博士論文

Dynamics of congested urban rail transit: A macroscopic model with demand and supply interaction

（需要供給統合化マクロモデルを用いた都市鉄道の混雑ダナミクス解析 ）

張 嘉華

Jiahua Zhang

Department of Civil Engineering, School of Engineering,

The University of Tokyo

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2020

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Jiahua Zhang
August 2020

Acknowledgements

Upon the completion of this dissertation, I would like to express my sincere gratitude to everyone who helped me on this study. Firstly, I would like to acknowledge Professor Oguchi for guiding me to tackle this research topic and giving me lots of constructive advice and comments during each of our meetings. Also, I would like to thank him for his kindly help to support my life in Japan in this five years. Secondly, I would like to thank Professor Wada for attentively following my research progress from the beginning to the end of this dissertation. He really provided me with so many valuable suggestions during our regular research meetings. Without his guidance, this dissertation cannot be successfully accomplished.

Then, I would also like to acknowledge the committee members of my dissertation. Professor Kato raised several important questions on the scope and assumptions of this study, and I would like to thank him for these questions that let me correctly position this study in a broader view. Professor Iryo also helped me so much by encouraging me to strengthen the theoretical depth of this study (especially regarding the first-best pricing which proves to be an important outcome from this research). Professor Fukuda gave me a lot of helpful advice on train fundamental diagram and recommended me to clearly summarize the theoretical new points of this study. Again, I truly appreciate all the efforts that each committee member has made on my research during this three years.

Next, I would like to thank the members of Oguchi laboratory. Research associate Toriumi provided me with many beneficial ideas and suggestions during the lab seminars and several rehearsals of my presentations. Her questions and comments broadened my viewpoint of this study. Dr. Satsukawa (now assistant professor in Tohoku University) also helped me a lot on this research during our daily discussions in the lab. As the other Chinese student in Oguchi lab, Han actively discussed with me and inspired me a lot on this study. Besides, I would also like to thank all the other members in Oguchi laboratory for creating such a good environment for research and discussion.

In addition, I would like to thank Mike Barry and Brian Card for providing the Boston subway system data on Github, see Barry and Card (2014).

Finally, I would like to thank my parents for supporting me from all aspects in this three years.

Abstract

Urban rail transit, with its high capacity and punctuality, serves as the major solution to commuters' travel demand during rush hours in most metropolises. However, travel experience of commuting by rail transit is frequently deteriorated by severe congestion and chronic delays. Moreover, passenger crowding both in-vehicle and on-platform and delays of trains can easily develop into a vicious circle especially in a high-frequency operated rail transit system. Therefore, a new solution to improve the level of service of congested rail transit is urgently needed.

On one hand, most of the efforts so far have been made to increase the supply of system, but these efforts are gradually reaching the ceiling due to financial and constructional limitations. On the other hand, practices of demand management for rail transit are still few and not successful to alleviate the congestion and delay issues. Most existing studies on rail transit optimization and management either treated passenger demand distribution as given information or ignored the influence of passenger demand concentration on railway operation. In order to derive insights into policy implications of management strategies, an appropriate model considering the interaction between demand and supply is indispensable for a congested rail transit.

The objectives of this study is threefold. First, to comprehensively understand influence of passenger demand concentration on rail transit operation by investigating a train fundamental diagram model and its variants using empirical data. Second, to macroscopically develop a dynamic model capable of describing equilibrium distribution of passenger arrivals in a congested rail transit system based on information of timetable and passenger preferences. Third, to derive insights into the design of probable management strategies and their combinations from a macroscopic view by applying the proposed model.

Through the analysis of Boston subway red line passenger flow and operation data, the influence of passenger demand concentration on rail transit operation is confirmed to be significant. By calibrating and evaluating the three variants (i.e., assumption of train dwell time constant, monotonically increasing or first constant then monotonically increasing with boarding passenger number) of rail transit fundamental diagram (train-FD) models using empirical data, the original train-FD model assuming monotonically increasing relation

between trains' dwell time and waiting passenger number is found to be adequate to describe influence of passenger demand concentration from a macroscopic point of view.

The equilibrium model to describe passenger arrivals in a congested rail transit system is developed by employing the train-FD model with a novel particle-based dynamic approach. The conceptual structure of the proposed equilibrium model is a two-level model different from the existing one-level one for road traffic. The particle-based approach is found to be tractable for deriving insights into management strategies. A microscopic simulation program is also developed to verify the results from the equilibrium model. Outputs from the simulation and the model are found to be consistent to a great extent.

By applying the proposed model, three issues are investigated. First, optimum peak/off-peak timetable which minimizes the sum of travel time cost (TTC) and schedule delay cost (SDC) is explored. An optimum peak/off-peak timetable in free-flow regime of train-FD is proposed and numerically compared with different timetable patterns. Second, the user equilibrium under first-best pricing scheme is theoretically derived and numerically compared with the user equilibrium without pricing. Third, because the first-best pricing scheme requires time varying pricing which is difficult to realize in real world, an appropriate single-step coarse pricing scheme is specified and its effectiveness is numerically evaluated. The coarse pricing scheme is compared by indicators such as the change of sum of TTC and SDC, and change of in-vehicle crowding.

If all dispatched trains operate in free-flow regime of train-FD, peak/off-peak timetable is found to be optimized when average train flow before and after desired departure time are the same, and the flow of train with maximum passenger concentration reaches the critical state of train-FD. It is proved that the user equilibrium under first-best pricing scheme can significantly reduce the sum of TTC and SDC by flattening the change rate of riding passenger number among dispatched trains compared with the user equilibrium without pricing. Riding passenger number difference between successive trains under first-best pricing is exactly half of that in the equilibrium without pricing. The effectiveness of single-step coarse pricing is found to strongly depend on the temporal design of the scheme. The appropriately designed coarse pricing scheme without breaking the equilibrium shows an acceptable reduction of sum of TTC and SDC compared to the first-best one. Finally, a complex pricing scheme by jointly giving reward and imposing surcharge is examined, and it is found that a Pareto-improvement situation may occur depending on the timetable pattern with significant reduction of sum of TTC and SDC.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

1.1.1 Severe congestion and chronic delays of urban rail transit

Urban rail transit, with its high capacity and punctuality, serves as the major solution to commuters' travel demand during rush hours in most metropolises (Vuchic, 2017). However, travel experience of commuting by rail transit is frequently deteriorated by severe congestion and chronic delays. In many metropolises, congestion and delay of rail transit have brought about tremendous psychological stress to commuters and considerable economic loss to the whole society. For example, Kariyazaki et al. (2015) estimated that the social cost due to the delay of trains in Japan exceeded 1.8 billion dollars per year. Xu et al. (2019) reported that around one third of subway stations in Beijing adopted ordinary passenger flow control to avoid extreme congestion on the platform. According to the New York Times, only 58% weekday trains arrived on time in January of 2018, mainly due to equipment failure and persons on the track (Hu, 2018).

One tricky issue for high-frequency operated urban rail transit is that passenger congestion (both on-platform and in-vehicle) and delay could easily develop into a vicious circle. More specifically, during rush hours, the dwell time of a train at one station is highly probable to be extended due to an unexpected longer time for passenger exchange. As a result, the following train is forced to stop ahead of the station to keep the safety clearance (this phenomenon is generally called as "knock-on delay" on the track, see Carey and Kwieciński (1994)). The extra stop time between stations disturb the regular operation headway and further exacerbate the congestion on the platform. In fact, most short delays within 5 minutes in Japan are caused by this viscous circle, and this kind of chronic delay due to passenger congestion happens almost everyday in many railway lines (Railway Policy Planning, 2020). Therefore,

any model aiming at preventing delay or relieving congestion for urban rail transit should take into consideration of this mechanism.

1.1.2 Necessity of flattening demand for urban rail transit

In order to provide a reliable and high-level rail transit service, most of efforts so far were made on physically increasing the supply of railway system. For example, to dispatch more trains during the peak hours, to add more cars for one train, or even to build multiple tracks for one direction. These efforts were undoubtedly effective, but were severely limited in many cases.

On the other hand, strategies to temporally flatten the travel demand, which has long been discussed and practiced for road traffic, was seldom implemented for urban rail transit. Some trials by differentiating peak/off-peak fares of metro system were carried out in Washington and London, although not only at the purpose of alleviating congestion. Melbourne introduced an "Early Bird" scheme to allow commuters to enjoy free travel by rail before 7 a.m (Currie, 2010). They reported that 23% of passengers have shifted the time of travel. A similar experiment to shift peak-hour ride was also done in Netherlands by rewarding passengers travelling during off-peak (Peer et al., 2016). Nevertheless, strategies to flatten the demand, especially when it comes to fare, are controversial since it relates to other issues like mobility and equity. Commuters in Japan's metropolitan areas have long suffered from jamming in the morning trains, but the strategies to flatten the demand was hardly implemented, the congestion so far was gradually relieved mainly due to increase of supply. Meanwhile, newly operated rail transit systems in metropolises of developing countries like China, India or Philippines have encountered or will surely encounter huge commuting demand over available supply. Since increasing the supply is costly and takes long time, to temporally flatten the commuting demand is probably the only choice to maintain an acceptable congestion level.

In fact, studies on flattening commuting travel demand (e.g., fare scheme design, flexible work time) for urban rail transit were relatively few compared to road traffic. Most studies just simply employed the framework for road traffic without considering passenger congestion influence. As the first step, it is necessary to deeply understand the characteristics of passenger demand for rail transit system.

1.1.3 Lack of appropriate model considering interaction between demand and supply

With regard to existing models for rail transit, most of them were either microscopic simulation or mathematical approach specifically for timetable optimization. A critical issue

for these models is that their discussions are based on a given passenger demand pattern, either static or dynamic. However, temporal distribution of demand is inevitably affected by supply condition. For example, if the seats for a portion of morning trains can be reserved, it can be expected that some commuters would depart from their home earlier or later to utilize these train services. In real-world situations, some travellers were also found to wait outside the station until they could ride the trains with off-peak fares (Winward, 2016). The change of demand distribution according to supply adjustment is self-evident since most commuters are rational to make a travel decision. So far, models considering the interaction between demand distribution and supply of rail transit were still very rare. Yet, the ability to properly describe this interaction would surely be crucial for any model to design or evaluate strategies of flattening the demand.

In addition, compared to most microscopic simulation models that tried to include practical settings of rail transit operation as mush as possible, a macroscopic model, if well-developed, is much more tractable and interpretable. Research works based on macroscopic modelling are again receiving a growing concern in recent years (see e.g., macroscopic fundamental diagram for road network by Daganzo 2007). Similar framework can also be considered for urban rail transit system.

1.2 Objective

This study aims to propose a macroscopic model capable of describing dynamics of passenger arrivals and rail transit operation to derive insights into management strategies from both demand and supply sides. More specifically, the objectives can be divided into three levels:

- To comprehensively understand passenger congestion influence on rail transit operation by investigating a train fundamental diagram model and its variants using empirical data.
- To build a dynamic model that can estimate equilibrium distribution of passenger arrivals at a station by inversely using train fundamental diagram.
- To derive insights into effectiveness of typical management strategies and their combinations by comparing indicators such as total travel cost, congestion level or duration of peak hours.

1.3 Composition of dissertation

The remaining chapters of this dissertation are structured as follows. Chapter 2 comprehensively reviews existing studies on modelling, optimization and management of rail transit system. The review starts from the basics of train operation, followed by representative macroscopic modelling frameworks of traffic dynamics. Further, researches on optimization, control and management strategies of rail transit are summarized from both supply and demand sides.

Chapter 3 introduces the train fundamental diagram (train-FD), a tractable model which can describe steady state macroscopic performance of rail transit considering passenger congestion influence in a simple manner. Based on different assumptions of dwell time, variants of train-FD are derived and compared. Then, by employing empirical data from the Boston subway system, train-FDs are calibrated and evaluated.

Chapter 4 expands the steady state train-FD into a dynamic model which can describe the temporal evolution of system performance with time-dependent passenger demand. Two approaches, namely the fluid-based one and the particle-based one are formulated with consideration of their advantages and limitations. Then, the dynamic models are verified by a microscopic simulation.

Chapter 5 derives the equilibrium distribution of passenger arrivals by solving the inverse problem of dynamic model introduced in chapter 4. The solution method and existence conditions of equilibrium are presented. The characteristics of equilibrium are also discussed by a series of numerical experiments.

Chapter 6 focuses on application of the proposed model. The optimization problem of a conventional peak/off-peak timetable pattern is investigated. User equilibrium under first-best pricing is theoretically derived and numerically evaluated. A single-step coarse pricing is also specified and numerically evaluated.

Chapter 7 summarizes the findings of dissertation and discusses the possible future works of this study.

Chapter 2

Literature review

The second chapter aims to summarize the existing studies which focused on either supply side (e.g., improving the service quality by optimizing timetable) or demand side (e.g., designing fare scheme to flatten the travel demand) of rail transit system to reveal the research gap that will be filled by this study.

Except for unexpected incidents (e.g., equipment failure, person on the tracks or extreme weather), the delay of trains are closely related to surging passenger demand during rush hours (Palmqvist et al., 2020). Especially in a high-frequency operated urban rail transit system, extra dwell time of one train can easily lead to the propagation of delay along the railway line. Furthermore, this kind of delay exacerbated the on-platform congestion and in-vehicle crowding, and finally developed into a vicious circle (Tirachini et al., 2013). To tackle this problem, strategies have long been considered from two aspects: one is to precisely predict the travel demand and then optimize the timetable. As a result, all the dispatched trains could operate punctually as the optimized timetable; Another one is to affect individual's time choice behavior by some incentives, which is expected to temporally redistribute the travel demand. As a result, the capacity could always meet the demand, and delay could hardly happen. In fact, strategies from both aspects have their advantages and limitations. The existing studies on both aspects would be reviewed in this chapter.

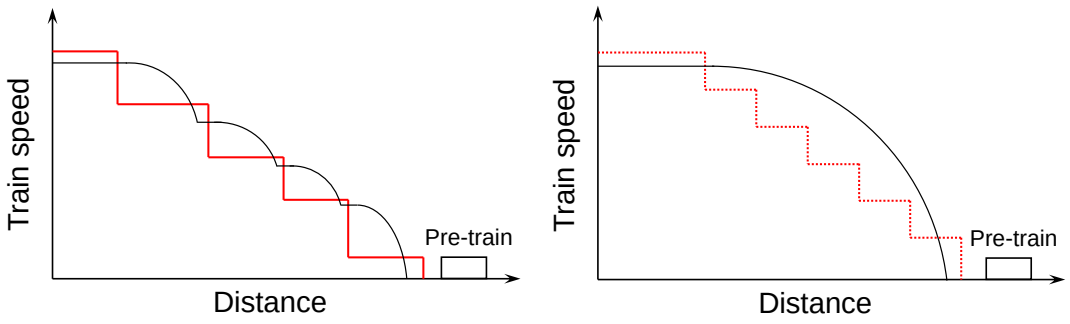
Firstly, the basic rules of train operation is introduced (i.e., cruising and dwelling behavior, prevailing control strategy). Secondly, studies on timetable optimization and dynamic control of trains are reviewed. Thirdly, studies on individual's departure time choice problem, together with studies on designing pricing scheme to flatten the travel demand, are widely reviewed. Although most of these studies are targeted at road traffic, their theoretical framework and implications serve as the foundation of this study. Finally, based on the reviewed works, the position of this study is briefly described.

2.1 Basics of train operation

Railway system, especially the urban rail transit, has its unique characteristics that differs from other transportation system. The basic operational characteristics of urban rail transit has to be first made clear before moving to the discussion of modelling and management. In an urban transit system, one train generally has three action patterns during the operation. The first one is cruising between adjacent stations, the second one is dwelling at stations, and the third one is shunting to turn around for the next run. We skip the discussion of shunting since it mainly serves as a constraint of dispatch frequency and does not have interaction with passengers.

2.1.1 Safely cruising on the track

Several decades ago, a cruising train kept safety spacing with the preceding train relying on confirming signal light along the track by the train driver. Thanks to the development of signal and control system, trains nowadays can run much faster with higher dispatch frequency than before. Nevertheless, the basic idea to keep a safety distance based on cruising speed never changed. Take the prevailing automatic train control (ATC) system (Takashige, 1999) as an example, the relation between distance from preceding train and speed can be illustrated as Fig. 2.1.



(a) Multi-break ATC system (b) One-break ATC system

Fig. 2.1 An illustration of prevailing ATC system.

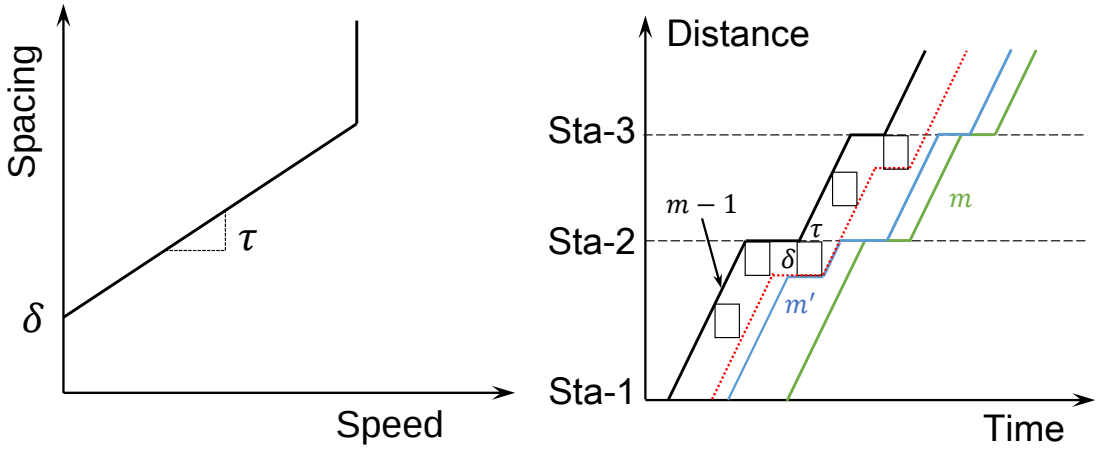
The red lines indicated the speed limit for a block (i.e., a section of track that can only be occupied by one train, either fixed or moving) and the black curves showed an example of speed reduction pattern of a train. The multi-break ATC shown in Fig. 2.1a is generally designed for fixed-block track circuit. Since the multi-break is not comfortable for passengers on the train, many of these old systems have been gradually replaced by the new one shown in Fig. 2.1b. The one-break ATC system (or digital ATC called by some railway

companies) judges the real-time position of preceding train and sends the information by wayside equipment to the control center and following train. In this way, the breaking can be optimized while keeping ride comfort at a high level. In recent years, more advanced automatic train operation (ATO) system has been put into use in many urban rail lines (Yin et al., 2017a), by which the driver is not necessary or is only responsible for handling the unexpected incidents.

Since the objective of this study is not to improve the operation system, we prefer a simple way to describe the cruising behavior of a train. In fact, Newell (2002) proposed a simplified car-following model which is also appropriate for railway system. This model is a special case of the well-known road traffic flow model, the Lighthill–Whitham–Richards (LWR) model (Lighthill and Whitham, 1955; Richards, 1956). The basic idea is that a train either travels at its desired speed or follows the preceding train while keeping the safety clearance. This can be described by the following formula:

$$x_{\mathrm{m}}(t+\Delta t) = \min\{x_{\mathrm{m}}(t) + v_{\mathrm{f}}\Delta t,\ x_{\mathrm{m}-1}(t+\Delta t-\tau) - \delta\}, \tag{2.1}$$

where $m-1$ refers to the preceding train of train m. $x_{\mathrm{m}}(t)$ is the position of train m at time t, v_{f} is the desired cruising speed, τ is the safety headway and δ is the minimum spacing. For a better understanding of τ and δ, the relation between spacing and cruising speed is drawn in Fig. 2.2a. In addition, an example of trajectories of two succeeding trains considering safety clearance is shown in Fig. 2.2b.



(a) Relation between spacing and speed (b) Trajectories of two succeeding trains

Fig. 2.2 An illustration of simplified train-following model.

One can easily find that Fig. 2.2a is equivalent to an linear approximation of ATC system shown in Fig. 2.1 (convert distance to spacing and exchange the two axes). Regarding

Fig. 2.2b, the red dotted line indicated the safety clearance of train $m-1$, which means the following train cannot enter the left side of this line. For example, the trajectory of train m (green line) always lies at the right side of red dotted line so its position is always obtained from the first term of Eq. (2.1). On the other hand, train m' (blue line) approached the safety clearance before station 2, so it has to stop to keep the safety clearance.

Although there are many high-order car-following models that might be more accurate to capture the realistic trajectory of a train, the parsimonious Newell's model are quite parsimonious and tractable to be employed in traffic flow models.

2.1.2 Dwell time

Unlike private car dominated road traffic, an unique characteristic of rail transit system is that all the operated trains have to stop at stations. The dwell time at stations is of crucial importance to the timetable planning, operational control of railway system. However, dwell time is one of the most susceptible components of railway operation, mostly due to the fluctuating number of alighting and boarding passengers. In general, dwell time at stations are determined by the following factors:

- Time required for door opening and closing.
- Time required for passengers alighting and boarding. This time is actually affected by many sub-factors like in-vehicle crowding or on-platform congestion level (Lam et al., 1999; Sun et al., 2014), re-open and re-close of doors due to passengers rushing into the train.
- Timetable. In most countries, trains are not allowed to depart ahead of schedule (Jiang et al., 2015).
- Stop signal ahead. Sometimes due to the on-track congestion or accident, the train dwelling at the station is not permitted to depart.

For a better understanding, Fig. 2.3 showed an example of dwell time components when a train arrived slightly later than the timetable. A/B is the abbreviation of alighting and boarding.

Researches on dwell time can roughly be classified into two categories: microscopic simulation models and data-driven statistical models. The former ones originated from applications of pedestrian behavior modelling, and usually aimed to reproduce or evaluate the detailed process of alighting and boarding behavior, but required complex setting of situation and behavior rule of passengers. For example, Zhang et al. (2008) proposed a cellular

Fig. 2.3 Example of dwell time components.

automata-based simulation model capable of reproducing the cooperation and negotiation of passengers at a microscopic level. As a by-product, dwell time in Beijing metro network was estimated. Yamamura et al. (2013) developed a multi-agent simulation model to estimate dwell time considering the in-vehicle crowding and on-platform congestion at the same time. They also compared the efficiency of possible measures to reduce the dwell time like using trains with more and wider doors or without seats.

The statistical models on dwell time have been conducted from several decades ago (Levinson, 1983; Weidmann, 1994). The common idea was to use simple or multivariate regressions to relate dwell time to factors mentioned above. Most studies define dwell time as the time for passengers to alight and board plus the time to open and close doors like the following formula:

$$T = a \times N + b, \tag{2.2}$$

where a is the average alighting or boarding time per passenger, N is the total number of alighting and boarding passengers, and b is the time for opening and closing the doors. The automated fare collection (AFC) system widely introduced in the 2000s made it possible to acquire huge amount of public transport data, by which influence of vehicle characteristics (e.g., door width, floor height) and passenger related factors (e.g., age, friction effects) were investigated (Dueker et al., 2004; Fernández et al., 2010; Sun et al., 2014; Tirachini, 2013). For example, Sun et al. (2014) used smart card data in Singapore to conduct a deeper analysis of bus boarding and alighting process, and they found a critical occupancy beyond which the process was slowed down due to the friction among passengers. Instead of deterministic description of dwell time, some studies added stochastic component into the dwell time since its random variation could not be neglected (Cornet et al., 2019; Larsen et al., 2014). For example, Cornet et al. (2019) proposed a method for estimating deterministic minimum dwell time depending on passenger flow while considering a random deviation due to stochastic influence such as inhomogeneous passenger distribution on the platform

or passenger carrying luggage. They included their model into a simulation platform and reported good performance using Paris real-world data set.

In addition, there are some implicit models to estimate dwell time without the knowledge of passenger flow (Kecman and Goverde, 2015; Li et al., 2016). For example, Li et al. (2016) predicted the dwell time of a train at a station using the information of dwell times of the same train at previous stations and dwell times of previous trains at the same stations.

2.2 Macroscopic modelling of traffic dynamics

With the knowledge of basic rules of train operation, how to describe the dynamics of railway system deserves a concise organization. The widely used macroscopic models for dynamic traffic representation will be introduced in this section. The modelling of any dynamic system should start from its basic component. Since most of the early studies on transportation system originated from road traffic, macroscopic modelling of the road link should be summarized first.

2.2.1 Exit-flow model

The exit-flow model first proposed by Merchant and Nemhauser (1978) discretely described the link volume $n(i+1)$ at the beginning of time period $i+1$ as:

$$\begin{cases} n_{i+1} = n_i + \lambda_i - \mu_i, \\ \mu_i = g\left(n_i\right), \end{cases} \tag{2.3}$$

where λ_i and μ_i are the number of vehicles entering into/departing from this link during time period i, respectively. $g(n)$ here is generally called the exit-function, and it was usually written as:

$$g(n) = \min\{\gamma_m n,\ G_m\}, \tag{2.4}$$

where γ_m is an constant parameter smaller than 1 and G_m represents the maximum number of vehicles can be discharged from the link during a time period. This model has been deeply discussed by Carey (1986), and widely extended by many following studies (e.g., Carey and McCartney, 2004; Friesz et al., 1989).

2.2.2 Delay function model

Instead of specifying the function of outflow, delay function model (Friesz et al., 1993; Nie and Zhang, 2005b; Wu et al., 1998) directly specified the link travel time when a vehicle enters the link. The most general form of this model using a linear delay function can be expressed as:

$$\begin{cases} \dot{n}(t) = \lambda(t) - \mu(t), \\ TT(t) = \alpha_d + \beta_d n(t), \\ \mu(t + TT(t)) = \frac{\lambda(t)}{1 + \dot{TT}(t)}, \end{cases} \tag{2.5}$$

where α_d and β_d are parameters describing a linear delay function, and $TT(t)$ is the travel time when an vehicle enters the link at time t. The third equation imposes First-In-First-Out (FIFO) principle on the dynamics. However, Nie and Zhang (2005a) pointed out that finding an appropriate delay function seems challenging and there exists probably no FIFO-consistent delay functions other than this linear one.

2.2.3 Point queue model

Point queue model (Kuwahara and Akamatsu, 1997; Smith, 1984) is also a widely used approach to describe the dynamics of the link. It has two basic assumptions: (1), the physical length of a vehicle is zero so if a queue is formed at the bottleneck, it also does not have a physical length; (2), vehicles travels along the link at free flow speed before they arrive at bottleneck. The derivative of queuing vehicle number $n_q(t)$ at time t, outflow $\mu(t)$, and link travel time $TT(t)$ can separately be described as:

$$\dot{n}_q(t) = \begin{cases} 0, & \text{if } n_q(t) = 0 \text{ and } \lambda(t - TT_0) < G_m, \\ \lambda(t - TT_0) - G_m, & \text{otherwise}, \end{cases} \tag{2.6}$$

$$\mu(t) = \begin{cases} \lambda(t - TT_0), & \text{if } n_q(t) = 0 \text{ and } \lambda(t - TT_0) < G_m, \\ G_m, & \text{otherwise}, \end{cases} \tag{2.7}$$

$$TT(t) = TT_0 + n_q(t + TT_0)/G_m, \tag{2.8}$$

where TT_0 is the free-flow/minimum travel time.

2.2.4 Cell transmission model

The cell transmission model (CTM) proposed by Daganzo et al. (1994) used discrete cells to approximate LWR (Lighthill and Whitham, 1955; Richards, 1956) hydrodynamic traffic flow flow. The number of vehicle in cell i at time $t+1$ equals its occupancy at time t, plus the inflow and minus the outflow, i.e.,

$$\begin{cases} n_i(t+1) = n_i(t) + y_i(t) - y_{i+1}(t), \\ y_i(t) = \min\{n_{i-1}(t),\ Q_i,\ N_i - n_i(t)\}. \end{cases} \tag{2.9}$$

where $y_i(t)$ represents number of vehicles entering cell i from cell $i-1$ during time interval t, Q_i is the capacity flow into cell i, N_i is the holding capacity of cell i, thus $N_i - n_i(t)$ is the amount of empty space in cell i at time t.

CTM has been widely employed in many dynamic traffic assignment models and showed good robustness although it was not tractable in most situations. Nie and Zhang (2005a) compared the performance of these four link models under various inflow profiles. He found that point queue model behaved identically as the CTM model except for cases when the link's exit boundary conditions also depended on traffic conditions of downstream links. While the delay function model systematically overestimated link traversal times. Exit-flow model was found to overestimate link traversal times when inflow rate decreased suddenly but underestimated link traversal times when inflow rate increased suddenly.

In recent years, modelling of dynamic traffic in network scale has received growing concern. As a representative of these studies, macroscopic fundamental diagram and its variants will also be briefly introduced.

2.2.5 Macroscopic fundamental diagram

Macroscopic fundamental diagram (MFD), with its theoretical foundation established by Daganzo (2007) and Geroliminis and Daganzo (2007), described the well-defined relation between travel production (i.e., the product of average flow and trip length) and accumulation in an aggregate manner. There are two important assumptions for the existence of an MFD: (1), the targeted area should be homogeneously congested; (2), the inflow into the area changes slowly compared with system relaxation time. When these two assumption hold or approximately hold, the dynamics of an area can be described by:

$$\dot{n}(t) = \lambda(t) - G(n(t)), \tag{2.10}$$

where $G(n)$ is a non-negative unimodal exit function indicating the relation between outflow/trip completion rate and accumulation. The existence of MFD has also been empirically confirmed by Geroliminis and Daganzo (2008).

Recently, some new studies (Arnott, 2013; Daganzo and Lehe, 2015; Lamotte and Geroliminis, 2016) proposed an trip-based MFD to overcome the drawback of strong assumptions of above-mentioned accumulation-based MFD although it is analytically difficult to solve. This type of MFD is generally expressed as:

$$L = \int_{t-TT(t)}^{t} V\left(n(s)\right) ds, \tag{2.11}$$

where L is the travel distance of a trip and $V(n)$ is the mean speed of travelers. Mariotte et al. (2017) made a thorough comparison of these two types of MFD and suggested that the trip-based one was a promising tool to account for heterogeneity in travel distance. There also emerges many MFD-like concepts for multi-modal transportation networks (see e.g., Chiabaut, 2015; Geroliminis et al., 2014) which will not be further introduced here.

2.3 Optimization and control from supply side

The macroscopic modelling is aimed at describing the system dynamics in an aggregate manner, but an crucial issue for an operator is how to optimally dispatch trains during a day to maximize its profit while guaranteeing an acceptable level of service. This section mainly introduces studies from supply side (i.e., timetable optimization and dynamic control) of railway system.

2.3.1 Timetable optimization

Timetable of railway system determines the arrival and departure times for each train at each station. The design and optimization of timetable should consider both the time-dependent passenger demand information and many practical constraints such as available number of rolling stocks and limited capacity of the train. In general, the timetables can be categorized into two types: (1) periodic timetable (i.e., a timetable that is constantly repeated, for example with departures at 0, 20 and 40 min every hour), which has the advantage of being easily memorized by passengers; (2) non-periodic timetable (or demand/service-oriented), which is aimed to meet the time-dependent demand of passengers to a greater degree. Most of studies (e.g., Goverde, 2007; Nachtigall, 1996; Odijk, 1996; Peeters, 2003) on periodic timetable originated from the Periodic Event Scheduling Problem (PESP) formulation first proposed by

Serafini and Ukovich (1989). For example, Odijk (1996) built a model consisting of periodic time window constraints by means of which arrival and departure time can be associated. Goverde (2007) studied the robustness and stability of periodic timetable by introducing a max-plus system theory and analyzed train delay propagation processes. A complete review of the PESP models' characteristics can be referred to Liebchen and Möhring (2007). An application of PESP timetable for large-scale networks can be found in Kroon et al. (2009). Moreover, periodic timetabling under an elastic demand for a railway network has also been investigated by Cordone and Redaelli (2011). However, periodic timetables generally cannot respond very well to time-dependent passenger demand, which may lead to long waiting time especially in a high-frequency operated railway system like urban rail transit.

One the other hand, another branch of studies focused on designing timetable depending on a given passenger demand pattern. Based on fluid traffic flow models, the early studies (Hurdle, 1973; Newell, 1971; Osuna and Newell, 1972) did many mathematical analysis on deriving an optimal dispatching frequency in a very idealized public transportation system. For example, Newell (1971) and Hurdle (1973) analytically minimized total passenger waiting time by using cumulative curves from queuing theory. Ceder (1984) firstly addressed the importance of ridership information and he pointed out that service frequency should correspond to temporal passenger demand. In a follow-up study, Ceder (1987) introduced an alternative approach for the transit timetable design problem, with the objective to synchronize vehicle-departure times under dynamic passenger demands. Further, Ceder (2016) provided a comprehensive modeling framework to adjust vehicle departure times and smooth the transitions between time periods by regular time headways or average passenger loads.

For any optimization problem, there are two key components that should be clarified, i.e., objective functions and constraints. Since the objective of optimization can be considered from two perspectives, namely the operator and passengers, prevailing objective functions can be categorized to minimize:

- Total/average waiting time of passengers,
- Total travel time (including transfer time) of passengers,
- Operation cost/energy consumption,
- Multi-objectives by combining the above objectives.

While the major constraints can be categorized as:

- Arrival/departure sequence for no over-taking system,

- Total passenger demand,
- Available number of rolling stocks,
- Limited capacity of train,
- Minimum headway of consecutive trains,
- Limited cruising speed or minimum travel time between stations.

Take the first objective function as an example, it can generally be written as:

$$\min \sum_{i \in S} \sum_{j \in M} \left(A_p^i(t_j^i) - D_p^i(t_j^i) \right) \left(t_j^i - t_{j-1}^i \right), \tag{2.12}$$

where $A_p^i(t)$ and $D_p^i(t)$ are cumulative passenger arrivals and departures at station i at time t, respectively. t_j^i is the departure time of train j at station i.

In recent two decades, most of timetable optimization studies, especially those targeted at handling dynamic passenger demand, are based on integer and mixed integer programming formulations including linear or nonlinear ones. In fact, binary variables are widely employed in the timetable optimization models since they can conveniently describe the usage of train or path, the loading condition of a specific train and so on. Carey (1994) and Higgins et al. (1996) were probably the first researchers to introduce binary variables to describe the order between train services. They developed models and heuristic algorithms to solve the optimization problem of train pathing and scheduling for a station and a single line rail corridor. Carey and Crawford (2007) extended the approach of Carey (1994) to a series of complex stations linked by multiple railway lines. The recent studies introduced in the remaining part of this sub-section mainly followed this formulation style of optimization problem and a brief summary of these studies are shown in Table 2.1.

Niu and Zhou (2013) focused on optimizing timetable under a dynamic demand scenario, especially they considered the case when the system is over-congested so that some passengers may not be able to board the next arriving train. Their objective function was to minimize total number of waiting passengers and weighted remaining passengers. Their nonlinear optimization problem was solved by a genetic algorithm through a special binary coding method. They showed the effectiveness of their model by using a real-world data set with 13 stations.

Barrena et al. (2014) developed two nonlinear optimization models to minimize the average passenger waiting time at stations, which generalized the non-periodic timetabling problem on a single line under a dynamic demand environment. They also proposed a fast adaptive large neighborhood search (ALNS) algorithm to solve large instances of the problem

within short computation times, and they reported the average reduction in passenger waiting times could reach 26% with their optimized timetable.

Canca et al. (2014) proposed a nonlinear integer programming model that aimed to optimize arrival and departure time of trains under dynamic passenger demand. Their objective function jointly considered passenger waiting times and operation cost. They tested their model on the C5 line of Madrid rapid transit system and discussed the trade-off between system supply and service quality.

Sun et al. (2014) formulated three optimization models to design demand-sensitive timetables by representing train operation using equivalent time (interval) concept. The three models had the same objective function to minimize the average waiting time, but the first model did not consider capacity constraints, the second one considered, and the third one further adopted a step-shape peak and off-peak timetable pattern. They compared the performance of the three models using data from a metro line in Singapore and they reported that the second model with time-varying headway was the best.

Niu et al. (2015) further proposed a quadratic integer programming to minimize total passenger waiting time with given dynamic passenger demand in a railway system with predetermined skip-stop patterns. They examined the effectiveness of their model through numerical experiments of real-world test cases.

Yin et al. (2017b) considered an integrated approach to jointly minimize the operational costs (i.e., energy consumption) and passenger waiting time. They developed a Lagrangian relaxation-based heuristic algorithm to decompose the primal problem into two sets of sub-problems and thus found the solution in short computational time. They verified the effectiveness of their approaches with operation data of Beijing metro system.

Wang et al. (2018b) studied the integrated optimization of timetable and rolling stock circulation plans under dynamic passenger demand for an urban rail transit line. They formulated a multi-objective mixed integer nonlinear programming model with three solution approaches. They compared the performance and computation cost of the three approaches with real-world data.

Shi et al. (2018) proposed a method to jointly optimize timetable and accurate passenger flow control strategies for an over-saturated metro line. They developed a hybrid algorithm to solve the problem and conducted numerical experiments for a small-scale case and a real-world instance.

Yan and Goverde (2019) combined the line planning and timetable optimization problem for a railway system where strongly heterogeneous train services are operated. They designed a multi-objective function to provide both periodic and non-periodic plans considering

timetable robustness, regularity and passenger travel time. They also verified the performance of their model by a case study.

All the above mentioned studies focused on uni-direction (single or multiple) or bi-direction railway lines. There are also a number of studies that considered the optimization problem for railway networks. Wong et al. (2008) presented a mixed integer programming optimization model for schedule synchronization problem with non-periodic timetables that minimized the interchange waiting times of all passengers. They tested their models for the mass transit railway system in Hong Kong and indicated that their approach significantly improved the synchronization of timetable compared to current ones. Wang et al. (2015) proposed an even-driven model to optimize timetable for a simple railway network with two railway lines. The routing and transfer behavior of passengers at transfers stations were also taken into consideration. Robenek et al. (2016) presented a timetable optimization framework to maximize train operator's profit while maintaining a certain level of passenger satisfaction. They found that non-periodic timetables could perform better under high demand by a sensitivity analysis using data from railway network of Switzerland. In their follow-up study Robenek et al. (2018), elastic demand was included by considering the mode choice behavior of passengers. Similar studies on the timetable optimization for railway networks can also be found in Canca et al. (2016) and Wang et al. (2020).

