum schedule delay cost difference between two service runs, namely, \( g(S)d > \gamma \cdot h \).

Assumption 1 ensures that the crowding cost is high enough to cause travelers to change to a farther service run when the current one is crowded. Empirical studies have also indicated that the crowding costs at full capacity are significant enough which can be greater than the schedule delay cost difference between two service runs (Tian et al., 2009; Wardman and Whelan, 2011). Moreover, assumption 1 implies that for a given \( \lambda \), the free-flow duration of SFI is greater than \( \left[(1-\lambda) \cdot N/s\right]^+ \), where \( \left[ \cdot \right]^+ \) is rounded up to the next integer. This is because commuters can reduce their equilibrium cost by shifting to a farther service run due to the high crowding cost of the current service run.

Under Assumption 1 and Proposition 1, we suppose all efficient reward ratios are feasible which satisfy the constraints of the optimization problem. We thus have Theorem 2 below.

**Theorem 2:** Suppose a set of efficient reward ratios, \( \Omega_i \), is in a large range such that \( \lambda_{\min} \) is close enough to zero \( (\lambda_{\min} \to 0) \) and \( \lambda_{\max} \) is close enough to 1 \( (\lambda_{\max} \to 1) \) and all reward ratios within the set are feasible, and \( \forall n \leq S, g'(n) \geq g''(n) \). Then there exists a unique optimal reward ratio \( \lambda_i^* \) such that the total equilibrium costs decrease with the reward ratio at \( \lambda_i \in (0, \lambda_i^*) \) and increase afterward at \( \lambda_i \in [\lambda_i^*, 1] \).

**Proof:** Proof of Theorem 2 is provided in Appendix 3.

Theorem 2 states that given the feasible efficient reward ratio set, the optimal reward ratio exists and is unique. Combining Proposition 1 and Theorem 1 implies that given an SFI duration, we can find a unique optimal reward ratio within an efficient reward set \( (0, 1) \).

In Theorem 2, the efficient reward ratio set is assumed to have a sufficiently large range
within \((0,1)\), whereas its actual range varies with the SFI duration according to Definition 2. We thus introduce Corollary 1 below.

**Corollary 1**: Given the SFI duration \(l\), efficient reward ratio set \(\Omega\), and optimal SFI location \(e_i^*\), suppose all efficient reward ratios are feasible, we have three cases:

a) The total equilibrium costs are decreasing with the efficient reward ratios, and the optimal reward ratio is \(\lambda_i^* = \max \Omega\).

b) The total equilibrium costs are increasing with the efficient reward ratios, and the optimal reward ratio is \(\lambda_i^* = \min \Omega\).

c) The total equilibrium costs are decreasing first with the efficient reward ratios until a threshold value \(\lambda_i = \overline{\lambda}_i\) and increasing afterward, and the optimal reward ratio is \(\lambda_i^* = \overline{\lambda}_i\).

Theorem 2 and Corollary 1 reveal that a unique optimal reward ratio will always exist despite the range of the efficient reward ratio set if all efficient reward ratios are feasible.

The analysis so far does not consider the incentive constraint and assumes all efficient reward ratios are feasible. For a given \(l\) with corresponding \(e_i^*\) and \(\Omega\), the feasible set of variables and the performance of the SRS are constrained by the incentive as stated in constraint (16), which leads to Theorem 3 below.

**Theorem 3**: Given an SFI duration \(l\) and corresponding \(\Omega\) and \(e_i^*\), there exists a lower bound on the surcharge \(\Delta s\) and a threshold value of reward ratio \(\lambda_i^{\Delta s}\) such that

a) When the surcharge is set such that \(\Delta s \geq \Delta s_i\), the constraint (16) is always satisfied and the feasible reward ratio set is the same as the efficient reward ratio set \(\Omega\).

b) When the surcharge is set such that \(\Delta s < \Delta s_i\), if \(\Delta s < \left(\nu(n^{\text{RFI}}) - \nu(n^{\text{SFI}})\right) \cdot \lambda_{\min}\), no feasible solutions exist, and if \(\Delta s \geq \left(\nu(n^{\text{RFI}}) - \nu(n^{\text{SFI}})\right) \cdot \lambda_{\min}\), the feasible reward ratio set is \(\lambda \in [\lambda_{\min}, \lambda_i^{\Delta s}]\).
Proof. Proof of Theorem 3 is provided in Appendix 4.

Theorem 3 shows that there exists a lower bound on the surcharge which achieves the system optimum. The theorem indicates that a mass transit system with a higher refundable surcharge and thus a higher reward is more flexible due to the larger set of feasible reward ratios and may perform better. Intuitively, it can be interpreted that a high surcharge and thus a high reward yields a high difference in the trip costs between SFI and RFI commuters, which strengthens the incentives. On the contrary, if the surcharge or reward is low, commuters would have less incentive to shift their departure times, and hence the queuing time is not reduced sufficiently.

From Theorem 2 and Theorem 3, we can readily obtain the feasible set and the optimum of the reward ratio given \((l, e^*_j, \Omega)\), which leads to Corollary 2 below.

