Train following model for urban rail transit performance analysis

Contents lists available at ScienceDirect

Transportation Research Part C

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

Train following model for urban rail transit performance analysis☆

Saeid Saidi[a,b,*], Haris N. Koutsopoulos[c], Nigel H.M. Wilson[d], Jinhua Zhao[e]

[a] *Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, Alberta, Canada*
[b] *Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA*
[c] *Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA*
[d] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
[e] *Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA*

ARTICLE INFO

Keywords:
Rail operations
Mesocopic modeling
Train following
Headway regularity
Rail performance
Train delays
Disruptions

ABSTRACT

In this paper we introduce a mesoscopic Train Following Model which accurately captures train interactions and predicts delays based on spacing between consecutive trains. The Train Following Model is applied recursively block by block estimating train trajectories given initial conditions (i.e. the trajectory of an initial train and dispatching headways of following trains from the terminal station). We validate the proposed model using data from the Red Line of the Massachusetts Bay Transportation Authority (MBTA). The results indicate that it accurately represents train operations under both normal and disrupted conditions. Based on the model developed, the impacts of factors such as service frequency, headway variations, passenger demand, and initial train delays on line performance (i.e. line throughput and train knock-on delays) are explored. The proposed Train Following Model is generic and can be developed based on readily available historical train tracking data. It is not as resource intensive as micro simulation models, while it can efficiently address the drawbacks of macro-scale analytical models and complex discrete algebraic models. The proposed model can be used to predict system performance either off-line or in real-time.

1. Introduction

1.1. Background and motivation

Heavy rail transit is a critical mode of transport in many large cities, and improving the performance of rail transit while ensuring efficient, reliable, and robust operations in real-time is a key objective for many transit agencies. This is especially critical when the line is operating near capacity. Operating near capacity often increases train delays due to increased interactions between trains, controlled by the signaling system. In addition, maximizing the scheduled capacity reduces the buffer time between trains and thus reduces the stability and robustness of operations when disruptions occur. This trade-off between scheduled capacity and the stability of the operations when disruptions occur is rarely considered in the literature (Bešinović and Goverde, 2019). While in operations planning, conflict-free operations between trains are typically assumed, near capacity operations during peak hours or as a result of service

☆ This article belongs to the Virtual Special Issue on “Advanced railway transport-Advanced artificial intelligence, optimization and control for railway transport”.

* Corresponding author.
E-mail address: ssaidi@ucalgary.ca (S. Saidi).

https://doi.org/10.1016/j.trc.2023.104037
Received 6 February 2022; Received in revised form 15 November 2022; Accepted 20 January 2023
Available online 8 February 2023

interruptions can result in delays.

The effective capacity or line throughput of an urban rail line is a function of various factors: the train control system, station spacing, dwell times (especially at bottleneck stations), and variability in both dwell times and headways. Line throughput can be greatly degraded under poor operating regimes such as high dispatching headway variability or frequent disruptions in train operation. Especially for older rail transit systems operating without automatic train control (ATC), the impact of headway and dwell time variability on performance can be severe Wolofsky et al. (2019).

With uncertainties in operations, including variability in operator behavior and equipment-related failures, disruptions or interruptions in train operations are inevitable (Huang et al., 2020). Many studies in the literature aim to develop tools capable of predicting the effect of disruptions to provide operators with decision-support tools (Lessan et al. 2019), in order to, for example, deploy the proper recovery strategy in response to a disruption or deviation from schedule. Thus, developing a flexible train operations model which could estimate the impact of different response strategies would be very helpful.

In this paper, we propose a Train Following Model, a mesoscopic model to capture train interactions based on spacing between consecutive trains. The proposed train following model can be used to predict system performance either off-line or in real-time, and to evaluate candidate corrective actions given initial conditions, information about passenger demand, and the type of disruption or deviation. Based on the model, line throughput can be determined considering frequency of service and variations in dwell times and dispatching headways, as well as disruptions in operations. The paper also assesses the importance of headway regularity on overall rail line performance. We validate the proposed model using data from the Red Line of the Massachusetts Bay Transportation Authority (MBTA).