2.3.2 Dynamic control

Once an timetable is put into operation, disturbances and disruptions are inevitable due to the stochastic nature of passenger demand and unpredictable accidents. A considerable number of dynamic control strategies have been proposed towards a robust operation. According to the type of strategy, a brief introduction is made on studies focusing on bunching control and rescheduling.

The bunching phenomenon, especially for bus system, has been known for nearly 60 years since the first explanation by Newell and Potts (1964). As pointed out by many studies, bus systems are naturally unstable due to a small disturbance. When one bus is delayed, it will encounter more passengers at stops ahead, these extra passengers delay it further. On the contrary, the following bus will encounter fewer passengers and tends to catch up the preceding bus. Although the cruising time of a train between two stations is generally not affected by external environment like buses being caught in a traffic jam, the high frequency operated rail transit system also suffers from such kind of bunching phenomenon, which is also called "knock-on delay" by some scholars in railway field (Carey and Kwieciński, 1994). Since the essence of bunching does not have significant difference between bus and railway systems, we hopefully expect that most of control strategies for bus bunching can also be

Table 2.1 Summary of recent studies on timetable optimization

Publications	Environment	Objective(s)	Model structure	Solution algorithm
Wong et al. (2008)	Network	Transfer waiting time	MILP	CPLEX
Niu and Zhou (2013)	Bi-direc	Total waiting time	Nonlinear	GA
Barrena et al. (2014)	Uni-direc	Average waiting time	MILP, nonlinear	ALNS
Canca et al. (2014)	Uni-direc	Average waiting time and average load factor	MINLP	GAMS
Sun et al. (2014)	Uni-direc	Average waiting time	MILP	CPLEX
Niu et al. (2015)	Uni-direc	Total waiting time	MINLP	GAMS
Wang et al. (2015)	Small network	Total travel time and energy consumption	Nonlinear	SQP, GA
Canca et al. (2016)	Network	Total travel time and operation cost	MINLP	Cutting plane
Robenek et al. (2016)	Network	Operator profit and passenger satisfaction	MILP	CPLEX
Yin et al. (2017b)	Bi-direc	Total waiting time and energy consumption	MILP	CPLEX, heuristic
Shi et al. (2018)	Bi-direc	Total waiting time with passenger flow control	ILP	CPLEX, heuristic
Wang et al. (2018b)	Bi-direc	Headway and load factor variations, number of depot	MINLP	Iterative approach, CPLEX
Yan and Goverde (2019)	Uni-direc-multi	Total travel time, timetable robustness and regularity	MILP	Iterative approach
Wang et al. (2020)	Network	Total waiting time, transfer failure	MILP	CPLEX, GA

Symbols in the table: bi-directional single line (Bi-direc); uni-directional single line (Uni-direc); uni-directional multiple lines (Uni-direc-multi); mixed integer linear programming (MILP); Cplex solver (CPLEX); genetic algorithm (GA); adaptive large neighborhood search (ALNS); mixed integer nonlinear programming (MINLP); Gams software (GAMS); sequential quadratic programming (SQP); integer linear programming (ILP).

effective for urban rail transit. Here, the prevailing bunching control can be categorized into three types:

- Headway-based holding strategies (e.g., Bartholdi III and Eisenstein, 2012; Daganzo, 2009; Xuan et al., 2011)
- Speed regulation strategies (e.g., Daganzo and Pilachowski, 2011; Li et al., 2019),
- Skip-stop strategies (e.g., Delgado et al., 2009; Fu et al., 2003; Suh et al., 2002; Sun and Hickman, 2005).

Here, the general idea of headway-based holding strategies is introduced since this type of strategy is the most applicable one also for rail transit system. Follow the paper by Xuan et al. (2011), consider a railway system in which trains are dispatched from the first station with a constant time interval H, the scheduled arrival time of train n at station $s+1$ can be written as:

$$T_{n,s+1} = T_{n,s} + \beta_s H + d_s + c_s, \tag{2.13a}$$

$$T_{n,s} = T_{n-1,s} + H, \tag{2.13b}$$

where β_s is a dimensionless parameter for demand level at station s, d_s is the amount of slack time in schedule at station s, c_s is the average cruising time from station s to $s+1$. Similarly, the actual arrival times can be described as:

$$t_{n,s+1} = t_{n,s} + \beta_s h_{n,s} + D_{n,s} + c_s + \nu_{n,s+1}, \tag{2.14}$$

where $h_{n,s}$ is the actual time headway between train n and $n+1$ at station s, $D_{n,s}$ is the holding time of train n at station s that should be designed, and $\nu_{n,s+1}$ is a random noise for the trip of train n from station s to $s+1$. By combining Eq. (2.13) and (2.14), the deviation from schedule can be written as:

$$\varepsilon_{n,s+1} = \varepsilon_{n,s} + \beta_s(\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \nu_{n,s+1} + (D_{n,s} - d_s). \tag{2.15}$$

Now, it is clear that the objective of holding strategies is to set a series of $D_{n,s}$ to make $\varepsilon_{n,s+1}$ converges to zero or a small enough value with the increase of n and s. However, when we consider the morning commute problem with this control strategy, one critical issue is that demand parameter β_s and slack time should be time-dependent, otherwise the dwell time would be either too long for off-peak and too short for the peak hours.

Regarding the speed regulation strategy, Daganzo and Pilachowski (2011) presented an adaptive control scheme that adjusted the cruising speed of a bus in real-time based on both,

its front and rear spacings much as if successive bus pairs were connected by springs. They proved that this scheme could yield regular headways with faster bus travel than existing control methods. A overview of dynamic control for bus systems can be found in Chapter 6 of Ibarra-Rojas et al. (2015)'s paper.

Recent studies tend to design the bunching control using hybrid strategies with more consideration of stochastic dynamic passenger demand. For example, Wu et al. (2017) allowed the overtaking of buses and modeled the distributed passenger boarding behavior to design a holding control scheme. Li et al. (2019) proposed a robust hybrid bus control scheme combining holding and speed regulation strategies to reduce bunching and improve the schedule adherence and headway regularity.

With regard to the rescheduling, it is designed to adjust the current timetable in an efficient way when a disturbance or disruption occurs. In general, strategies can be divided into two categories, one is to recover the original timetable and the other is to redesign a new scheduling plan. For disturbances that affected limited segment of the railway line in a short period, to recover the original timetable might be a good choice. In this case, control strategies are more or less similar to bunching control. However, when a large disruption occurs, some scheduled trains have to be cancelled and a new timetable might be more appropriate to re-assign available rolling stocks and crews. Cacchiani et al. (2014) did a thorough review of rescheduling models. According to his classification, the existing studies can be categorized as shown in Table 2.2.

Table 2.2 Classification of rescheduling models

Target	Model level	Standpoint
Disturbance (small delay)	Microscopic (precise modelling)	Operator-centric (minimize loss for operators)
Disruption (large accident)	Macroscopic (high-level modelling)	Passenger-centric (minimize loss for passengers)

For example, Corman et al. (2010) developed a series of tabu search algorithms to solve alternative graph model for rescheduling and rerouting trains from disturbances. Binder et al. (2017) formulated a multi-objective model to deal with rescheduling problem from a macroscopic perspective in case of large disruptions. Li et al. (2017) systematically investigated a joint optimization of train regulation and passenger flow control for a metro system.

2.4 Modelling and management from demand side

The previous section summarized studies that focused on how to optimally dispatch the train services and how to dynamically control the operation. However, the passenger demand, either static or dynamic, is usually treated as given information. In fact, to estimate the temporal or spatial distribution of demand itself remains as a challenging task especially when management strategies to flatten the demand are implemented. This section briefly summarizes studies on the modelling and management of travel demand, more specifically, departure time choice problem and fare scheme design.

2.4.1 Departure time choice

The essence of departure time choice problem is to understand how an individual choose his/her departure time from home when he/she has a desired time to arrive the destination. Departure time choice problem has a premise that if all travelers tried to arrive at their desired time, the total travel time they experienced would significantly increase due to the limited capacity of transportation infrastructure (either road or railway system). Therefore, travellers trade off their total travel time and schedule delay (i.e., time difference between desired and actual arrival time at destination) to decide their departure time from home.

Economists like Vickrey (1969) and Henderson (1974) might be the first to consider individuals' departure time decisions for commuting problem by private cars. They considered the existence of a bottleneck on the road and assumed that passengers minimized their total travel costs to determine their departure time. When the total travel cost is minimized for any individual, no one would have the incentive to change his/her departure time, which is generally referred to as an equilibrium state. The early studies (e.g., Hendrickson and Kocur, 1981; Hurdle, 1981) employed deterministic queuing theory to derive the time-dependent demand at a bottleneck under equilibrium state. Hurdle (1981) studied two different situations, i.e., a morning case where travellers had their desired arrival time for work and a afternoon case where travellers had desired departure time after work, but he did not consider the schedule delay cost. Hendrickson and Kocur (1981) considered the linear schedule delay cost and found that the equilibrium arrival rate at a bottleneck for a scheduled event should be two constant values related to bottleneck capacity and users' value of time. Further, they discussed the system optimum with time-dependent tolls to eliminate the queue. They concluded that no increase in capacity was capable of eliminating peak period congestion with many workers starting work at the same time. Smith (1984) and Daganzo (1985) further strengthened the theoretical foundation of deterministic departure time choice problem at a bottleneck by proving the existence and uniqueness of the solution.

A review study of departure time choice problem for bottleneck based road traffic can be found in Kuwahara (1998). Fig. 2.4 illustrates the conventional framework of this problem by cumulative curves. $W(\cdot)$ is the cumulative desired departures from the bottleneck, $A(\cdot)$ and $D(\cdot)$ are cumulative arrivals/departures at/from the bottleneck, respectively. μ is the bottleneck capacity and $\lambda(t)$ is the arrival rate of car that should be solved. In general,

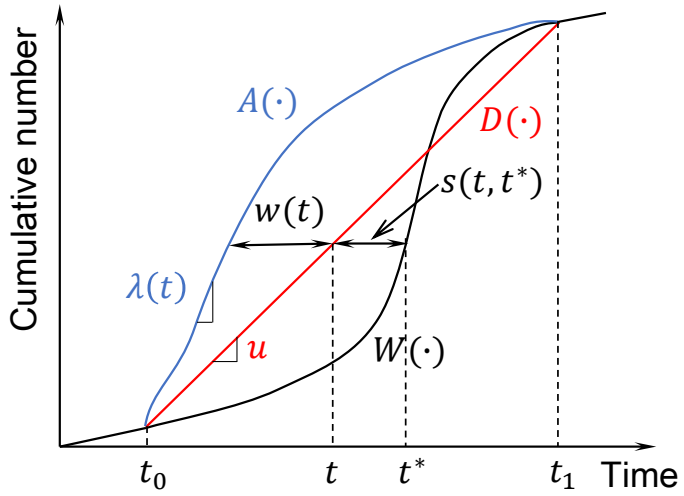


Fig. 2.4 Cumulative curves for single bottleneck departure time choice problem.

deterministic model solve $\lambda(t)$ by assuming the total travel cost of an user like:

$$\begin{aligned} TC\{t,t^*,\vec{x}\} &= f_m\{m(\vec{x})\} + f_w\{w(t)\} + f_s\{s(t,\ t^*)\} \\ &= f_m\{m(\vec{x})\} + p\{t,\ t^*\}. \end{aligned} \tag{2.16}$$

$f(\cdot)$ represent functions converting time to cost. $m(x)$ is the constant part of travel time which is only determined by the distance vector $\vec{x}$ from home to workplace. $w(t)$ is the queuing time at the bottleneck and $s(t,\ t^*)$ is the schedule delay when a user departs from bottleneck at t with a desired departure time t^*. $p(t,\ t^*)$ concisely represents the dynamic cost of the travel. Then, an equilibrium state refers to the situation when the total travel cost of any user is minimized, which can be written as:

$$\frac{\partial TC(t,t^*,\vec{x})}{\partial t} = \frac{\partial p(t,t^*)}{\partial t} = 0. \tag{2.17}$$

Now, if $f_s(\cdot)$ and $f_w(\cdot)$ are given, equilibrium queuing time $w(t)$ can be solved using Eq. (2.17). Finally, according to the conservation law, $w(t)$ is determined by $\lambda(t)$ and μ, thus equilibrium $\lambda(t)$ can also be solved.

With this basic framework, subsequent researches expanded the original problem in many different dimensions. For example, some studies analyzed multiple bottlenecks situations with different configurations (see "Y-shape" by Arnott et al. 1993a or along one corridor by Kuwahara 1990 and Akamatsu et al. 2015). Some studies discussed the conditions when commuters are heterogeneous (i.e., they have different values of time, see e.g., Newell 1987, Lindsey 2004, Takayama and Kuwahara 2017). Some studies considered the distributed/staggered work start time cases (see e.g., Henderson, 1981; Takayama, 2015). There are also a considerable number of studies that introduced random utility theory to departure time choice problem (see e.g., Ben-Akiva et al., 1986; De Palma et al., 1983; Mahmassani and Chang, 1986). Elastic demand issue (Arnott et al., 1993b; Ben-Akiva et al., 1986) or route choice problem (Arnott et al., 1990a) has also been considered for departure time choice problem.

All the studies mentioned above are targeted at road traffic. In fact, studies on departure time choice problem for rail transit are relatively few. Sumi et al. (1990) proposed a stochastic model to estimate commuters' decisions of departure time and route choice, they tested their model with survey data and concluded that their model could well represent departure time choice within mass transit systems. Tabuchi (1993) analyzed the commuting congestion when there is mass transit parallel to a road. They investigated the optimality and efficiency of several railroad fare and road toll regimes and provided practical rules for attaining the social optimum. Alfa and Chen (1995) proposed an equilibrium model to estimate temporal demand distribution along a transit corridor with multiple origins and destinations, where they assumed a random boarding order of passengers. Kraus and Yoshida (2002) considered the time-of-use decision into a model of optimal pricing and service in urban mass transit. Their model is closely related to the highway bottleneck model, with difference arising in the intermittent nature of mass transit capacity provision. Tian et al. (2007) considered a many-to-one transit system where passengers were assumed to trade off in-vehicle crowding cost and schedule delay cost to decide their departure time. They analyzed the equilibrium properties and derived plausible insights such as: (1), trains arriving the destination at the desired time by everyone were utilized by commuters from all stations; (2) the duration of commute was longer for stations farther from the destination. Recent studies on departure time choice for rail transit usually took in-vehicle crowding or seat availability as the main factor for decision making process. For example, Sumalee et al. (2009) proposed a stochastic dynamic transit assignment model with an explicit seat allocation process for a general transit network. Xie and Fukuda (2014) presented an empirical study which examined the theoretical time-varying marginal utility model introduced by Vickrey (1973), using data on urban rail commuters in Tokyo. De Palma et al. (2015) discussed the formulation of

crowding in public transport and its implications for pricing, seating capacity and optimal scheduling.

Probable limitations of these studies are: (1), most studies considered a constant travel time between the origin and destination. This assumption is acceptable when operation is totally independent of passenger congestion condition. However, as discussed in Chapter 1, the operation was found to be frequently disturbed by the extended dwell time due to passenger congestion. Therefore, the constant travel may not be an appropriate assumption for a high-frequency operated congested rail transit system; (2), the departure time choice in most studies is considered as a beforehand decision making process, while more practically, it is also affected by real-time operation information available for commuters (Noursalehi et al., 2018).

2.4.2 Fare scheme design

As a major demand management strategy to alter users' travel behavior, to collect surcharges or to provide monetary incentives has long been studied for road traffic. For example, based on the bottleneck model introduced in last subsection, time-dependent optimal toll and its approximated form, step tolls have been thoroughly investigated (Arnott et al., 1990b; Laih, 1994; Lindsey et al., 2012). However, discussions of fare scheme design for rail transit are still emerging.

Kraus and Yoshida (2002) analyzed the optimal fare and service frequency in urban mass transit by analogously treating waiting time at the station as queuing time of bottleneck model. They showed that increased ridership led to higher average user cost, and relationship between optimal service frequency and ridership did not conform to the square root principle.

Zhang et al. (2014) examined the impact of incentives on commuters' travel behavior based upon a questionnaire survey conducted in Beijing subway system. They confirmed that offering monetary incentives such as fast-food coupons or reduced fares were effective to encourage some passengers to avoid travel during the morning peak.

Rey et al. (2016) proposed a lottery-based revenue-neutral incentive mechanism to reduce the congestion in urban rail transit systems. They tested the validity of the proposed mechanism through monetized laboratory experiments and they concluded that risk attitudes played a significant role in explaining behavior in lottery-based incentive mechanisms.

De Palma et al. (2017) did an economic analysis of time decisions of rail transit users who trade off the crowding and schedule delay cost. They pointed out that without fare or a uniform fare led to concentrated ridership on timely trains. The time-dependent fares based on marginal-cost-pricing could smooth the unbalanced load of trains and generate more revenue than an optimal uniform fare.

Yang and Tang (2018) designed a revenue-neutral fare-reward scheme to incentivize a shift in departure time to the shoulder periods of the peak hours by rewarding a commuter with one free trip during shoulder periods after a certain number of paid trips during the peak hours. Their studies indicated that depending on the original fare, the proposed fare scheme resulted in an optimal reward ratio up to 50% and yielded a reduction of system total time costs and average equilibrium trip costs by at least 25% and 20%, respectively. Tang et al. (2020) further extended this fare scheme to a hybrid one by offering alternative options to heterogeneous commuters with different scheduling flexibility. They showed that the hybrid fare scheme was not only revenue-neutral but also Pareto-improving.

Wang et al. (2018a) considered a fare scheme to spatially redistribute the commute demand by encouraging an appropriate number of passengers to utilize the neighboring uncongested stations. They proved the existence of the unique Pareto-optimal solution in the absence of explicit demand functions. A trial-and-error fare design scheme was proposed to identify the Pareto-optimal solution.

Facilitated by the availability of smart card (AFC) data, Ma and Koutsopoulos (2019) proposed a general framework for the optimal design of promotion based demand management strategies in urban rail systems. Their framework consisted of two major components: network performance and optimization. They demonstrated the applicability of their framework through a case study focusing on the design of a pre-peak discount promotion scheme for the Hong Kong MTR subway system.

Zhang et al. (2020) developed a model for congested urban rail transit with heterogeneous users under monopoly. They considered both travel time delays and crowding externalities and characterized the user heterogeneity by a three-dimensional space of value of time, value of crowding, and willingness to pay. They proved the existence of a unique user demand equilibrium. They also derived the operator's fare and frequency decisions using a bi-level equilibrium approach.

2.5 Position of this study

Based on the literature review in previous four sections, the major research gaps can be concluded as two issues:

First, most studies on optimization and control from supply side ignored the interaction between demand and supply by taking passenger demand, either static or dynamic, as a given information or obeyed a given distribution. This is acceptable if the modification of supply is so insignificant that passengers do not notice and accordingly change their commute behavior. But as suggested by both theoretical studies (De Palma et al., 2017; Zhang et al., 2020) and

empirical practices (Currie, 2010; Peer et al., 2016; Zhang et al., 2014) from demand side, passengers, especially those with high flexibility, would probably adjust their departure time for a travel if they could expect some benefits (i.e., available seats or monetary reward) from such adjustments. Therefore, for any model aiming at improving the service level of rail transit by management strategies, the interaction of supply and demand cannot be ignored.

Secondly, very few studies on modelling and management from demand side considered the dynamics of rail transit system, especially passenger congestion influence on trains' operation. Most of them assumed a constant travel time between two stations, which means the chronic delays frequently happened during peak hours were ignored. In fact, many empirical studies (Kariyazaki et al., 2015; Kato et al., 2012; Tirachini et al., 2013) reported that total travel time by rail transit during peak hours were significantly longer than that of off-peak hours. Therefore, this congestion influence should be considered when modelling passengers' time choice behavior.

To fill these gaps, this study focused on the equilibrium demand distribution for a congested rail transit and its application for management strategies with the consideration of demand and supply interaction dynamically. An illustration of position of this study is shown in Fig. 2.5.

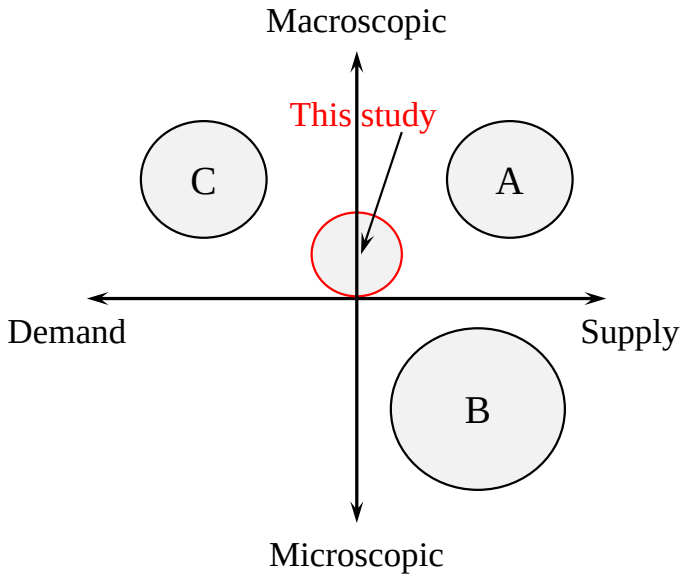


Fig. 2.5 An illustration of position of this study.

As can be seen from this figure, most existing studies lay in the three areas: (A), macroscopic approach for supply analysis (most early studies on transit optimal frequency belong to this area, see e.g., Hurdle 1973; Mohring 1972; Newell 1971); (B), microscopic approach for supply analysis (recent studies on timetable optimization or dynamic control introduced

in Section 2.3 belong to this area); (C), macroscopic approach for demand analysis (relatively few studies for rail transit introduced in Section 2.4). This study took the position at the intersection of demand, supply and macroscopic approach. By considering the dynamic interaction of demand and supply in a congested urban rail transit system using a macroscopic approach, this study is expected to provide insights on management strategies to relieve congestion and prevent the occurrence of delay.

Chapter 3

Fundamental diagram of urban rail transit

As explained in Chapter 1, to properly describe passenger congestion influence on trains' operation is crucial for any model targeted at urban rail transit system. This chapter aims to comprehensively understand this influence by investigating a macroscopic model (i.e., fundamental diagram of urban rail transit by Seo et al. 2017, abbreviated to train-FD hereinafter) with empirical data from Boston subway system. This chapter is structured as follows:

The first section introduced the nature of data source and derived the empirical train-FD from the data. The second section analytically derived the train-FD and its variants based on different dwell time assumptions. The third section calibrated the variant train-FDs and evaluated them using the Akaike information criterion (abbreviated to AIC, Akaike, 1998).

3.1 Analysis of Boston red line operation data

To understand passenger congestion influence on railway operation, the data set should include both the movements of trains and the arrival of passengers. Fortunately, the Massachusetts Bay Transportation Authority (MBTA) recently published a substantial amount of required data through its APIs. The raw data includes per minute turnstile entry counts at each station, as well as subway operation conditions in Google's GTFS format (Barry and Card, 2014). Here the busiest section of the red line (from Alewife to JFK/Umass with 13 stations, shown in Fig. 3.1) was chosen as the analysis target. The flow and density of the railway system are calculated by employing Edie's definition (Edie, 1963) of traffic flow written as:

$$q(A) = \frac{\sum_{n \in N} d_n}{|A|}, \tag{3.1a}$$

$$k(A) = \frac{\sum_{n \in N} \tau_n}{|A|}, \tag{3.1b}$$

where A is the measurement time-space area and $|A| = L \times \Delta t$. d_n and τ_n are the total travel distance and travel time of vehicle n in A, respectively. The total length of the selected railway line section $L = 14.4\ km$ and the time unit $\Delta t = 10\ min$. This implies that one data point in the FD represents the 10 min average flow and density of the railway system. Accordingly, the per minute passenger entry data is also aggregated into 10 min average entries at each station, and is then converted to arrivals per hour (pax/h). The calculation utilizes data from 18 normal weekdays from 6:00 to 24:00. Fig. 3.2 shows the time evolution of train flow (southbound) and the passenger arrival rate within-day, where the curve represents the mean values and the shadow indicates the variation. It can be observed that during the rush hours, the train flow declines after the peak of the passenger arrival rate, which implies that passenger congestion influences railway operation.

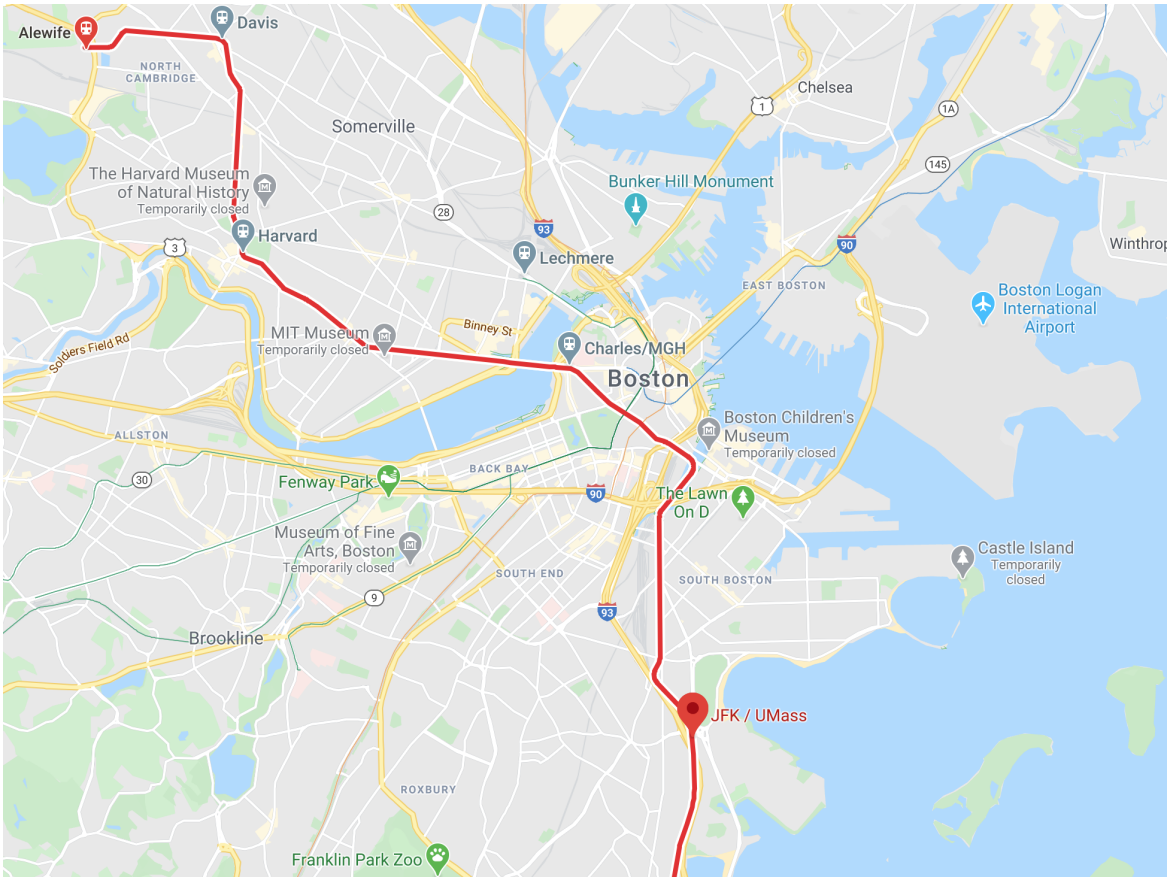


Fig. 3.1 Map of selected section of Boston red line

To obtain relatively steady state data, the unsteady data was filtered out by judging the adjacent train flow change over 20%. Finally, the train-FD of the Boston red line is depicted

in Fig. 3.3. The color used represents the value of the passenger arrival rate, as illustrated in the color bar (pax/h). As can be seen from Fig. 3.3, the approximately linear relation between train density and flow became more insignificant under high passenger arrival rate (i.e., more yellow and red points lay lower for a given train density). This results also implied that high passenger volume affected the operation of trains.

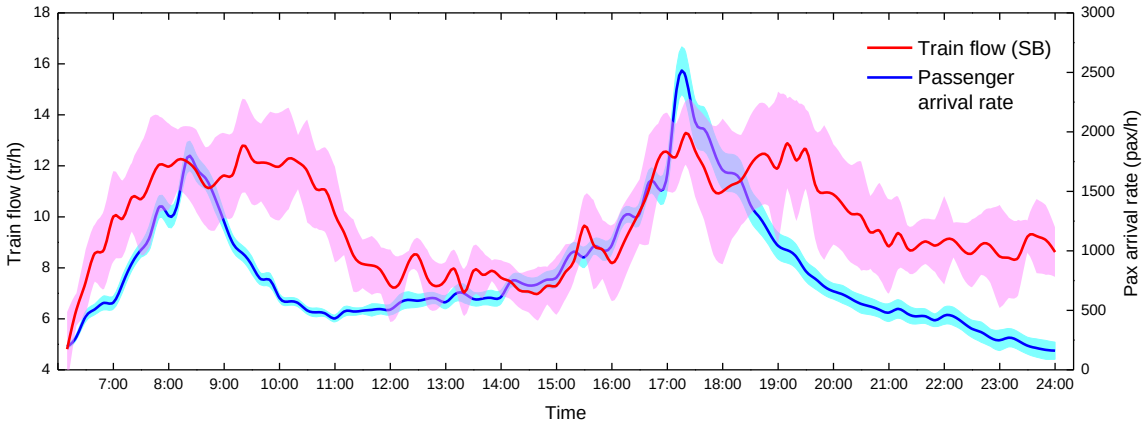


Fig. 3.2 Train and passenger flow evolution during one day

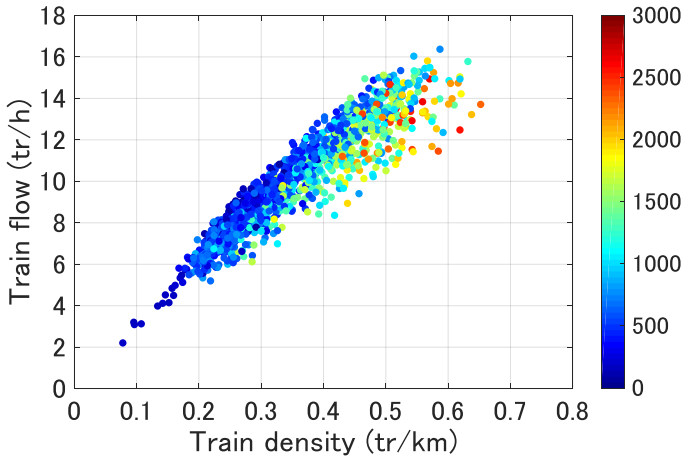


Fig. 3.3 Weekday train-FD of the Boston red line (southbound)

3.2 Model formulation

In this section, three different train-FD models are derived. One is proposed by Seo et al. 2017 and the others are slightly modified versions of the former, based on different train dwell time assumptions. The first subsection explains the assumptions of railway operation,

the second subsection derives the train-FD, which is almost the same contents referred from Seo et al. 2017.

3.2.1 Assumptions on railway operation

The operation of the railway system basically depends on the dwelling and cruising behaviors of each train. With regard to the dwelling behavior at stations, we assume that the dwell time t_b is determined by the number of boarding passengers N_p on the platform [1]. Here, three assumptions for dwell time are considered:

1. t_b keeps constant at t_bcon regardless of N_p, which indicates that passenger congestion does not affect railway operation;

2. t_b monotonically increases with N_p from a minimum value t_b0 (buffer time), which is the same assumption made by Seo et al. 2017;

3. t_b keeps constant until a critical passenger number N_0 is reached. It then starts increasing. This idea is inspired by the empirical work on passenger boarding by Kariyazaki et al. 2015.

It is also assumed that all of the waiting passengers can always board the next approaching train, which means $N_\mathrm{p} = a_\mathrm{p} H$, where a_p is the passenger arrival rate at stations and H is the time headway of consecutive trains. Now, the three assumptions of t_b can be expressed as Eq. (3.2) - (3.4), respectively. μ here can be interpreted as the dwell time growth rate with increase of boarding passenger number.

$$t_\mathrm{b} = t_\mathrm{bcon}, \tag{3.2}$$

$$t_\mathrm{b} = t_\mathrm{b0} + \mu a_\mathrm{p} H, \tag{3.3}$$

$$t_\mathrm{b} = \begin{cases} t_\mathrm{b0}, & \text{if } a_\mathrm{p} H \leq N_0 \\ t_\mathrm{b0} + \mu (a_\mathrm{p} H - N_0). & \text{if } a_\mathrm{p} H > N_0 \end{cases} \tag{3.4}$$

The cruising behavior of a train is modeled using Newell's simplified car-following model (Newell, 2002). More specifically, the position $x_\mathrm{m}(t)$ of train m at time t is expressed as:

$$x_\mathrm{m}(t) = \min\{x_\mathrm{m}(t-\tau) + v_\mathrm{f}\tau,\ x_{\mathrm{m}-1}(t-\tau) - \delta\}, \tag{3.5}$$

[1] N_p can also be interpreted as the total number of alighting and boarding passengers. However, this will complicate the following discussions. Even though alighting and boarding passengers are carefully distinguished, the following derivations are still valid and this would not affect the final results

where $m-1$ indicates the preceding train, τ is the minimum time headway of consecutive trains, and δ is the minimum spacing. The first term represents the free-flow regime where the train cruises with desired speed v_{f}. The second term indicates the congested regime where the train decreases its speed to maintain the minimum headway and spacing. For clarity, hereinafter, we refer to each model, based on assumptions 1, 2 and 3, as models A, B, and C, respectively.

3.2.2 Derivation

To derive a train-FD, we consider railway operation under the steady state (also known as equilibrium state). Specifically, the following conditions are considered:

- Parameters t_{bcon}, t_{b0}, μ, a_{p}, N_0, τ, δ are time-independent.
- Headway H and desired cruising speed v_{f} are also time-independent.

Also, for simplicity, a homogeneous railway system is considered, which indicates:

- Trains stop at each station.
- a_{p} for each station is the same.
- Distance between any two adjacent stations is the same, referred to as l.

Now, the train-FD as expressed in Eq. (3.6) can be separately derived in free-flow and congested regime by combining the above-mentioned assumptions.

$$q = Q(k,\ a_{\mathrm{p}}) = k\bar{v}, \tag{3.6}$$

where q is the steady state train flow (tr/h) and $q = 1/H$, k (tr/km) is the average density of the railway line, and $\bar{v}$ is the average traveling speed of a train (or system), which can be described by Eq. (3.7).

$$\bar{v} = \frac{l}{t_{\mathrm{b}} + l/v}, \tag{3.7}$$

where v is the average cruising speed of a train. In the free-flow regime, $v = v_{\mathrm{f}}$ so that the explicit expression of q for models A, B and C can be easily derived by substituting Eq. (3.2) - (3.4) and Eq. (3.7) into Eq. (3.6).

In the congested regime, the headway H should satisfy:

$$H \geq t_{\mathrm{b}} + \frac{\delta + v\tau}{v}. \tag{3.8}$$

By taking the equal boundary condition of Eq. (3.8) and employing $q = 1/H$, Eq. (3.2) - (3.4) can be substituted into Eq. (3.8) so that q can be described as a function of v and a_p:

$$q = f_1(v, a_\mathrm{p}). \tag{3.9}$$

Then, by inserting Eq. (3.7) and Eq. (3.9) into Eq. (3.6), we can also obtain k as a function of v and a_p:

$$k = f_2(v, a_\mathrm{p}). \tag{3.10}$$

By using Eq. (3.9) and Eq. (3.10), the slope of the train-FD in the congested regime dq/dk can be derived since $dq/dk = (dq/dv) \cdot (dv/dk)$. Finally, employing the critical state train flow q^*:

$$q^* = \frac{1}{t_\mathrm{b} + \delta/v_\mathrm{f} + \tau}, \tag{3.11}$$

as a boundary condition, the train-FDs of models A, B and C can be formulated in Eq. (3.12) and (3.13) (see also, Seo et al. 2017, for the details of the derivation of the train-FD).
When $a_\mathrm{p}/N_0 < q_1^{*c}$,

$$Q(k, a_\mathrm{p}) = \begin{cases} \dfrac{kl - \mu a_\mathrm{p}}{t_{\mathrm{b}0} + l/v_\mathrm{f} - \mu N_0}, & \text{if } \mu a_\mathrm{p}/l \le k < k_1^c \\ \dfrac{kl}{t_{\mathrm{b}0} + l/v_\mathrm{f}}. & \text{if } k_1^c \le k < k_1^{*c} \end{cases} \tag{3.12a}$$

$$Q(k, a_\mathrm{p}) = \begin{cases} -\dfrac{\delta l}{(l-\delta)t_{\mathrm{b}0} + \tau l}(k - k_1^{*c}) + q_1^{*c}, & \text{if } k_1^{*c} \le k < k_2^c \\ -\dfrac{\delta l}{(l-\delta)(t_{\mathrm{b}0} - \mu N_0) + \tau l}(k - k_2^{*c}) + q_2^{*c}. & \text{if } k \ge k_2^c \end{cases} \tag{3.12b}$$

When $a_\mathrm{p}/N_0 \ge q_1^{*c}$,

$$Q(k, a_\mathrm{p}) = \begin{cases} \dfrac{kl - \mu a_\mathrm{p}}{t_{\mathrm{b}0} + l/v_\mathrm{f} - \mu N_0}, & \text{if } \mu a_\mathrm{p}/l \le k < k_2^{*c} \\ -\dfrac{\delta l}{(l-\delta)(t_{\mathrm{b}0} - \mu N_0) + \tau l}(k - k_2^{*c}) + q_2^{*c}. & \text{if } k \ge k_2^{*c} \end{cases} \tag{3.13}$$

Where

$$q_1^{*c} = \frac{1}{t_{\mathrm{b}0} + \delta/v_\mathrm{f} + \tau}, \tag{3.14a}$$

$$k_1^{*c} = \frac{t_{\mathrm{b}0} + l/v_\mathrm{f}}{(t_{\mathrm{b}0} + \delta/v_\mathrm{f} + \tau)l}, \tag{3.14b}$$

$$q_2^{*c} = \frac{1-\mu a_\mathrm{p}}{t_{\mathrm{b}0}+\delta/v_\mathrm{f}+\tau-\mu N_0}, \tag{3.14c}$$

$$k_2^{*c} = \frac{(1-\mu a_\mathrm{p})(t_{\mathrm{b}0}+l/v_\mathrm{f}-\mu N_0)}{(t_{\mathrm{b}0}+\delta/v_\mathrm{f}+\tau-\mu N_0)l} + \frac{\mu a_\mathrm{p}}{l}, \tag{3.14d}$$

$$k_1^c = \frac{a_\mathrm{p}}{N_0 l}(t_{\mathrm{b}0}+l/v_\mathrm{f}), \tag{3.14e}$$

$$k_2^c = \left(q_1^{*c} + \frac{\delta l k_1^{*c}}{(l-\delta)t_{\mathrm{b}0}+\tau l} - \frac{a_\mathrm{p}}{N_0}\right) \cdot \frac{(l-\delta)t_{\mathrm{b}0}+\tau l}{\delta l}. \tag{3.14f}$$

Model A is expressed by Eq. (3.13) by taking $\mu = 0$, $N_0 = 0$ and $t_{\mathrm{b}0} = t_{\mathrm{bcon}}$. Model B is also expressed by Eq. (3.13) by taking $N_0 = 0$. For model C, the equation has to be separately written, depending on the relation between a_p/N_0 and q_1^{*c}.