**Corollary 2:** Given an SFI duration \(l\) and corresponding \(\Omega\) and \(e^*_j\), the optimal reward ratio is determined such that

a) when \(\Delta s \geq \Delta s^*_j\), the optimal reward ratio is the same as in Corollary 1.

b) when \(\Delta s < \left( v\left(n^{\text{RFI}}\right) - v\left(n^{\text{SFI}}\right) \right) \cdot \lambda_{\text{min}}, \) there is no feasible reward ratio.

c) when \(\left( v\left(n^{\text{RFI}}\right) - v\left(n^{\text{SFI}}\right) \right) \cdot \lambda_{\text{min}} \leq \Delta s < \Delta s^*_j\),

   i. if \(\lambda_{\text{min}} \leq \lambda_{j}^{\Delta s} \leq \lambda_{\text{max}}\), the optimal reward ratio is \(\lambda_{j}^{\Delta s}\);
   
   ii. if \(\lambda_{\text{min}} \leq \lambda_{j} \leq \lambda_{j}^{\Delta s}\), the optimal reward ratio is \(\lambda_{j}\);
   
   iii. for \(\lambda_{\text{min}} \leq \lambda_{j} \leq \lambda_{j}^{\Delta s}\), the optimal reward ratio is \(\lambda_{\text{min}}\).

where \(\lambda_{\text{min}}\) is the minimum efficient reward ratio, \(\Delta s^*_j\) and \(\lambda_{j}^{\Delta s}\) correspond to the lower bound on the surcharge and the threshold value of the reward ratio as indicated in Theorem 2, and \(\lambda_{j}\) is the threshold value of the reward ratio as in Corollary 1.

Corollary 2 implies that the feasible and optimal reward ratios are affected by the surcharge. The existence of an optimal solution also depends on the surcharge. The aforementioned theorems and corollaries state the optimal conditions of the SFI location, reward ratio and surcharge given the SFI duration. Since the feasible set of the SFI duration is limited by
the free-flow duration such that \( l \leq l_N \), we can find the global optimal solution for the bilevel problem with a sequential iterative algorithm.

4. Solution algorithm

The proposed SRS can be implemented in the system with an existing differential fare strategy or in a new fare system. In the case of implementation with existing pricing strategies, the surcharge is given. Hence the decision variables of the upper-level problem reduce to \( l, e \), and \( \lambda \) with a given surcharge \( \Delta s \). The values of the three variables can be determined sequentially, as indicated in Section 4.1 below. On the other hand, when the SRS is implemented together with a new fare system, we can find the lower bound on the surcharge in which the SRS achieves the system optimum, as indicated in Theorem 3 (a) and based on the proposed algorithm with different stopping criteria, as described in Section 4.2.

4.1 Algorithm with a given surcharge

The algorithm consists of three main steps. In step 1, the original departure time equilibrium without a SRS is obtained by solving the lower-level model. The upper bound on the SFI duration and the original total equilibrium costs are derived from the equilibrium departure time choice. The step size of the reward ratio is determined for the iteration. In step 2, given an SFI duration, the algorithm calculates the optimal solution of the SFI location and obtains the optimal solution of the reward ratio as proved in the above analysis. Finally, step 3 adopts the global optimal SRS design for the mass transit system. The algorithm is outlined below.

Algorithm 1: Solve the bilevel optimal fare incentive model with a given surcharge

Step 1: Compute the initial equilibrium departure time choice

1.1. For the given demand and transit service settings, solve the equilibrium departure time choice model and obtain the original total equilibrium costs \( TEC_0 \) and free-flow duration \( l_N \).

1.2. Set \( TEC_{\min}^{(0)} = TEC_0 \).

1.3. Set \( l^{(0)} = 0 \) and \( l_{\max} = l_N \).

1.4. Set a tolerance number \( \Delta \lambda \) as the step size for updating the reward ratio.
1.5. Let $i = 1$.

Step 2: Compute the optimal solution.

2.1. Update the SFI duration by setting $l^{(i)} = l^{(i-1)} + 1$, and compute the optimal starting service run $e_{(i)}^*$ as indicated in Theorem 1.

2.2. Set the initial minimum reward ratio as in the full capacity condition, i.e. $\lambda^{(i,0)} = \max \left\{ \left( 1 - \frac{l^{(i)}}{s/N} \right) , 0 \right\}$.

2.3. Solve the lower-level problem with $l^{(i)}, e_{(i)}^*$ and $\lambda^{(i,0)}$.

2.4. If $v^{\text{SFL}(i,0)} + \frac{\Delta \lambda}{\lambda^{(i,j)}} < v^{\text{RFL}(i,0)}$, set $i = i + 1$ and go to Step 2.1, otherwise update the total equilibrium costs $TEC_{\text{min}}^{(i)} = TEC^{(i,0)}$ and continue.

2.5. Compute the optimal reward ratio at $l^{(i)}$ (includes an inner loop with index $j$).

2.5.1. Let $j = 1$.

2.5.2. Update the reward ratio and set $\lambda^{(i,j)} = \lambda^{(i,j-1)} + \Delta \lambda$.

2.5.3. Solve the lower-level problem and obtain $n^{\text{SFL}(i,j)}, n^{\text{RFL}(i,j)}, v^{\text{SFL}(i,j)}, v^{\text{RFL}(i,j)}$ and total equilibrium costs $TEC^{(i,j)}$.

2.5.4. Check termination criteria. If one of the following stopping criteria is reached, end the loop and proceed to Step 2.6:

i. $v^{\text{SFL}(i,j)} + \frac{\Delta \lambda}{\lambda^{(i,j)}} < v^{\text{RFL}(i,j)}$ (optimal reward ratio property as indicated in Corollary 2)

ii. $v^{\text{SFL}(i,j)} + \frac{\Delta \lambda}{\lambda^{(i,j)}} \geq v^{\text{RFL}(i,j)}$ and $TEC^{(i,j)} > TEC_{\text{min}}^{(i)}$ (optimal solution property as indicated in Corollary 2)

iii. $n^{(i,j)}_{e_{(j)}^*} \leq 0$ or $n^{(i,j)}_{e_{(j)}^* + l^{(i,j-1)}} \leq 0$ (efficient reward ratio requirement as in Proposition 1)

2.5.5. Update $TEC_{\text{min}}^{(i)} = TEC^{(i,j)}$, set $j = j + 1$ and go to step 2.5.2.
2.6. Update the global optimal solution: If $TEC_{\min}^{(i)} \leq TEC_{\min}$, set $TTC_{\min} = TTC^{(i)}$,

$$l^* = l^{(i)}, e^* = e^{(i)}_{\min}, \text{ and } \lambda^* = \lambda^{(i,j)}.$$ 

2.7. If $l^{(i)} \leq l_{\min} - 1$, set $i = i + 1$ and go to step 2.1. Otherwise, proceed to step 3.

Step 3. Output $\left(l^*, e^*, \lambda^* \right)$ as the optimal solution.