1.2. Literature review

Performance analysis and optimization models for train operations have been developed using both detailed (microscopic) and aggregate (macroscopic) models. Microscopic models consider the detailed infrastructure such as block lengths, curvatures, gradients and the restrictions imposed by the signalling system, as well as passenger demand, while macroscopic models consider nodes and links based on stations and junctions (Bešinović and Goverde, 2019). Microsimulation models explicitly represent train dynamics, in order to estimate the effects of possible disruptions, and evaluate the performance of a given schedule, or to develop operational solutions (Nash and Huerlimann, 2004; Quaglietta, 2014; Koutsopoulos and Wang, 2007; Zhou et al., 2020, and Zhou 2021). However, preparing and evaluating different scenarios can be time/resource-intensive because of the required detailed representation of the system, and usually cannot be run in real-time. Macroscopic models (Wendler, 2007; Hansen and Pachl, 2014; and Büker and Seybold, 2012), in contrast, employ mathematical models based on more aggregate representation of the network and train interactions. They lend themselves for use in optimization models (Bešinović et al., 2016). While computationally efficient, macroscopic models may not be sensitive to factors needed for the analysis of specific disruption events. To address these concerns, several studies combine microscopic and macroscopic models (Bešinović et al., 2016; Goverde et al., 2016; and Schlechte et al., 2011).

Another approach often used for rail timetable analysis is max-plus algebra (Braker, 1993). This approach is particularly useful for the analysis of a schedule, including stability and impact of delays (Goverde, 2007). Büker and Seybold (2012) enhanced the initial analytical models by employing cumulative density functions to capture the randomness in operations and estimate delay propagation. More recently, this approach has also been applied to real-time control of metro lines (Schanzenbächer et al., 2020; Farhi et al., 2017). The model can effectively identify critical processes and gauge the stability of rail network performance, but cannot capture details such as trains slowing down due to proximity to the lead train, a major factor affecting delay and line capacity in urban rail transit systems operating close to capacity.

Similar to Büker and Seybold (2012), several data-driven methods have also been developed. Huang et al. (2020) present a comprehensive review of data-driven methods for timetable rescheduling under disruptions. These models aim at estimation and prediction of delays using a variety of methods such as regression, Markov chains, neural networks, deep learning, clustering models, and Bayesian networks. For instance, Corman and Kecman (2018) developed a Bayesian inference network which uses information about running trains to generate the probability distribution of future train delays. While these studies contribute to the state of the art for prediction or estimation of delays, they do not capture the physics of the interactions among trains as governed by the control system.

Block occupancy has been used in the rail operations literature to capture train schedule conflicts and for capacity analysis of fixed block signaling systems (Pachl, 2002, Landex et al., 2008). It is defined as the total time each section of the track (i.e. block in a fixed block signaling system) is exclusively allocated to a train movement and hence is blocked for other trains (Pachl, 2002). Occupancy ends when the train clears the section. At that point permission can be granted to another train to occupy the section. Analyzing blocking times on all sections of a line for a group of trains can determine the minimum line headway and capacity utilization. Several studies have looked at optimizing train schedules based on maximizing infrastructure utilization using block occupancy models (Goverde et al., 2013; Jensen et al., 2017).

Many studies also examined the impact of heterogeneity of train types (slow and fast moving trains) on system performance (Jensen et. al 2017; UIC, 2013). These studies mostly relate to intercity rail traffic. Weik & Nieβen (2020) found that train running and stopping time variability can have substantial impacts on the stability of operations in dense rail corridors. They report that theoretical capacity obtained from deterministic models overestimates the actual throughput. For urban rail systems, where all trains on a line are typically of the same type, the main sources of variability in performance are dispatching headways and station dwell times. The impact of variability on line throughput and delays have rarely been considered.

Some researchers have proposed train movement models based on car-following theory to test control algorithms and design solutions for train operations. Carey and Kwieciński (1994) developed a stochastic simulation of the interaction between trains and found that following train travel times and expected knock-on delays, increase under shorter scheduled headways. Li and Gao (2013) proposed a modified car-following model to simulate train movements with a fixed block signaling system. Xun et al. (2013) developed a cellular automata model for railway traffic for moving block signaling systems. Corman et al. (2021) proposed extending existing traffic flow models to rail using stochastic process models. Ketphat et al. (2022) proposed a distance and velocity difference approach for trains operating under a virtual coupling system. Liu et al. (2021) indicated some of the proposed train-following models in the literature ignore train dynamics which may reduce the applicability of the models in real world scenarios. Most studies presented here highlight how an optimized control system can improve the performance of train operations. However, they have used analytical or simulation-based models for theoretical railway systems not necessarily based on a real transit system.