Eq. (3.12) actually describes the situation when a_p is not large enough to force a condition in which the dwell time is always larger than $t_{\mathrm{b}0}$. More specifically, when $k_1^c \le k \le k_2^c$, the dwell time $t_\mathrm{b} = t_{\mathrm{b}0}$, which implies that operation under this condition can guarantee the dwell time is not extended due to passenger influence. While out of this range, dwell time would be extended either because trains in operation are insufficient or abundant (train bunching). On the contrary, Eq. (3.13) describes the situation when a_p is large enough so that dwell time is always larger than $t_{\mathrm{b}0}$.

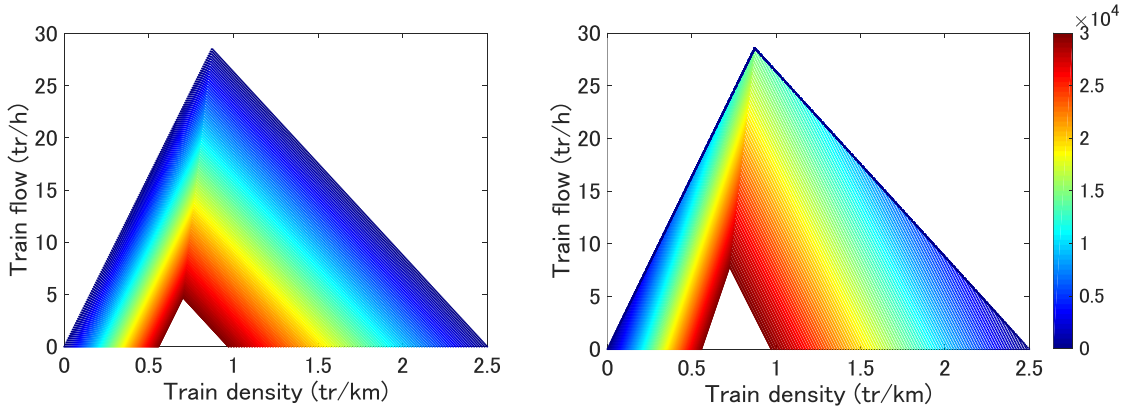
For a better understanding of passenger influence on train flow, two numerical examples of the train-FDs for model B and C are given, as shown in Fig. 3.4, based on the parameters in Table 3.1. From the comparison of Fig. 3.4a and 3.4b, it can be observed that under the same a_p, model C can achieve higher train flow due to a relatively short dwell time.

Table 3.1 Parameters used in the numerical example

Parameter	Value
$t_{\mathrm{b}0}$, N_0, μ	$30/3600\ h$, $500\ pax$, $0.1/3600\ h/pax$
l, v_f, δ, τ	$1.5\ km$, $40\ km/h$, $0.4\ km$, $1/60\ h$
a_p	$[0,\ 30000]\ pax/h$

3.3 Model calibration and evaluation

In this section, the proposed models will be calibrated and evaluated employing data from the Boston red line.



(a) Model B (b) Model C

Fig. 3.4 Numerical examples of the train-FDs.

3.3.1 Calibration based on enumeration method

The three models are calibrated using the enumeration method. Specifically, by pre-selecting the variation range for each parameter, a parameter set is built by accounting for all possible combinations. The parameter set is then sorted based on root mean square error (RMSE), which is calculated by Eq. (3.15):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(q_i^m - q_i^e)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_i^m - Q(k_i^m,\, a_{p,i}^m)\right)^2}, \tag{3.15}$$

where q_i^m, k_i^m, $a_{p,i}^m$ are, respectively, train flow, density, and passenger arrival rate, measured from the Boston red line train-FD (southbound), while q_i^e is the train flow estimated from the three models (calculated by Eq. (3.12) and (3.13)).

Finally, the parameter vector with the smallest RMSE will be regarded as the best fitting one, denoted by $\hat{\boldsymbol{\beta}}$. Since almost all of the data points from the Boston red line were found to lie in the free-flow regime of the train-FD (as shown in Fig. 3.2), parameters δ and τ cannot be calibrated using this data source. In addition, the average distance between adjacent stations can be measured from Google maps as $l = 1.2\ km$. Therefore, there are two (t_{bcon}, v_{f}), three (t_{b0}, μ, v_{f}), and four (t_{b0}, N_0, μ, v_{f}) parameters that need to be calibrated for models A, B, and C, respectively. The variation range and number of combinations for the three models are listed in Table 3.2, where Δ is the increment.

As can be seen from Table 3.2, the number of combinations increased from around eight thousand, to nearly twenty million, when two more parameters were added. The smallest RMSE and $\hat{\boldsymbol{\beta}}$ for the three models are listed in Table 3.3. From Table 3.3 it can be found that model A has a larger RMSE than both model B and model C. Also, in model A, the

Table 3.2 Variation range of the calibrated parameters

Model A: $116 \times 71 = 8236$	Model B: $56 \times 71 \times 100 \approx 4 \times 10^5$	Model C: $56 \times 71 \times 100 \times 50 \approx 2 \times 10^7$
$t_{\text{bcon}} \in [5,\ 120], \Delta = 1\ s$ $v_{\text{f}} \in [10,\ 80], \Delta = 1\ km/h$	$t_{\text{b0}} \in [5,\ 60], \Delta = 1\ s$ $v_{\text{f}} \in [10,\ 80], \Delta = 1\ km/h$ $\mu \in [0.01,\ 1],$ $\Delta = 0.01\ s/pax$	$t_{\text{b0}} \in [5,\ 60], \Delta = 1\ s$ $v_{\text{f}} \in [10,\ 80], \Delta = 1\ km/h$ $\mu \in [0.01,\ 1],$ $\Delta = 0.01\ s/pax$ $N_0 \in [5,\ 250], \Delta = 5\ pax$

values of $\hat{t}_{\text{bcon}}$ and $\hat{v}_{\text{f}}$ appear to be rather unrealistic, considering the actual travel experience of the authors, in a practical subway systems. On the other hand, identical RMSE and $\hat{\mu}$, and similar values of $\hat{v}_{\text{f}}$ and $\hat{N}_0$ (model B can be considered as a special case of model C with $N_0 = 0$) all imply that model B and C perform quite similarly, yet better than model A.

Table 3.3 Calibration results

Model A: $RMSE = 1.041$	Model B: $RMSE = 0.940$	Model C: $RMSE = 0.940$
$\hat{t}_{\text{bcon}} = 101\ s$ $\hat{v}_{\text{f}} = 80\ km/h$	$\hat{t}_{\text{b0}} = 27\ s$ $\hat{v}_{\text{f}} = 38\ km/h$ $\hat{\mu} = 0.16\ s/pax$	$\hat{t}_{\text{b0}} = 18\ s$ $\hat{v}_{\text{f}} = 35\ km/h$ $\hat{\mu} = 0.16\ s/pax$ $\hat{N}_0 = 5\ pax$

3.3.2 Sensitivity analysis and performance evaluation

To show the sensitivity of the parameters, the contour maps of the RMSE with respect to the variation of parameters for the three models were drawn in Fig. 3.5 - 3.6. On the contour maps, color is used to represent the value of the RMSE.

In Fig. 3.5, low RMSE values appear as a blue strip. Because t_{bcon} and l/v_{f} in Eq. (3.13) are linearly combined, the calibration just ensures the sum of these two terms remains constant. In other words, only two parameters cannot be separately determined by minimizing the RMSE of the train flow. However, for model B and C, RMSEs derived from the relations between $\mu - t_{\text{b0}}$ and $l/v_{\text{f}} - \mu$ minimized in the enclosed area (blue ellipses in Fig 3.6b and 3.6c, Fig 3.7e and 3.7f), which yielded realistic estimates for the parameters. Here, other parameters in Fig 3.6 and 3.7 took the corresponding values of $\hat{\boldsymbol{\beta}}$ in Table 3.3. From Fig 3.7b it can be observed that the RMSE is not sensitive to the change of N_0, since l/v_{f} dominates the denominator of Eq. (3.12a) when $t_{\text{b0}} = \hat{t}_{\text{b0}}$ and $\mu = \hat{\mu}$. In addition, the RMSE in the

blank area of Fig 3.7c is not available because the product of μ and N_0 approaches $\hat{t}_{b0} + l/\hat{v}_f$, causing the denominator of Eq. (3.12a) to become zero.

To compare the performance of the three models, AIC (Akaike, 1998) was adopted to assess the trade-off between goodness of fit and parsimony. The fitness of a model generally increases with the number of free parameters. However, a simple form of the model is always desirable and over-fitting should be avoided. A model with a smaller AIC value is better, and the AIC is generally defined as Eq. (3.16):

$$AIC = -2\ln(ML) + 2P, \tag{3.16}$$

where ML is the maximum likelihood (MLE) and P is the number of estimated parameters. Here we assume that the measured train flow q^m obeys the normal distribution of $N(\hat{q}^e, \sigma^2)$. Then, ML can be derived by Eq. (3.17) - (3.19).

$$ML\left(\hat{\boldsymbol{\beta}}\right) = \prod_{i=1}^{n} f\left(q_i^m | \hat{\boldsymbol{\beta}}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2\right). \tag{3.17}$$

$$\hat{q}_i^e = Q\left(k_i^m, a_{p,i}^m | \hat{\boldsymbol{\beta}}\right). \tag{3.18}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2. \tag{3.19}$$

Finally, submitting Eq. (3.17) - (3.19) into Eq. (3.16), the AIC can be obtained by Eq. (3.20). The calculation results for the AIC of the three models are listed in Table 3.4. The calculation results show that model B produced the smallest AIC, although it was quite similar to model C. The AICs of both models were much smaller than that of model A. Therefore, it can be concluded that models that consider the influence of passenger congestion on railway operation perform much better.

$$AIC = n\ln(2\pi) + n\ln\left(\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2/n\right) + n + 2P. \tag{3.20}$$

Table 3.4 Calculation results of model evaluation

	Model A	Model B	Model C
$n\ln\left(\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2/n\right)$	116.63	-180.58	-180.62
$2P$	4	6	8
AIC	4286.64	3991.42	3993.39

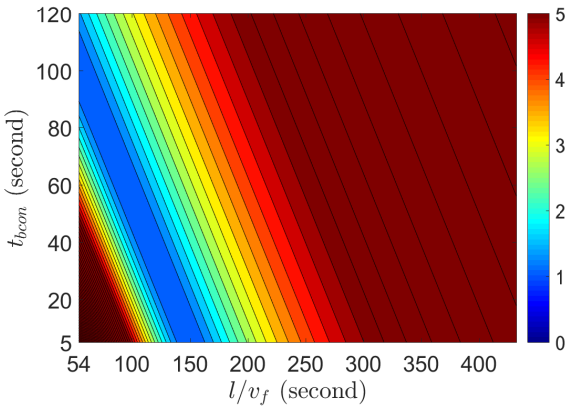
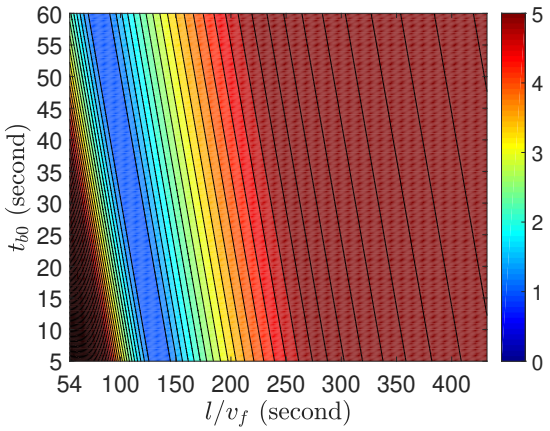
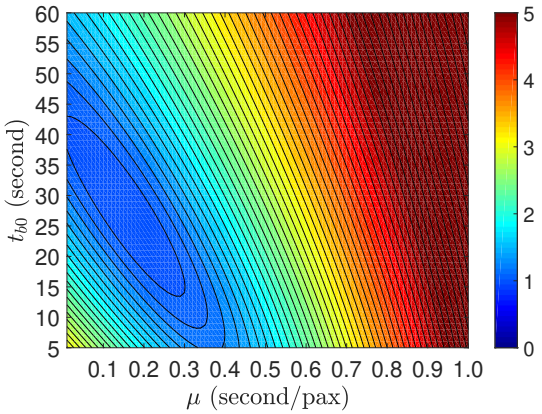


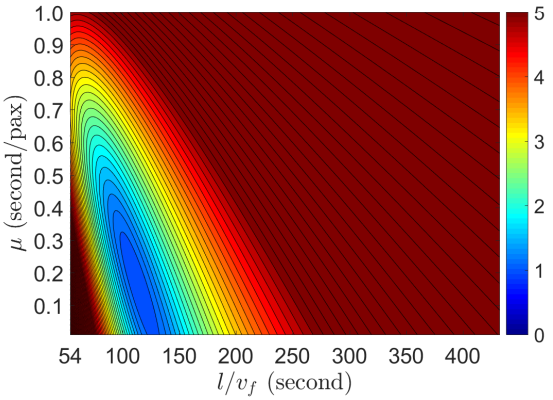
Fig. 3.5 Parameter sensitivity of model A



(a) RMSE: l/v_{f} and t_{b0}

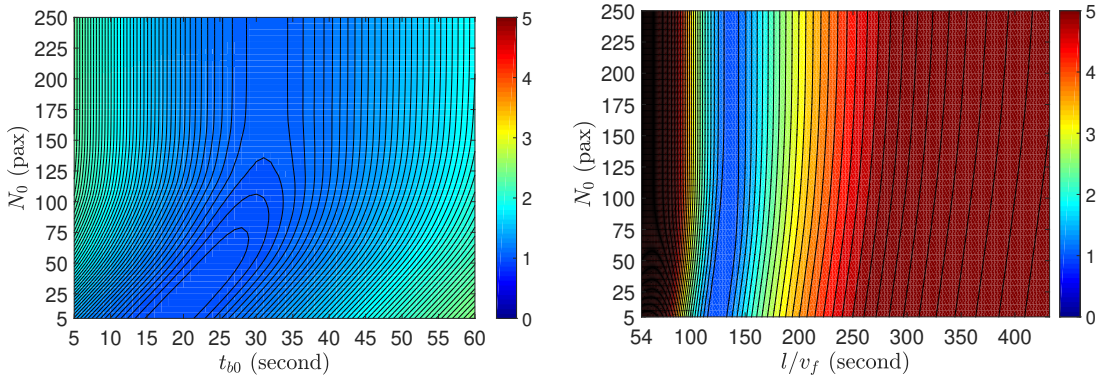


(b) RMSE: μ and t_{b0}



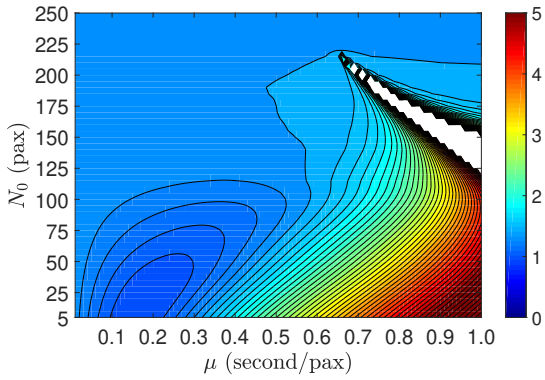
(c) RMSE: l/v_{f} and μ

Fig. 3.6 Parameter sensitivity of model B.

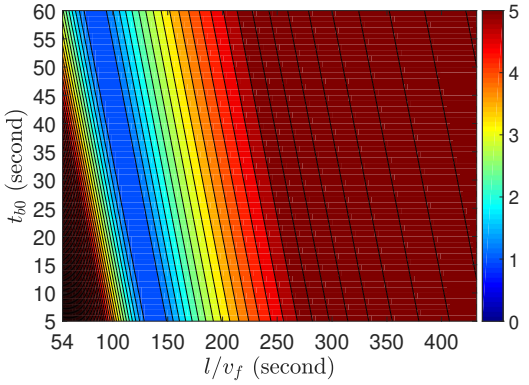


(a) RMSE: t_{b0} and N_0

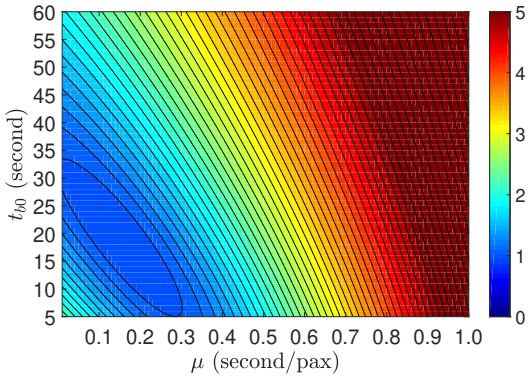
(b) RMSE: l/v_{f} and N_0



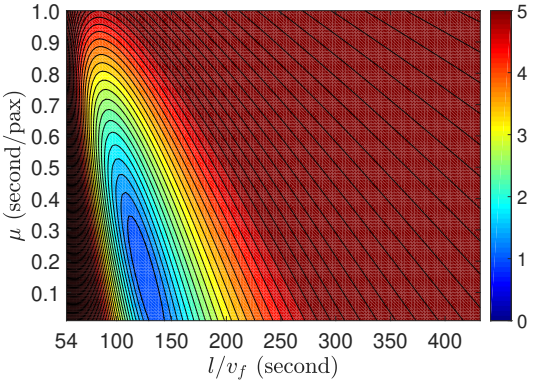
(c) RMSE: μ and N_0



(d) RMSE: l/v_{f} and t_{b0}



(e) RMSE: μ and t_{b0}



(f) RMSE: l/v_{f} and μ

Fig. 3.7 Parameter sensitivity of model C.

3.4 Brief summary

This chapter investigated three variants of train-FDs, which are based on different assumptions of train dwell time, by employing operation data from the Boston red line. The conclusion indicates that passenger congestion influence on operation of urban rail transit was significant, and that train dwell time monotonically increasing with the number of boarding passengers, might be sufficient to describe this influence. However, since the Boston subway system lacked data in high-frequency operation, open questions remain for future work on the verification of congested regime of the train-FD.

Chapter 4

Dynamics of urban rail transit

The train-FD introduced in Chapter 3 only described the steady state operation of trains. However, during the morning commute period, passenger arrivals are undoubtedly not evenly distributed, the conventional dispatch frequencies of trains are also not a constant value. Therefore, rail transit system during peak hours are essentially under an unsteady state which means the density and flow of trains along the railway line are time-dependent and space-dependent. This chapter aims to describe the dynamic performance of rail transit system with two approaches, one is an fluid-based approach based on the widely used exit-flow model (Carey and McCartney, 2004; Merchant and Nemhauser, 1978), and the other one is a newly developed particle-based approach. The two approaches are verified by a micro-simulation.

4.1 Fluid-based approach

Fluid-based approach in this study describes the dynamic performance of the railway system as a function of continuous time, and it is essentially an expansion of exit-flow model applied in rail transit system. In fact, Seo et al. (2017) first proposed a dynamic model based on train-FD exit-flow function. This study considered the more general form of their model. The advantage of this approach is that it is mathematically tractable for analysis of a dynamic, large-scale, and complex transportation systems. There have been some applications of this approach to avoid congestion (Daganzo, 2007) or to design pricing scheme (Fosgerau, 2015; Geroliminis and Levinson, 2009) for road traffic.

4.1.1 Formulation

Consider the railway system as an input-output system, trains and passengers enter the system to finish the trip and leave the system. The dynamic performance of the system are mainly

described by its instantaneous density $k(t)$ and flow $q(t)$. The input and output of the system and their cumulative numbers are summarized as follows:

- $a(t)$: train inflow rate entering the system at time t,
- $A(t)$: cumulative number of trains that have entered the system at time t,
- $a_{\mathrm{p}}(t)$: passengers' arrival rate at their origin station at time t,
- $A_{\mathrm{p}}(t)$: cumulative number of passengers that have entered the origin station at time t,
- $d(t)$: train outflow rate leaving the system at time t,
- $D(t)$: cumulative number of trains that have left the system at time t,
- $d_{\mathrm{p}}(t)$: passengers' departure rate from their destination station at time t,
- $D_{\mathrm{p}}(t)$: cumulative number of passengers that have left the destination station at time t,
- $T(t)$: travel time of a train/passenger from the origin station to the destination station when leaving the system at time t. Thus, the time entering the system becomes $t - T(t)$.

According to the above definitions, the cumulative variables can be written as:

$$A(t) = \int_0^t a(s)\,ds, \tag{4.1a}$$

$$D(t) = \int_0^t d(s)\,ds, \tag{4.1b}$$

$$A_{\mathrm{p}}(t) = \int_0^t a_{\mathrm{p}}(s)\,ds, \tag{4.1c}$$

$$D_{\mathrm{p}}(t) = \int_0^t d_{\mathrm{p}}(s)\,ds. \tag{4.1d}$$

If denote the total length of railway line as L, according to the conservation law, the density and its derivative with respect to time can be respectively written as:

$$k(t) = \frac{A(t) - D(t)}{L}, \tag{4.2}$$

$$\frac{dk(t)}{dt} = \frac{1}{L}\left[a(t) - d(t)\right]. \tag{4.3}$$

Meanwhile, to express the instantaneous average train flow $q(t)$ as a function of inflow $a(t)$ and outflow $d(t)$, the spatial heterogeneity of railway system is simplified so that for

any $x \in [0,L]$, either $q(0,t) \leq q(x,t) \leq q(L,t)$ or $q(L,t) \leq q(x,t) \leq q(0,t)$ is met. Since $q(0,t) = a(t)$ and $q(L,t) = d(t)$, the $q(t)$ can therefore be represented as:

$$q(t) = \frac{1}{L}\int_0^L q(x,t)dx = (1-\zeta)a(t) + \zeta d(t) \ (0 \leq \zeta \leq 1). \tag{4.4}$$

Where ζ is a dimensionless parameter between zero and one to express the weight of inflow and outflow when deriving the mean value.

Now, the task of a dynamic rail transit model is to solve the system performance (i.e, $k(t)$ and $q(t)$) and output ($d(t)$, $d_{\mathrm{p}}(t)$) when inputs ($a(t)$, $a_{\mathrm{p}}(t)$) are given. The original exit-flow model assumed that the exit function describing the steady state situation could roughly predict outflow in the dynamic case (see Daganzo, 2007; Geroliminis and Levinson, 2009; Seo et al., 2017). By employing the train-FD as the exit function, the train outflow $d(t)$ can be approximated by

$$d(t) = Q\left(k(t), a_{\mathrm{p}}(t)\right). \tag{4.5}$$

However, this study assumed that steady state train-FD could roughly predict the system average flow in the dynamic case as:

$$q(t) = Q\left(k(t), a_{\mathrm{p}}(t)\right). \tag{4.6}$$

Then, by substituting Eq. (4.6) and (4.4) into Eq. (4.3), the dynamics of system is described by:

$$\frac{dk(t)}{dt} = \frac{1}{L\zeta}\left[a(t) - Q\left(k(t), a_{\mathrm{p}}(t)\right)\right]. \tag{4.7}$$

Here, Eq. (4.7) is a first order ordinary differential equation and can be solved if $a(t)$, $a_{\mathrm{p}}(t)$ and initial condition $k(0)$ are given.

In addition, since FIFO is assumed in this study, the cumulative departures of trains and passengers can also be written as:

$$D(t) = A\left(t - T(t)\right), \tag{4.8}$$

$$D_{\mathrm{p}}(t) = A_{\mathrm{p}}\left(t - T(t)\right). \tag{4.9}$$

When $d(t)$ is solved from Eq. (4.7), $D(t)$ can be integrated from $d(t)$ and $T(t)$ can then be calculated by Eq. (4.8). Finally, $D_{\mathrm{p}}(t)$ and $d_{\mathrm{p}}(t)$ can be obtained by Eq. (4.9) and its derivative form if $T(t)$ and $A_{\mathrm{p}}(t)$ are known.

So far, the dynamic model based on fluid approach was built. Given the environmental parameters (such as L and parameters used in train-FD) and time-dependent input information

(i.e, $a(t)$ and $a_{\mathrm{p}}(t)$), the dynamic performance and output of the system can be obtained using the proposed model. In addition, the dynamic model by Seo et al. is a special case of this formulation when $\zeta = 1$.

4.1.2 Numerical examples

To intuitively grasp the characteristics of the proposed dynamic model, numerical examples are presented under two types of train inflow $a(t)$ pattern. The first one is the simplest constant case, and the second one is a conventional two-step case. The passenger arrival rate a_{p} is set to obey a uni-modal distribution. The train-FD in Eq. (4.7) employed the model B introduced in the previous chapter. The parameter settings are listed in Table 4.1, and input information of $a(t)$ and $a_{\mathrm{p}}(t)$ are shown in Fig. 4.1, or equivalently written as:

$$a_1(t) = a_0 \tag{4.10a}$$

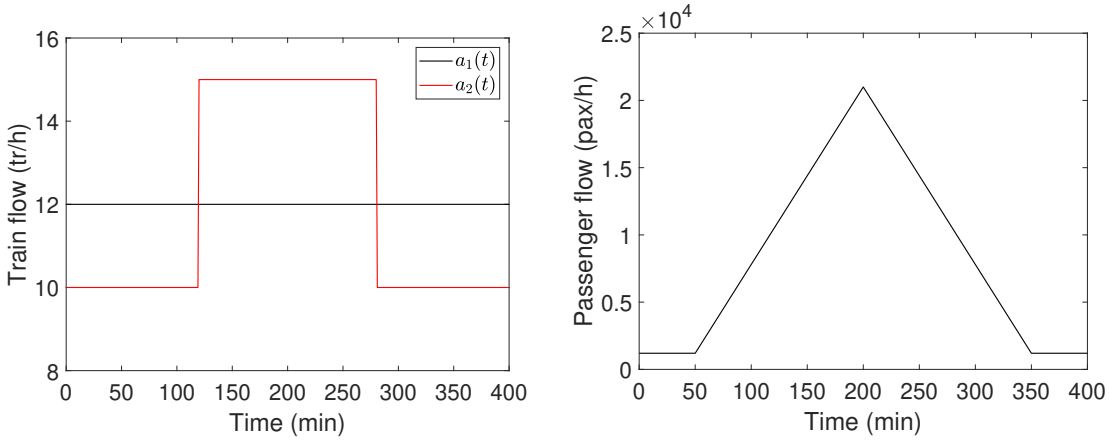
$$a_2(t) = \begin{cases} a_l, & \text{if } t \in [120, 280], \\ a_h, & \text{Otherwise}, \end{cases} \tag{4.10b}$$

$$a_{\mathrm{p}}(t) = \begin{cases} a_{\mathrm{p}}^0 + \dfrac{a_{\mathrm{p}}^m - a_{\mathrm{p}}^0}{150}(t-50), & \text{if } t \in [50, 200], \\ a_{\mathrm{p}}^0 + \dfrac{a_{\mathrm{p}}^m - a_{\mathrm{p}}^0}{150}(350-t), & \text{if } t \in (200, 350], \\ a_{\mathrm{p}}^0, & \text{Otherwise}, \end{cases} \tag{4.10c}$$

Table 4.1 Parameter settings for numerical examples (fluid).

Parameter	Value	Parameter	Value
l	1.2 km	ζ	0.5
L	18 km	a_0	12 tr/h
v_{f}	40 km/h	a_l	10 tr/h
$t_{\mathrm{b}0}$	20 sec	a_h	15 tr/h
μ	0.1 sec/pax	a_{p}^0	1200 pax/h
δ	0.4 km	a_{p}^m	21000 pax/h
τ	1.0 min		

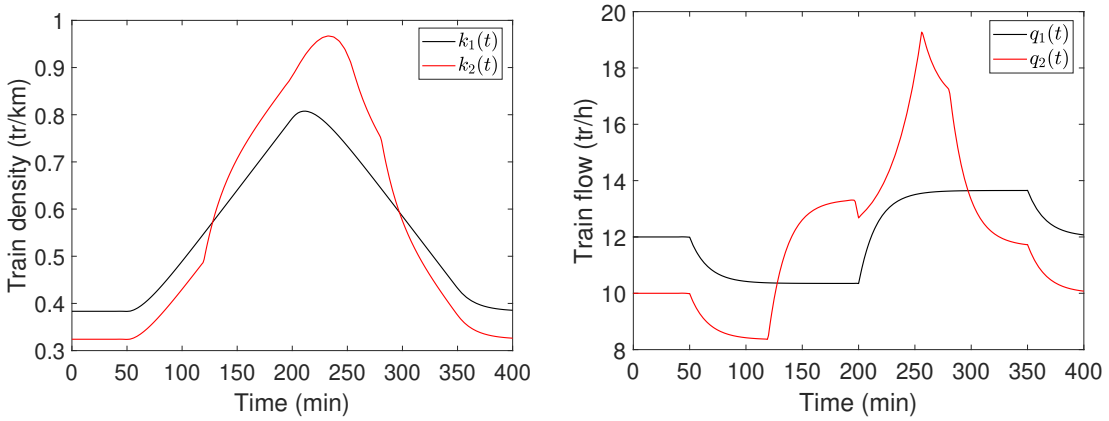
In fact, one may find that cumulative dispatched trains in the two cases are the same at the end of experiment time. This setting aims to fairly check the rationality of employing two-step timetable in practice during morning commute. Fig. 4.2 presented the dynamic performance of railway system calculated from the proposed model. It can be observed from



(a) Train inflow $a(t)$ (b) Passenger arrival rate $a_{\mathrm{p}}(t)$

Fig. 4.1 Input information for numerical examples (fluid).

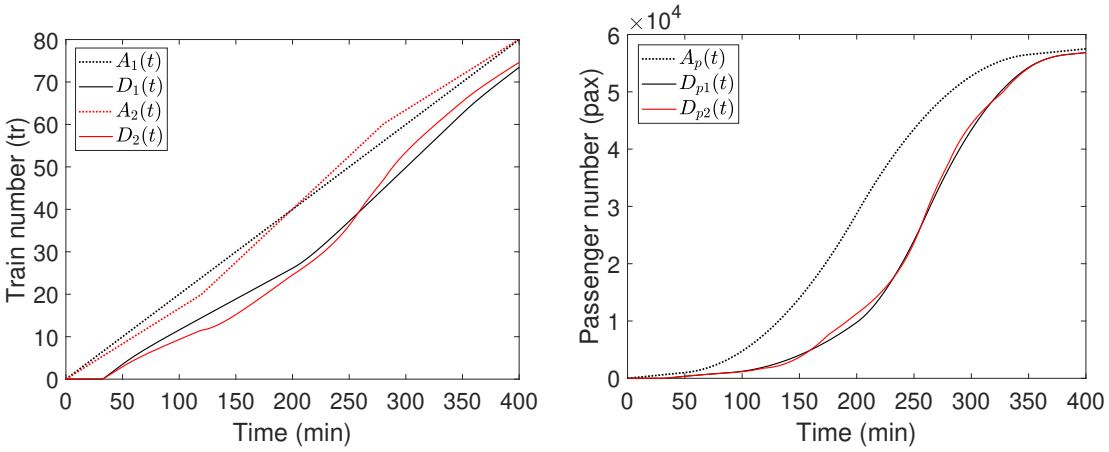
Fig. 4.2a that the evolution of train density in both cases had the unique peak appearing at almost the same time with the peak of $a_{\mathrm{p}}(t)$. Meanwhile, due to sudden change of dispatch frequency, the change rate of density was significant higher in the two-step case. Regarding the system average flow $q(t)$ in Fig. 4.2b, it is easier to see from the first case that the sudden changes of $dq(t)/dt$ were exactly simultaneous with those of $da_{\mathrm{p}}(t)/dt$. Further, the change direction of this two variables are opposite, which means when the derivatives of $a_{\mathrm{p}}(t)$ started to increase (e.g., at 50 *min* and 350 *min*), derivatives of $q(t)$ would start to decrease, or vice versa. When sudden changes emerged in $a_2(t)$, $q(t)$ also quickly reacted to these sudden changes. In addition, the increase of $q_2(t)$ around 250 *min* was mainly due to entering the congested regime of train-FD.



(a) Density of railway system $k(t)$ (b) Average flow of railway system $q(t)$

Fig. 4.2 Dynamic performance of railway system (fluid).

Fig. 4.3 showed the cumulative curves of trains and passengers. It can again be confirmed from Fig. 4.3a that the total dispatched trains at 400 *min* are the same for the two cases. The density $k(t)$ and travel time $T(t)$ are the vertical and horizontal distance between $A(t)$ and $D(t)$, respectively. Similarly, the vertical distance between $A_{\mathrm{p}}(t)$ and $D_{\mathrm{p}}(t)$ in Fig. 4.3b represents the passenger number in the railway system. It can be found that passenger numbers around 200 *min* and 300 *min* are slightly smaller in the second case. This confirmed that compared to a constant dispatch frequency, two-step timetable might be more efficient to carry passengers arriving with a single peak.



(a) Cumulative curves of trains (b) Cumulative curves of passengers

Fig. 4.3 Cumulative curves of trains and passengers (fluid).

The travel time $T(t)$ for two cases are depicted in Fig. (4.4). Under a uni-modal a_{p}, it can be seen that when $a(t)$ is constant, $T(t)$ is also uni-modal. However, when $a(t)$ has sudden changes, $T(t)$ becomes rather unstable with multiple small peaks.

4.2 Particle-based approach

Particle-based approach in this study describes the dynamic performance of railway system by each cruising train. This approach is developed based on two assumptions:

- FIFO principle is strictly complied.
- The whole trip of each train can be approximated as a steady state [1].

[1]The ground of this assumption is that by taking the macroscopic average variables like density and flow, their sudden changes during the trip can be flattened, and the traffic states represented by the averaged variables can approximately be regarded as the steady state.

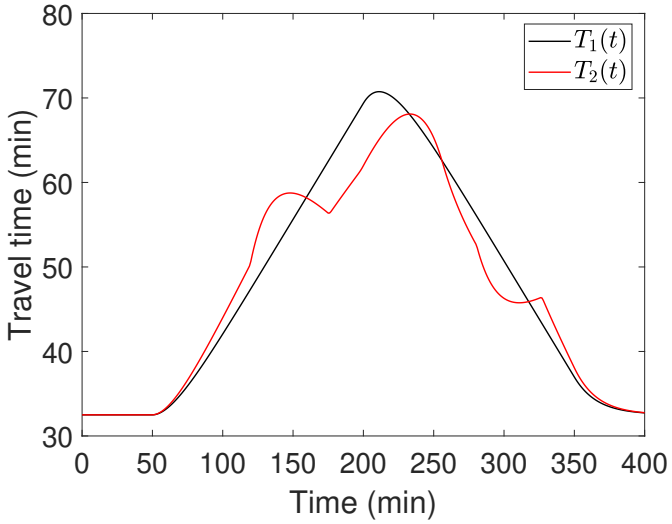


Fig. 4.4 Travel time of trains and passengers (fluid).