4.2 The lower bound on the optimal surcharge

In the case of the system optimum with the corresponding optimal solution of surcharges, reward ratio, SFI duration and location, the lower bound on the optimal surcharge is set such that incentive constraint (12) is binding, as demonstrated in Theorem 3. Therefore, the solution algorithm, in this case, does not include the stopping criteria in steps 2.4, 2.5.4.i and 2.5.4.ii which involve the incentive constraint. Alternatively, the surcharge can be set to a sufficiently large number so that the stopping criteria related to the incentive constraint are inapplicable. Once the optimal solution $\left(l^*, e^*, \lambda^* \right)$ is obtained, the lower bound on the surcharge is set such that $\Delta s = \left(v^{RFI} - v^{SFI}\right) \cdot \lambda^*.$

5. Numerical examples

5.1 System settings and characteristics

In this section, we turn to numerical examples to investigate the performance of the SRS and its effects on the commuter departure time choice pattern. We consider a typical commuting trip in Copenhagen, from Holte Station to Hellerup Station via the S-train. The S-train is an urban-suburban rail service for the Copenhagen metropolitan area in Denmark which consists of 85 stations on six lines that facilitating 350,000 passenger journeys on a busy weekday (Danish Ministry of Transport, 2011). Currently, a temporally differential fare strategy is applied in the Copenhagen metropolitan area where the peak surcharge is 4.1 DKK during the morning peak while the base fare is 16.4 DKK during off-peak. The SRS can be implemented with the current surcharge and we also propose two additional regimes for comparison as shown in Table 1. The cost unit is the Danish Krone (DKK) hereinafter.

Table 1. Existing and proposed surcharges for a typical commuting trip in Copenhagen

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destination</th>
<th>Base fare</th>
<th>SRS regimes</th>
<th>Refundable surcharge</th>
</tr>
</thead>
</table>

22
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holte Station</td>
<td>Hellerup Station</td>
<td>16.4</td>
<td>4.1 (current)</td>
<td>20</td>
</tr>
</tbody>
</table>

The transit service and characteristics of the commuting trip are listed as follows:

\[ h = 3 \text{ min}, \quad d = 12 \text{ min}, \quad N = 2000, \quad S_{\text{seat}} = 340, \quad S_{\text{std}} = 360, \quad s = 100, \quad \beta = 60 \text{ DKK/h}, \quad \gamma = 180 \text{ DKK/h}, \quad t^* = 7:30 \text{ am}, \]

where \( S_{\text{seat}} \) and \( S_{\text{std}} \) are the full seating and standing capacities for one service run respectively and the total full capacity \( S = S_{\text{seat}} + S_{\text{std}} = 700 \). Meanwhile, \( s \) is the available capacity of each service run between the origin and destination. The unit cost of the early arrival time and late arrival time are given on the basis of Fosgerau et al. (2007) and Danish Ministry of Transport (2011). The schedule delay cost function is set to be linear as in equation (2). Moreover, we consider both the cases of linear and nonlinear crowding cost functions to represent different transit system settings as follows:

**Case (a) Nonlinear\(^2\):** 
\[
g(n) = 100 \cdot \left( \frac{n}{340} \right)^2 \quad \forall n \leq 340
\]
\[
g(n) = 50 \cdot \left( \frac{n-340}{360} \right)^2 \quad \forall n > 340
\]

**Case (b) Linear:**  
\[
g(n) = 400 \cdot \frac{n}{(340 + 360)}
\]

The cost functions are set such that \( g(S) d > \gamma \cdot h \) as required in assumption 1. The two crowding cost functions with respect to the number of commuters on a service run are shown in Figure 5.

---

\(^2\) The crowding cost function is formulated with due consideration of the empirical study (Wardman and Whelan, 2011) which indicate a smooth increase in crowding cost with the number of passengers, and the theoretical study (de Palma et al., 2015) which proposed an exponential term. Here we provide a twice differentiable convex function for nonlinear crowding costs.
The original equilibrium departure time without the SRS and the corresponding time costs are given in Table 2 and the equilibrium departure time distributions are shown in Figure 6. The departure time distribution (the number of boarding passengers on each service run) is shown in Appendix 5. In case (a), the departure time period is from 6:36 am – 7:48 am with 25 service runs while in case (b), the departure time period is from 6:42 am – 7:45 am with 22 service runs. Moreover, it can be seen that the number of full service runs in case (b) is greater than that in case (a) because of the higher crowding cost at full capacity in case (a). The original endogenous departure time distribution provides the upper bound for SFI duration determination in the optimization of the SRS as stated above.