2. Objectives and contributions

This paper presents an urban rail operations model, the Train Following Model, which captures interactions between consecutive trains enforced by the train signaling system. The model, which can be calibrated using historical data from the track circuit system, is the building block for the proposed mesoscopic train operations performance model. The model outputs overall performance metrics such as travel times, delays, and headways at any station on the line, and can be used for both operations planning and real time applications such as prediction of disruption impacts and evaluation of different operational strategies. The model can serve as a quick response tool to screen actions, such as dispatching interventions, dwell time control and other operational strategies, both in real-time and offline. Through a detailed case study, the model is shown to accurately predict the line performance. Because of its recursive structure the model is implemented in a spreadsheet and is not as time- and resource- intensive as more detailed simulation models.

The remainder of the paper is organized as follows. The next section presents the theory underlying the Train Following Model, and the resulting line performance model. Section 3 presents an approach to calibrate the model using data from the case study and validates the model using different performance measures. Section 4 presents several applications of the proposed model. The final section summarizes the key results and contributions of the paper.

3. Methodology

3.1. Train following model

In rail transit, the signaling system is used to maintain safe separation between trains and to protect paths through interlockings (i.e. at junctions and crossovers). Other functions include automatic train stops when a train travels through a stop signal, and speed control to protect trains approaching junctions, sharp curves, and stations (Kittelson et al., 2013). In a fixed block signaling system, the rail tracks are divided into blocks. The signaling system detects a train when it occupies a block. Thus, presence of a train in a block is measured by block occupancy and not the exact position of the train within the block.

Fig. 1 shows a simple example of a fixed signal block system. When a train is within a block, the occupied block and its preceding block will have a stop signal. In addition, other upstream blocks may also receive lower speed codes. The speed code depends on the track segment geometry, signaling system, fleet specifications, and the spacing between consecutive trains. Block colors indicates blocks that are occupied by train k-1 and the additional block(s) required for safe stopping. Red blocks receive a 'stop' speed code and yellow blocks receive a 'slow' speed code for the following train k at the time it enters these blocks due to the presence of a lead train a block ahead. For a train to receive a green aspect (i.e. the train can maintain its design speed) when entering a block j, the preceding train should be at least Δ_j blocks ahead of block j entrance. We call this track section (blocks j to j + Δ_j-1) the 'interaction zone' of block