4.2.1 Formulation

Consider one trip of a train i from the origin station to the destination station. Denote $H_{\mathrm{a}}(i)$ and $H_{\mathrm{d}}(i)$ are the time headways of train $i-1$ and train i (simply referred to as the headway of train i hereinafter) at the origin station and the destination station, respectively. $T(i)$ is the travel time of train i and $a_{\mathrm{p}}(i)$ represents the average passenger arrival rate during $T(i)$. In fact, one can imagine that if $H_{\mathrm{a}}(i)$, $H_{\mathrm{d}}(i)$ and $T(i)$ are known for any dispatched train $i \in I$, the dynamics of the whole railway system has been figured out from a macroscopic view. One can also understand that the travel time $T(i)$ is mainly determined by the dwell time (dwell time further determined by $a_{\mathrm{p}}(i)$ and headway with its preceding train $i-1$) and bunching deceleration (if any) if the cruising speed is pre-given and fixed. Then, it is natural to employ train-FD developed from microscopic operation rule to relate travel time, headway and passenger arrival rate. To do this, one has to first express the density and flow by headway and travel time. This is not difficult if macroscopically considering the whole trip of a train. More specifically, the average flow $q(i)$ experienced by train i (or equivalently average headway $H(i)$) can be written as:

$$q(i) = \frac{1}{H(i)} = \left(\frac{H_{\mathrm{a}}(i) + H_{\mathrm{d}}(i)}{2} \right)^{-1}. \tag{4.11}$$

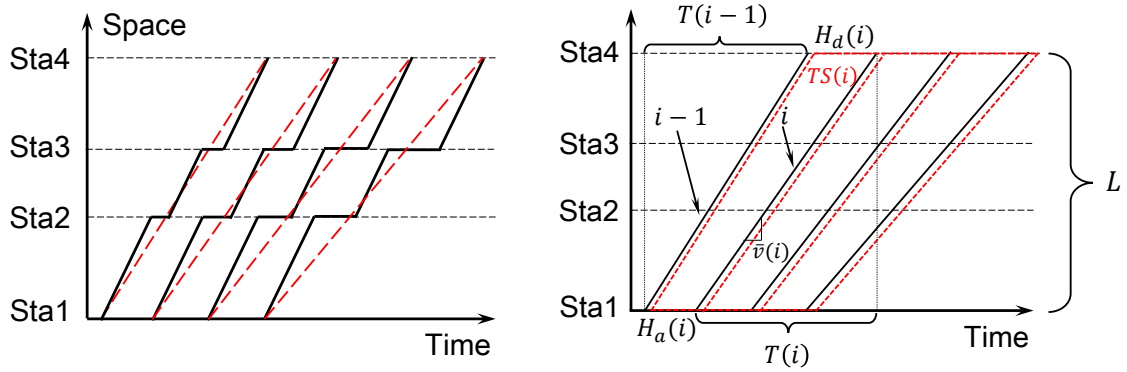
Similarly, the average density $k(i)$ (or equivalently average spacing $s(i)$) can be written as:

$$k(i) = \frac{1}{s(i)} = \left(\frac{s_{\mathrm{a}}(i) + s_{\mathrm{d}}(i)}{2} \right)^{-1}, \tag{4.12}$$

$$s_{\mathrm{a}}(i) = \bar{v}(i)H_{\mathrm{a}}(i) = \frac{L}{T(i)}H_{\mathrm{a}}(i), \tag{4.13a}$$

$$s_{\mathrm{d}}(i) = \bar{v}(i)H_{\mathrm{d}}(i) = \frac{L}{T(i)}H_{\mathrm{d}}(i). \tag{4.13b}$$

Where $\bar{v}(i)$ is the average travelling speed of train i. For a more intuitive understanding, the definitions of these variables are illustrated in Fig. 4.5.



(a) An example of trajectories of trains

(b) Simplified trajectories of trains

Fig. 4.5 An example of variable definitions in particle-based approach.

Fig. 4.5a gives an example of trajectories of trains operating on a railway line with four stations from "Sta1" to "Sta4". When the morning commute starts, passenger arrival rate increase with time so that the dwell time of trains is gradually extended at each station as can be understood from Fig. 4.5a. To describe the average density and flow, the trajectories are simplified by connecting the start and end point on the time-space diagram. Then, the whole trip of a train i can macroscopically be described by its average travelling speed $\bar{v}(i)$ or travel time $T(i)$ as shown in Fig. 4.5b[2].

Now, according to the first assumption, there is a relation between headways and travel time for two succeeding trains $i-1$ and i:

$$H_{\mathrm{d}}(i) - H_{\mathrm{a}}(i) = T(i) - T(i-1). \tag{4.14}$$

Meanwhile, according to the second assumption, the average flow $q(i)$ can be approximated by

$$q(i) = Q\left(k(i), a_{\mathrm{p}}(i)\right). \tag{4.15}$$

[2]In addition, if the time-space diagram is delimited as the trapezium in red dotted line like $TS(i)$ in Fig. 4.5b, one may find that the definitions of $q(i)$ and $k(i)$ in Eq. (4.11) and Eq. (4.12) are consistent with Edie's definition in a time-space diagram.

Then, substitute Eq. (4.11 - 4.13) into Eq. (4.15) and employ Model B of train-FD, Eq. (4.15) can be explicitly rewritten as two equations corresponding to free-flow and congested regime of train-FD, respectively:

$$T(i) = \frac{L}{l}\left[\frac{(H_{\mathrm{a}}(i)+H_{\mathrm{d}}(i))}{2}\cdot \mu a_{\mathrm{p}}(i) + t_{\mathrm{b0}} + l/v_{\mathrm{f}}\right], \tag{4.16a}$$

$$T(i) = \frac{L}{l}\left[\frac{(H_{\mathrm{a}}(i)+H_{\mathrm{d}}(i))}{2}\cdot \left(\frac{l}{\delta} - \mu a_{\mathrm{p}}(i)\left(\frac{l}{\delta}-1\right)\right) - \left(\frac{l}{\delta}-1\right)t_{\mathrm{b0}} - \frac{l}{\delta}\tau\right]. \tag{4.16b}$$

The implication of Eq. (4.16a) derived from free-flow regime of train-FD is very straightforward: the travel time of a train is determined by the sum of dwell time (minimum dwell time plus average dwell time increase for passenger exchanging) and cruising time l/v_{f}. Finally, substitute Eq. (4.14) into Eq. (4.16), the travel time $T(i)$ (or $H_{\mathrm{d}}(i)$) can be expressed as a function of $T(i-1)$, $H_{\mathrm{a}}(i)$ and $a_{\mathrm{p}}(i)$. This indicates that when $H_{\mathrm{a}}(i)$ and $a_{\mathrm{p}}(i)$ are given for all $i \in I$, if an initial travel time $T(1)$ is known, the $H_{\mathrm{d}}(i)$ and $T(i)$ for all the other trains can be solved.

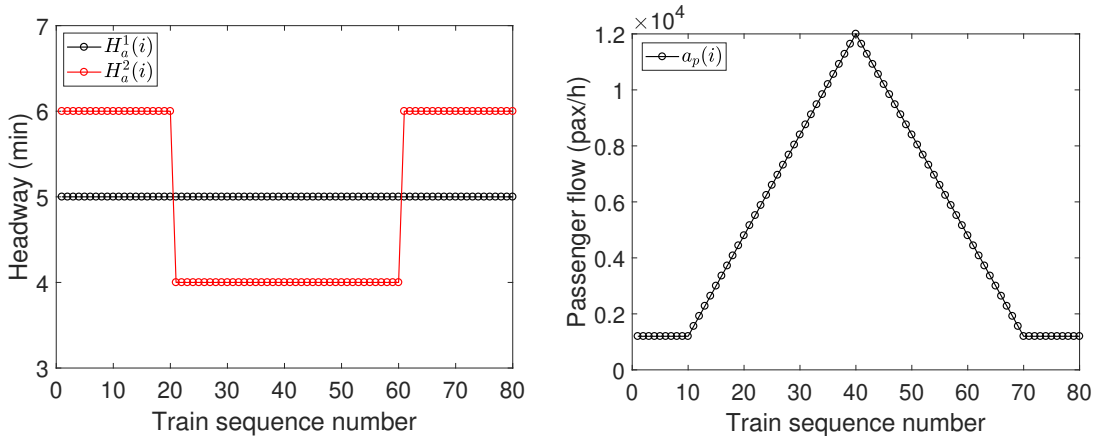
So far, the dynamic model based on particle-based approach was built. Given the environmental parameters (including initial conditions) and time-dependent input information (i.e, $H_{\mathrm{a}}(i)$ and $a_{\mathrm{p}}(i)$), the dynamics of the railway system can be solved using the proposed model.

4.2.2 Numerical examples

Similar as the fluid-based approach, the parameters used for numerical examples are listed in Table 4.2. Since the second assumption requires an approximate steady state, the trip length was shortened and the passenger congestion influence was also weakened for the following numerical examples. The input information is given by Fig. 4.6.

Table 4.2 Parameter settings for numerical examples (particle).

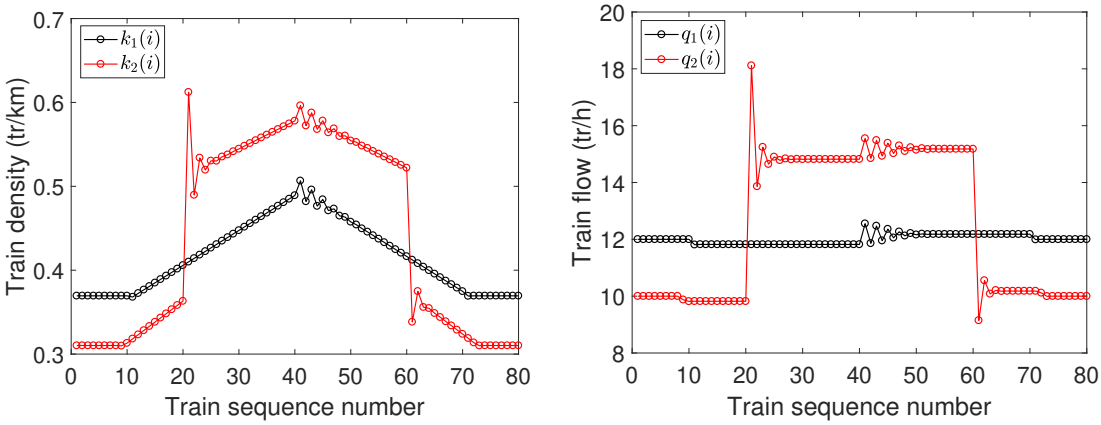
Parameter	Value	Parameter	Value
l	1.2 km	H_{a1}	5 min
L	7.2 km	$\min\{H_{\mathrm{a}}^1\}$	4 min
v_{f}	40 km/h	$\max\{H_{\mathrm{a}}^2\}$	6 min
t_{b0}	20 sec	H_{a}^2 peak number	$[21,40]$
μ	0.05 sec/pax	$\min\{a_{\mathrm{p}}(i)\}$	1200 pax/h
δ	0.4 km	$\max\{a_{\mathrm{p}}(i)\}$	12000 pax/h
τ	1.0 min	$a_{\mathrm{p}}(i)$ peak period	$[50,350]$



(a) Train headway $H_{\mathrm{a}}(i)$ (b) Passenger arrival rate $a_{\mathrm{p}}(i)$

Fig. 4.6 Input information for numerical examples (particle).

The horizontal axis of Fig. 4.6 was replaced by the train sequence number i while the total dispatched train numbers are kept the same with the previous numerical examples for fluid approach. Fig. 4.7 shows the density and flow evolution for all dispatched trains. It can be observed that the combination of high and low dispatch frequency in the second case led to significant change of both density and flow for the corresponding trains. In addition, compared to fluid-based approach, the particle-based approach is relatively vulnerable to the change of input as can be understood from the fluctuations at the time when input suddenly changed.

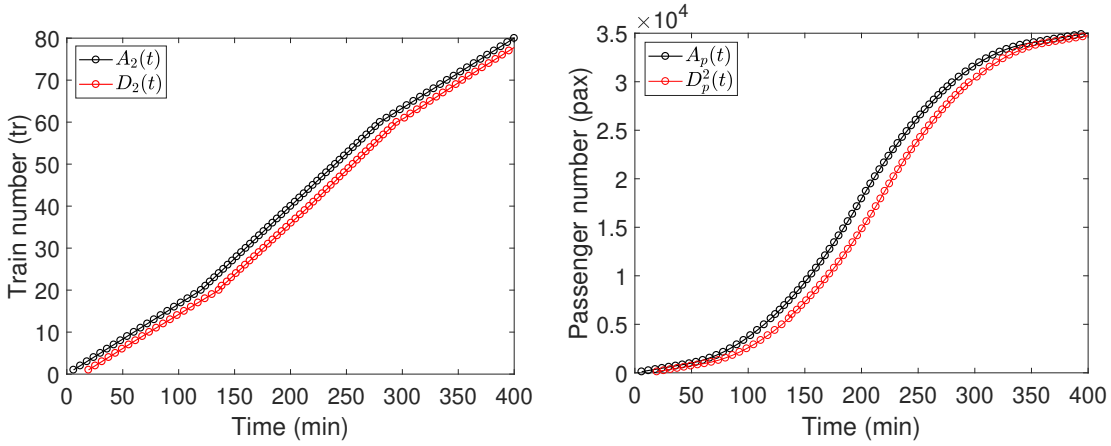


(a) Train density $k(i)$ (b) Train flow $q(i)$

Fig. 4.7 Dynamic performance of railway system (particle).

Fig. 4.8 gives an example of cumulative trains and passengers with time only for the second case since the difference of the two cases are rather small. Although the calculation

of particle-based approach was based on each train i, the entering and departure time to/from the system can easily be inferred from $H_{\mathrm{a}}(i)$ and $H_{\mathrm{d}}(i)$. The travel time of two cases are depicted in Fig. 4.9. It can be found that different from fluid approach, travel time under high dispatch frequency is continuously smaller than that of the constant case. This result is plausible since high dispatch frequency reduced the waiting passenger number on the platform and under the same arrival rate. Therefore, the dwell time decreased, which led to the shorter travel time.



(a) Cumulative curves of trains (b) Cumulative curves of passengers

Fig. 4.8 Cumulative curves of trains and passengers (particle).

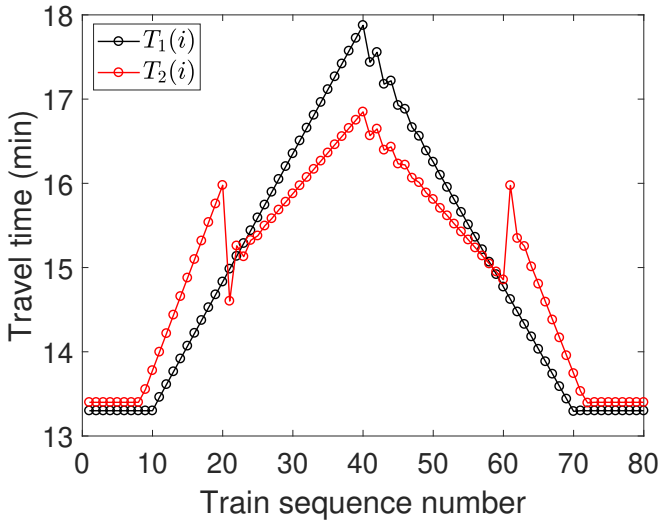


Fig. 4.9 Travel time of trains and passengers (particle).

4.3 Verification by micro-simulation

This section aims to verify the proposed dynamic models by a simple micro-simulation of urban rail transit. The operation rules of the simulation will first be introduced. Then, the simulation will be used to reproduce the situations under the same settings with the numerical examples of the two dynamic models. The results from the simulation and the dynamic models will be compared to verify whether the models can well describe the dynamics of the railway system.

4.3.1 Operation rules

The micro-simulation in this study is built based on the two rules. One determines how long a train should dwell at the stations and the other specifies how a train should cruise between the stations. Here, the simplest dwell strategy is considered: try to carry all the waiting passengers with a maximum dwell time limit. This strategy can be described as the following equation:

$$TD_i^s = \min\{TD_{\max},\ t_{b0} + \int_{TD_{i-1}^s}^{TD_i^s} a_p(t)dt\}, \tag{4.17}$$

where TD_i^s is the dwell time of train i at station s, $TD_{\max}$ is the maximum dwell time limit and takes the value of 5 *min*.

The cruising behavior again employs the Newell's simplified car-following model (Newell, 2002) as introduced in Chapter 2. For convenience, equation ruling the cruising behavior is again written here:

$$x_i(t+\Delta t) = \min\{x_i(t) + v_f \Delta t,\ x_{i-1}(t+\Delta t-\tau) - \delta\}, \tag{4.18}$$

where $i-1$ refers to the preceding train of train i. $x_i(t)$ is the position of train i at time t. Now, for a given rail transit system (i.e., parameters such as L, l, t_{b0}, v_f, δ and τ are given), input $a_p(t)$ and $H_{\mathbf{a}}(i)$ for all $i \in I$, the trajectories of all dispatched trains can be produced from the simulation.

Based on the trajectories, the instantaneous average density and flow of whole railway system can be derived as:

$$k(t) = \frac{n(t)}{L}, \tag{4.19}$$

$$q(t) = \frac{1}{L}\sum_{i=1}^{n(t)} \bar{v}_i, \tag{4.20}$$

where $n(t)$ is the number of trains within the railway system at time t, and $\bar{v}_i$ here takes the average travelling speed of train i, which means $\bar{v}_i = L/T(i)$.

4.3.2 Verification of fluid-based approach

As an example, the trajectories produced under the first case of settings in Table 4.1 is shown in Fig. 4.10. It can be observed that with the increase of a_{p} from 50 *min*, the dwell time of trains also grew. Meanwhile, the bunching phenomenon became severer and trains arriving the destination "in groups" during $[150, 300]$ *min*. This is largely due to the no-control dwell strategy (Daganzo, 2009; Xuan et al., 2011). The bunching phenomenon disappeared from around 350 *min* and system returned the initial condition in the end.

Then, comparisons of density and flow evolution are made between the dynamic model and the simulation in Fig. 4.11. It can be confirmed from the both graphs that before the peak of a_{p} (i.e, 200 *min*), the results from dynamic model and simulation fitted very well with each other. However, due to the severe bunching after the peak of passenger arrival, density $k(t)$ produced from micro-simulation was significantly higher than that from the dynamic model; flow $q(t)$ from micro-simulation first lower and then higher than that from dynamic model until the bunching phenomenon disappeared.

With regard to the departure curves $D(t)$ in Fig. 4.12a and travel time $T(t)$ in Fig. 4.12b, the results form dynamic model and simulation were differed in a similar way: due to the bunching phenomenon, trains leaving the system after the peak of a_{p} experienced longer travel time in the micro-simulation. Therefore, it can be conclude that when bunching phenomenon is significant, the results from fluid-based dynamic model are not consistent with micro-simulation.

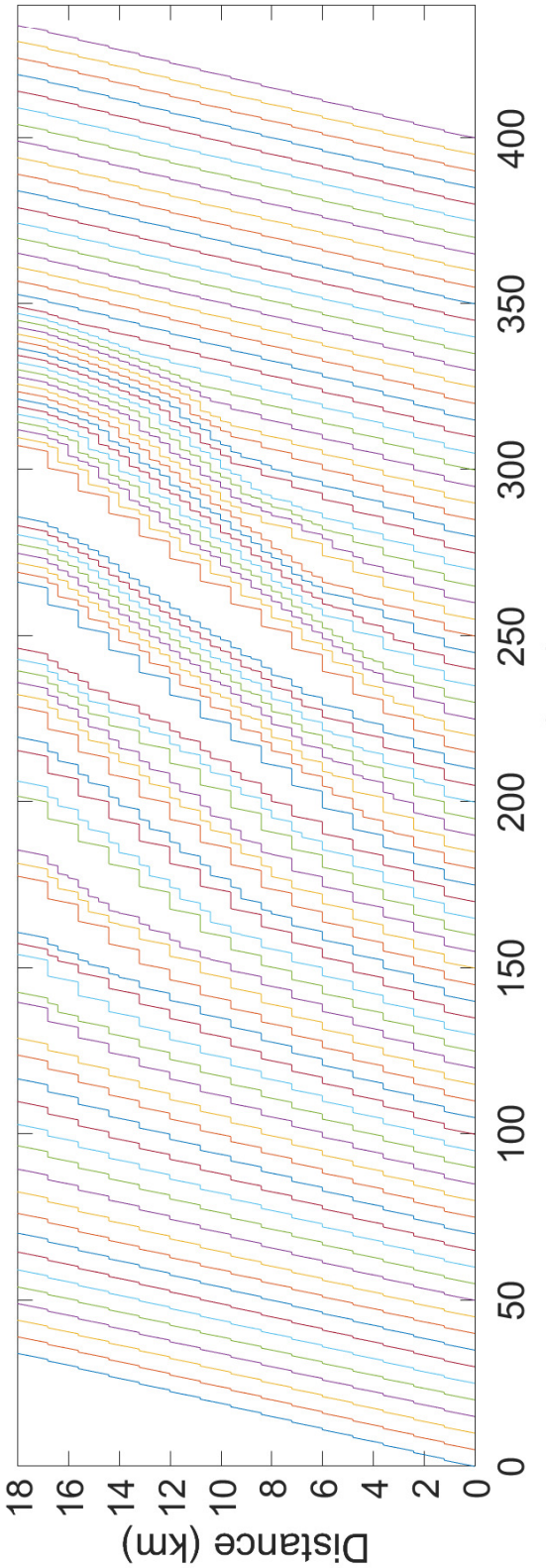
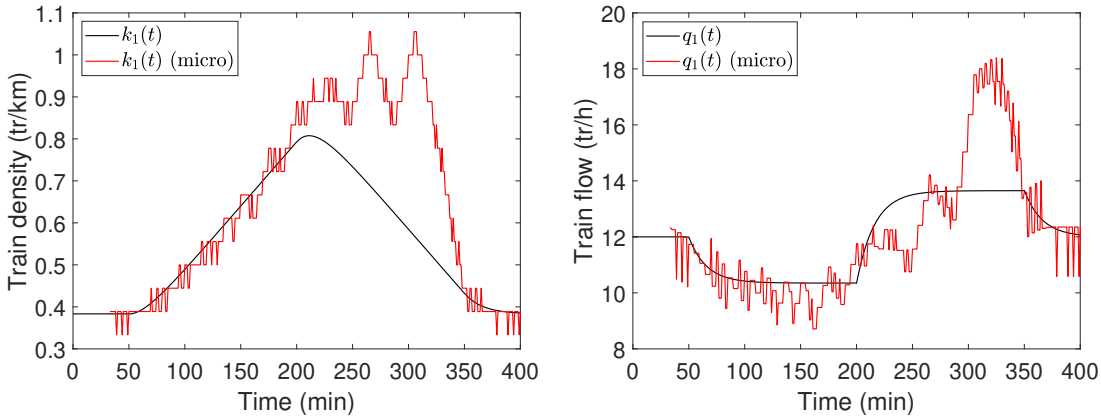
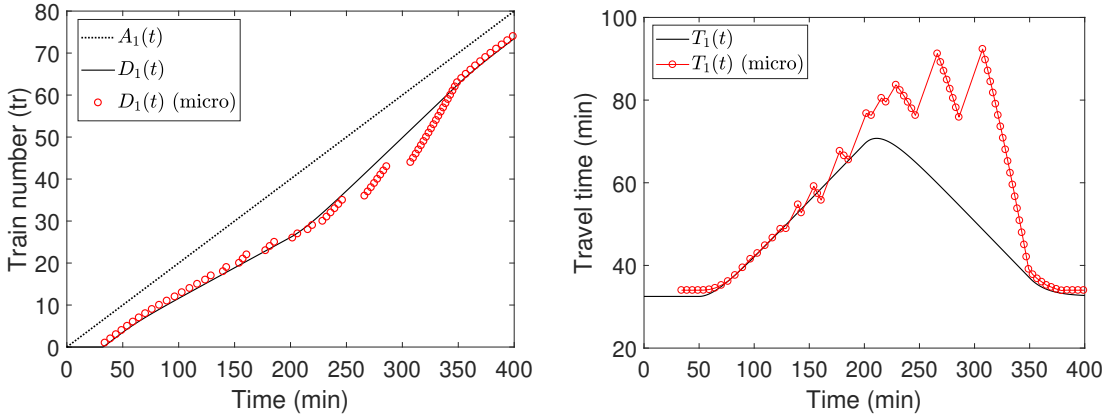


Fig. 4.10 An example of trains' trajectories from micro-simulation



(a) Comparison of train density $k(t)$ (b) Comparison of train flow $q(t)$

Fig. 4.11 Comparison of dynamic performance of railway system (fluid).



(a) Comparison of cumulative departures $D(t)$ (b) Comparison of travel time $T(t)$

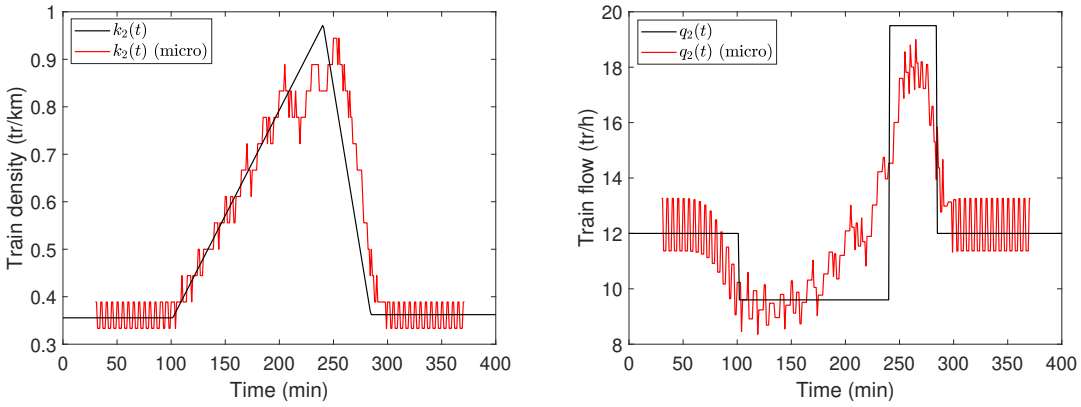
Fig. 4.12 Comparison of cumulative departures and travel time (fluid).

4.3.3 Verification of particle-based approach

Similarly, comparisons between dynamic model and simulation are also made for particle-based approach. Here, the passenger arrival information takes the output value from equilibrium model example in Chapter 5. From Fig. 4.13a it can be observed that the difference of density between dynamic model and simulation mainly appears around the peak of $T(t)$. This is also due to the bunching phenomenon as explained in the previous subsection. From Fig. 4.13b it can be seen that the different from dynamic model, sudden change of flow can not be achieved by micro-simulation. Besides, the flow produced from micro-simulation fluctuates

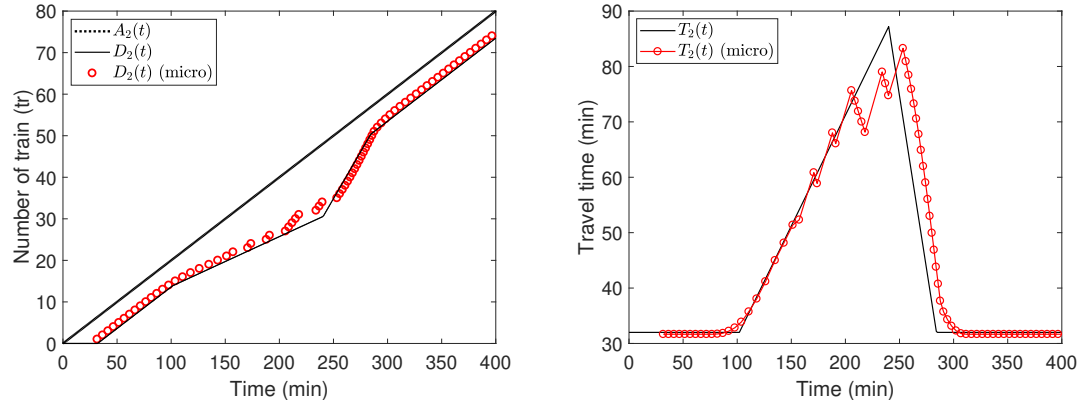
even when the value from dynamic model is constant. This is due to the discontinuous counting required by Eq. (4.19).

The comparison of trains' cumulative curve and travel time is shown in Fig. 4.14. It can be seen that except for the time period during the peak of travel time (bunching phenomenon), results from particle-based dynamic model are also consistent with those from micro-simulation.



(a) Comparison of train density $k(t)$ (b) Comparison of train flow $q(t)$

Fig. 4.13 Comparison of dynamic performance of railway system (particle).



(a) Comparison of cumulative departures $D(t)$ (b) Comparison of travel time $T(t)$

Fig. 4.14 Comparison of cumulative departures and travel time (particle).

4.4 Brief summary

This chapter introduced two dynamic models to describe the operation performance of rail transit under time-dependent system inputs. The fluid-based approach is a more general form of exit-flow model by Seo et al. (2017), which assumes that system instantaneous average flow used for train-FD is a liner combination of inflow and outflow. The particle-based approach introduces a strong assumption that the whole trip of a train can be approximated as a steady state. The density and flow are therefore time averaged values for one train during the trip. Through the numerical experiments and verification by microscopic simulation, the fluid-based model is found to be robust under most situations while particle-based model is rather fragile due to the strong assumption. However, particle-based model can works well together with equilibrium model introduced in the next chapter.

Chapter 5

Equilibrium distribution of passenger arrivals

The dynamic models proposed in Chapter 4 takes the passenger arrival distribution as a given information. However, as explained in Section 2.5, passengers would temporally adjust their travel plans if they could expect some benefits (e.g., fare discount, available seats or shorter travel time) or contrarily receive some penalties (e.g., surcharge, severe in-vehicle crowding or longer travel time). Therefore, it is important to understand the interaction between passengers' time choice behavior and operational information of rail transit system (e.g., timetable and congestion condition, pricing scheme). Especially for models designed to derive insights on management strategies, this interaction cannot be ignored. This chapter aims to estimate equilibrium distribution of passenger arrivals at a station and analyze the characteristics of the equilibrium.

5.1 Passenger travel cost

Before defining the travel cost, four assumptions are made here.

- Travel time in rail transit system changes due to both passenger congestion and train congestion on the track.
- Passengers are considered not to change their travel mode or give up their travel (inelastic travel demand).
- Individuals mainly trade off among the travel time, schedule delay and monetary reward/penalty to determine their departure time from home.

- The travel time from home to the nearest station for any individual is assumed to be constant and homogeneous, so it should not be considered in the cost function.

For the sake of clarity, "passenger arrival" means arrival at the origin station of the trip, "passenger departure" refers to the departure from the destination station hereinafter. Under the above assumptions, the total travel cost (TC) of individual i when he or she departs the destination station at time t is defined as

$$TC(t,t_i^*) = \alpha\,(T(t) - T_0) + s(t,t_i^*) + p(t), \tag{5.1}$$

where t_i^* is the desired departure time from the destination station for individual i,[1] α ($\alpha > 0$) is value of time for extra travel time, $T(t)$ is the travel time of an individual when he or she leaves the rail transit system at time t,[2] T_0 is the minimum travel time before the morning commute starts. For the sake of clarity, the first term of Eq.(5.1) is called travel time cost (TTC). The second term, $s(t,t_i^*)$, is the schedule delay cost (SDC) for individual i when he or she departs the system at time t. When all passengers have the same fixed desired departure time or work start time t^*, $t_i^* = t^*$ for any i. $p(t) > 0$ represents the surcharge a passenger should pay when he or she leaves the system at time t, and $p(t) < 0$ represents the reward. The fare scheme will be discussed in detail in Chapter 6, so discussions in this chapter takes $p(t) \equiv 0$.

Here, the widely used piece-wise linear schedule delay function (e.g., De Palma et al., 2017; Geroliminis and Levinson, 2009; Hendrickson and Kocur, 1981; Tian et al., 2007; Yang and Tang, 2018) was employed in this study. Schedule delay cost $s(t,t_i^*)$ is defined as

$$s(t,t_i^*) = \begin{cases} \beta(t_i^* - t), & \text{if } t < t_i^*, \\ \gamma(t - t_i^*), & \text{if } t \geq t_i^*, \end{cases} \tag{5.2}$$

where β and γ ($\gamma > \beta > 0$) are the value of time for early and late departures, respectively. For simplicity, it is assumed that all individuals have the same time cost α, β and γ.

Before moving to the next section, it should be addressed that the travel time increase in Eq. (5.1) are mainly caused by two reasons. One is the passenger congestion which leads to longer dwell time of train at stations, the other one is on-track congestion which leads to

[1] Here, t_i^* is also equivalent to the work start time for individual i if the time from the destination station to workplace is assumed as a constant value.

[2] For simplicity, $T(t)$ is the travel time of both trains and passengers if the boundary of the input-output system is set by the origin and the destination station for those commuting passengers. In this sense, this model can be applied to either one section or the whole rail transit line. But more practically, it is better to pick up the most congested section of a rail transit line and then consider the largest OD demand in this section.

the deceleration or stop of trains between the stations. The congestion will be described by train-FD model introduced in Chapter 3.

5.2 User equilibrium

In this section, the passenger arrival rate under user equilibrium (UE) is first derived. Then, solution method is given, and finally, the existence conditions of the equilibrium are briefly discussed.

5.2.1 Derivation

The equilibrium state is reached when the total travel cost of any individual can not be reduced. Denote t_i as the departure time of individual i from rail transit under equilibrium, the following equation holds

$$\frac{\partial TC}{\partial t}(t_i, t_i^*) = 0. \tag{5.3}$$

The derivative of travel time $T(t)$ can therefore be obtained by substituting Eq. (5.1) and Eq. (5.2) into Eq. (5.3) as

$$\frac{dT}{dt}(t_i) = \begin{cases} \beta/\alpha, & \text{if } t_i < t_i^*, \\ -\gamma/\alpha, & \text{if } t_i \geq t_i^*. \end{cases} \tag{5.4}$$

It can be observed that the values of travel time derivative do not depend on the desired departure time t_i^*. This is, however, not a general conclusion but a special result when piece-wise linear schedule delay function is employed. Eq. (5.4) implies that for passengers who depart the last station earlier than their desired departure time, the travel time they have experienced is increasing. On the contrary, for passengers who depart late, the travel time is decreasing. Since the derivatives of travel time in Eq. (5.4) are two constant values, travel time evolution under the equilibrium will be a piece-wise linear function with a single peak. If FIFO and First-In-First-Work (FIFW) hold, the travel time $T(t)$ maximizes at $t_{\max}$ when the schedule delay is zero. When all passengers have the same fixed desired departure time t^*, $t_{\max} = t^*$. When the demand is distributed, $t_{\max}$ is denoted by $\hat{t}$. Finally, the travel time $T(t)$ under user equilibrium can be described as

$$T(t) = \begin{cases} T_0 + \dfrac{\beta}{\alpha}(t - t_0), & \text{if } t_0 \leq t < t_{\max}, \\ T_0 + \dfrac{\beta}{\alpha}(t_{\max} - t_0) - \dfrac{\gamma}{\alpha}(t - t_{\max}), & \text{if } t_{\max} \leq t \leq t_{\text{ed}}, \end{cases} \tag{5.5}$$

where t_0 and t_{ed} represent the start and end time of the equilibrium, respectively. Meanwhile, t_0 and t_{ed} are also the time when commuters start and finish departing the rail transit system, respectively. The results so far are almost the same with most previous studies (Arnott et al., 1990b; De Palma et al., 2017; Geroliminis and Levinson, 2009; Hendrickson and Kocur, 1981; Tian et al., 2007; Yang and Tang, 2018).

In order to derive the equilibrium distribution of passenger arrivals, the dynamic model of rail transit system introduced in Chapter 4 has to be employed. The basic idea is to assume that average density and flow of trains follow the steady state train-FD introduced in Chapter 3. Thus, the equilibrium passenger arrival rate can be calculated by the inverse function of train-FD as:

$$a_{\mathrm{p}} = Q^{-1}\left(\bar{k},\ \bar{q}\right). \tag{5.6}$$

The ground of this assumption is that by aggregating the density and flow, their sudden change with time can be flattened, and the traffic states represented by the averaged variables can approximately be regarded as the steady state.

Due to an incompatibility problem of fluid-based dynamic model with equilibrium state, the particle-based dynamic model introduced in Chapter 4 is adopted. Since equilibrium travel time in Eq. (5.5) is defined in continuous time period, it has to be converted to particle-based variables in order to obtain headways. First, train inflow $a(t)$ (or equivalently $H_{\mathrm{a}}(i)$) should be treated as given information, then the outflow $d(t)$ can be obtained by differentiating the both sides of Eq. (4.8) (FIFO) as:

$$d(t) = a\left(t - T(t)\right)\left(1 - \frac{dT(t)}{dt}\right). \tag{5.7}$$

Substitute the given $a(t)$ and equilibrium travel time in Eq. (5.5), $d(t)$ under equilibrium can also be obtained from Eq. (5.7). Next, according to the definition, $H_{\mathrm{a}}(i)$ and $H_{\mathrm{d}}(i)$ can be derived from

$$H_{\mathrm{a}}(i) = \frac{d\left(t - T(t)\right)}{di} = \frac{1}{a\left(t - T(t)\right)}, \tag{5.8a}$$

$$H_{\mathrm{d}}(i) = \frac{dt}{di} = \frac{1}{d(t)}. \tag{5.8b}$$

Again, rewrite Eq. (4.14) in Chapter 4 to obtain travel time $T(i)$ for each train i:

$$T(i) = T(i-1) + H_{\mathrm{d}}(i) - H_{\mathrm{a}}(i). \tag{5.9}$$

With $H_{\mathrm{a}}(i)$, $H_{\mathrm{d}}(i)$ and $T(i)$, the density and flow can be calculated from Eq. (4.12) and (4.11). For convenience, these two equations are re-written here:

$$k(i) = \frac{T(i)}{L \times \left(H_{\mathrm{a}}(i) + H_{\mathrm{d}}(i)\right)/2}, \tag{5.10}$$

$$q(i) = \frac{1}{L \times \left(H_{\mathrm{a}}(i) + H_{\mathrm{d}}(i)\right)/2}. \tag{5.11}$$

Finally, passenger arrival rate under equilibrium can be derived by substituting the $q(i)$ and $k(i)$ into Eq. (5.6) as:

$$a_{\mathrm{p}}(i) = Q^{-1}\left(k(i), q(i)\right). \tag{5.12}$$

For an intuitive understanding, the explicit expression of inverse function of Model B train-FD are given as:

$$Q^{-1}(k,q) = \begin{cases} \dfrac{kl - (t_{\mathrm{b0}} + l/v_{\mathrm{f}})q}{\mu}, & \text{if } k < k^*(q), \\ \dfrac{l - \delta lk - [(l-\delta)t_{\mathrm{b0}} + \tau l]\,q}{(l-\delta)\mu}, & \text{if } k \geq k^*(q), \end{cases} \tag{5.13}$$

where

$$k^*(q) = \frac{1 + [(l-\delta)/v_{\mathrm{f}} - \tau]\,q}{l}. \tag{5.14}$$

5.2.2 Boundary conditions

For a specific morning commute situation, boundary conditions have to be introduced to get feasible results. Two types of passenger demands are considered in this study, one is that all passengers have the same desired departure time (or work start time), referred to as Case S hereinafter; another one is that the desired departure time is distributed within a time period, referred to as Case D hereinafter.