Table 2. Equilibrium characteristics in the original transit system without the SRS

<table>
<thead>
<tr>
<th>Case</th>
<th>Equilibrium cost</th>
<th>Total equilibrium costs $\times 10^4$</th>
<th>Total crowding costs $\times 10^4$</th>
<th>Total queuing costs $\times 10^4$</th>
<th>Total schedule delay costs $\times 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>125.82</td>
<td>25.16</td>
<td>18.39</td>
<td>2.16</td>
<td>4.61</td>
</tr>
<tr>
<td>b</td>
<td>119.31</td>
<td>23.86</td>
<td>15.90</td>
<td>3.45</td>
<td>4.51</td>
</tr>
</tbody>
</table>
5.2 Benchmark and algorithm validation

In the benchmark, we assume that the refundable surcharge is sufficiently large. Hence the incentive constraint is always binding, and all variables are feasible inclusive of SFI duration, location, and efficient reward ratio. Therefore, the system optimum is expected to be achieved with the proposed algorithm. The validation of the proposed algorithm is demonstrated by comparing the iterations with the entire feasible solution space considering the transit system characteristics as stated in Section 5.1 and the linear crowding cost function. The step size for updating the reward ratio is set to $\Delta \lambda = 0.001$. The service run at the preferred arrival time $t^* = 7:30$ is indexed as service run 29. Figure 7 shows the system total equilibrium costs in the three- and two-dimensional space with respect to the reward ratio and the SFI starting service run index. As shown in Figure 7(a), the bilevel problem has multiple local optimal solutions. The proposed algorithm starts with a high reward ratio and converges to the global optimal solutions in 10 iterations. The iterative results are shown in Table 3. The system optimum is achieved with 10 service runs in the SFI from the 22nd service run to the 31st service run and the optimal reward ratio is 0.514.
Table 3. Iterative results with the global optimization method

<table>
<thead>
<tr>
<th>Optimal reward ratio</th>
<th>Iteration number</th>
<th>Starting service run index</th>
<th>Ending service run index</th>
<th>Total equilibrium costs $\times 10^5$ (DKK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.950</td>
<td>1</td>
<td>29</td>
<td>29</td>
<td>2.346</td>
</tr>
<tr>
<td>0.900</td>
<td>2</td>
<td>28</td>
<td>29</td>
<td>2.313</td>
</tr>
<tr>
<td>0.850</td>
<td>3</td>
<td>27</td>
<td>29</td>
<td>2.286</td>
</tr>
<tr>
<td>0.800</td>
<td>4</td>
<td>26</td>
<td>29</td>
<td>2.265</td>
</tr>
<tr>
<td>0.750</td>
<td>5</td>
<td>26</td>
<td>30</td>
<td>2.234</td>
</tr>
<tr>
<td>0.707</td>
<td>6</td>
<td>25</td>
<td>30</td>
<td>2.221</td>
</tr>
<tr>
<td>0.663</td>
<td>7</td>
<td>24</td>
<td>30</td>
<td>2.211</td>
</tr>
<tr>
<td>0.613</td>
<td>8</td>
<td>23</td>
<td>30</td>
<td>2.206</td>
</tr>
<tr>
<td>0.557</td>
<td>9</td>
<td>23</td>
<td>31</td>
<td>2.193</td>
</tr>
<tr>
<td>0.514</td>
<td>10</td>
<td>22</td>
<td>31</td>
<td>2.185</td>
</tr>
</tbody>
</table>

5.3 Optimal solutions and performance of the SRS

In this section, we examine and compare the performance of the SRS with three regimes shown in Table 1. The global optimal solution is obtained via the algorithm developed in Section 4. As shown in Table 4 and Figure 8, the optimal SFI duration, location and reward ratio vary across SRS regimes. In both case (a) and case (b), regimes I and II achieve a suboptimal solution with smaller refundable surcharges and rewards. The equilibrium departure time distributions of both cases under the three regimes are presented in Appendix 5.

At the system optimum in case (b) with the lower bound value of surcharge, the refundable surcharge is 11.33 DKK while the reward ratio is 0.514 and the corresponding SFI is from
7:09 am – 7:36 am. This indicates that approximately half of the number of commuters travel in the RFI s on an average day. It also means that commuters on average take approximately the same number of trips in the SFI and RFIs under the system optimal regime. Moreover, for regime II with a current surcharge of 4.1 DKK, the optimal reward is 9.04 DKK (reward ratio 0.314) and the corresponding SFI is from 7:00 am – 7:39 am. Since commuters are eligible to use the reward if they have enough balance from previous surcharges, in this case one can accumulate three sets of refundable surcharges and use the 9.04 DKK reward for the next trip. Overall, commuters on average take 31.4% of trips in the RFIs, while the other 68.6% trips are in the SFI. Although in regime II the value of the reward is lower and the number of RFI trips is greater than the case of the system optimum, the corresponding optimal surcharge period (SFI duration) is shorter which indicates commuters enjoy greater time flexibility in using the rewards.

Furthermore, as shown in Figure 8, there exists a discontinuity of full capacity conditions at the time when the RFI or SFI transitions into the other interval. For instance, the last RFI service run before the preferred arrival time is at full capacity while the first service run in the SFI is not full. This is because the reward scheme reduces the fare cost in the RFI as well as the trip cost. Therefore, taking a full service run with a reward in the RFI is better than taking an non-full service run in the SFI with a surcharge. The SFI and RFI commuters have different equilibrium costs. The unused capacity in the SFI indicates the potential of the SRS to optimize and reduce the capacity for long-term infrastructure planning in the transit system.

Table 4. The optimal solution for different transit systems

<table>
<thead>
<tr>
<th>Case</th>
<th>Regime</th>
<th>Δs (DKK)</th>
<th>Δr (DKK)</th>
<th>Reward ratio</th>
<th>SFI duration</th>
<th>SFI equilibrium cost (DKK)</th>
<th>RFI equilibrium cost (DKK)</th>
<th>Average equilibrium cost (DKK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>System optima</td>
<td>13.34 (lower bound)</td>
<td>11.13</td>
<td>0.545</td>
<td>7:09 – 7:36 am (10 service runs)</td>
<td>114.67</td>
<td>131.82</td>
<td>118.77</td>
</tr>
<tr>
<td>I</td>
<td>3</td>
<td>11.22</td>
<td>0.211</td>
<td></td>
<td>6:51 – 7:42 am (18 service runs)</td>
<td>116.30</td>
<td>130.44</td>
<td>119.28</td>
</tr>
<tr>
<td>II</td>
<td>4.1</td>
<td>13.05</td>
<td>0.239</td>
<td></td>
<td>6:51 – 7:42 am (18 service runs)</td>
<td>114.67</td>
<td>131.82</td>
<td>118.77</td>
</tr>
<tr>
<td>III</td>
<td>20</td>
<td>16.70</td>
<td>0.545</td>
<td></td>
<td>7:09 – 7:36 am (10 service runs)</td>
<td>103.27</td>
<td>127.74</td>
<td>116.61</td>
</tr>
</tbody>
</table>

27
Table 5 shows the relative reduction in costs for three different regimes under the optimal solution. Implementing the SRS with a current surcharge of 4.1 DKK, the total equilibrium costs can be reduced by 5.70% and 6.94% in case (a) and case (b) respectively. The relative reduction is lower if the refundable surcharge and reward are set lower as in regime I which is consistent with the analysis in Theorem 3. At the system optimum, the total equilibrium cost reduction is 7.46% in case (a) and 8.42% in case (b).
Table 5. Performance of the optimal SRS

<table>
<thead>
<tr>
<th>Case</th>
<th>Regime</th>
<th>Total equilibrium costs reduction</th>
<th>Total crowding costs reduction</th>
<th>Total queuing costs reduction</th>
<th>Total schedule delay costs reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>I</td>
<td>5.29%</td>
<td>4.60%</td>
<td>34.05%</td>
<td>-5.35%</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>5.70%</td>
<td>4.76%</td>
<td>43.06%</td>
<td>-7.94%</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>7.46%</td>
<td>3.32%</td>
<td>69.34%</td>
<td>-4.90%</td>
</tr>
<tr>
<td></td>
<td>(System optima)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>I</td>
<td>5.97%</td>
<td>0.22%</td>
<td>41.82%</td>
<td>-1.19%</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>6.94%</td>
<td>0.16%</td>
<td>48.60%</td>
<td>-0.98%</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>8.42%</td>
<td>0.16%</td>
<td>59.56%</td>
<td>-1.54%</td>
</tr>
<tr>
<td></td>
<td>(System optima)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusions

This paper has developed a bilevel optimization model of a fare incentive strategy (surcharge-reward scheme) for public transport services and analyzed the equilibrium properties for the global optimal solution to manage mass transit peak-hour queuing and crowding. In the proposed fare incentive strategy, commuters incur surcharges during a central period. These surcharges are refundable in the shoulder periods and can be accumulated in the commuter’s personal account for a valid period. Commuters are accordingly eligible to use the monetary rewards from their account balance if traveling in shoulder periods.

A bilevel problem is developed to optimize the overall performance of the surcharge-reward scheme. The upper-level model of the surcharge-reward scheme optimization determines the duration and location of central period, the surcharge and the reward, while the lower-level model solves the commuter departure time equilibrium problem which is based on commuters’ eligibility for taking trips using rewards and the trade-offs between the schedule delay cost, crowding cost and queuing time cost. The novelty of this approach is as follows:

1) This paper included crowding effects in a general cost function within a traveler’s choice model, for which a network-like equilibrium was solved by an optimization model. This is different from the bottleneck model where equilibrium patterns can be solved analytically. As the crowding cost becomes increasingly important in
metropolitan public transport networks, it is a novel research question on how to involve demand management strategies with crowding cost in modeling and how to investigate its effects on commuters’ travel patterns. This can provide important analysis and information for congestion management in mass transit systems.

2) In previous work on bottleneck models, the system performance depended on a single variable which is the reward ratio because of the analytical structure of the bottleneck model that considers only queuing time and schedule delay. In this study, all four variables affect the system performance inclusive of starting time, the duration, the reward values and the reward ratio, which provides more flexible and realistic insights for cost reduction and real-world implementation. Moreover, the multiple control variables can also give transit operators or policymakers more control freedom.

3) The framework introduces a personal account system for congestion management in mass transit systems, which fits with the current smartcard and mobile technology.

We have analyzed the general properties of the equilibrium departure time distribution with the SRS and developed a sequential iterative algorithm to solve the bilevel problem. We found that the existence of an optimal solution depends on the scheme design. The performance of the surcharge-reward scheme varies with the refundable surcharge or reward, which also reflects the strength of the incentives for commuters to shift their departure times. The transit system achieves the optimum if the surcharge or reward is higher than a threshold value depending on the transit system settings.

The surcharge-reward scheme can be implemented in transit smartcard systems with or without existing peak surcharge strategies. The SRS can incorporate the reward mechanism, which would improve the practicality and the public’s acceptance of the current demand management strategies. The SRS can also be introduced in the monthly commuting pass as it reflects the average shifting trips of commuters within a valid period. Commuters don’t need to use the reward immediately but can use it in a valid period at their convenience which increases the flexibility of usage for commuters. Hence it can be integrated into monthly tickets or commuting tickets (maybe at a lower price). Further research can take into account the heterogeneous commuters with different degrees of flexibility and different working hours. It will be useful for transit operators or policymakers to analyze the demographics of commuters and their flexibility for better implementation. Future studies can also incorporate demand elasticity and examine the long-term impacts of the implementation of the SRS by considering transit operators’ cost, which is related to the
train capacity and the number of service runs.