First, with regard to Case S, the inverse problem can be solved if the start and end time (i.e., t_0 and t_{ed}) of the morning commute are determined. Therefore, two boundary conditions should be introduced. The first one is that the travel time when commuters finish departing the rail transit system should be the same with the initial travel time before morning commute starts, as described by Eq. (5.15). The second one is that there is a total travel demand denoted by N_{p} for commuters as described by Eq. (5.16).

$$T(t_{\mathrm{ed}}) = T_0 = L/l\left(t_{\mathrm{b0}} + l/v_{\mathrm{f}}\right). \tag{5.15}$$

$$\int_1^I a_{\mathrm{p}}(i)\,di = N_{\mathrm{p}}. \tag{5.16}$$

In addition, $a_{\mathrm{p}}(n)$ calculated from Eq. (5.12) should be adjusted so that commuters only arrive during $[t_0, t_{\mathrm{ea}}]$ (i.e., Eq. (5.17)), where t_{ea} is the time when final commuters arrive the rail transit system as explained in Eq. (5.18).

$$\begin{cases} a_{\mathrm{p}}(i) \geq 0, & \text{if } t_0 \leq t(i) \leq t_{\mathrm{ea}} \\ a_{\mathrm{p}}(i) = 0, & \text{otherwise} \end{cases} \tag{5.17}$$

$$t_{\mathrm{ea}} = t_{\mathrm{ed}} - T(t_{\mathrm{ed}}) = t_{\mathrm{ed}} - T_0 \tag{5.18}$$

Then, when the demand is distributed, a demand curve $W(t)$ has to be introduced which indicates the cumulative number of passengers whose desired departure time from system (or work start time) is equal or before t. Although commuters in Case D have different work start time, there is a unique time instant $\hat{t}$ when the travel time $T(t)$ is maximized and schedule delay becomes zero (Daganzo, 1985; Smith, 1984). Therefore, at $\hat{t}$, $W(\hat{t})$ should be equal to $D_{\mathrm{p}}(\hat{t})$ and further equal to $A_{\mathrm{p}}(\hat{t} - T(\hat{t}))$ as described by Eq. (5.19).

$$D_{\mathrm{p}}(\hat{t}) = A_{\mathrm{p}}(\hat{t} - T(\hat{t})) = W(\hat{t}). \tag{5.19}$$

Similarly, when the morning commute ends at t_{ed}, the demand curve $W(\cdot)$ should again intersects with departure curve $D_{\mathrm{p}}(\cdot)$ and coincides with each other after t_{ed}, as described by Eq. (5.20).

$$D_{\mathrm{p}}(t_{\mathrm{ed}}) = A_{\mathrm{p}}(t_{\mathrm{ed}} - T(t_{\mathrm{ed}})) = A_{\mathrm{p}}(t_{\mathrm{ea}}) = W(t_{\mathrm{ed}}). \tag{5.20}$$

Together with another two boundary conditions as shown in Eq. (5.15) and Eq. (5.18), the critical time instant (i.e., t_0, $\hat{t}$, t_{ea}, and t_{ed}) to determine the equilibrium can also be solved for Case D.

5.2.3 Solution method

The calculation process to solve Case S is shown in Algorithm 1, where Δ is the step size of time, and N_ε is the tolerance limit of passenger number.

The calculation process for Case D is very similar to Algorithm 1, just to add another step to calculate $\hat{t}$ by Eq. (5.19) after Step 1, and replace Eq. (5.16) in Step 8 with Eq. (5.20). This algorithm converges when two conditions are met. The first one is that the demand is appropriate with regard to the operational parameters. The second one is that the setting of Δ and N_ε should be matching which means large setting of Δ requires larger N_ε.

Algorithm 1 Solution to Case S

Input: Operational parameters l, L, t_{b0}, μ, v_f, δ, τ; cost function $TC(t,t^*)$, train inflow $a(t)$, total travel demand N_p;

Output: Train flow $q(i)$, train density $k(i)$, and passenger arrival rate $a_p(i)$;

1: Set an initial t_0;
2: Calculate t_{ed} and t_{ea} by Eq. (5.15) and Eq. (5.18);
3: Calculate $T(t)$ and $d(t)$ by Eq. (5.5) and Eq. (5.7);
4: Derive $H_a(i)$ and $H_d(i)$ by Eq. (5.8);
5: Calculate $k(i)$ and $q(i)$ by Eq. (5.10) and (5.11);
6: Calculate $a_p(i)$ by Eq. (5.12);
7: Adjust $a_p(i)$ according to Eq. (5.17);
8: Calculate $LHS - RHS$ of Eq. (5.16), denoted as *error*;
9: **if** $error < -N_\varepsilon$, **then**
10: $t_0 = t_0 - \Delta$, repeat Step 2 to 8;
11: **else if** $error > N_\varepsilon$, **then**
12: $t_0 = t_0 + \Delta$, repeat Step 2 to 8;
13: **else**
14: Calculation converges, t_0, t_{ea} and t_{ed} determined;
15: **end if**
16: Outputs are derived from Step 5 and Step 6 when the calculation converges.

5.2.4 Existence conditions of equilibrium

When the travel time under the equilibrium and train-FD are jointly employed, several constraints have to be imposed to ensure the calculation results are physically feasible. In fact, train-FD is a bounded set in which all feasible traffic states (i.e., sets of $\{k, a_p, q\}$ that conform to Eq. (3.13)) are included. In other words, any traffic state outside of train-FD cannot be reached due to the physical limits of a rail transit system.

The first constraint is that the outflow of trains should always be positive,[3] which means

$$d(t) = a\left(t - T(t)\right)\left(1 - \frac{dT(t)}{dt}\right) > 0, \quad \forall t. \tag{5.21}$$

Train inflow $a(t)$, in this study, is treated as a positive given input so that $1 - dT(t)/dt > 0$ should hold all the time. Substitute Eq. (5.4) into Eq. (5.21), the following formula can be obtained:

$$\alpha > \beta. \tag{5.22}$$

This constraint is consistent with most equilibrium models for road traffic (e.g., Arnott et al., 1990b; Hendrickson and Kocur, 1981).

The second constraint is that passenger arrival rate $a_p(i)$ calculated from Eq. (5.12) should not be negative as written in Eq. (5.23).

$$a_p(i) \geq 0, \quad \forall i. \tag{5.23}$$

In fact, this constraint is equivalent to that the set of $(k(i), q(i))$ calculated from equilibrium travel time should not exceed the boundary of train-FD. However, the evolution of $(k(i), q(i))$ depends on not only the constant operational parameters, but also the settings of train inflow and total travel demand. To obtain an explicit expression of Eq. (5.23), a simplified case when inflow $a(t) \equiv a_c$ is considered. From the discussion in Section 5.3.1 it can be understood that as long as the upper right corner of $(k(i), q(i))$ loop does not exceed the congested regime boundary of train-FD, the whole loop of $(k(i), q(i))$ will lie inside the train-FD. According to the equilibrium condition, the upper right corner of $(k(i), q(i))$ loop refers to the state when the travel time is just maximized and starts to decrease. Therefore, the right corner $k_{rc}(i)$ and $q_{rc}(i)$ can be written as:

$$k_{rc}(i) = \frac{T_{max}}{\left(\frac{1}{a_c} + \frac{1}{a_c(1+\gamma/\alpha)}\right) L/2}, \tag{5.24}$$

[3] The conditions when trains are stopped by an accident and when the rail transit is not in operation are not considered

$$q_{\mathrm{rc}}(i) = \frac{1}{\left(\frac{1}{a_{\mathrm{c}}} + \frac{1}{a_{\mathrm{c}}(1+\gamma/\alpha)}\right)/2}. \tag{5.25}$$

Then, the constraint that $k_{\mathrm{rc}}(i)$ and $q_{\mathrm{rc}}(i)$ does not exceed congested regime boundary of train-FD can be derived by substituting $k_{\mathrm{rc}}(i)$ and $q_{\mathrm{rc}}(i)$ into lower branch of Eq. (3.13) as:

$$\frac{\alpha+\gamma}{\alpha+\gamma/2} a_{\mathrm{c}} \left[\frac{T_{\max}}{L} + \frac{(l-\delta)t_{\mathrm{b}0} + \tau l}{\delta l}\right] \leq \frac{1}{\delta}. \tag{5.26}$$

Where $T_{\max}$ is an indicator of total travel demand since t_0 is negatively correlated with total travel demand as can be understood from Eq. (5.5).

Finally, Eq. (5.22), Eq. (5.23) can be considered as the existence conditions of the equilibrium while Eq. (5.26) gives an explicit expression of Eq. (5.23) for a simplified case when train inflow is constant.

5.3 Characteristics of the equilibrium

In this section, the characteristics of the equilibrium by the proposed model are discussed by a series of numerical examples. The first subsection introduces the parameter settings for the numerical examples, interprets the dynamics of rail transit and passenger arrival distribution under equilibrium for both high and low demand conditions. The second subsection describes the train timetable (i.e., $a(t)$ or $H_{\mathrm{a}}(i)$) impact by a comparison of different train dispatch frequency plans.

5.3.1 Rail transit and passenger arrival distribution under equilibrium

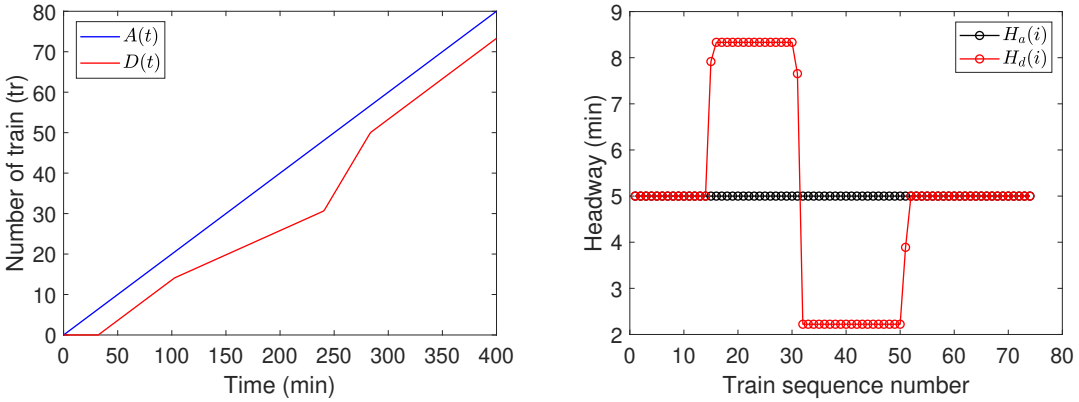
The parameter settings used in the following numerical examples are listed in Table 5.1. For simplicity, the train inflow $a(t)$ is set as constant and "Z" shape demand curve $W(t)$ is adopted. The slope of $W(t)$ is denoted as w_{p}, and time period for the increase of $W(t)$ is $[210, 270]$ *min*.

Since dynamics of rail transit system is almost the same for Case S and Case D, Fig. 5.1 takes Case S as an example. It can be seen from Fig. 5.1a that $D(t)$ will first diverge from $A(t)$ from t_0 to t^* and then approach to $A(t)$ from t^* to t_{ed}. In this way, the travel time is under user user equilibrium as described in Eq. (5.5). Fig. 5.1b shows the headways of trains $H_{\mathrm{a}}(i)$ and $H_{\mathrm{d}}(i)$. Fig. 5.2 shows the costs for the two cases. For Case S, the total travel cost under the equilibrium (i.e., during $[t_0, t_{\mathrm{ed}}]$) is minimized and constant so that no one has the incentive to change his time choice behaviour. For Case D, an example for a commuter whose $t_i^* < \hat{t}$ is taken. It can be observed that his total travel cost is minimized before t_i^*

Table 5.1 Parameter settings for numerical example (equilibrium).

Parameter	Value	Parameter	Value
l	1.2 km	α	20 $\$/h$
L	18 km	β	8 $\$/h$
v_{f}	40 km/h	γ	25 $\$/h$
t_{b0}	20 sec	t^*	240 min
μ	0.1 sec/pax	$a(t)$	12 tr/h
δ	0.4 km	w_{p}	500 pax/min
τ	1.0 min	N_{p}	30000 pax
Δ	1.0 min	N_{ε}	100 pax

which means he will depart the rail transit system no later than t_i^* (in other words, he will not be late for his work). Under FIFW, his actual departure time from the system can be obtained by jointly confirming the Fig. 5.2b and Fig. 5.3b. More specifically, find t_i^* on $W(t)$ and draw a horizontal line towards the left, the intersection point with D_{p} will be his actual departure time.

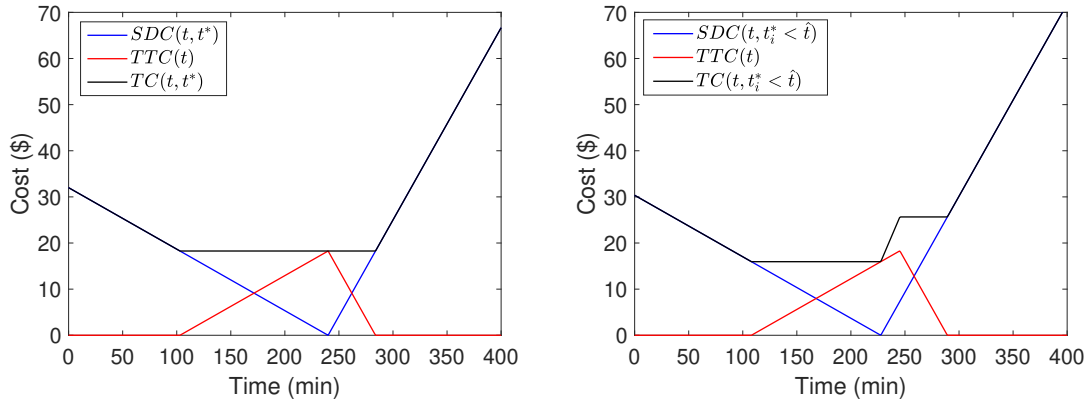


(a) Cumulative number of trains

(b) Headways of trains

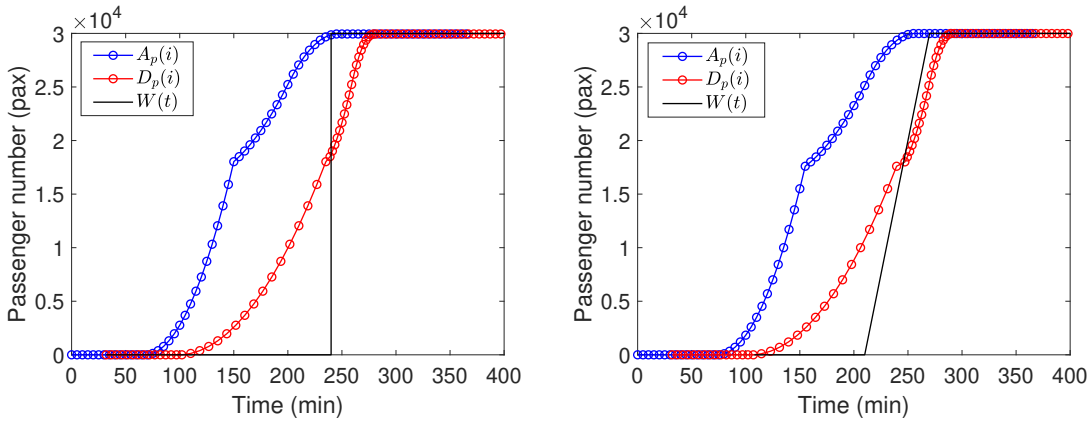
Fig. 5.1 Dynamics of rail transit system.

Fig. 5.3 shows the cumulative arrival and departure of passengers, the difference of two adjacent circles on the curve indicates the number of passengers carried by the preceding train. Since the travel demand for the two cases is set the same and the time midpoint of $W(t)$ slope is equal to t^*, characteristics of A_{p} and D_{p} for the two cases are almost the same. However, compared to Case S, the schedule delay in Case D (i.e., distance between D_{p} and $W(t)$) becomes significantly smaller so that the total travel cost differs among passengers. As can be understood from Fig. 5.2b, passengers whose t_i^* equalled to $\hat{t}$ endured the highest total travel cost $TC(t,\hat{t})$ while TC decreased with the difference between t_i^* and $\hat{t}$ in case D.



(a) Case S (b) Case D with $t_i^* < \hat{t}$

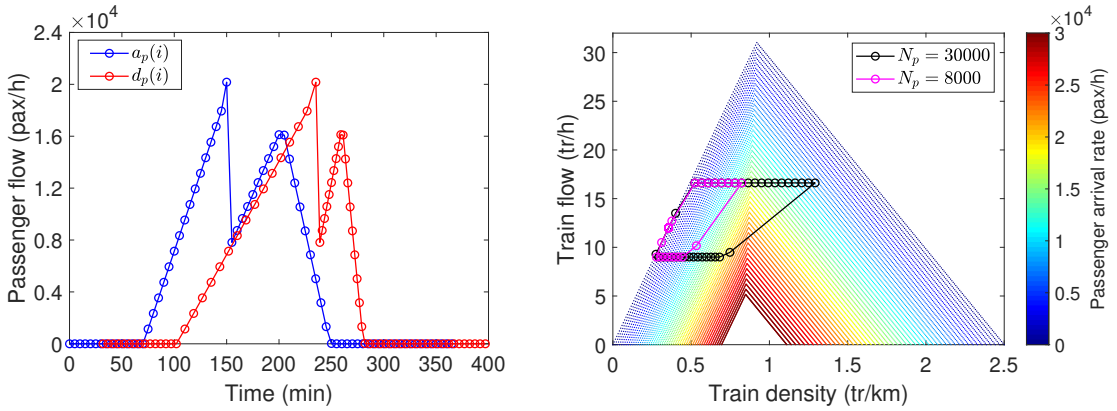
Fig. 5.2 Travel cost for the two cases.



(a) Case S (b) Case D

Fig. 5.3 Cumulative number of passengers.

Passenger arrival rate $a_{\mathrm{p}}(i)$ and departure rate $d_{\mathrm{p}}(i)$ are depicted in Fig. 5.4a. It can be observed that a_{p} has two peaks in this example, the large one corresponds to the arrival of commuters who experienced the longest travel time but departed the system just before t^* while the small one corresponds to the arrival of commuters who experienced shorter travel time at the cost of departing the system after t^*. Since a_{p} is derived from the inverse function of train-FD, how $k(i)$ and $q(i)$ evolve on train-FD is also drawn in Fig. 5.4b. It can be seen that the evolution of $(k(i),q(i))$ during the morning commute starts from the left boundary of train-FD and formulates a counter-clockwise closed loop within the train-FD. Divided by $a(t) = 12\,tr/h$, the lower half of the loop represents the average density and flow experienced by trains departing the system from t_0 to t^*, while the upper half represents the dynamics from t^* to t_{ed}. The maximum of a_{p} is reached at the lower right corner of the loop. In fact, if train inflow $a(t)$ is constant and late departure in the schedule delay function (i.e., the condition $t > t_i^*$ in Eq. (5.2) is valid) is permitted, passenger arrival rate with two-peaks appear when high travel demand forces $(k(i),q(i))$ to enter the congested regime of train-FD.



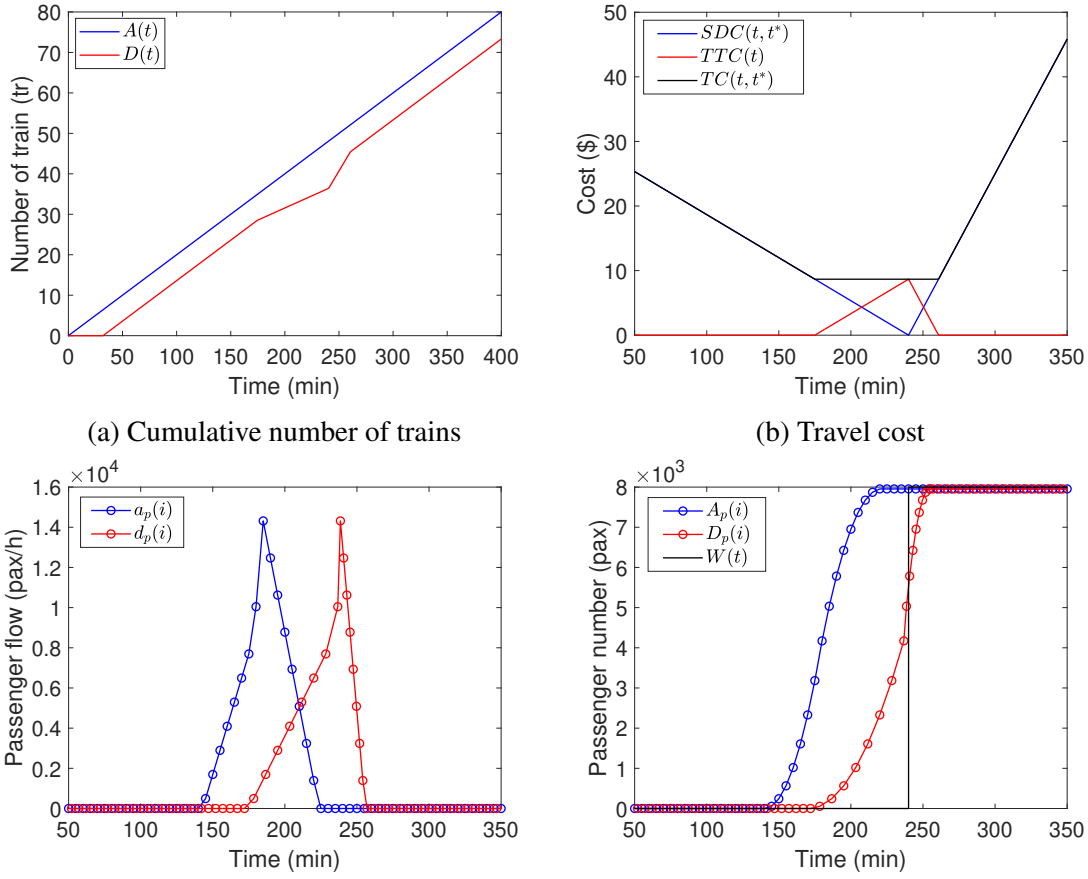
(a) Passenger flow ($N_{\mathrm{p}} = 30000$) (b) Dynamics of density and flow on train-FD

Fig. 5.4 An example of passenger flow and dynamics of trains on train-FD.

On one hand, if late departure is not permitted (i.e., only the condition $t \leq t_i^*$ in Eq. (5.2) is valid), the travel time under the equilibrium will continue increasing until t^* so that $k(i)$ calculated from Eq. (5.10) also keeps increasing. As a result, $(k(i),q(i))$ goes deeper into the train-FD from the left side so that a_{p} increases in the free-flow regime and decreases in the congested regime. Then, after t^*, no commuter departs from the system (i.e., $a_{\mathrm{p}} = 0$) so that $(k(i),q(i))$ suddenly jumps up to the upper boundary of train-FD and finally returns to the initial point along the boundary of the train-FD. In this way, a_{p} will only have one peak.

On the other hand, if the late departure is permitted but the total travel demand is rather low, a_{p} may also have only one peak. Fig. 5.5 gives an example when a_{p} has one peak under the travel demand $N_{\mathrm{p}} = 8000$. The comparison of $(k(i),q(i))$ evolution between high and

low demand is also shown in Fig. 5.4b. It can be understood that if $(k(i), q(i))$ after t^* (upper right of the loop) does not enter the congested regime of train-FD, there will be only one peak for a_{p}. In other words, commuters who departed the system after t^* are represented in the decline side of a_{p} if total travel demand is low. On the contrary, when the demand is high, arrivals of early and late commuters are separated on the two peaks. This is a rather new and interesting characteristic of equilibrium distribution of passenger arrivals for a rail transit system.



(a) Cumulative number of trains

(b) Travel cost

(c) Passenger flow

(d) Cumulative number of passengers

Fig. 5.5 An example under the travel demand $N_{\mathrm{p}} = 8000$.

5.3.2 Impact of train timetable

So far, $a(t)$ or $H_{\mathrm{a}}(i)$ in all above-mentioned examples are set as constant. However, as can be experienced in our daily life, the dispatch frequency of trains are higher during the peak hours and lower during the off-peak. This strategy is undoubtedly reasonable and empirically

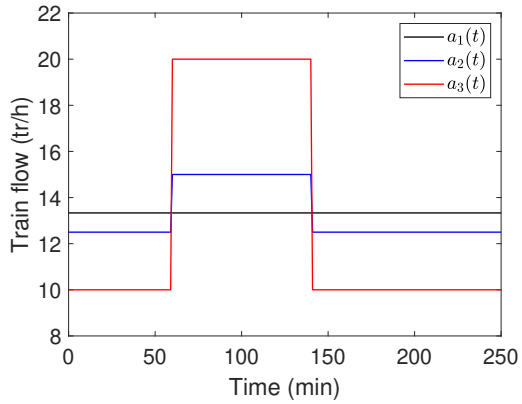
correct. While in fact, we have very limited knowledge about the interaction between the time choice behaviour of commuters and the change of train timetable, especially from a theoretical point of view. As mentioned in Section 2.5, most existing studies focused on the optimization of train timetable under a given passengers' arrival pattern. But some passengers may adjust their time choice when the timetable is updated, which again makes the optimized timetable not optimized under the new arrival pattern. In this sense, the value of the proposed model lies in its capability to estimate an equilibrium distribution of passenger arrivals when the train timetable is updated. Here, influence of train timetable is discussed by a series of numerical experiments comparing three train inflow patterns.

To make a fair comparison, it is ensured ensured that the cumulative arrivals of trains before t^* are the same for the three cases (other parameters kept the same with Table 5.1 for Case S, $N_\mathrm{p} = 30000$). As shown in Fig. 5.6a, $a_1(t)$ is still set as constant, referred to as Case H1; $a_2(t)$ makes a step shape to simulate the conventional train timetable during morning rush hours, referred to as Case H2; $a_3(t)$ sets a larger gap between the peak and off-peak compared to $a_2(t)$, referred to as Case H3. Here, compared to average passenger arrival rate $a_\mathrm{p}(i)$, the boarding passenger number for each dispatched train $n_\mathrm{p}(i)$ defined in Eq. (5.27) can more intuitively compare the duration of morning commute and in-vehicle crowding , so $n_\mathrm{p}(i)$ are depicted for the three cases in Fig. 5.6b.

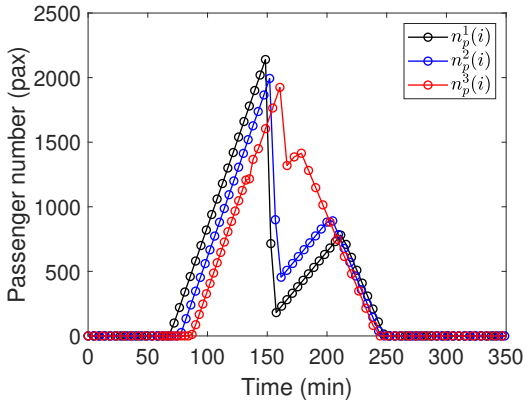
$$n_\mathrm{p}(i) = a_\mathrm{p}(i) \cdot \frac{H_\mathrm{a}(i) + H_\mathrm{d}(i)}{2}. \tag{5.27}$$

It can be observed that $n_\mathrm{p}(i)$ calculated from Case H1 has the earliest start time and the highest peak while Case H3 has the latest start time and the lowest peak. This is reasonable because on one hand, when larger train inflow (or high dispatch frequency) matches with the period of high passenger arrival rate, passenger numbers carried by each train will decrease at the ascending side of n_p. On the other hand, smaller train inflow after $t^* - T(t^*)$ will relieve the bunching phenomenon of trains, which again makes it possible for trains to carry more passengers at the descending side of n_p. As s result, trains are more efficiently used so that the in-vehicle crowding is relieved and the morning commute period (i.e., $t_\mathrm{ed} - t_0$) is shortened. In addition, since t_0 increased from Case H1 to H3, equilibrium travel cost $TC(t_0, t^*) = \beta(t^* - t_0)$ decreased so that overall system cost $N_\mathrm{p} \cdot TC(t_0, t^*)$ also decreased. In this sense, timetable modification can serve as an effective measure to approach system optimal. As a supplementary material, the cumulative curves of trains and passengers for Case H2 and H3 are also shown in Fig. 5.7. As can be understood from 5.7c, by a combination of high and low dispatch frequencies, $d(t)$ increased before t^* and decreased after t^* compared to Case H1, thus average headway $(H_\mathrm{a}(i) + H_\mathrm{d}(i))/2$ was shortened before t^* and extended

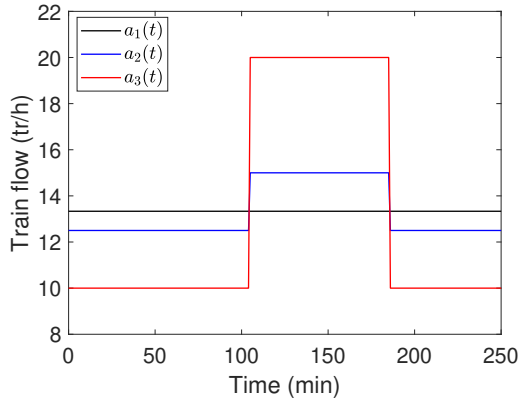
after t^*. In this way, in-vehicle crowding was alleviated before the peak of n_p and bunching phenomenon was relieved after t^*.



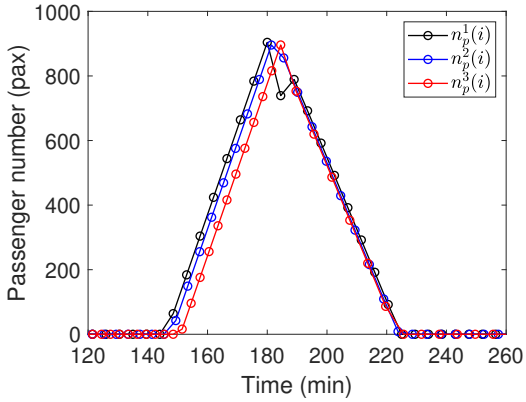
(a) Train inflow ($N_\mathrm{p} = 30000$)



(b) Number of boarding passenger ($N_\mathrm{p} = 30000$)



(c) Train inflow ($N_\mathrm{p} = 8000$)

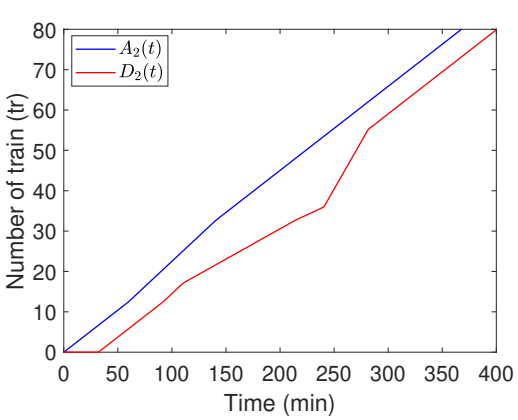


(d) Number of boarding passenger ($N_\mathrm{p} = 8000$)

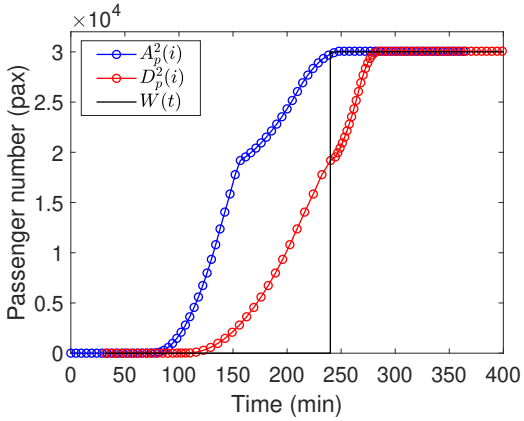
Fig. 5.6 A comparison of boarding passenger numbers between different train inflow patterns.

Moreover, a similar comparison was done also for the low demand situations (i.e., $N_\mathrm{p} = 8000$) as shown in Fig. 5.6c and 5.6d. Since the start time of passenger arrivals will be later when the demand is low, peak period of $a(t)$ is accordingly adjusted to match a_p, but the values are kept unchanged. This three cases can be referred to as L1, L2, and L3, respectively. It can again be confirmed from Fig. 5.6d that similar to the high demand cases, the step size train inflow can shorten the morning commute period and reduce the in-vehicle crowding, although the difference is relatively small compared to high demand cases. In this numerical example, the travel demand and system supply (i.e., total number of dispatched trains) are the same, but boarding passenger number (ascending side) at a given time decreased around 50% and morning commute period shortened around 14% from Case H1 to Case H3. From Case L1 to Case L3, these two indices reduced 40% and 8%, respectively. Although the efficiency

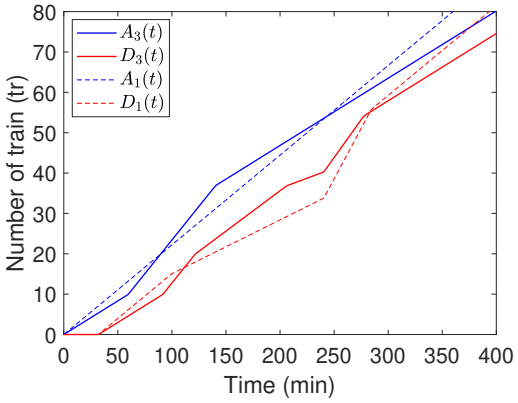
improvement by modifying the timetable in this example is a matter of course from our travel experience, the capability to show passengers' interaction to the updated timetable can be regarded as a remarkable achievement of the proposed model.



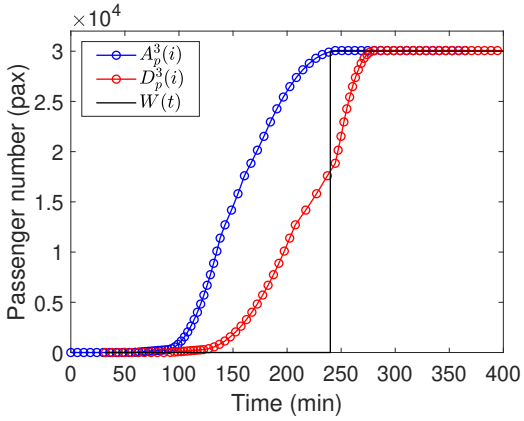
(a) Cumulative number of trains (Case H2)



(b) Cumulative number of passengers (Case H3)



(c) Cumulative number of trains (Case H3)



(d) Cumulative number of passengers (Case H3)

Fig. 5.7 Cumulative number of trains and passengers for Case H2 and Case H3.

5.4 Incompatibility of fluid-based approach with equilibrium model

This section briefly explains why fluid-based approach to describe the railway dynamics cannot work together with the equilibrium model mentioned in Section 5.2.1. In one word, the reason lies in the fact that dynamics described by spatial instantaneous average variables (such as fluid-based dynamic model) inevitably moves out of train-FD under user equilibrium,

which yields physically infeasible condition of density, flow and passenger arrival rate for the railway system.

More specially, the difference between train-FD and road traffic macroscopic fundamental diagram (MFD) should be clarified. MFD is basically a two-dimensional relation between the aggregated instantaneous spatial average density and flow in a road network. Both the inside and outside of MFD can be interpreted as unsteady traffic state from a macroscopic view. However, the three-dimensional train-FD with an additional passenger arrival rate dimension does not have convincing interpretation of traffic state out of train-FD. The microscopic derivation of train-FD determines that the left boundary of train-FD is the maximum achievable density and flow when no passenger affects the operation of railway system. Therefore, the traffic state out of train-FD cannot be feasibly interpreted unless it is modified considering the interpretation of traffic state out of train-FD.

Next, why dynamic traffic state described by fluid-based model would move out of train-FD under user equilibrium is explained. To simplify the explanation, consider a constant train inflow (denoted as a_c) case.[4] In general, it is considered that the traffic state at the start and end of equilibrium should be the same which means $k(t_0) = k(t_{\mathrm{ed}})$ (or equivalently $T(t_0) = T(t_{\mathrm{ed}})$ from the cumulative curve of trains). If the passenger arrival rate at t_0 starts from zero which means $a_{\mathrm{p}}(t_0) = Q^{-1}(k(t_0), q(t_0)) = 0$, since equilibrium flow approaching t_{ed} larger than that at t_0 (i.e., $q(t_{\mathrm{ed}}) > q(t_0)$), passenger arrival rate at the end of equilibrium state would inevitably be smaller than zero on free-flow regime of train-FD, which means:

$$a_{\mathrm{p}}(t_{\mathrm{ed}}) = Q^{-1}\left(k(t_{\mathrm{ed}}), q(t_{\mathrm{ed}})\right) = Q^{-1}\left(k(t_0), q(t_{\mathrm{ed}})\right) < 0. \tag{5.28}$$

The explanation so far can be illustrated by the evolution of density and flow on train-FD in Fig. 5.8. Even if one assumes the existence of captive users at the start of equilibrium like $a_{\mathrm{p}}(t_0) = Q^{-1}(k(t_0), q(t_0)) = a_{\mathrm{pc}}$ (equivalent as moving the loop to rightward), $a_{\mathrm{p}}(t_{\mathrm{ed}}) < a_{\mathrm{pc}}$ is also not interpretable.