Acknowledgment

The authors wish to express their thanks to the editors and two anonymous reviewers, whose useful comments have benefited the study immensely. This paper was written when Yili Tang was a guest PhD at the Technical University of Denmark. The paper was supported by a grant from the Hong Kong Research Grants Council (HKUST16211218) and also partly funded by the Innovation Fund Denmark (IFD) under File No. 4109-00005.

Appendix 1: Proof of Proposition 1

Proof: Suppose a reward ratio $\lambda$ is not an efficient reward ratio for an SFI duration $l_a$, then there exists an empty service in the SFI. Without loss of generality, suppose the empty service run is denoted by $m_a$, which is the starting service run in the SFI. The SFI commuters thus achieve a free-flow equilibrium occupying service run from $m_{a+1}$ onwards.

Since the service runs from $m_1$ to $m_{a-1}$ in the first period of the RFI are non-empty, the equilibrium generalized trip cost in the SFI can be reduced if commuters take one service run closer to the preferred arrival time while keeping the same equilibrium distribution, namely, the first period of the RFI becomes $m_2$ through $m_a$. In the meantime, the equilibrium generalized trip cost in the SFI is unchanged because the free-flow duration is unchanged. Hence the total equilibrium trip cost will be reduced. In this case, the reward ratio $\lambda$ is an efficient reward ratio for the SFI duration $l_{a-1}$ with the SFI location $m_{a+1}$, while it is a non-efficient reward ratio for the SFI duration $l_a$ with the SFI location $m_a$. Therefore, an efficient reward ratio introduces a lower total equilibrium cost than a non-efficient reward ratio such that an optimal reward ratio must be an efficient reward ratio. This completes the proof. □

Appendix 2: Proof of Theorem 1

Proof: The equilibrium costs in the SFI are first analyzed. Since commuters at the two ends of the SFI do not incur the queuing time cost, the equilibrium generalized trip cost is determined by the crowding cost and the schedule delay cost of the starting service run $e_i$ or the ending service run $e_i + l - 1$, whichever is greater. This can be expressed by

$$v_i^{SFI} = \max \left\{ \delta(t_{e_i}) + g(n_{e_i}^{SFI})d, \delta(t_{e_i+l-1}) + g(n_{e_i+l-1}^{SFI})d \right\}$$
where $n_{e_i}^{SFI}$ is the number of SFI commuters on the starting service run $e_i$. Meanwhile, without loss of generality, suppose the equilibrium generalized trip cost of the SFI commuters is determined by the schedule delay and crowding costs of the starting service run such that

$$v_i^{SFI} = \delta(t_{e_i}) + g(n_{e_i}^{SFI})d = \delta(t_{e_i+1}) + g(n_{e_i+1}^{SFI})d + q(t_{e_i+1})$$

where the first term is the equilibrium trip costs of service run $e_i$ and the second term is the equilibrium trip costs of the last service run $e_i + l - 1$ under the SRS design with starting service run $e_i$.

There exist two cases of schedule delay cost for starting service run $e_i$ as shown in Figure 9, such that

1. $\max \{\delta(t_{e_i}), \delta(t_{e_i+1})\} = \delta(t_{e_i})$, $\max \{\delta(t_{e_i+1}), \delta(t_{e_i+2})\} = \delta(t_{e_i+1})$, and $\delta(t_{e_i}) > \delta(t_{e_i+1})$

2. $\max \{\delta(t_{e_i}), \delta(t_{e_i+1})\} = \delta(t_{e_i})$, $\max \{\delta(t_{e_i+1}), \delta(t_{e_i+2})\} = \delta(t_{e_i+1})$, and $\delta(t_{e_i}) < \delta(t_{e_i+1})$

For the first case, if all commuters on service run $e_i$ shift to service run $e_i + l$, they incur lower trip costs with less schedule delay and the same crowding costs. In this non-equilibrium condition, the starting service run becomes $e_i + 1$ and the trip costs on service run $e_i + 1$ are equal to the previous equilibrium trip costs since the number of boarding passengers does not change but the trip costs are higher than the trip costs on the service run.
run $e_i + 1$. To achieve the equilibrium, the number of commuters on service runs will be balanced such that the trip costs for all service runs are equal. Hence, overall, the equilibrium trip costs with starting service run $e_i + 1$ are lower than those with starting service run $e_i$. Thus setting the SFI location to $e_i + 1$ is always better than setting it to $e_i$ in this case.

Similarly, for the second condition, commuters on service run $e_i + 1$ can shift to $e_i$ and the equilibrium trip costs will be reduced. Thus setting the SFI location to $e_i$ is always better than setting it to $e_i + 1$ in this case.

Therefore, the optimal starting service run for SFI commuters is set to minimize the schedule delay cost such that

$$e_i = \arg \min \left\{ \max \left\{ \delta(t_m), \delta(t_{m+1}) \right\} \right\}.$$ 

On the other hand, for RFI commuters, without loss of generality, equilibrium cost can be expressed as the schedule delay cost and the crowding cost of the first service run such that

$$v_{t_i}^{RFI} = \delta(t_i) + g(n_{t_i}^{RFI})d.$$ 

Denote by $\Delta \delta = \delta(t_i) - \delta(t_{e_i})$ the schedule delay cost difference between the first service run of the RFI and the SFI location $e_i$. The above equation can be reformulated as

$$v_{t_i}^{RFI} = \delta(t_i) + g(n_{t_i}^{RFI})d = \delta(t_i) - \delta(t_{e_i}) + \delta(t_{e_i}) + g(n_{t_i}^{RFI})d
\begin{align*}
\quad = \Delta \delta + \delta(t_{e_i}) + g(n_{t_i}^{RFI})d.
\end{align*}$$

In the above equations, although both $\delta(t_i)$ and $\delta(t_{e_i})$ are affected by the value of $e_i$, the schedule delay cost difference $\Delta \delta$ remains unchanged, by the fact that the value of $e_i$ does not change the relative difference in the schedule delay cost between the two consecutive service runs. As a result, the number of RFI commuters on each service run (departure time distribution) will not change based on the trade-offs between schedule delay, queuing time and crowding. Hence the number of non-empty service runs in the RFI is the same and the schedule delay cost difference $\Delta \delta$ is unchanged, which is determined by the departure time distribution. In this case, minimizing the SFI equilibrium cost is the same as minimizing the RFI equilibrium cost.
In summary, the optimal SFI location should be set to minimize the schedule delay cost of the SFI. Otherwise, commuters can always reduce their equilibrium cost by shifting to the other service run with the same number of commuters. This completes the proof. □

Appendix 3: Proof of Theorem 2

Proof: For the total equilibrium costs in the upper-level model, we have

$$\frac{\partial TEC}{\partial \lambda_i} = \left( v(n^{RFI}) - v(n^{SFI}) + \frac{\partial v(n^{SFI})}{\partial \lambda_i} (1 - \lambda_i) + \frac{\partial v(n^{RFI})}{\partial \lambda_i} \lambda_i \right) N. \quad (29)$$

In the equilibrium departure time choice model, the equilibrium generalized trip cost of the SFI or the RFI depends on the schedule delay cost and crowding cost of non-full service runs. Hence, we use the trip cost of the first service runs in the SFI and RFI to represent their equilibrium generalized trip costs and, without loss of generality, we assume that the two service runs remain non-full under all efficient reward ratios. Therefore, we have

$$\frac{\partial v(n^{SFI})}{\partial \lambda_i} = \frac{\partial v(n^{SFI})}{\partial n_{e_i}^{SFI}} \frac{\partial n_{e_i}^{SFI}}{\partial \lambda_i} = \frac{\partial g(n_{e_i}^{SFI})d + \delta^{SFI}(t_i)}{\partial \lambda_i} \frac{\partial n_{e_i}^{SFI}}{\partial \lambda_i} = g'(n_{e_i}^{SFI})d \cdot \frac{\partial n_{e_i}^{SFI}}{\partial \lambda_i} \quad (30)$$

$$\frac{\partial v(n^{RFI})}{\partial \lambda_i} = \frac{\partial v(n^{RFI})}{\partial n_{l_i}^{RFI}} \frac{\partial n_{l_i}^{RFI}}{\partial \lambda_i} = \frac{\partial g(n_{l_i}^{RFI})d + \delta^{SFI}(t_i)}{\partial \lambda_i} \frac{\partial n_{l_i}^{RFI}}{\partial \lambda_i} = g'(n_{l_i}^{RFI})d \cdot \frac{\partial n_{l_i}^{RFI}}{\partial \lambda_i}. \quad (31)$$

Since $g'(\cdot) > 0$ by the definition, and the number of SFI commuters decreases with the reward ratio while the number of RFI commuters increases with the reward ratio. We thus have

$$\frac{\partial n_{e_i}^{SFI}}{\partial \lambda_i} < 0 \text{ and } \frac{\partial n_{l_i}^{SFI}}{\partial \lambda_i} > 0. \quad (32)$$

Hence

$$\frac{\partial v(n^{SFI})}{\partial \lambda_i} < 0 \text{ and } \frac{\partial v(n^{RFI})}{\partial \lambda_i} > 0. \quad (33)$$

It is clear that if $\lambda_i = 0$, then $v(n^{RFI}) = 0$ and $\frac{\partial TEC}{\partial \lambda_i} < 0$ and if $\lambda_i = 1$, then $v(n^{SFI}) = 0$. 

34
and \( \frac{\partial TEC}{\partial \lambda_i} > 0 \).

With the second derivative of the equation (29), we obtain

\[
\frac{\partial^2 TEC}{\partial \lambda_i^2} = \left( \frac{\partial^2 v(n^{RFI})}{\partial \lambda_i^2} + \frac{\partial^2 v(n^{SFI})}{\partial \lambda_i^2} \cdot (1-\lambda_i) \right) - \left( \frac{\partial v(n^{SFI})}{\partial \lambda_i} + \frac{\partial v(n^{RFI})}{\partial \lambda_i} + \frac{\partial^2 v(n^{RFI})}{\partial \lambda_i^2} \cdot \lambda_i \right) \cdot N \quad (34)
\]

Note that

\[
\frac{\partial^2 v(n^{SFI})}{\partial \lambda_i^2} = \left[ g'(n^{SFI}_i) d + g^*(n^{SFI}_i) \right] d \cdot \frac{\partial^2 v(n^{SFI})}{\partial \lambda_i^2} \quad \text{and} \quad \frac{\partial^2 v(n^{RFI})}{\partial \lambda_i^2} = \left[ g'(n^{RFI}_i) d + g^*(n^{RFI}_i) \right] d \cdot \frac{\partial^2 v(n^{RFI})}{\partial \lambda_i^2} \quad (35)
\]

The second derivative of the equilibrium cost in the equation (35) represents the trends of \( \frac{\partial n^{SFI}_i}{\partial \lambda_i} \) and \( \frac{\partial n^{RFI}_i}{\partial \lambda_i} \) with respect to the reward ratio. As the reward ratio increases, \( \frac{\partial n^{SFI}_i}{\partial \lambda_i} \) increases. This can be proved that at \( \lambda_{mn} \to 0 \), where all passengers traveling within the SFI, increasing the reward ratio first results in decreasing in the number of passengers in a single non-full service run. Without loss of generality, suppose the non-full service run is the service run \( e_i \). Hence \( \frac{\partial n^{SFI}_i}{\partial \lambda_i} = -\Delta \lambda_i N/\Delta \lambda = -N \). As the reward ratio increases, the number of non-full service runs in the SFI increases because it is always better for the commuters with a full service run to shift to the non-full service run given the assumption 1 that \( g(S) d > \gamma \cdot h \). In this case, the number of SFI commuters decreases less and becomes \( \frac{\partial n^{SFI}_i}{\partial \lambda_i} > -N \), and thus \( \frac{\partial^2 n^{SFI}_i}{\partial \lambda_i^2} > 0 \).