One may try to change the boundary condition of $k(t_0) = k(t_{\mathrm{ed}})$ to $k(t_{\mathrm{ed}}) = (1 + \zeta\gamma/\alpha)k(t_0)$ to avoid jumping out of train-FD like shown in Fig. 5.9. This boundary condition can avoid traffic state out of train-FD, but it leads to a new contradictory that for $a_{\mathrm{p}}(t^* < t < t_{\mathrm{ed}}) > 0$, some of these passengers may leave the railway system after t_{ed}. Since the travel time decline on the left boundary of train-FD in Fig. 5.9 is slower than the horizontal line on Fig. 5.8, their travel time cost would be larger than the equilibrium cost, which means these passengers still

[4] By ingeniously set a large inflow before t^* and a small inflow after t^*, dynamics described by fluid-based model may avoid jumping out of train-FD, but this kind of setting will lose the generality and makes the model only available under specific setting.

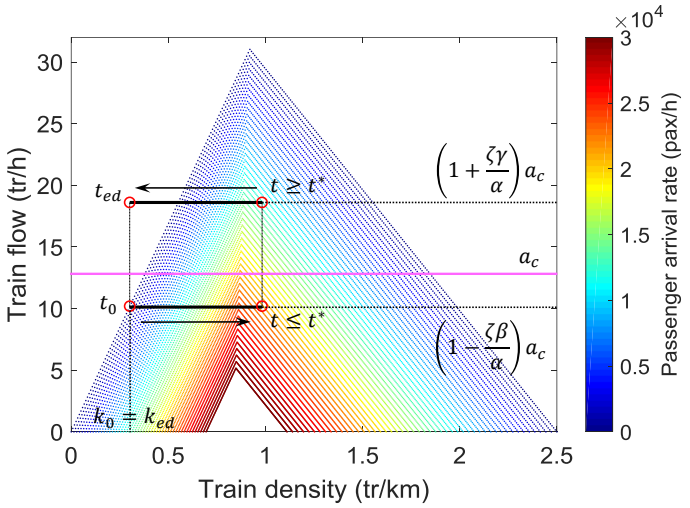


Fig. 5.8 Density and flow evolution on train-FD for fluid-based approach ($k(t_0) = k(t_{\mathrm{ed}})$).

can reduce their travel costs by arriving earlier at the station. This breaks the equilibrium state and is also not acceptable.

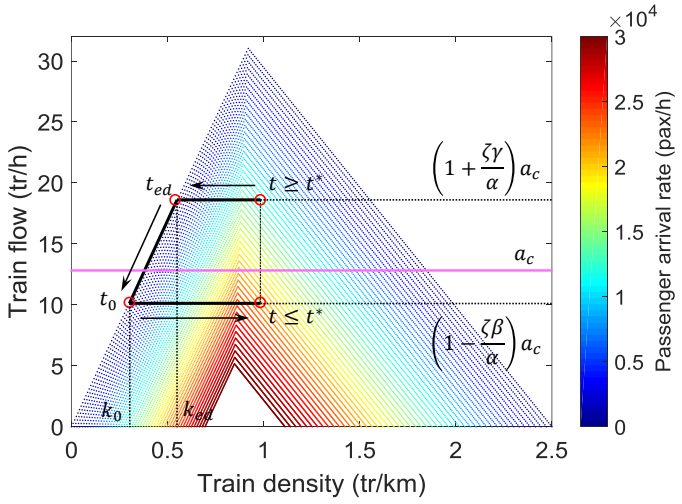


Fig. 5.9 Density and flow evolution on train-FD for fluid-based approach ($k(t_0) \neq k(t_{\mathrm{ed}})$).

5.5 Brief summary

This chapter derived the equilibrium distribution of passenger arrivals for the morning commute problem from the particle-based dynamic model by employing the inverse function of train-FD. The solution method and existence conditions of the equilibrium are given, and characteristics of the equilibrium are discussed based on a series of numerical experiments.

Through the numerical examples, it is found that total travel demand and train timetable all had significant influence on the characteristics of the passenger arrival distributions. More specifically, if all passengers have to depart the rail transit system no later than their desired departure time, there will be only one peak for the equilibrium distribution of passenger arrivals, otherwise, there may exist two peaks depending on the travel demand. When the travel demand is low so that train density and flow do not enter the congested regime of train-FD, the equilibrium distribution will also be unimodal. In addition, the proposed model numerically confirmed that compared to a constant dispatch frequency, the combination use of high and low dispatch frequencies can shorten the morning commute period and relieve the in-vehicle at the same time since this strategy improved the efficiency use of trains. At the end of this chapter, why fluid-based approach is not used in equilibrium model is briefly discussed.

Chapter 6

Model applications

By applying the equilibrium model developed in Chapter 5, this chapter aims to evaluate management strategies from both demand and supply sides. First, as a further investigation of supply management briefly discussed Section 5.3.2, conventional peak/off-peak timetable is optimized based on inference of dynamics evolution on train-FD. Second, user equilibrium under first-best pricing is analytically derived. Finally, a demand management strategy-coarse pricing scheme (including surcharge, reward and their combinations) is numerically evaluated based on indicators such as user cost, social cost and in-vehicle crowding.

6.1 Peak/off-peak timetable optimization

In this section, the optimization problem of a conventional peak/off-peak timetable without pricing is discussed. Peak/off-peak timetable here refers to a simplified timetable pattern that

- only two types of dispatch frequencies (or inflow) are employed,
- the high dispatch frequency lasts for an uninterrupted period enclosed by the low dispatch frequency.

Since this timetable pattern is widely employed in practice, it is desirable to explore the optimal setting of peak/off-peak timetable by the proposed model. This issue can be generalized as a bi-level optimization problem described as:

$$\begin{cases} \min F(N) \\ F(N) = \min\limits_{a_1,a_2\in\Omega} f(a_1,a_2|N). \end{cases} \tag{6.1}$$

Where N represents the number of dispatched train number required to carry N_p passengers, a_1 and a_2 are peak and off-peak train inflow ($a_1 \geq a_2$), respectively. $F(N)$ is the function to

minimize individual equilibrium cost C^e with relate to N, which is referred to the upper-level optimization. The lower-level optimization in the second formula means to minimize $F(N)$ when N is determined by designing a_1 and a_2. The upper-level optimization is still not solved at the completion time of this dissertation, so the next subsection discusses a specific case for the lower-level optimization problem.

6.1.1 Lower-level optimization in a specific case

An ideal temporal setting of a_1 and a_2 can be described as the situation that trains which carry commute passengers and leave the system before t^* are exactly dispatched with high inflow a_1, and other trains are dispatched with low inflow a_2. This is reasonable because under equilibrium before t^*, train outflow is smaller than inflow, while after t^*, outflow is larger than inflow. This setting can maintain a relatively high average flow before t^* and avoids a extremely high outflow after t^*. Fig. 6.1 describes this situation.

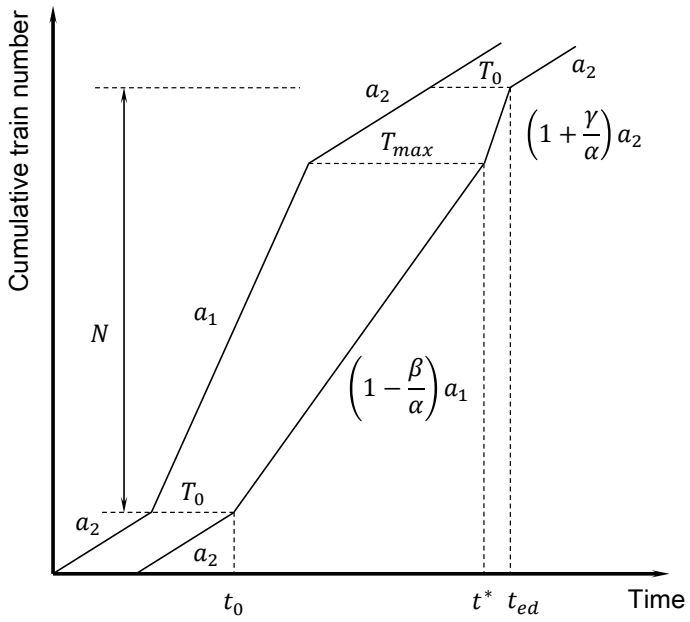


Fig. 6.1 An specific case of train cumulative curve for peak/off-peak timetable pattern.

Under this specific case, the lower-level optimization can be formularized. More specifically, individual equilibrium cost C^e can be written as the SDC at the start or end of the equilibrium period:

$$C^e = \beta(t^* - t_0) = \gamma(t_{\text{ed}} - t^*). \tag{6.2}$$

According to the description in Fig. 6.1, the number of trains N that carried these N_{p} passengers can be derived as:

$$a_1\left[t^* - T_{\max} - (t_0 - T_0)\right] + a_2\left[t_{\mathrm{ed}} - T_0 - (t^* - T_{\max})\right] = N. \tag{6.3}$$

Meanwhile, according to the equilibrium travel time defined in Eq. (5.5), the maximum travel time increase can be described as:

$$T_{\max} - T_0 = \frac{\beta}{\alpha}(t^* - t_0). \tag{6.4}$$

Next, substitute Eq. (6.2) and (6.4) into Eq. (6.3), C^e can be re-written as:

$$C^e = \frac{N}{\frac{\alpha-\beta}{\alpha\beta}a_1 + \frac{\alpha+\gamma}{\alpha\gamma}a_2}. \tag{6.5}$$

The lower-level optimization problem can therefore be considered as a problem to minimize the following objective function with four constraints:

$$\min_{a_1,a_2\in\Omega} f(a_1,a_2|N) = \min_{a_1,a_2\in\Omega}\left\{\frac{N}{\frac{\alpha-\beta}{\alpha\beta}a_1 + \frac{\alpha+\gamma}{\alpha\gamma}a_2}\right\} \tag{6.6}$$

$$s.t.\ 0 < a_2 \le a_1 \le 1/\tau, \tag{6.7a}$$

$$s.t.\ \omega a_1 + (1-\omega)a_2 \le a_0, \tag{6.7b}$$

$$s.t.\ a_1 \le \frac{\alpha-\beta/2}{\alpha-\beta}\left(\frac{\delta}{L}T_{\max} + \left(1-\frac{\delta}{l}\right)t_{\mathrm{b}0} + \tau\right)^{-1} = a_1^*. \tag{6.7c}$$

$$s.t.\ a_2 \le \frac{\alpha+\gamma/2}{\alpha+\gamma}\left(\frac{\delta}{L}T_{\max} + \left(1-\frac{\delta}{l}\right)t_{\mathrm{b}0} + \tau\right)^{-1} = a_2^*. \tag{6.7d}$$

The Ω is the enclosed area set by the four constraints. The first constraint clarifies that train inflow can not be larger than the reciprocal of minimum time headway τ. The second constraint limits the achievable train inflow (due to available train fleet) as a_0. The third and fourth constraints comes from the feasible density and flow within train-FD, which is related to demand and has been explained in Eq. (5.26). In fact, ω can be explicitly derived from the duration ratio for a_1 and a_2 as:

$$\omega = \frac{\gamma(\alpha-\beta)}{\beta(\alpha+\gamma)+\gamma(\alpha-\beta)}, \tag{6.8a}$$

$$1-\omega=\frac{\beta(\alpha+\gamma)}{\beta(\alpha+\gamma)+\gamma(\alpha-\beta)}. \tag{6.8b}$$

Now, to minimize the objective function of lower-optimization problem is equivalent to maximize the denominator of Eq. (6.6). This problem can be discussed by Fig. 6.2.

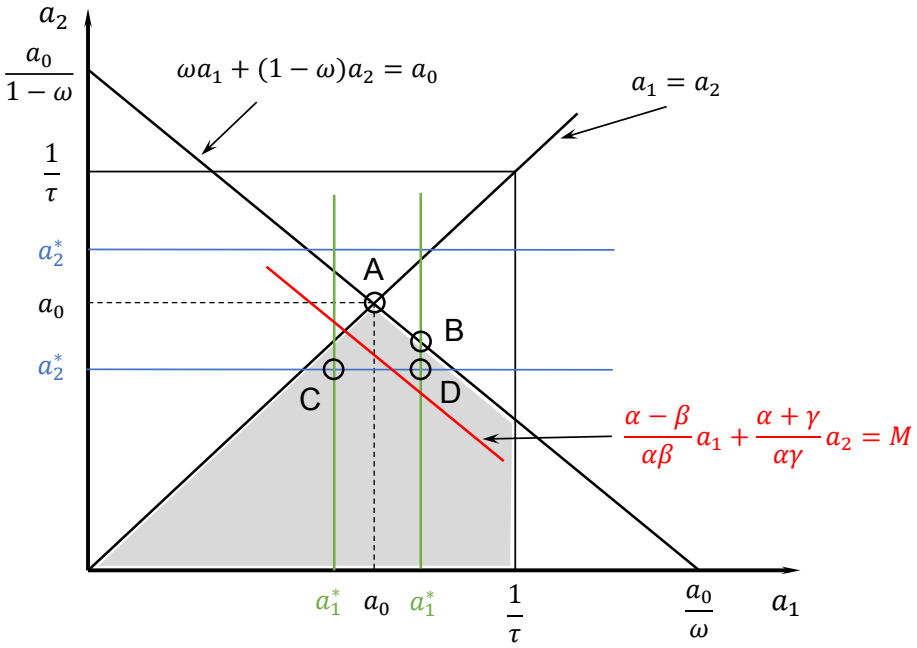


Fig. 6.2 Optimization of train inflow for ideal peak/off-peak timetable pattern.

It can be understood from Fig. 6.2 that the first and second constraint enclose an feasible area (i.e., shadowed area) for a_1 and a_2. The denominator of objective function can be expressed as the red line with the same slope with the second constraint. Let its value as M, then to maximize M is to find the largest intercept on a_2 axis for the red line within the shadowed area. However, the position of a_1^* a_2^* (green and blue vertical lines) depends on the total travel demand. Therefore, the optimal setting of a_1 and a_2 can only be determined when the relation between a_0, a_1^* and a_2^* is clarified.

Nevertheless, the optimal conditions can be separately discussed. Fig. 6.2 shows three possible combinations of the third and the fourth constraint ($a_1^* > a_2^*$).

- When $a_2^* < a_1^* < a_0$, the optimum is reached at point C which means $a_1 = a_1^*$ and $a_2 = a_2^*$. This case refers to the situation that the capability to dispatch trains is abundant compared to train-FD constraint.

- When $a_2^* < a_0 < a_1^*$, the optimum is reached at point D, similar as the first case. In fact, depending on the relation between a_1^* and a_2^* (a_1^* further larger), the optimum may also be reached on the line of the second constraint.

- When $a_1^* > a_2^* > a_0$, the optimum is reached at any point on line AB. This case refers to the situation that capability to dispatch trains is relatively insufficient. Under this situation, to adopt the constant timetable can also reach the optimum (i.e., point A).

6.1.2 Optimum in free-flow regime of train-FD

In fact, the discussion in the previous subsection avoids the upper-level optimization. To solve upper-level optimization is difficult because it inevitably relates to regime-dependent utilization of train-FD (only free-flow or free-flow + congested). Therefore, a more conservative optimum only in free-flow regime of train-FD is presented in this subsection.

The optimization objective is still to minimize the total travel cost defined as the sum of travel time cost (TTC) and schedule delay cost (SDC) under the constraint of total passenger travel demand N_p and total available trains N. Since the description of total travel cost by fluid-based model involves tedious integration which is difficult for analytical derivation, only the description by particle-based model is discussed hereinafter.

$$\min\{\sum_{i=1}^{N}\alpha\left(T(i)-T_0\right)n_\mathrm{p}(i)+\sum_{i=1}^{N_1}\beta\left(t^*-t_i\right)n_\mathrm{p}(i)+\sum_{i=N_1+1}^{N}\gamma(t_i-t^*)\,n_\mathrm{p}(i)\}, \tag{6.9}$$

$$s.t.\sum_{i=1}^{N}n_\mathrm{p}(i)=N_\mathrm{p}, \tag{6.10}$$

where $n_\mathrm{p}(i)$ is the riding passenger number in train i, t_i is the time when train i exits the system, and N_1 is the number of train that exited the system before t^*. TTC written in the first term of Eq. (6.9) can be further simplified by substituting the relation between travel time $T(i)$ and riding passenger number $n_\mathrm{p}(i)=a_\mathrm{p}(i)\cdot(H_\mathrm{a}(i)+H_\mathrm{d}(i))/2$ (i.e., Eq.(4.16) in Chapter 4). As a result, TTC for train i can be described by the following two equations in free-flow and congested regime of train-FD, respectively.

$$TTC(i)=\alpha\mu\frac{L}{l}(n_\mathrm{p}(i))^2, \tag{6.11a}$$

$$TTC(i)=-\alpha\mu\left(\frac{L}{\delta}-\frac{L}{l}\right)(n_\mathrm{p}(i))^2+\alpha\frac{L}{\delta}\left(q(i)^{-1}-t_{\mathrm{b}0}-\delta/v_\mathrm{f}-\tau\right)n_\mathrm{p}(i). \tag{6.11b}$$

On one hand, it can be understood from Eq. (6.10) and (6.11a) that if all dispatched trains operated in free-flow regime of train-FD, the minimum of total TTC (i.e., the first term of Eq. (6.9)) is reached when all $n_\mathrm{p}(i)$ are the same and equal to N_p/N. However, without pricing interference, travel time $T(i)$ inevitably first increase from T_0 and then decrease back to T_0 under equilibrium. As a result, $n_\mathrm{p}(i)$ also increase from zero to the maximum and then

decrease back to zero. On the other hand, the maximum of Eq. (6.11b) is reached when $n_{\mathrm{p}}(i)$ takes the value:

$$n_{\mathrm{p}}^{\max}(q(i)) = \frac{q(i)^{-1} - t_{\mathrm{b0}} - \delta/v_{\mathrm{f}} - \tau}{2\mu(1-\delta/l)}. \tag{6.12}$$

Meanwhile, at the boundary of free-flow and congested regime, critical riding passenger number $n_{\mathrm{p}}^{*}(i)$ can be calculated as:

$$n_{\mathrm{p}}^{*}(i) = a_{\mathrm{p}}^{*}(i) \cdot \frac{H_{\mathrm{a}}(i) + H_{\mathrm{d}}(i)}{2} = \frac{a_{\mathrm{p}}^{*}(i)}{q^{*}(i)} = \frac{a_{\mathrm{p}}^{*}(i)(t_{\mathrm{b0}} + \delta/v_{\mathrm{f}} + \tau)}{1 - \mu a_{\mathrm{p}}^{*}(i)}. \tag{6.13}$$

Now, if substitute $q^{*}(i)$ into Eq. (6.12), it can be found that when $1/(1-\delta/l) < 2$,

$$n_{\mathrm{p}}^{\max}(q^{*}(i)) < n_{\mathrm{p}}^{*}(i). \tag{6.14}$$

From practical experience, minimum safety distance $\delta < l/2$ should be a general situation. Thus, Eq. (6.14) indicates that if $q(i)$ remains constant, when $n_{\mathrm{p}}(i)$ or equivalently $a_{\mathrm{p}}(i)$ enters into the congested regime from the critical state, $TTC(i)$ would increase. Therefore, under optimal situation, traffic state on train-FD should not enter the congested regime.

Next, let's express the SDC in a more explicit way. For simplicity, the sum of the second and the third terms in Eq. (6.9) can be approximated by the two right-angled triangular area between cumulative passenger departure curve D_{p} and demand curve $W(t)$. Thus, SDC can be approximately written as:

$$SDC \approx \frac{\beta(t^{*} - t_0)N_{\mathrm{p1}}}{2} + \frac{\gamma(t_{\mathrm{ed}} - t^{*})N_{\mathrm{p2}}}{2}, \tag{6.15}$$

where N_{p1} is the total number of passengers who exited the railway system before t^{*} and N_{p2} is the passenger number departing after t^{*}. Thus, the following equation hold.

$$N_{\mathrm{p1}} + N_{\mathrm{p2}} = N_{\mathrm{p}}. \tag{6.16}$$

Note, different from previous subsection, if the passenger congestion influence is considered, N should be a increasing function of N_{p}, say for example, $N = f(N_{\mathrm{p}})$. Then, substitute individual equilibrium cost C^{e} (i.e., Eq. (6.2) and (6.5)) and Eq. (6.16) into Eq. (6.15), SDC

can be written as:

$$\begin{aligned} SDC &\approx \frac{\beta(t^*-t_0)N_{\mathrm{p1}}}{2}+\frac{\gamma(t_{\mathrm{ed}}-t^*)N_{\mathrm{p2}}}{2}=\frac{C^e(a_1,a_2,N_{\mathrm{p}})N_{\mathrm{p}}}{2} \\ &= \frac{f(N_{\mathrm{p}})N_{\mathrm{p}}}{2\left(\frac{\alpha-\beta}{\alpha\beta}a_1+\frac{\alpha+\gamma}{\alpha\gamma}a_2\right)}. \end{aligned} \tag{6.17}$$

Now, it can be understood from Eq. (6.17) that even though passenger congestion influence is considered, the optimization problem of train inflow does not change essentially. The conclusion is still the same as the previous subsection from the perspective of reducing the travel cost.

However, there is a paradox between adopting high dispatch frequency and stability of operation. This paradox can be understood from the density and flow dynamics on train-FD under the equilibrium. Basically, the evolution of $(k(i),q(i))$ on train-FD horizontally (because equilibrium train flow are separately constant for the peak/off-peak timetable before or after t^*) first moves rightward and then moves leftward back. In this way, the travel time $T(i)$ would linearly increase and then decrease as required by the equilibrium. If a quite high inflow is adopted, the number of passengers that can be carried by trains operating in the free-flow regime of train-FD would decrease (this is because the upper part of train-FD is narrower). As a result, the operation condition quickly enters the congested regime, the system becomes unstable and bunching phenomenon occurs. Therefore, the high dispatch frequency which leads to lower SDC should be balanced with the stability of operation. Here, an approximate optimum situation is considered for the peak/off-peak timetable pattern:

- All the dispatched trains operate in free-flow regime of train-FD and only the train experienced the longest travel time reaches the critical state on train-FD.
- Average equilibrium flow before and after t^* should be the same.

Let's explain why this is an approximate optimum situation. First, it is necessary to prove that total TTC should be approximate minimized in the described situation. This can be done by jointly considering the characteristic of free-flow regime train-FD (i.e., Eq. (4.16a)) and relation between headway and travel time (i.e., Eq. (4.14)) under equilibrium.

$$n_{\mathrm{p}}(i)-n_{\mathrm{p}}(i-1)=\frac{T(i)-T(i-1)}{\mu L/l}=\frac{H_{\mathrm{d}}(i)-H_{\mathrm{a}}(i)}{\mu L/l}. \tag{6.18}$$

Meanwhile, according to FIFO under equilibrium, departure headway can further be written as functions arrival headway:

$$H_{\mathrm{d}}(i) = \begin{cases} \dfrac{1}{1-\beta/\alpha} \cdot H_{\mathrm{a}}(i), & \text{if } i \in [1, N_1], \\ \dfrac{1}{1+\gamma/\alpha} \cdot H_{\mathrm{a}}(i), & \text{if } i \in [N_1+1, N], \end{cases} \tag{6.19}$$

Substitute Eq. (6.19) into Eq. (6.18) and use inflow to replace H_{a}, Eq. (6.18) can be further written as:

$$n_{\mathrm{p}}(i) - n_{\mathrm{p}}(i-1) = \begin{cases} \dfrac{\beta l}{(\alpha-\beta)\mu L} \cdot \dfrac{1}{a_1}, & \text{if } i \in [1, N_1], \\ -\dfrac{\gamma l}{(\alpha+\gamma)\mu L} \cdot \dfrac{1}{a_2}, & \text{if } i \in [N_1+1, N], \end{cases} \tag{6.20}$$

Now, it is clear that the difference of riding passenger number between two successive trains is inversely proportional to the inflow. Also notice that total TTC is a sum of Eq. (6.11a) under the constraint Eq. (6.10). Therefore, a small difference of n_{p} is preferable to reduce the total TTC. This means an sufficiently larger train inflow can reduce the total TTC to the maximum extent. Meanwhile, higher inflow would also reduce SDC as can be understood from Eq. (6.17). Thus, it is always beneficial to adopt inflow as high as possible. However, when inflow is too high that traffic state enters the congested regime, TTC would contrarily increase (see discussion after Eq. (6.14)), and the operation becomes unstable due to the bunching phenomenon. Therefore, to adopt a as high as possible inflow while keeping traffic state within the free-flow regime of train-FD should be an approximate optimum situation. This situation is described by the previous two items.

6.1.3 Solution and numerical example

This subsection further solves the approximate optimum situation introduced in the previous subsection. A numerical example of system performance under optimum situation is given and compared with other situations. First, average flow under equilibrium before t^* calculated from particle-based dynamic models can be written as:

$$q_1 = \left[\frac{H_{\mathrm{a}} + H_{\mathrm{d}}}{2}\right]^{-1} = \left(\frac{(a_1)^{-1} + ((1-\beta/\alpha)a_1)^{-1}}{2}\right)^{-1} = \frac{\alpha-\beta}{\alpha-\beta/2} a_1. \tag{6.21}$$

Similarly, average equilibrium flow after t^* can be derived as:

$$q_2 = \frac{\alpha+\gamma}{\alpha+\gamma/2} a_2. \tag{6.22}$$

Then, on one hand, from the second requirement of optimum situation $q_1 = q_2 = q$, the ratio of inflow a_1 and a_2 can be determined from Eq. (6.21) and (6.22) as:

$$q_1 = \frac{\alpha - \beta}{\alpha - \beta/2} a_1 = q_2 = \frac{\alpha + \gamma}{\alpha + \gamma/2}, \tag{6.23}$$

$$\frac{a_1}{a_2} = \frac{(\alpha + \gamma)(\alpha - \beta/2)}{(\alpha - \beta)(\alpha + \gamma/2)}. \tag{6.24}$$

It can be seen that the optimum ratio of peak/off-peak train inflow only relates to the passengers' value of time. This is a valuable insight for railway operator to verify and refer.

On the other hand, from the first requirement of optimum situation, the maximum density $k_{\max}$ should not be larger than the critical density $k^*(q)$, which can be written as:

$$k_{\max} = \frac{T_{\max}}{L} \cdot q \leq k^*(q) = \frac{1 + [(l - \delta)/v_{\mathrm{f}} - \tau] q}{l}. \tag{6.25}$$

Substitute $T_{\max} = T_0 + \beta/\alpha(t^* - t_0)$ into the above formula, use q_1 as an example:

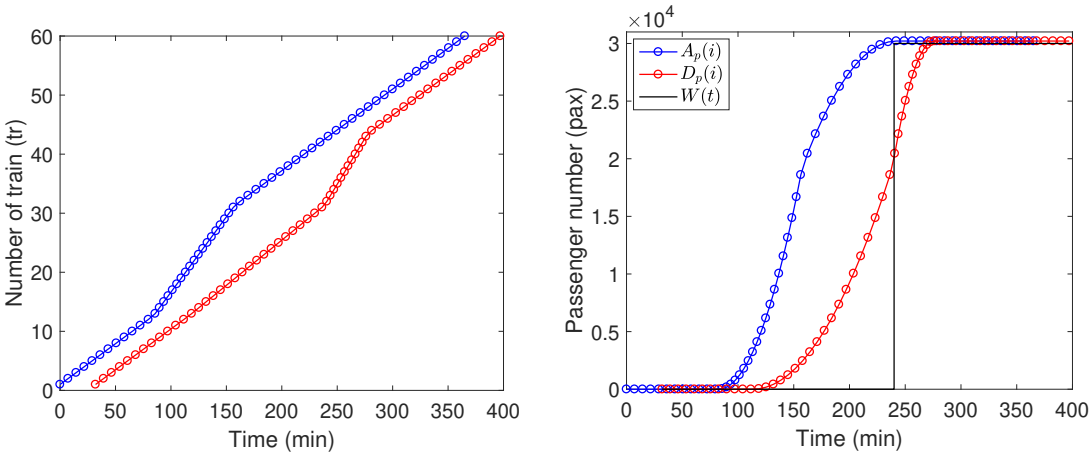
$$\left[\frac{T_0 + \beta/\alpha(t^* - t_0)}{L} - \frac{(l - \delta)/v_{\mathrm{f}} - \tau}{l} \right] \frac{\alpha - \beta}{\alpha - \beta/2} a_1 \leq \frac{1}{l}. \tag{6.26}$$

Now, take the equal condition of Eq. (6.26), optimum inflow a_1 and also a_2 can be expressed as a function of t_0. Therefore, the solution problem becomes the same as the one described in Algorithm 1 in Chapter 5. The only difference is that inflow is not pre-determined but as a function of t_0.

The numerical example here still use the same parameters as shown in Table 5.1. The cumulative curves of trains and passengers are drawn in the following figure. By comparing the cumulative curves of trains under optimum situation with constant case (i.e., Fig. 5.1a), it can be observed that the change of travel time is realized mainly by adjusting train inflow. In this way, the outflow is always kept at a relatively high level, and schedule delay cost will be reduced compared to the constant case. The cumulative passenger curve under optimum situation also becomes smoother compared to constant case.

The evolution of density and flow is also depicted on train-FD as shown in Fig. 6.4. It can be seen that under this optimum situation, the average flow of railway system is always kept constant during the equilibrium period. The maximum density reached the critical state on train-FD. Therefore, the result of numerical experiment is consistent with the first description of optimum situation in the previous subsection.

In order to further understand the advantage of this approximate optimum situation over other timetable situations, the costs (including TTC and SDC) are compared among the



(a) Cumulative number of trains (b) Cumulative number of passengers

Fig. 6.3 Cumulative curves for optimal peak/off-peak timetable pattern.

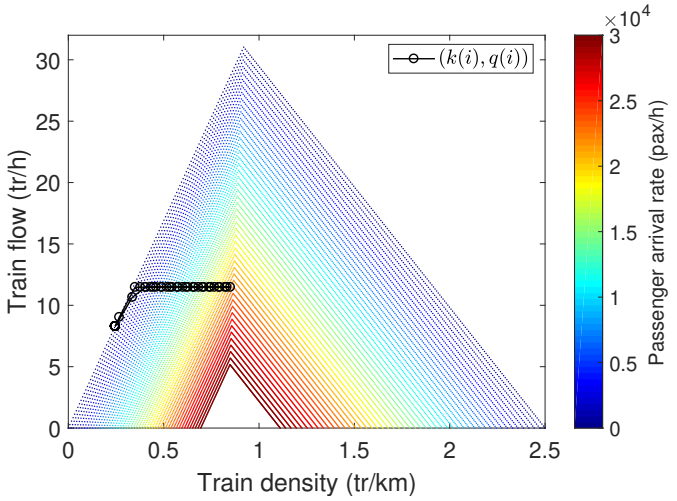


Fig. 6.4 Dynamics of density and flow on train-FD for optimal peak/off-peak timetable pattern.

optimum case (OPM), constant case (CON) and three practical cases (PR1 to PR3) in Table 6.1.

Table 6.1 Comparison of passenger costs for different timetable patterns.

Scenario	a_1 (tr/h)	a_2 (tr/h)	**Duration** (a_1, min)	$\sum TTC$ $(10^4$ \$)	$\sum SDC$ $(10^4$ \$)	$\sum TC$ $(10^4$ \$)	**TC Change (%)**
OPM	15.3	8.3	74	33.34	16.09	49.43	-
CON	10.3	10.3	-	38.09	17.12	55.20	+10
PR1	20.0	7.5	60	33.09	18.51	51.60	+4
PR2	24.0	6.0	60	36.85	21.55	58.40	+15
PR3	20.0	9.5	30	34.88	16.72	51.60	+4

The number of dispatched train that carry $N_p = 30000$ passengers is kept constant as $N = 43$ in all scenarios. This is realized by adjusting train inflow, especially a_2. The last column presents the TC change compared to OPM case.

As can be confirmed from Table 6.1, compared to OPM case, the TC in CON case increased 10%. This is mainly due to the increase of TTC since the operation of some trains after t^* entered congested regime of train-FD as can be understood from comparison of Fig. 5.4b and Fig. 6.4. As discussed in the previous subsection, operation under congested regime would raise the TTC since more time is wasted on cruising with lower speed between stations. Regarding the three practical cases, the increase of TC in PR1 and PR3 are relatively small while PR2 using a quite high a_1 yields a 15% increase of TC. This can be explained from two aspects: 1, adopting extremely high peak inflow leads to a vulnerable operation that trains easily enter the congested regime of train-FD; 2, due to the total available train fleet N constraint, high peak inflow makes the available train number for off-peak frequency less, these trains have to carry more passengers since trains dispatched with peak frequency did not carry enough passengers in congested regime. As a result, the capacity of trains are not efficiently used, so both TTC and SDC significantly increased for PR2. This can also be understood from the density and flow evolution on train-FD as shown in Fig. 6.5.

From the comparison in Table 6.1, two insights for trains operation without pricing scheme can be concluded:

- A moderate peak/off-peak timetable is preferable from the perspective of both reducing total travel cost and robust operation.
- Adopt a relatively small difference between peak/off-peak dispatch frequencies and extends the duration of peak frequency might be a good direction towards the optimum situation.

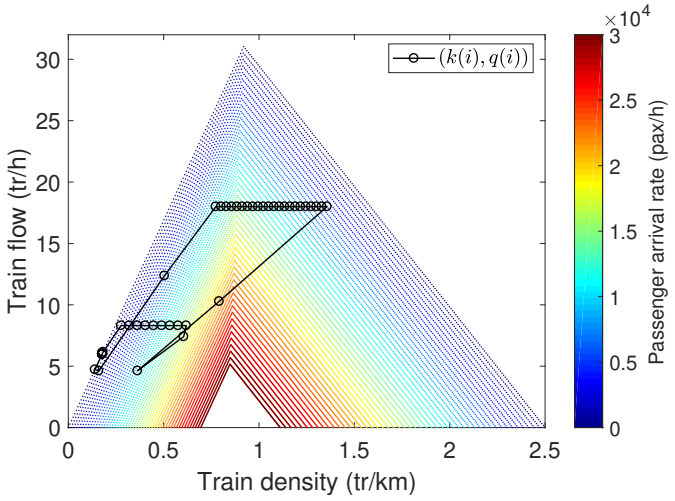


Fig. 6.5 Dynamics of density and flow on train-FD for scenario PR2.

6.2 User equilibrium with pricing

In this section, user equilibrium under first-best pricing is firstly derived and compared with no pricing condition both theoretically and numerically. Then, a more practical (or second-best) one-step pricing scheme is proposed and evaluated. Policy implication is derived from the evaluation of this one-step pricing scheme.

6.2.1 First-best pricing

The first best pricing generally describes the pricing scheme which is able to minimize social cost (SC). In this study, SC consists of TTC and SDC, thus the objective is still to minimize the sum of TTC and SDC as written in Eq. (6.9). Here, the user equilibrium under first-best pricing is described as the following condition.

Lemma 1. *Assume that (1), $n_{\mathrm{p}}(i) > 0$ for any available train $i \in [1, N]$ that carry N_{p} commute passengers, and (2), any train i operate in free-flow regime of train-FD. Then, time boundary (i.e., t_0 and t_{ed}) is determined by N_{p} and N, and the riding passenger number is monotonically increasing for trains leaving the system before t^*, and decreasing for trains leaving after t^*. The increments are respectively:*

$$n_{\mathrm{p}}(i) - n_{\mathrm{p}}(i-1) = \begin{cases} \dfrac{\beta H_{\mathrm{a}}(i)}{2(\alpha-\beta)\mu L/l}, & \text{if } i \in [2, N_1], \\ -\dfrac{\gamma H_{\mathrm{a}}(i)}{2(\alpha+\gamma)\mu L/l}, & \text{if } i \in [N_1+2, N]. \end{cases} \tag{6.27}$$

Meanwhile, the relation between arrival and departure headway before and after t^* can respectively be derived as:

$$H_{\mathrm{d}}(i) = \begin{cases} \dfrac{2\alpha-\beta}{2(\alpha-\beta)} H_{\mathrm{a}}(i), & \text{if } i \in [2, N_1], \\ \dfrac{2\alpha+\gamma}{2(\alpha+\gamma)} H_{\mathrm{a}}(i), & \text{if } i \in [N_1+2, N], \end{cases} \tag{6.28}$$

Proof. As a preparation work to prove the proposed lemma, the social cost described in Eq. (6.9) can be re-written as:

$$\begin{aligned} SC =& \alpha \sum_{i=1}^{N} \left(T(i) - T_0\right) n_{\mathrm{p}}(i) + \beta \sum_{i=1}^{N_1} \left[t^* - \left(t_0 - T_0 + \sum_{j=1}^{i} H_{\mathrm{a}}(j) + T(i) \right) \right] n_{\mathrm{p}}(i) \\ &+ \gamma \sum_{i=N_1+1}^{N} \left[\left(t_0 - T_0 + \sum_{j=1}^{i} H_{\mathrm{a}}(j) + T(i) \right) - t^* \right] n_{\mathrm{p}}(i). \end{aligned} \tag{6.29}$$

where N_1 is the number of trains that leave the system before t^* and $N_2 = N - N_1$ is the number of trains after t^*. Since all dispatched trains are assumed to operate in free-flow regime of train-FD, inspired by Eq. (6.11a), Eq. (6.29) can be further simplified as:

$$\begin{aligned} SC =& (\alpha-\beta)\mu \frac{L}{l} \sum_{i=1}^{N} \left(n_{\mathrm{p}}(i)\right)^2 + \beta \sum_{i=1}^{N_1} \left(t^* - t_0 - \sum_{j=1}^{i} H_{\mathrm{a}}(j) \right) n_{\mathrm{p}}(i) \\ &+ (\alpha+\gamma)\mu \frac{L}{l} \sum_{i=N_1+1}^{N} \left(n_{\mathrm{p}}(i)\right)^2 + \gamma \sum_{i=N_1+1}^{N} \left(t_0 + \sum_{j=1}^{i} H_{\mathrm{a}}(j) - t^* \right) n_{\mathrm{p}}(i). \end{aligned} \tag{6.30}$$

Now, it can be observed from Eq. (6.30) that SC is basically determined by the combination of linear and quadratic function of $n_{\mathrm{p}}(i)$ with given information such as α, β, γ, L/l, μ, t^* and $H_{\mathrm{a}}(i)$, and t_0 determined by N_{p}. Under the total available train fleet constraint N and total travel demand constraint N_{p} (i.e., Eq. (6.10)), the minimum of SC can be derived by the method of Lagrange multiplier. Introduce a Lagrange function as:

$$\mathscr{L}(n_{\mathrm{p}}(i), \lambda) = SC - \lambda \left(\sum_{i=1}^{N} n_{\mathrm{p}}(i) - N_{\mathrm{p}} \right). \tag{6.31}$$

If the minimum of SC exists, the partial derivative of Lagrange function with any $n_{\mathrm{p}}(i)$ should be zero, that is:

$$\frac{\partial \mathscr{L}}{\partial n_{\mathrm{p}}(i)} = 0, \ \forall i. \tag{6.32}$$

More specifically, Eq. (6.32) can be written as:

$$\begin{cases} 2(\alpha-\beta)\mu\dfrac{L}{l}n_{\mathrm{p}}(1)+\beta\left(t^*-t_0-\sum\limits_{j=1}^{1}H_{\mathrm{a}}(j)\right)-\lambda=0, \\ \cdots, \\ 2(\alpha-\beta)\mu\dfrac{L}{l}n_{\mathrm{p}}(N_1)+\beta\left(t^*-t_0-\sum\limits_{j=1}^{N_1}H_{\mathrm{a}}(j)\right)-\lambda=0, \\ 2(\alpha+\gamma)\mu\dfrac{L}{l}n_{\mathrm{p}}(N_1+1)+\gamma\left(t_0+\sum\limits_{j=1}^{N_1+1}H_{\mathrm{a}}(j)-t^*\right)-\lambda=0, \\ \cdots, \\ 2(\alpha+\gamma)\mu\dfrac{L}{l}n_{\mathrm{p}}(N)+\gamma\left(t_0+\sum\limits_{j=1}^{N}H_{\mathrm{a}}(j)-t^*\right)-\lambda=0. \end{cases} \tag{6.33}$$

Now, the difference of two successive equations of Eq. (6.33) can be easily obtained as:

$$n_{\mathrm{p}}(i)-n_{\mathrm{p}}(i-1)=\begin{cases} \dfrac{\beta H_{\mathrm{a}}(i)}{2(\alpha-\beta)\mu L/l}, & \text{if } i\in[2,N_1], \\ -\dfrac{\gamma H_{\mathrm{a}}(i)}{2(\alpha+\gamma)\mu L/l}, & \text{if } i\in[N_1+2,N]. \end{cases} \tag{6.34}$$

The first statement of the lemma is therefore proved.