Similarly, for RFI commuters, as the reward ratio increases, the duration increases and the number of non-full service runs increases. Hence the number of RFI commuters increases less on the service run \( m_i \), namely, \( \frac{\partial^2 n^{RFI}_i}{\partial \lambda_i^2} < 0 \) and the absolute value of \( \frac{\partial^2 n^{RFI}_i}{\partial \lambda_i^2} \) is less than \( \frac{\partial n^{RFI}_i}{\partial \lambda_i} \) on the same service run. Therefore, we have

\[
\frac{\partial^2 v(n^{SFI})}{\partial \lambda_i^2} > 0 \quad \text{and} \quad \frac{\partial^2 v(n^{RFI})}{\partial \lambda_i^2} < 0, \quad \left| \frac{\partial^2 n^{RFI}_i}{\partial \lambda_i^2} \right| < \left| \frac{\partial n^{RFI}_i}{\partial \lambda_i} \right| . \quad (36)
\]

As a result, we can obtain

\[
\frac{\partial^2 TEC}{\partial \lambda_i^2} > 0 . \quad (37)
\]
With the properties of the equation (37) and (29), the total equilibrium costs will decrease first and then increase with the reward ratio if the efficient reward ratio set is sufficiently large within $(0,1)$. Therefore, we can conclude that there exists a unique optimal reward ratio $\lambda^*_i$ such that the total equilibrium costs decrease with the reward ratio at $\lambda \in (0, \lambda^*_i]$ and increase afterward at $\lambda \in [\lambda^*_i, 1)$. This completes the proof.

**Appendix 4. Proof of Theorem 3**

**Proof:** Since $v\left(n^{\text{RFI}}\right)$ is increasing while $v\left(n^{\text{SFI}}\right)$ is decreasing with the reward ratio, the equilibrium cost difference, $v\left(n^{\text{SFI}}\right) - v\left(n^{\text{RFI}}\right)$, is a monotonically decreasing function with respect to the reward ratio. Meanwhile, the fare difference $\Delta s/\lambda_i$ is also a monotonically decreasing function with respect to the reward ratio.

Hence, the left-hand side of the incentive constraint (16), namely $v\left(n^{\text{SFI}}\right) + \Delta s/\lambda_i - v\left(n^{\text{RFI}}\right)$, is monotonically decreasing with the reward ratio. Since $\lambda_i \in \Omega_i$, denote $\lambda_{\text{min}} = \min \Omega_i$ and $\lambda_{\text{max}} = \max \Omega_i$. Then there exists a surcharge value such that $v\left(n^{\text{SFI}}\right) + \Delta s/\lambda_{\text{max}} - v\left(n^{\text{RFI}}\right) = 0$, and if $\Delta s \geq \Delta s_i$, we have

$$\forall \lambda_i \in \Omega_i, \quad \left(v\left(n^{\text{SFI}}\right) + \Delta s/\lambda_i - v\left(n^{\text{RFI}}\right)\right) \geq 0$$

namely, the incentive constraint is always satisfied. Any efficient reward ratio is feasible. On the other hand, if $\Delta s < \Delta s_i$, we have

$$\exists \lambda_i \in \Omega_i, \quad \left(v\left(n^{\text{SFI}}\right) + \Delta s/\lambda_i - v\left(n^{\text{RFI}}\right)\right) < 0.$$  

Then if $v\left(n^{\text{SFI}}\right) + \Delta s_i/\lambda_{\text{min}} - v\left(n^{\text{RFI}}\right) > 0$, there exists a reward ratio such that

$$v\left(n^{\text{SFI}}\right) + \Delta s_i/\lambda_{\text{max}} - v\left(n^{\text{RFI}}\right) = 0$$

and the feasible reward ratios are $\lambda \in [\lambda_{\text{min}}, \lambda^*_i]$. Otherwise, there is no feasible solution within the efficient reward ratio set. This proves Theorem 3. □
Appendix 5. Departure time distributions

Figure 10. Departure time distribution for each service run under different regimes

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Case (a)</th>
<th>Case (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Origin</td>
<td>regime 1</td>
</tr>
<tr>
<td>6:30 AM</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>6:33 AM</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>6:36 AM</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>6:39 AM</td>
<td>26</td>
<td>46</td>
</tr>
<tr>
<td>6:42 AM</td>
<td>41</td>
<td>57</td>
</tr>
<tr>
<td>6:45 AM</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>6:48 AM</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>6:51 AM</td>
<td>76</td>
<td>45</td>
</tr>
<tr>
<td>6:54 AM</td>
<td>86</td>
<td>56</td>
</tr>
<tr>
<td>6:57 AM</td>
<td>94</td>
<td>67</td>
</tr>
<tr>
<td>7:00 AM</td>
<td>100</td>
<td>78</td>
</tr>
<tr>
<td>7:03 AM</td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>7:06 AM</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>7:09 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:12 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:15 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:18 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:21 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:24 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:27 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:30 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:33 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:36 AM</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7:39 AM</td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>7:42 AM</td>
<td>86</td>
<td>56</td>
</tr>
<tr>
<td>7:45 AM</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>7:48 AM</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>7:51 AM</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

References