Next, the second statement can be proved by employing the relation between travel time difference and headway difference (i.e., Eq. (4.14)). In free-flow regime of train-FD, this relation can be further written as:

$$\begin{aligned} H_{\mathrm{d}}(i) &= T(i)-T(i-1)+H_{\mathrm{a}}(i) \\ &= \mu\frac{L}{l}(n_{\mathrm{p}}(i)-n_{\mathrm{p}}(i-1))+H_{\mathrm{a}}(i). \end{aligned} \tag{6.35}$$

Substitute Eq. (6.34) into Eq. (6.35), the relation between departure headway $H_{\mathrm{d}}(i)$ and arrival headway $H_{\mathrm{a}}(i)$ can be derived as:

$$H_{\mathrm{d}}(i)=\begin{cases} \left(1+\dfrac{\beta}{2(\alpha-\beta)}\right)H_{\mathrm{a}}(i)=\dfrac{2\alpha-\beta}{2(\alpha-\beta)}H_{\mathrm{a}}(i), & \text{if } i\in[2,N_1], \\ \left(1-\dfrac{\gamma}{2(\alpha+\gamma)}\right)H_{\mathrm{a}}(i)=\dfrac{2\alpha+\gamma}{2(\alpha+\gamma)}H_{\mathrm{a}}(i), & \text{if } i\in[N_1+2,N], \end{cases} \tag{6.36}$$

The second statement of the lemma is therefore proved. □

Under the situation described in Lemma 1, the riding passenger number on each train i is actually determined since combining Eq. (6.33) with N_{p} constraint, there are $N+1$ unknown variables and $N+1$ linearly independent equations. Therefore, λ and all $n_{\mathrm{p}}(i)$ can be uniquely solved if $H_{\mathrm{a}}(i)$ for any i is given. Note, t^*-t_0 here can be obtained as:

$$t^*-t_0=\sum_{i=2}^{N_1}H_{\mathrm{d}}(i)=\frac{2\alpha-\beta}{2(\alpha-\beta)}\sum_{i=2}^{N_1}H_{\mathrm{a}}(i). \tag{6.37a}$$

$$t_{\mathrm{ed}}-t^*=\sum_{i=N_1+2}^{N}H_{\mathrm{d}}(i)=\frac{2\alpha+\gamma}{2(\alpha+\gamma)}\sum_{i=N_1+2}^{N}H_{\mathrm{a}}(i). \tag{6.37b}$$

Since solving the general case of Eq. (6.33) with N_{p} constraint is tedious, a special case when $H_{\mathrm{a}}(i)\equiv H_{\mathrm{a}}$ is picked up as an example to show the explicit calculation result. When H_{a} is constant, N_1 and N_2 can be solved by jointly using the equal boundary cost (i.e., Eq. (6.2)) and Eq. (6.37) as:

$$N_1=\frac{\gamma(2\alpha+\gamma)(\alpha-\beta)}{(\beta+\gamma)\left[\alpha(2\alpha-\beta+\gamma)-\beta\gamma\right]}N, \tag{6.38a}$$

$$N_2=\frac{\beta(2\alpha-\beta)(\alpha+\gamma)}{(\beta+\gamma)\left[\alpha(2\alpha-\beta+\gamma)-\beta\gamma\right]}N, \tag{6.38b}$$

By summing all the equations in Eq. (6.33), the Lagrange multiplier λ under constant H_{a} case can be solved as:

$$\begin{aligned}\lambda=&\frac{\beta\gamma[\alpha\gamma(2\alpha+\gamma)+(\alpha(2\alpha-\beta+\gamma)-\beta\gamma)(2\alpha-\beta)]}{2(\beta+\gamma)(\alpha(2\alpha-\beta+\gamma)-\beta\gamma)(2\alpha-\beta+\gamma)}NH_{\mathrm{a}}\\&-\frac{\beta\gamma}{2(2\alpha-\beta+\gamma)}H_{\mathrm{a}}+2\mu\frac{LN_{\mathrm{p}}}{lN}\left(\alpha-\frac{\beta\gamma}{2\alpha-\beta+\gamma}\right).\end{aligned} \tag{6.39}$$

In order to reach the minimum SC, the first-best (or continuous) pricing scheme can be derived considering the total travel cost (TC), or user cost (UC) hereinafter as:

$$UC^{EP}(t,t^*)=\alpha\left(T^{EP}(t)-T_0\right)+s(t,t^*)+p(t). \tag{6.40}$$

Here, the superscript "EP" means equilibrium with first-best pricing. Since directly solving $T^{EP}(t)$ relates to Eq. (6.33) and makes the derivation difficult, it is easier to first rewrite Eq.

(6.36) using inflow and outflow as:

$$d(t)=\begin{cases}\dfrac{2(\alpha-\beta)}{2\alpha-\beta}a(t-T(t))=\left(1-\dfrac{\beta}{2\alpha-\beta}\right)a(t-T(t)), & \text{if } t\in[t_0,t^*],\\ \dfrac{2(\alpha+\gamma)}{2\alpha+\gamma)}a(t-T(t))=\left(1+\dfrac{\gamma}{2\alpha+\gamma}\right)a(t-T(t)), & \text{if } t\in(t^*,t_{\mathrm{ed}}].\end{cases}\tag{6.41}$$

Meanwhile, notice the FIFO principle that $D(t)=A(t-T(t))$, differentiate this equation and compare with Eq. (6.41), it can be understood that the derivatives of T^{EP} can be obtained as:

$$\frac{dT^{EP}(t)}{dt}=\begin{cases}\dfrac{\beta}{2\alpha-\beta}, & \text{if } t\in[t_0,t^*],\\ -\dfrac{\gamma}{2\alpha+\gamma}, & \text{if } t\in(t^*,t_{\mathrm{ed}}].\end{cases}\tag{6.42}$$

Then, since the individual cost does not change under equilibrium, $\partial UC^{EP}(t,t^*)/\partial t=0$, the derivatives of first-best pricing can be derived as:

$$\frac{dp(t)}{dt}=\begin{cases}\dfrac{(\alpha-\beta)\beta}{2\alpha-\beta}, & \text{if } t\in[t_0,t^*],\\ -\dfrac{(\alpha+\gamma)\gamma}{2\alpha+\gamma}, & \text{if } t\in(t^*,t_{\mathrm{ed}}].\end{cases}\tag{6.43}$$

Finally, consider the boundary condition of $n_{\mathrm{p}}(N_1)$ and $T(N_1)$, the first-best pricing $p(t)$ can be expressed as:

$$p(t)=\begin{cases}\dfrac{(\alpha-\beta)\beta}{2\alpha-\beta}(t-t^*)+p_{\mathrm{m}}, & \text{if } t\in[t_0,\ t^*],\\ -\dfrac{\gamma}{2\alpha+\gamma}(t-t^*)+p_{\mathrm{m}}, & \text{if } t\in(t^*,\ t_{\mathrm{ed}}],\end{cases}\tag{6.44}$$

where

$$\begin{aligned}p_{\mathrm{m}}&=\left(1+\frac{\alpha}{2(\alpha-\beta)}\right)\beta(t^*-t_0)-\frac{\alpha\left(\lambda+\beta\sum_{i=1}^{N_1}H_{\mathrm{a}}(i)\right)}{2(\alpha-\beta)}\\&=\left(1+\frac{\alpha}{2(\alpha-\beta)}\right)\cdot\frac{2\alpha-\beta}{2(\alpha-\beta)}\sum_{i=2}^{N_1}H_{\mathrm{a}}(i)-\frac{\alpha\left(\lambda+\beta\sum_{i=1}^{N_1}H_{\mathrm{a}}(i)\right)}{2(\alpha-\beta)}.\end{aligned}\tag{6.45}$$

For a better understanding, the UC^{EP}, SDC and $p(t)$ under first-best pricing are depicted in Fig. 6.6. $p(t_0)$ and $p(t_{\mathrm{ed}})$ can be determined when N is given or by vertically move $p(t)$ upward or downward, N can be raised or reduced. But since $n_{\mathrm{p}}(1)$ and $n_{\mathrm{p}}(N)$ are undoubtedly

positive according to the assumption, $p(t_0) < 0$ and $p(t_{\text{ed}}) < 0$ should hold even in the general case. This indicates that giving time-dependent reward to a portion of passengers departing at the shoulder of the rush hours might be necessary to reach the minimum social cost for a urban rail transit system. This is a rather novel insight from this study.

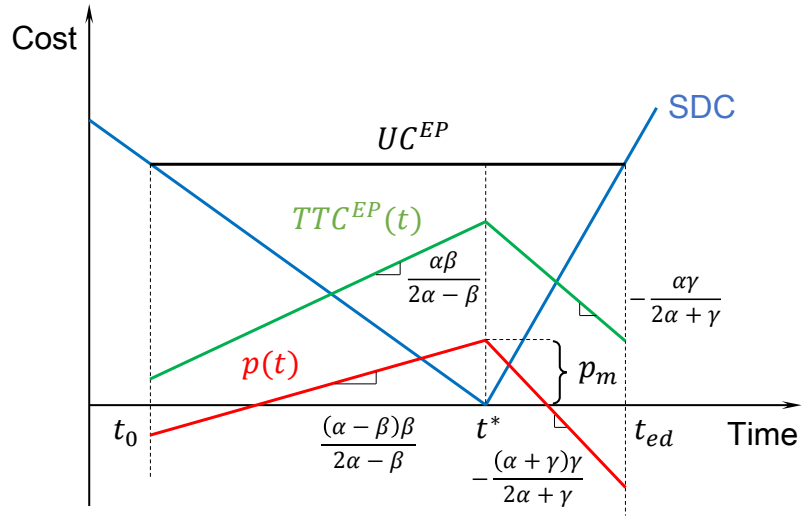


Fig. 6.6 Costs under first-best pricing.

6.2.2 Comparison of UE with/without pricing

This subsection first theoretically compares the change of travel time and riding passenger number under user equilibrium with or without first-best pricing. Then, a numerical example is given to compare the cost change when the first-best pricing is applied.

The derivative of travel time in UE and EP have been derived in Chapter 5 and previous subsection. Here, it is clearly compared again as:

$$\frac{dT^{UE}(t)}{dt} = \begin{cases} \dfrac{\beta}{\alpha}, & \text{if } t \in [t_0, t^*], \\ -\dfrac{\gamma}{\alpha}, & \text{if } t \in (t^*, t_{\text{ed}}]. \end{cases} \tag{6.46a}$$

$$\frac{dT^{EP}(t)}{dt} = \begin{cases} \dfrac{\beta}{2\alpha-\beta}, & \text{if } t \in [t_0, t^*], \\ -\dfrac{\gamma}{2\alpha+\gamma}, & \text{if } t \in (t^*, t_{\text{ed}}]. \end{cases} \tag{6.46b}$$

It can be understood from Eq. (6.46) that the absolute value of travel time derivatives for EP are always smaller than those for UE as long as $\alpha > \beta$ holds. This means passenger

arrival influence is weakened in EP compared to UE. Besides, the riding passenger difference between successive two trains for UE and EP can also be summarized as:

$$n_{\mathrm{p}}^{UE}(i)-n_{\mathrm{p}}^{UE}(i-1)=\begin{cases}\dfrac{\beta H_{\mathrm{a}}(i)}{(\alpha-\beta)\mu L/l}, & \text{if } i\in[2,N_1],\\ -\dfrac{\gamma H_{\mathrm{a}}(i)}{(\alpha+\gamma)\mu L/l}, & \text{if } i\in[N_1+2,N].\end{cases} \tag{6.47a}$$

$$n_{\mathrm{p}}^{EP}(i)-n_{\mathrm{p}}^{EP}(i-1)=\begin{cases}\dfrac{\beta H_{\mathrm{a}}(i)}{2(\alpha-\beta)\mu L/l}, & \text{if } i\in[2,N_1],\\ -\dfrac{\gamma H_{\mathrm{a}}(i)}{2(\alpha+\gamma)\mu L/l}, & \text{if } i\in[N_1+2,N].\end{cases} \tag{6.47b}$$

It can be seen from Eq. (6.47) that the increment/decrement of n_{p} for EP is exactly half of that for UE. This indicates that passengers are more evenly distributed in the dispatched trains when first-best pricing is applied.

Next, UE and EP are numerically compared by the change of total TTC, SDC and UC. The environmental parameters still take the same values as in Table 5.1. Train inflow $a(t)\equiv 12\ tr/h$ (i.e., $H_{\mathrm{a}}\equiv 5\ min$). A comparison when number of trains used to carry N_{p} is the same (i.e., $N^{EP}=N^{UE}$) are shown in Table 6.2. It can be confirmed from Table 6.2

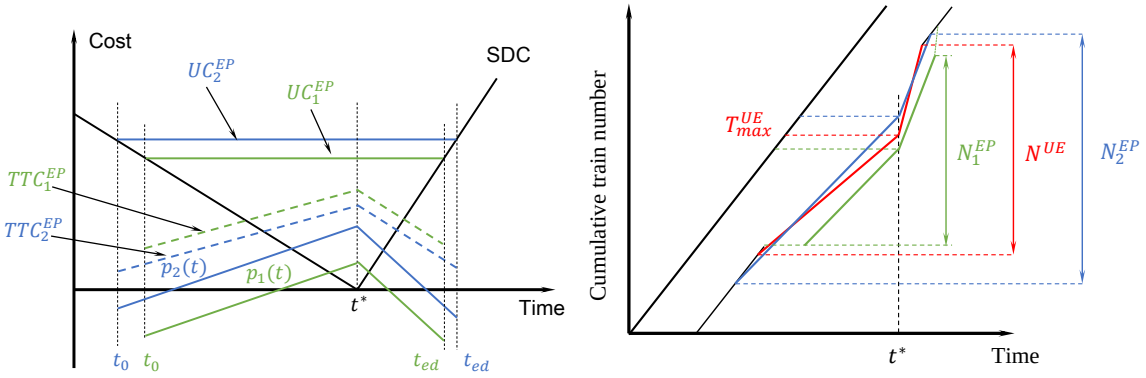
Table 6.2 Comparison of total costs between UE and EP when $N^{EP}=N^{UE}$.

	$\sum TTC$ (10^4 \$)	$\sum SDC$ (10^4 \$)	SC (10^4 \$)	$\sum UC$ (10^4 \$)
UE	33.69	19.59	53.29	53.29
EP	26.28	16.53	42.80	58.67
Change	-22.0%	-15.7%	-19.7%	+10.1%

the sum of TTC and SDC decreased 22% and 15.7% respectively from UE to EP when $N^{EP}=N^{UE}$. As a result, the sum of social cost declined around 20%. This is a remarkable reduction from the comparison. To realize the change from UE to EP, a maximum $p_{\mathrm{m}}=5.3$\$ is charged from passengers who want to leave the system on time. Due to this first-best pricing, the user cost increased around 10%.

Further, the sensitivity of pricing amount (or equivalently N^{EP}) is numerically analyzed. First, it has to be clarified that by changing the setting of boundary p, the equilibrium period may be extended or shortened as can be understood from Fig. 6.7. If $p(t)$ is raised up or pulled down as a whole, the TTC would act in an opposite way because their sum should be zero at the t_0 and t_{ed} (otherwise equilibrium cost is not continuous at the boundary). When TTC is raised up or pulled down as a whole, $n_p(i)$ in each train increases or decreases

accordingly. As a result, the number of train N^{EP} needed to carry N_{p} passengers decreases or increases.



(a) Travel costs (b) Cumulative curves of trains

Fig. 6.7 An illustration of pricing amount sensitivity.

Moreover, from the perspective of N^{EP}, the travel costs under different N^{EP} are compared with the no pricing equilibrium in Fig. 6.8. First, it can be observed that social cost is not sensitive to the change of N^{EP} because sum of TTC and SDC almost neutralizes as long as the first-best pricing is implemented. In fact, SC slightly decreases with the increase of N^{EP} until $n_{\mathrm{p}}(1) = 0$ or $n_{\mathrm{p}}(N^{EP}) = 0$. Second, user cost can be found to change sensitively with N^{EP}, this is reasonable because when more trains are used to carry the same amount of passengers, the pricing value should be higher to force the passengers to use earlier or later trains. As a result, UC increases. Third, on the contrary, to excessively reduce N^{EP} by shifting $p(t)$ lower leads to heavier in-vehicle crowding. Therefore, governor or operator should balance the social cost, user cost and in-vehicle crowding when setting the amount of first-best pricing.

6.2.3 Coarse pricing

In this section, a more practical pricing scheme: coarse pricing is discussed and evaluated. Through the numerical experiments, Pareto-improvement is found to exist by appropriately implementing the coarse pricing scheme.

Since the first-best pricing derived in previous section involves practical difficulties in collecting time-dependent fare or giving time-dependent reward, an approximation, step-tolling, has been widely discussed for road traffic context (Arnott et al., 1990b; Laih, 1994; Lindsey et al., 2012). With regard to rail transit, studies on fare scheme design and optimization are still at the beginning stage compared to road traffic (De Palma et al., 2017; Yang and

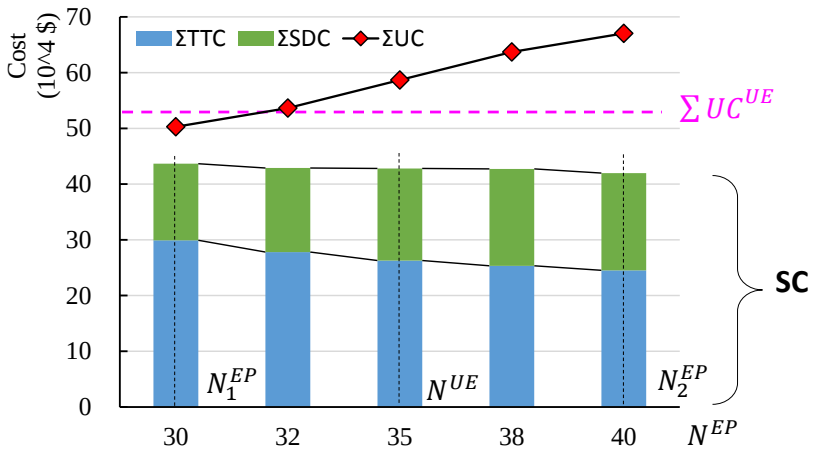


Fig. 6.8 Sensitivity of N^{EP} under first-best pricing

Tang, 2018; Zhang et al., 2020). Inspired by the first-best pricing derived above and previous step-tolling studies, a single-step pricing scheme, i.e., coarse pricing is evaluated in this subsection.

When a coarse pricing scheme is introduced, user cost (UC) can be written as the following equation (for simplicity, only consider the case when $t_i^* = t^*$ for any i):

$$UC(t,t^*) = \alpha\left(T(t) - T_0\right) + s(t,t^*) \pm p, \text{ if } t \in [t^+,t^-], \tag{6.48}$$

where $+p > 0$ refers to the single-step surcharge, $-p < 0$ refers to the single-step reward. t^+ and t^- are the start and end time of the pricing period, respectively. In order to keep UC unchanged during $[t^+,t^-]$, sudden change of TTC will inevitably appear at t^+ and t^-. An illustration of costs when coarse pricing is applied is shown in Fig. (6.9).

Further, since TTC is proportional to travel time $T(t)$, the change of travel time at time boundary of pricing scheme $\Delta T = p/\alpha$ has to be appropriately interpreted. For road traffic, Arnott et al. (1990b) proposed the concept of mass departures, Laih (1994) assumed the existence of additional lanes for drivers to wait. Lindsey et al. (2012) considered a more practical braking model to describe this phenomenon. For rail transit, mass departures are physically not possible since trains are cruising on the same track. Also, multiple trains ($\Delta n \geq 3$) departing with minimum headway to realize ΔT (i.e., $\Delta n(H_\mathrm{d} - \tau) = p/\alpha$) will break the equilibrium requirement since at least two trains would appear at one side of time boundary with minimum headway τ, which is different from H_d. Note, the coarse pricing is not valid for the situation when trains are departing with minimum headway. The only feasible situation is that the change of travel time due to pricing is carried out by the two trains departing just before and after the time boundary. An example of surcharge is

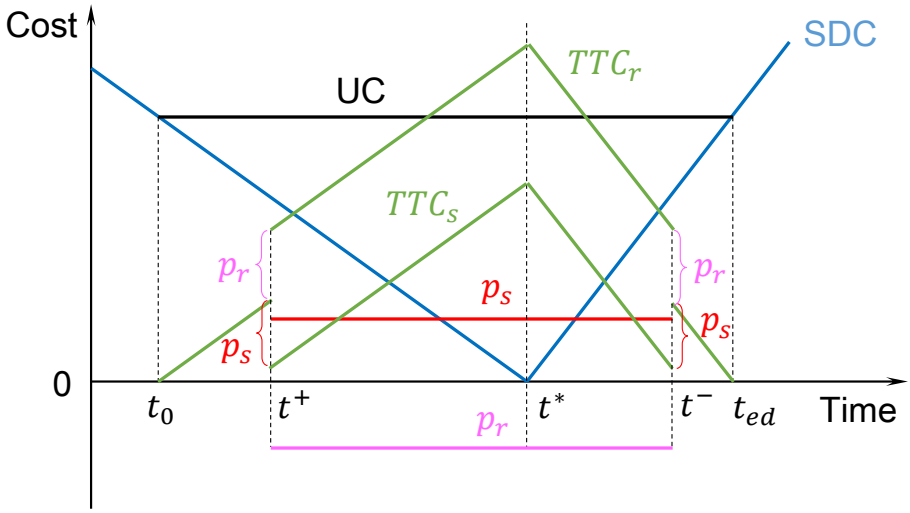


Fig. 6.9 Travel costs when coarse pricing applied

shown in Fig. 6.10. It can observed that the headway of train $m+1$ and $n+1$ departing just after t^+ and t^- decreases/increases p/α. In this way, the sudden change of TTC is realized without breaking the equilibrium. Meanwhile, a constraint of p/α exists due to the FIFO and minimum headway as:

$$\frac{p}{\alpha} \leq H_{\mathrm{d}} - \tau. \tag{6.49}$$

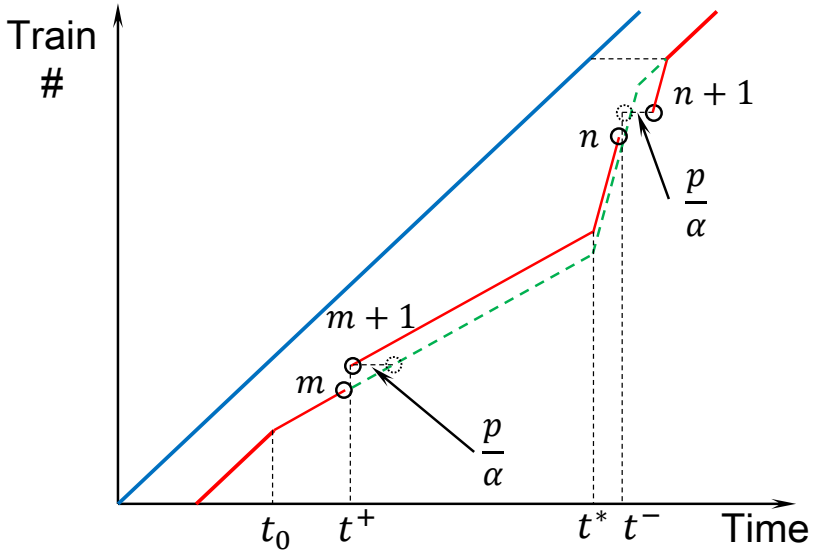


Fig. 6.10 Cumulative curves when surcharge imposed during $[t^+, t^-]$

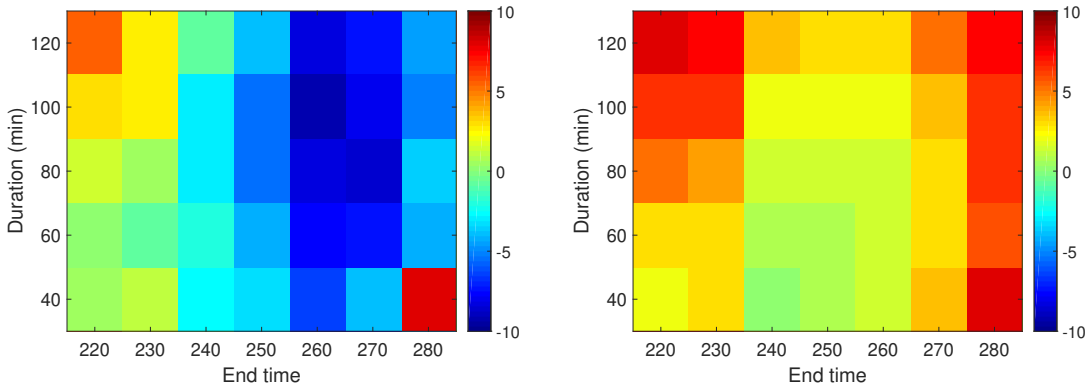
Then, when an appropriate pricing value p is set, solution to the equilibrium model is basically the same with original problem, but only modifies the travel time with $\Delta T = p/\alpha$ and corresponding $k(i)$ for trains departing between $[t^+, t^-]$. The effectiveness of coarse pricing are evaluated by the following three indicators in this study:

- Social cost (SC) change (%) compared to no pricing situation;
- User cost (UC) change (%) compared to no pricing situation;
- Standard deviation (STDEV) of riding passenger number change (%) compared to no pricing situation.

First, SC change evaluates whether the social welfare is improved or not since the monetary circulation between passengers and the operator is neutral from a social perspective. Second, UC change assess whether passengers benefit or not when the pricing scheme is implemented. Third, the change of np STDEV is employed to check whether the in-vehicle crowding is relieved or not because STDEV of np reflects the evenness of passenger distribution among dispatched trains. The decrease of this STDEV indicates the difference of np between congested and uncongested trains is narrowed.

Since the effect of pricing value is obvious and limited by Eq. (6.49), it is more interesting to examine how the temporal settings affect the effectiveness of the coarse pricing. The duration (i.e., $t^+ - t^-$) and end time of the coarse pricing are picked here as the indicators of temporal settings. Fig. 6.11 shows a series of numerical experiments of SC and UC change for the surcharge case $p = 2$ \$. The train inflow in these scenarios takes the constant value $a(t) = 12\ tr/h$. The color represents the percentage of cost change as illustrated by the legend. On one hand, it can be observed from Fig. 6.11a that SC change strongly depends on the end time of surcharge. When the surcharge covers the both side of t^* and ends at an appropriate time after t^* (in this example 20 *min*), its effect is maximized. This is reasonable because when surcharge ends before t^*, the most congested trains carrying passengers who want to be on time are not surcharged, thus the reduction of TTC is weakened. The first-best pricing $p(t)$ in Fig. 6.6 also suggests similar insight that the surcharge should be higher around t^*. On the other hand, the user cost in Fig. 6.11b basically increases as long as the surcharge is imposed. The longer the duration, the earlier or later it ends, the higher user cost would be.

To further interpret the interior reason of UC and SC change, UC is decomposed to the sum of TTC, SDC and total pricing cost (TP). Fig. 6.12 shows an example of each cost component change under a given duration (i.e., $t^- - t^+ = 100$ *min* in Fig. 6.12a) or under a given end time (i.e., $t^- - t^* = 20$ *min* in Fig. 6.12b). On one hand, it can be observed



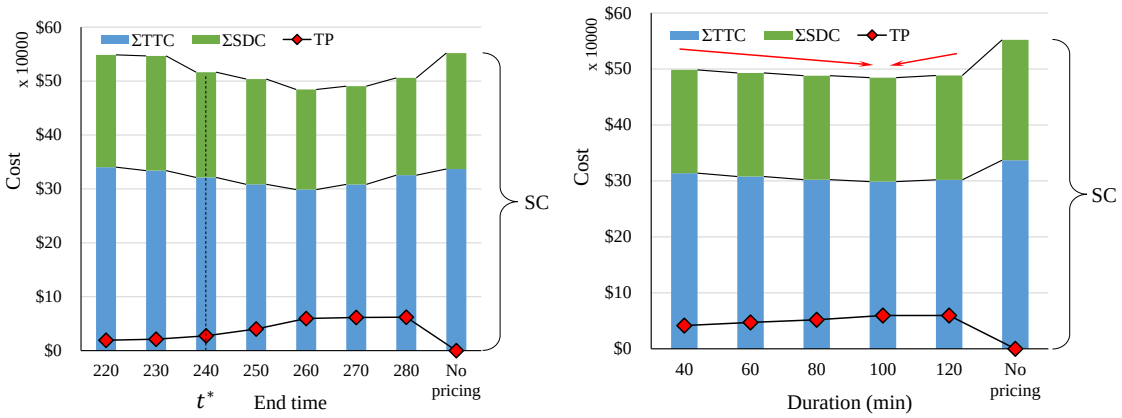
(a) Social cost change (%)

(b) User cost change (%)

Fig. 6.11 Cost sensitivity to temporal settings of surcharge.

from Fig. 6.12a that when the end time of 100 *min* surcharge moves from the left side to the right side of t^*, TP significantly increases while TTC first decreases until 260 *min* (i.e., 20 *min* after t^*) and then inversely increases. The increase of TP means more passengers are surcharged. The change of TTC indicates that there may exist an optimal end time of the surcharge, if it ends further later, the surcharge will cover all late departed commuters and forces some of them to use earlier trains including those that have already been heavily congested. As a result, riding passenger number difference is not efficiently reduced and TTC inversely increases. SDC in this comparison slightly increases or decreases and does not show a consistent result. On the other hand, when the end time of the surcharge is fixed to cover the both sides of t^* and the duration increases as shown in Fig. 6.12b, it can be observed that the TTC slightly declines and then rises with duration extension while SDC almost remains unchanged. This indicates that when the surcharge coves an appropriate period of both sides of t^*, it is not efficient to excessively extend the duration of surcharge (in this example, a duration around 100 *min* is optimal).

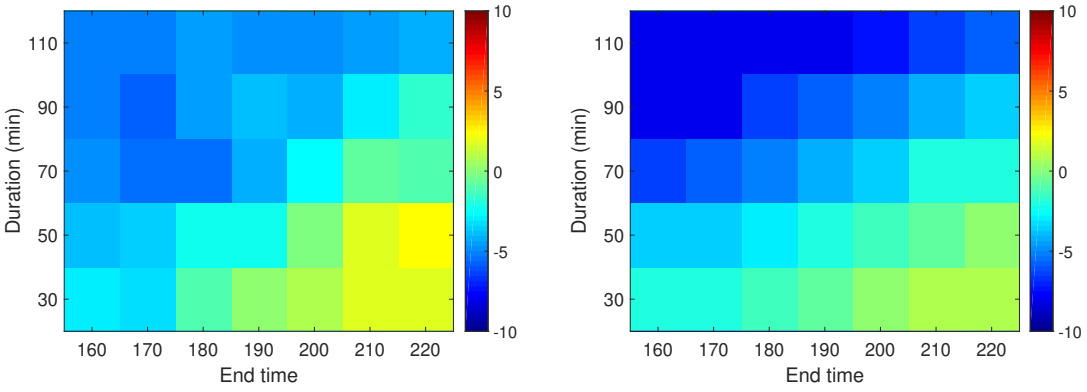
Similarly, the sensitivity of cost change is also conducted for the reward case as shown in Fig. 6.13 (Reward $p = 2$ \$ only given before t^* and train inflow still takes the constant value $a(t) = 12\ tr/h$). On one hand, it can be understood from Fig. 6.13a that SC change varies in a relatively small range compared to surcharge depending on both end time and duration, and there is an optimal end time and duration around (170 *min*, 90 *min*). When the end time of reward approaches t^*, SC inversely increase. This is because giving reward to passengers who should be surcharged according to first-best pricing would encourage passengers to utilize congested trains, as a result, TTC and SC increase. On the other hand, UC generally decreases as long as the reward is given as can be confirmed from Fig. 6.13b. The longer the duration, the earlier it ends, the lower user cost would be. However, there also exists a limit



(a) Cost breakdown when $t^- - t^+ = 100$ *min* (b) Cost breakdown when $t^- - t^* = 20$ *min*

Fig. 6.12 An example of cost breakdown for surcharge case.

for the reduction of UC around the left top of Fig. 6.13b since giving reward before the start of commute does not have effect.

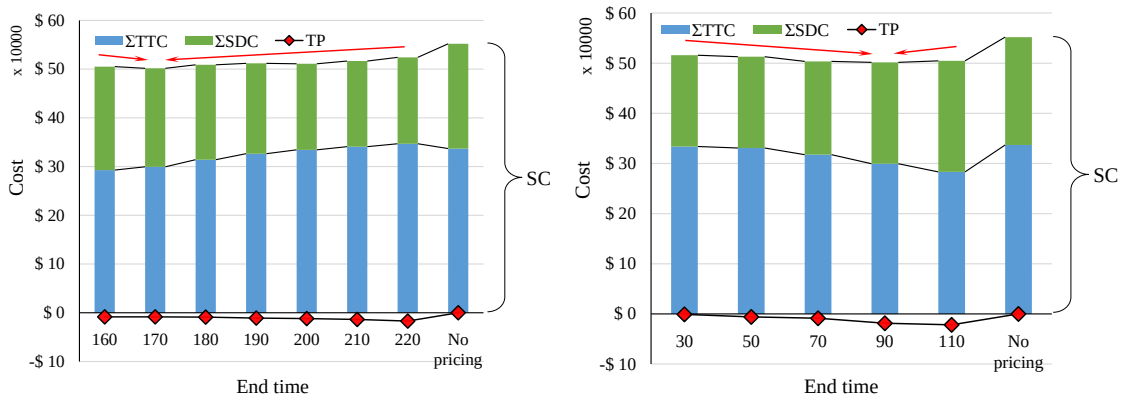


(a) Social cost change (%) (b) User cost change (%)

Fig. 6.13 Cost sensitivity to temporal settings of reward.

An example of cost breakdown for reward case is shown in Fig. 6.14 to further understand the reward effect on different components of the cost. On one hand, when the duration of the reward is fixed, it can be seen from Fig. 6.14a that TTC significantly decreases when the reward ends earlier before t^*. This is because the reward encourages some of the passengers to use the earlier trains, as a result, the difference of riding passenger number between congested and uncongested trains narrows, thus TTC declines. In addition, it can also be observed that when the reward ends earlier, SDC increases. This is reasonable because when more passengers are encouraged to ride on earlier trains, their schedule delay undoubtedly raise. Therefore, when the increase of SDC just exceeds the decrease of TTC, SC is minimized

(in this example around 170 *min*). On the other hand, when the end time of reward is fixed as 70 *min* before t^*, TTC decreases with the extension of the reward period. The reason is the same as explained before. However, the minimum of SDC appears at 70 *min*, which is also the minimum of SC. This result indicates that the use of reward is not the longer the better. If the reward period is too long, the increase of SDC may exceed the decrease of TTC. Therefore, an appropriate period of reward sufficiently earlier before t^* is the most desirable situation. This conclusion is also consistent with the implication of first-best pricing in Fig. 6.6.



(a) Cost breakdown when $t^- - t^+ = 90$ *min*

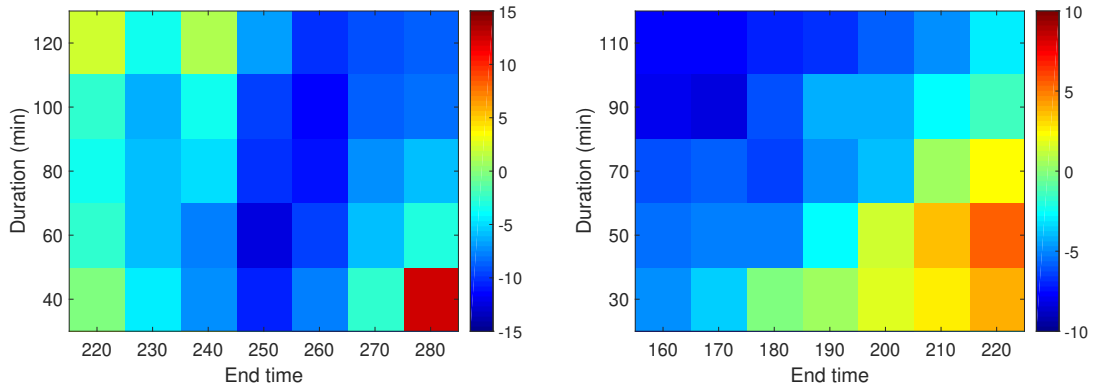
(b) Cost breakdown when $t^- - t^* = -70$ *min*

Fig. 6.14 An example of cost breakdown for reward case.

In addition, change of n_p STDEV for both surcharge and reward cases are shown in Fig. 6.15. It can be observed from Fig. 6.15a that when surcharged is imposed, the change of n_p STDEV behaves similar as that of SC in Fig. 6.11a. The decrease of n_p STDEV indicates that the in-vehicle crowding is more evenly distributed. Differently, n_p STDEV change for reward case seems sensitive to both the pricing duration and end time as can be understood from Fig. 6.15b. n_p STDEV even significantly increases when the reward lasts 30 *min* and ends 20 *min* before t^*. To conclude, the change of n_p STDEV is basically consistent with SC change, which implies that to reduce the SC basically will also alleviate in-vehicle crowding.

The implications from the above numerical experiments can be concluded as:

- By implementing one-step coarse pricing without breaking the user equilibrium, social cost can be reduced up to 10%.
- The surcharge period should cover both sides of t^*, while its duration is not necessary to be excessively long.



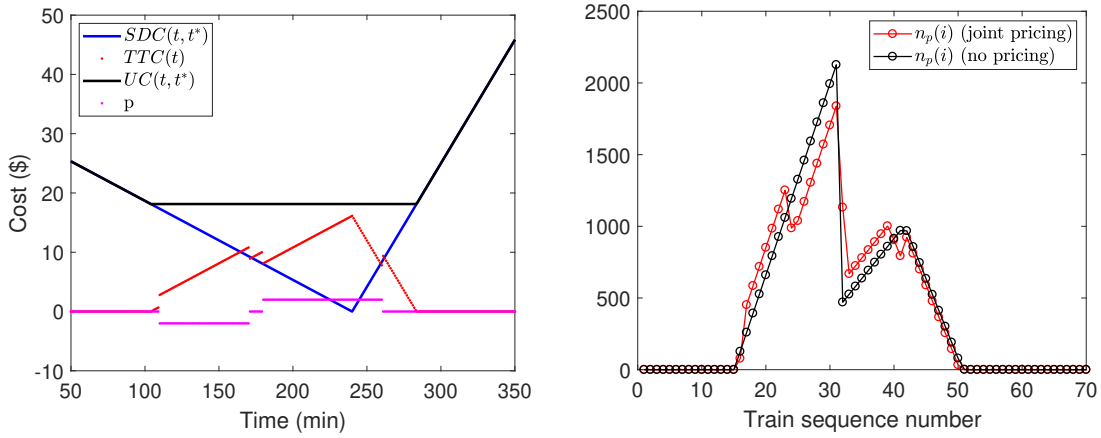
(a) STDEV of n_p change for surcharge case (%) (b) STDEV of n_p change for reward case (%)

Fig. 6.15 Sensitivity of riding passenger number standard deviation.

- The reward period should end before t^* with an appropriate time interval, while its duration has an optimal length from the perspective of social cost.

6.3 Joint pricing for different timetable patterns

This subsection discusses the effectiveness of coarse pricing scheme that jointly imposes surcharge and gives reward. Inspired by the first-best pricing and previous numerical examples that surcharge raises the UC but reward reduces the UC, if giving the reward at the very beginning of the rush hours and then imposing the surcharge, both UC and SC may decrease while $UC - SC > 0$. This indicates a Pareto-improvement situation since both passengers and the operator benefit from the implementation of the pricing. Fig. 6.16a presents such a situation that 2 \$ reward is given to passengers who depart during $[110, 170]$ *min*, and 2 \$ surcharge is imposed on passengers who depart during $[180, 260]$ *min*. The train inflow takes the constant value $a(t) = 12\ tr/h$. Under this pricing scheme, the comparison of riding passenger number on each train with no pricing situation is drawn in Fig. 6.16b. As can be seen from Fig. 6.16b, motivated by the pricing, a portion of passengers gave up utilizing the most congested trains and rode on the trains at the both shoulders of the peak hours. As a result, SC decreased 12.3%, TPN decreased 15.5% compared to no pricing situation. Meanwhile, although very slight, UC also reduced 1.5%. Since there are more passengers been surcharged than been rewarded, the operator gained from total surcharge minus total reward, equals around 5.4×10^4 \$ in this example. This result suggested that both passengers and the operator benefited from the implementation of this pricing scheme, which is a Pareto-improvement.



(a) Travel costs

(b) Comparison of riding passenger number

Fig. 6.16 Cost and riding passenger number comparison when surcharge and reward jointly implemented.

Further, the joint pricing experiments were also conducted for different timetable patterns listed in Table 6.1. Due to the constraint of FIFO principle under equilibrium explained in Eq. (6.49), feasible value of surcharge/reward varies according to the scenario. This value is calculated to approach the constraint, and for simplicity, the value of surcharge and reward are set the same. The duration of reward is set for those who depart during $[120, 160]$ *min* and the surcharge during $[180, 260]$ *min*. Train inflow takes the same setting as in Table 6.1. The comparison results of various costs for different timetable patterns under joint pricing is listed in Table 6.3.

Table 6.3 Comparison of costs for different timetable patterns under joint coarse pricing.

Scenario	p ($)	$\sum TTC$ (10^4 $)	$\sum SDC$ (10^4 $)	SC (10^4 $)	$\sum UC$ (10^4 $)	TP (10^4 $)
OPM	1.9	30.27 (−9.2%)	17.04 (+5.9%)	47.31 (−4.3%)	52.38 (+6.0%)	5.07
CON	2.0	31.17 (−18.2%)	16.71 (−2.3%)	47.88 (−13.3%)	53.40 (−3.3%)	5.52
PR1	1.3	31.88 (−3.7%)	19.04 (+2.8%)	50.91 (−1.3%)	54.40 (+5.4%)	3.49
PR2	1.0	32.01 (−13.1%)	19.84 (−8.0%)	51.84 (−11.2%)	54.80 (−6.2%)	2.96
PR3	1.3	32.72 (−6.2%)	16.82 (+0.6%)	49.55 (−4.0%)	52.80 (+2.3%)	3.25

$N_p = 30000$ in all scenarios. The percentages in the parentheses indicate the change compared to no pricing case.

First, it can be seen that TTC in all scenarios decrease as long as the coarse pricing is jointly implemented. The maximum reduction comes from constant timetable scenario, and this may due to the largest decrements of riding passenger number difference among the dispatched trains. However, the change of SDC does not show a consistent result when coarse pricing is implemented. Next, social cost, which is a sum of TTC and SDC, also declines in all scenarios. This further confirms that the decrease of TTC usually exceeds the increase of SDC (if any) as long as the coarse pricing is appropriately designed, and as a result, social cost will be reduced. Then, similar like SDC, user cost may increase or decrease depending on the setting of the timetable. For scenarios that UC decrease, Pareto-improvement is realized since positive TP will be the additional revenue for the operator, and if UC also drops, both the passengers (as a group) and the operator benefit from the implication of coarse pricing. The sum of pricing value, TP, is always positive since passengers who are surcharged are significantly more than those who are given the reward. In addition, the amount of TP can be observed to be approximately proportional to the amount of surcharge/reward (i.e., p).

It can also be concluded from the above comparison that coarse pricing mainly works on the reduction of TTC. If TTC is originally not absurdly high due to reasonable timetable setting (e.g., OPM and PR1 scenarios), the effect of coarse pricing is relatively insignificant (SC only declines around 4% and 1% for OPM and PR1 scenarios). In these situations, multi-steps pricing scheme can be further considered.

To conclude, by the proposed model, this section numerically applied a single-step pricing scheme including surcharge, reward and their combinations to derive the policy implication for the demand management. It is found that the effectiveness of the pricing scheme is strongly dependent on the temporal settings of the scheme, and if appropriately designed, Pareto-improvement situation exists which benefits both passengers and the operator.

6.4 Brief summary

By applying the proposed model, three issues are investigated in this chapter. First, optimal peak/off-peak timetable which minimizes the social cost (defined as the sum of travel time cost and schedule delay cost) is explored. An optimum peak/off-peak timetable in free-flow regime of train-FD is proposed and numerically compared with different timetable patterns. Second, the user equilibrium under first-best pricing scheme is theoretically derived and numerically compared with the user equilibrium without pricing. Third, because the first-best pricing scheme requires time varying pricing which is difficult to realize in real world, an appropriate single-step coarse pricing scheme is specified and its effectiveness is numerically evaluated.

If all dispatched trains operate in free-flow regime of train-FD, peak/off-peak timetable is found to be optimized when average train flow before and after desired departure time are the same, and the flow of train with maximum passenger concentration reaches the critical state of train-FD. It is proved that the user equilibrium under first-best pricing scheme can significantly reduce the social cost by flattening the change rate of riding passenger number among dispatched trains compared with the user equilibrium without pricing. Riding passenger number difference between successive trains under first-best pricing is exactly half of that in the equilibrium without pricing. The effectiveness of single-step coarse pricing is found to strongly depend on the temporal design of the scheme. The appropriately designed coarse pricing scheme without breaking the equilibrium shows an acceptable reduction of social cost compared to the first-best one. Finally, a complex pricing scheme by jointly giving reward and imposing surcharge is examined, and it is found that a Pareto-improvement situation may occur depending on the timetable pattern with significant reduction of social cost.

Chapter 7

Conclusions and future work

This chapter summarizes the findings and conclusions from the previous chapters and presents the probable directions to be improved for the future work.

7.1 Findings and conclusions

This study proposed a macroscopic model capable of describing distribution of passenger arrivals and rail transit operation given their dynamic interaction based on equilibrium theory and a train fundamental diagram traffic model. Through the literature review on basics of train operation, macroscopic modelling, optimization and management from both supply and demand sides, the major research gap are defined as the lack of an appropriate model that simultaneously considers the passenger congestion influence on railway operation and dynamic interaction between demand distribution and supply information. The findings and conclusions from Chapter 3 to Chapter 6 are separately summarized as follows:

Chapter 3 investigated a traffic flow model and its variants for urban rail transit (i.e., train fundamental diagram by Seo et al. 2017) with empirical data from Boston subway system to comprehensively confirm the passenger congestion influence on railway operation. The conclusions can be summarized as:

- The assumption of train dwell time at stations plays a key role in model's description of passenger congestion influence.
- By analyzing the operation data of Boston subway red line, passenger congestion influence to deteriorate the performance of railway system was confirmed to be significant.

- By comparing three variants of train-FD models, the original one which assumes train dwell time monotonically increasing with boarding passenger number, was found to be sufficient to describe passenger congestion influence from a macroscopic view.

Chapter 4 expanded the steady state train-FD to the dynamic case. Two approaches: fluid-based one developed from exit-flow model and particle-based one originally developed in this study, are considered to describe the dynamics of railway operation when inputs are time-dependent. Numerical examples are given to show the performance of these two dynamic models. A micro-simulation is also developed to verify the results calculated from the two dynamics models. The findings can be summarized as:

- Both approaches can acceptably approximate the dynamics of rail transit system when their assumptions hold.
- Fluid-based model is more robust for the forward problem (i.e., when train and passenger inflows are given information) compared to particle-based one, but it has intrinsic incompatibility when jointly employed with equilibrium theory.
- Particle-based model has a strong assumption that the whole trip of a train can be approximated as a steady-state, so it is only applicable to a short section of railway line. However, it works well with equilibrium theory and is more tractable for the analysis of management strategies.

Chapter 5 introduced the travel cost assumption and further derived the equilibrium distribution of passenger arrivals for the morning commute problem from the particle-based dynamic model by employing the inverse function of train-FD. The solution method and existence conditions of the equilibrium are given, and characteristics of the equilibrium are discussed based on a series of numerical experiments. The findings on characteristics of the equilibrium passenger arrivals can be concluded as:

- Cost function assumption, total travel demand and train timetable all have significant influence on the characteristics of the passenger arrival distributions.
- There may exist one or two peaks of passenger arrivals depending on the cost assumption and total travel demand.
- Conventional peak/off-peak timetable pattern is numerically confirmed to be effective in shortening the morning commute period and relieving the in-vehicle crowding due to more efficient use of dispatched trains.

Chapter 6 applied the proposed equilibrium model to evaluate management strategies from both demand and supply sides. The optimization problem of peak/off-peak timetable pattern is first discussed in a general situation. Then, a more conservative optimum condition only in free-flow regime of train-FD is proposed and numerically compared with other timetable patterns. Next, user equilibrium under first-best pricing is analytically derived and compared with equilibrium without pricing. Besides, a one-step coarse pricing (including surcharge, reward and their combinations) is proposed and numerically evaluated based on indicators such as social cost, user cost and in-vehicle crowding. Finally, the coarse pricing is also applied to different timetable patterns. The findings can be summarized as:

- The optimal peak/off-peak timetable which minimizes the social cost exists and its values mainly depend on the relation among total travel demand, available train fleet and coefficients of passengers' values of time.
- Optimal peak/off-peak timetable only operated in free-flow regime of train-FD produces shows an acceptable reduction of social cost compared to other timetables, and it also avoids the unstable operation in congested regime of train-FD which may lead to bunching phenomenon.
- User equilibrium under first-best pricing scheme can significantly reduce the social cost by flattening the change rate of riding passenger number among dispatched trains. Riding passenger number difference between successive trains under first-best pricing is exactly half of that in the equilibrium without pricing.
- The effectiveness of single-step coarse pricing is found to strongly depend on the temporal design of the scheme. The appropriately designed coarse pricing scheme without breaking the equilibrium shows an acceptable reduction of social cost compared to the first-best one.
- The surcharge period of coarse pricing should cover both sides of desired departure time from the railway system, while its duration is not necessary to be excessively long. The reward period should end before desired departure time with an appropriate time interval, while its duration has an optimal length from the perspective of social cost.
- By jointly giving reward and imposing surcharge, a Pareto-improvement situation may occur depending on the timetable pattern. The social cost reduction under joint coarse pricing is significant.

Nevertheless, the findings and conclusions derived above still deserve further investigation since the proposed model made several very ideal assumptions like homogeneity of railway

system and passengers, unlimited capacity of train and station, steady state during the whole trip of a train. When some of these assumptions do not hold, the quantitative findings and conclusions may change to some extent, but qualitative results should still hold.

7.2 Future work

To the end of this study, there are still many issues that can be improved to further strengthen the reliability of the findings above and make the insights more applicable to practice. Here, three directions of future work can be considered:

- First, improvement and further investigation of the train-FD should be conducted. On one hand, traffic flow model used in this study assumes a homogeneous railway system, which can be improved to be more realistic by developing a train-FD model also applicable for heterogeneous railway system. On the other hand, the newly proposed train-FD variant which assumes a first constant and then increasing relation between dwell time and boarding passenger number deserves further investigation by station specific data.

- Second, the equilibrium can be extended from two aspects. One is to consider the heterogeneity of commuting passengers which means the passenger's value of time should be various and group-dependent. Another aspect is to consider including the in-vehicle crowding cost into the cost function. These improvements may improve the reliability of the proposed model.

- Third, two probable issues can be considered to improve the practicality of the proposed model. One is to expand the single railway line system to simple network since passengers' transfer at major stations plays a non-negligible role in morning commute. Another one is to conduct case studies on timetable optimization and pricing schemes with the proposed model.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Akamatsu, T., Wada, K., and Hayashi, S. (2015). The corridor problem with discrete multiple bottlenecks. *Transportation Research Part B: Methodological*, 81:808–829.

Alfa, A. S. and Chen, M. (1995). Temporal distribution of public transport demand during the peak period. *European journal of operational research*, 83(1):137–153.

Arnott, R. (2013). A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110–121.

Arnott, R., de Palma, A., and Lindsey, R. (1990a). Departure time and route choice for the morning commute. *Transportation Research Part B: Methodological*, 24(3):209–228.

Arnott, R., De Palma, A., and Lindsey, R. (1990b). Economics of a bottleneck. *Journal of urban economics*, 27(1):111–130.

Arnott, R., De Palma, A., and Lindsey, R. (1993a). Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering. *Transportation science*, 27(2):148–160.

Arnott, R., De Palma, A., and Lindsey, R. (1993b). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, pages 161–179.

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

Barry, M. and Card, B. (2014). Visualizing mbta data. [online] http://mbtaviz.github.io/.

Bartholdi III, J. J. and Eisenstein, D. D. (2012). A self-coördinating bus route to resist bus bunching. *Transportation Research Part B: Methodological*, 46(4):481–491.

Ben-Akiva, M., De Palma, A., and Kanaroglou, P. (1986). Dynamic model of peak period traffic congestion with elastic arrival rates. *Transportation Science*, 20(3):164–181.

Binder, S., Maknoon, Y., and Bierlaire, M. (2017). The multi-objective railway timetable rescheduling problem. *Transportation Research Part C: Emerging Technologies*, 78:78–94.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., and Wagenaar, J. (2014). An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B: Methodological*, 63:15–37.

Canca, D., Barrena, E., Algaba, E., and Zarzo, A. (2014). Design and analysis of demand-adapted railway timetables. *Journal of Advanced Transportation*, 48(2):119–137.

Canca, D., Barrena, E., De-Los-Santos, A., and Andrade-Pineda, J. L. (2016). Setting lines frequency and capacity in dense railway rapid transit networks with simultaneous passenger assignment. *Transportation Research Part B: Methodological*, 93:251–267.

Carey, M. (1986). A constraint qualification for a dynamic traffic assignment model. *Transportation Science*, 20(1):55–58.

Carey, M. (1994). A model and strategy for train pathing with choice of lines, platforms, and routes. *Transportation Research Part B: Methodological*, 28(5):333–353.

Carey, M. and Crawford, I. (2007). Scheduling trains on a network of busy complex stations. *Transportation Research Part B: Methodological*, 41(2):159–178.

Carey, M. and Kwieciński, A. (1994). Stochastic approximation to the effects of headways on knock-on delays of trains. *Transportation Research Part B: Methodological*, 28(4):251–267.

Carey, M. and McCartney, M. (2004). An exit-flow model used in dynamic traffic assignment. *Computers & Operations Research*, 31(10):1583–1602.

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research Part A: General*, 18(5-6):439–453.

Ceder, A. (1987). Methods for creating bus timetables. *Transportation Research Part A: General*, 21(1):59–83.

Ceder, A. (2016). *Public transit planning and operation: Modeling, practice and behavior*. CRC press.

Chiabaut, N. (2015). Evaluation of a multimodal urban arterial: The passenger macroscopic fundamental diagram. *Transportation Research Part B: Methodological*, 81:410–420.

Cordone, R. and Redaelli, F. (2011). Optimizing the demand captured by a railway system with a regular timetable. *Transportation Research Part B: Methodological*, 45(2):430–446.

Corman, F., D'Ariano, A., Pacciarelli, D., and Pranzo, M. (2010). A tabu search algorithm for rerouting trains during rail operations. *Transportation Research Part B: Methodological*, 44(1):175–192.

Cornet, S., Buisson, C., Ramond, F., Bouvarel, P., and Rodriguez, J. (2019). Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas. *Transportation Research Part C: Emerging Technologies*, 106:345–359.

Currie, G. (2010). Quick and effective solution to rail overcrowding: Free early bird ticket experience in melbourne, australia. *Transportation Research Record*, 2146(1):35–42.

Daganzo, C. F. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation science*, 19(1):29–37.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F. (2009). A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921.

Daganzo, C. F. et al. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28(4):269–287.

Daganzo, C. F. and Lehe, L. J. (2015). Distance-dependent congestion pricing for downtown zones. *Transportation Research Part B: Methodological*, 75:89–99.

Daganzo, C. F. and Pilachowski, J. (2011). Reducing bunching with bus-to-bus cooperation. *Transportation Research Part B: Methodological*, 45(1):267–277.

De Palma, A., Ben-Akiva, M., Lefevre, C., and Litinas, N. (1983). Stochastic equilibrium model of peak period traffic congestion. *Transportation Science*, 17(4):430–453.

De Palma, A., Kilani, M., and Proost, S. (2015). Discomfort in mass transit and its implication for scheduling and pricing. *Transportation Research Part B: Methodological*, 71:1–18.

De Palma, A., Lindsey, R., and Monchambert, G. (2017). The economics of crowding in rail transit. *Journal of Urban Economics*, 101:106–122.

Delgado, F., Munoz, J. C., Giesen, R., and Cipriano, A. (2009). Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record*, 2090(1):59–67.

Dueker, K. J., Kimpel, T. J., Strathman, J. G., and Callas, S. (2004). Determinants of bus dwell time. *Journal of public transportation*, 7(1):21–40.

Edie, L. C. (1963). *Discussion of traffic stream measurements and definitions*. Port of New York Authority.

Fernández, R., Zegers, P., Weber, G., and Tyler, N. (2010). Influence of platform height, door width, and fare collection on bus dwell time: laboratory evidence for santiago de chile. *Transportation research record*, 2143(1):59–66.

Fosgerau, M. (2015). Congestion in the bathtub. *Economics of Transportation*, 4(4):241–255.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L., and Wie, B.-W. (1993). A variational inequality formulation of the dynamic network user equilibrium problem. *Operations research*, 41(1):179–191.

Friesz, T. L., Luque, J., Tobin, R. L., and Wie, B.-W. (1989). Dynamic network traffic assignment considered as a continuous time optimal control problem. *Operations Research*, 37(6):893–901.

Fu, L., Liu, Q., and Calamai, P. (2003). Real-time optimization model for dynamic scheduling of transit operations. *Transportation Research Record*, 1857(1):48–55.

Geroliminis, N. and Daganzo, C. (2007). Macroscopic modeling of traffic in cities. In *Transportation Research Board 86th Annual Meeting*.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770.

Geroliminis, N. and Levinson, D. M. (2009). Cordon pricing consistent with the physics of overcrowding. In *Transportation and Traffic Theory 2009: Golden Jubilee*, pages 219–240. Springer.

Geroliminis, N., Zheng, N., and Ampountolas, K. (2014). A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. *Transportation Research Part C: Emerging Technologies*, 42:168–181.

Goverde, R. M. (2007). Railway timetable stability analysis using max-plus system theory. *Transportation Research Part B: Methodological*, 41(2):179–201.

Henderson, J. V. (1974). Road congestion: a reconsideration of pricing theory. *Journal of Urban Economics*, 1(3):346–365.

Henderson, J. V. (1981). The economics of staggered work hours. *Journal of Urban Economics*, 9(3):349–364.

Hendrickson, C. and Kocur, G. (1981). Schedule delay and departure time decisions in a deterministic model. *Transportation science*, 15(1):62–77.

Higgins, A., Kozan, E., and Ferreira, L. (1996). Optimal scheduling of trains on a single line track. *Transportation research part B: Methodological*, 30(2):147–161.

Hu, W. (2018). New york subway's on-time performance hits new low. *The New York Times*.

Hurdle, V. (1973). Minimum cost schedules for a public transportation route—i. theory. *Transportation Science*, 7(2):109–137.

Hurdle, V. (1981). Equilibrium flows on urban freeways. *Transportation Science*, 15(3):255–293.

Ibarra-Rojas, O. J., Delgado, F., Giesen, R., and Muñoz, J. C. (2015). Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, 77:38–75.

Jiang, Z., Xie, C., Ji, T., and Zou, X. (2015). Dwell time modelling and optimized simulations for crowded rail transit lines based on train capacity. *Promet-Traffic&Transportation*, 27(2):125–135.

Kariyazaki, K., Hibino, N., and Morichi, S. (2015). Simulation analysis of train operation to recover knock-on delay under high-frequency intervals. *Case Studies on Transport Policy*, 3(1):92–98.

Kato, H., Kaneko, Y., and Soyama, Y. (2012). Departure-time choices of urban rail passengers facing unreliable service: Evidence from tokyo. In *Proceedings of the International Conference on Advanced Systems for Public Transport 2012*.

Kecman, P. and Goverde, R. M. (2015). Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3):295–319.

Kraus, M. and Yoshida, Y. (2002). The commuter's time-of-use decision and optimal pricing and service in urban mass transit. *Journal of Urban Economics*, 51(1):170–195.

Kroon, L., Huisman, D., Abbink, E., Fioole, P.-J., Fischetti, M., Maróti, G., Schrijver, A., Steenbeek, A., and Ybema, R. (2009). The new dutch timetable: The or revolution. *Interfaces*, 39(1):6–17.

Kuwahara, M. (1990). Equilibrium queueing patterns at a two-tandem bottleneck during the morning peak. *Transportation science*, 24(3):217–229.

Kuwahara, M. (1998). Review of theory on departure time choice for highway traffic. *Seisan-kenkyu*, 50(4):161–168.

Kuwahara, M. and Akamatsu, T. (1997). Decomposition of the reactive dynamic assignments with queues for a many-to-many origin-destination pattern. *Transportation Research Part B: Methodological*, 31(1):1–10.

Laih, C.-H. (1994). Queueing at a bottleneck with single-and multi-step tolls. *Transportation Research Part A: Policy and Practice*, 28(3):197–208.

Lam, W. H., Cheung, C.-Y., and Lam, C. (1999). A study of crowding effects at the hong kong light rail transit stations. *Transportation Research Part A: Policy and Practice*, 33(5):401–415.

Lamotte, R. and Geroliminis, N. (2016). The morning commute in urban areas: Insights from theory and simulation. In *Transportation Research Board 95th Annual Meeting*.

Larsen, R., Pranzo, M., D'Ariano, A., Corman, F., and Pacciarelli, D. (2014). Susceptibility of optimal train schedules to stochastic disturbances of process times. *Flexible Services and Manufacturing Journal*, 26(4):466–489.

Levinson, H. S. (1983). Analyzing transit travel time performance. *Transportation Research Record*, 915:1–6.

Li, D., Daamen, W., and Goverde, R. M. (2016). Estimation of train dwell time at short stops based on track occupation event data: A study at a dutch railway station. *Journal of Advanced Transportation*, 50(5):877–896.

Li, S., Dessouky, M. M., Yang, L., and Gao, Z. (2017). Joint optimal train regulation and passenger flow control strategy for high-frequency metro lines. *Transportation Research Part B: Methodological*, 99:113–137.

Li, S., Liu, R., Yang, L., and Gao, Z. (2019). Robust dynamic bus controls considering delay disturbances and passenger demand uncertainty. *Transportation Research Part B: Methodological*, 123:88–109.

Liebchen, C. and Möhring, R. H. (2007). The modeling power of the periodic event scheduling problem: railway timetables—and beyond. In *Algorithmic methods for railway optimization*, pages 3–40. Springer.

Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):317–345.

Lindsey, C. R., Van den Berg, V. A., and Verhoef, E. T. (2012). Step tolling with bottleneck queuing congestion. *Journal of Urban Economics*, 72(1):46–59.

Lindsey, R. (2004). Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transportation science*, 38(3):293–314.

Ma, Z. and Koutsopoulos, H. N. (2019). Optimal design of promotion based demand management strategies in urban rail systems. *Transportation Research Part C: Emerging Technologies*, 109:155–173.

Mahmassani, H. S. and Chang, G.-L. (1986). Experiments with departure time choice dynamics of urban commuters. *Transportation Research Part B: Methodological*, 20(4):297–320.

Mariotte, G., Leclercq, L., and Laval, J. A. (2017). Macroscopic urban dynamics: Analytical and numerical comparisons of existing models. *Transportation Research Part B: Methodological*, 101:245–267.

Merchant, D. K. and Nemhauser, G. L. (1978). A model and an algorithm for the dynamic traffic assignment problems. *Transportation science*, 12(3):183–199.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review*, 62(4):591–604.

Nachtigall, K. (1996). Periodic network optimization with different arc frequencies. *Discrete Applied Mathematics*, 69(1-2):1–17.

Newell, G. F. (1971). Dispatching policies for a transportation route. *Transportation Science*, 5(1):91–105.

Newell, G. F. (1987). The morning commute for nonidentical travelers. *Transportation Science*, 21(2):74–88.

Newell, G. F. (2002). A simplified car-following theory: a lower order model. *Transportation Research Part B: Methodological*, 36(3):195–205.

Newell, G. F. and Potts, R. B. (1964). Maintaining a bus schedule. In *Australian Road Research Board (ARRB) Conference, 2nd, 1964, Melbourne*, volume 2(1).

Nie, X. and Zhang, H. M. (2005a). A comparative study of some macroscopic link models used in dynamic traffic assignment. *Networks and Spatial Economics*, 5(1):89–115.

Nie, X. and Zhang, H. M. (2005b). Delay-function-based link models: their properties and computational issues. *Transportation Research Part B: Methodological*, 39(8):729–751.

Niu, H. and Zhou, X. (2013). Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. *Transportation Research Part C: Emerging Technologies*, 36:212–230.

Niu, H., Zhou, X., and Gao, R. (2015). Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. *Transportation Research Part B: Methodological*, 76:117–135.

Noursalehi, P., Koutsopoulos, H. N., and Zhao, J. (2018). Real time transit demand prediction capturing station interactions and impact of special events. *Transportation Research Part C: Emerging Technologies*, 97:277–300.

Odijk, M. A. (1996). A constraint generation algorithm for the construction of periodic railway timetables. *Transportation Research Part B: Methodological*, 30(6):455–464.

Osuna, E. and Newell, G. (1972). Control strategies for an idealized public transportation system. *Transportation Science*, 6(1):52–72.

Palmqvist, C.-W., Tomii, N., and Ochiai, Y. (2020). Explaining dwell time delays with passenger counts for some commuter trains in stockholm and tokyo. *Journal of Rail Transport Planning & Management*, page 100189.

Peer, S., Knockaert, J., and Verhoef, E. T. (2016). Train commuters' scheduling preferences: Evidence from a large-scale peak avoidance experiment. *Transportation Research Part B: Methodological*, 83:314–333.

Peeters, L. (2003). *Cyclic Railway Timetable Optimization*. PhD thesis, Erasmus Research Institute of Management.

Railway Policy Planning, M. (2020). The visualization of railway lines' delay in tokyo metropolitan area. [online] https://www.mlit.go.jp/report/press/content/001328948.pdf.

Rey, D., Dixit, V. V., Ygnace, J.-L., and Waller, S. T. (2016). An endogenous lottery-based incentive mechanism to promote off-peak usage in congested transit systems. *Transport Policy*, 46:46–55.

Richards, P. I. (1956). Shock waves on the highway. *Operations research*, 4(1):42–51.

Robenek, T., Azadeh, S. S., Maknoon, Y., de Lapparent, M., and Bierlaire, M. (2018). Train timetable design under elastic passenger demand. *Transportation research Part B: methodological*, 111:19–38.

Robenek, T., Maknoon, Y., Azadeh, S. S., Chen, J., and Bierlaire, M. (2016). Passenger centric train timetabling problem. *Transportation Research Part B: Methodological*, 89:107–126.

Seo, T., Wada, K., and Fukuda, D. (2017). A macroscopic and dynamic model of urban rail transit with delay and congestion. In *TRB 96th Annual Meeting Compendium of Papers*.

Serafini, P. and Ukovich, W. (1989). A mathematical model for periodic scheduling problems. *SIAM Journal on Discrete Mathematics*, 2(4):550–581.

Shi, J., Yang, L., Yang, J., and Gao, Z. (2018). Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach. *Transportation Research Part B: Methodological*, 110:26–59.

Smith, M. J. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation science*, 18(4):385–394.

Suh, W., Chon, K.-S., and Rhee, S.-M. (2002). Effect of skip-stop policy on a korean subway system. *Transportation Research Record*, 1793(1):33–39.

Sumalee, A., Tan, Z., and Lam, W. H. (2009). Dynamic stochastic transit assignment with explicit seat allocation model. *Transportation Research Part B: Methodological*, 43(8-9):895–912.

Sumi, T., Matsumoto, Y., and Miyaki, Y. (1990). Departure time and route choice of commuters on mass transit systems. *Transportation Research Part B: Methodological*, 24(4):247–262.

Sun, A. and Hickman, M. (2005). The real–time stop–skipping problem. *Journal of Intelligent Transportation Systems*, 9(2):91–109.

Sun, L., Tirachini, A., Axhausen, K. W., Erath, A., and Lee, D.-H. (2014). Models of bus boarding and alighting dynamics. *Transportation Research Part A: Policy and Practice*, 69:447–460.

Tabuchi, T. (1993). Bottleneck congestion and modal split. *Journal of Urban Economics*, 34(3):414–431.

Takashige, T. (1999). Signalling systems for safe railway transport. *Japan railway & Transport review*, 21:44–50.

Takayama, Y. (2015). Bottleneck congestion and distribution of work start times: The economics of staggered work hours revisited. *Transportation Research Procedia*, 7:499–518.

Takayama, Y. and Kuwahara, M. (2017). Bottleneck congestion and residential location of heterogeneous commuters. *Journal of Urban Economics*, 100:65–79.

Tang, Y., Yang, H., Wang, B., Huang, J., and Bai, Y. (2020). A pareto-improving and revenue-neutral scheme to manage mass transit congestion with heterogeneous commuters. *Transportation Research Part C: Emerging Technologies*, 113:245–259.

Tian, Q., Huang, H.-J., and Yang, H. (2007). Equilibrium properties of the morning peak-period commuting in a many-to-one mass transit system. *Transportation Research Part B: Methodological*, 41(6):616–631.

Tirachini, A. (2013). Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers. *Transportmetrica A: Transport Science*, 9(1):28–49.

Tirachini, A., Hensher, D. A., and Rose, J. M. (2013). Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation research part A: policy and practice*, 53:36–52.

Vickrey, W. (1973). Pricing, metering, and efficiently using urban transportation facilities. *Highway Research Board*, pages 36–48.

Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, pages 251–260.

Vuchic, V. R. (2017). *Urban transit: operations, planning, and economics*. John Wiley & Sons.

Wang, S., Zhang, W., and Qu, X. (2018a). Trial-and-error train fare design scheme for addressing boarding/alighting congestion at cbd stations. *Transportation Research Part B: Methodological*, 118:318–335.

Wang, Y., D'Ariano, A., Yin, J., Meng, L., Tang, T., and Ning, B. (2018b). Passenger demand oriented train scheduling and rolling stock circulation planning for an urban rail transit line. *Transportation Research Part B: Methodological*, 118:193–227.

Wang, Y., Li, D., and Cao, Z. (2020). Integrated timetable synchronization optimization with capacity constraint under time-dependent demand for a rail transit network. *Computers & Industrial Engineering*, 142:106374.

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

Weidmann, U. (1994). *Der Fahrgastwechsel im öffentlichen Personenverkehr*. PhD thesis, ETH Zurich.

Winward, S. (2016). On metro, not all off-peak discounts are created equal. *Greater Greater Washington*.

Wong, R. C., Yuen, T. W., Fung, K. W., and Leung, J. M. (2008). Optimizing timetable synchronization for rail mass transit. *Transportation Science*, 42(1):57–69.

Wu, J. H., Chen, Y., and Florian, M. (1998). The continuous dynamic network loading problem: a mathematical formulation and solution method. *Transportation Research Part B: Methodological*, 32(3):173–187.

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

Xie, C. and Fukuda, D. (2014). Scheduling preference modeling of rail passengers in the tokyo metropolitan area and evaluation of time-varying fare policy for a congested urban railway line. *Collection of papers of Civil Engineering D3 (Infrastructure planning and management)*, 70(5):535–548.

Xu, X., Li, H., Liu, J., Ran, B., and Qin, L. (2019). Passenger flow control with multi-station coordination in subway networks: algorithm development and real-world case study. *Transportmetrica B: Transport Dynamics*, 7(1):446–472.

Xuan, Y., Argote, J., and Daganzo, C. F. (2011). Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B: Methodological*, 45(10):1831–1845.

Yamamura, A., Koresawa, M., Aadchi, S., Inagi, T., and Tomii, N. (2013). Dwell time analysis in urban railway lines using multi agent simulation. In *13th World Conference on Transportation Research (WCTR13), Rio de Janeiro*.

Yan, F. and Goverde, R. M. (2019). Combined line planning and train timetabling for strongly heterogeneous railway lines with direct connections. *Transportation Research Part B: Methodological*, 127:20–46.

Yang, H. and Tang, Y. (2018). Managing rail transit peak-hour congestion with a fare-reward scheme. *Transportation Research Part B: Methodological*, 110:122–136.

Yin, J., Tang, T., Yang, L., Xun, J., Huang, Y., and Gao, Z. (2017a). Research and development of automatic train operation for railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 85:548–572.

Yin, J., Yang, L., Tang, T., Gao, Z., and Ran, B. (2017b). Dynamic passenger demand oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. *Transportation Research Part B: Methodological*, 97:182–213.

Zhang, J., Yand, H., Lindsey, R., and Li, X. (2020). Modeling and managing congested transit service with heterogeneous users under monopoly. *Transportation Research Part B: Methodological*, 132:249–266.

Zhang, Q., Han, B., and Li, D. (2008). Modeling and simulation of passenger alighting and boarding movement in beijing metro stations. *Transportation Research Part C: Emerging Technologies*, 16(5):635–649.

Zhang, Z., Fujii, H., and Managi, S. (2014). How does commuting behavior change due to incentives? an empirical study of the beijing subway system. *Transportation Research Part F: Traffic Psychology and Behaviour*, 24:17–26.