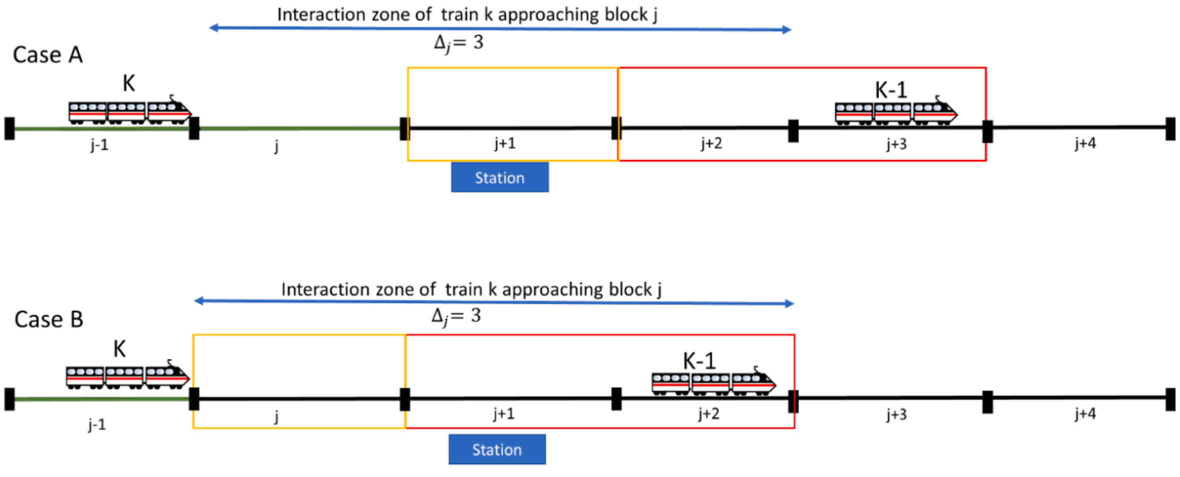

Interaction zone of train k approaching block j
$\Delta_j = 3$
Case A
K
K-1
j-1
j
j+1
j+2
j+3
j+4
Station
Interaction zone of train k approaching block j
$\Delta_j = 3$
Case B
K
K-1
j-1
j
j+1
j+2
j+3
j+4
Station

Fig. 1. Schematic of train interaction zone to determine interaction overlap time.

j. If a trailing train enters a block when the lead train is within this ‘interaction zone’, it will travel at a slower speed or will need to stop.

The Train Following Model (TFM) establishes the relationship between the maximum permitted speed of the train over the rail segment or block (here travel time on the block) and the location of the lead train. The proposed Train Following Model is inspired by car following models developed in traffic flow theory, where theoretical and empirical methods are used to model how vehicles maintain minimum space and time gaps from the lead vehicle. Given the spacing between two consecutive trains, we can determine the maximum allowed speed of the trailing train. For instance, as shown in Fig. 1 (Case A), when the following train k approaches block j, the lead train k-1 is outside the interaction zone of block j. In this case, the following train k receives its maximum (i.e. design) speed code. However, if the lead train k-1 is within the interaction zone of block j when the following train k approaches block j (Case B), the following train receives a lower speed code.

Let $\mathrm{X}_j(k)$ denote the time train k enters block j (this information is typically recorded in the Operations Control System (OCS) database for rail transit systems with fixed block signaling while the actual trajectory of trains within each block is not recorded). Let us assume that we are interested in estimating the travel time of train k to traverse block j when train k approaches block j at time $\mathrm{X}_j(k)$. If the position of train k-1 is outside the interaction zone of block j (blocks j to j + Δ_j −1), train k-1 has already entered block j + $\Delta_j$meaning that $X_{j+\Delta_j}(k-1) \le X_j(k)$. As such, train k should receive the maximum speed code at j (with a short time lag required for train k-1 to clear block j + Δ_j −1). Otherwise, if $X_{j+\Delta_j}(k-1) > X_j(k)$, train k-1 has not yet cleared j + Δ_j-1 (or not yet entered) and train k should receive a lower speed code or stop code when entering j.

Fig. 2 shows the train trajectories and block occupancies of 3 consecutive trains in a short section of a rail line on a time–space diagram. The total time that each line section is exclusively allocated to a train movement, and hence is blocked, is shown as a rectangle on the diagram. Train k receives a lower speed code at block j and thus has a longer travel time on block j while trains k-1 and k + 1 travel at the design speed as their lead trains are not within the interaction zone of block j when train k is approaching it. Let us define the overlap time of train k and train k-1, $OL_j(\mathrm{k})$, as the difference of $X_{j+\Delta_j}(k-1)$ and $\mathrm{X}_j(k)$:

$$OL_j(k) = X_{j+\Delta_j}(k-1) - X_j(k) \tag{1}$$

The overlap time measures the interaction time between the two trains. As described earlier, if $OL_j(\mathrm{k})$ is negative, the travel time of train k at j is the minimum travel time since there is no interaction between trains k and k-1 when train k enters block j. Otherwise, if $OL_j(\mathrm{k})$ is positive, a longer travel time is required for train k to clear block j. As a result, the travel time over block j is a function of the overlap time as shown in the example on Fig. 3.

As discussed, the travel time of a train on a block includes a minimum ‘conflict free’ travel time when the train is traveling at its design speed on the block and a possible additional travel time due to interaction with the lead train. The travel time of train k on block j can be expressed as:

$$\mathrm{T}_j(\mathrm{k}) = T_j^{min}(\mathrm{k}) + T_j^d(\mathrm{k}) \tag{2}$$

Where $T_j^{min}(\mathrm{k})$ is the minimum travel time of train k on block j under conflict-free operation, and $T_j^d(\mathrm{k})$ is the delay for train k on

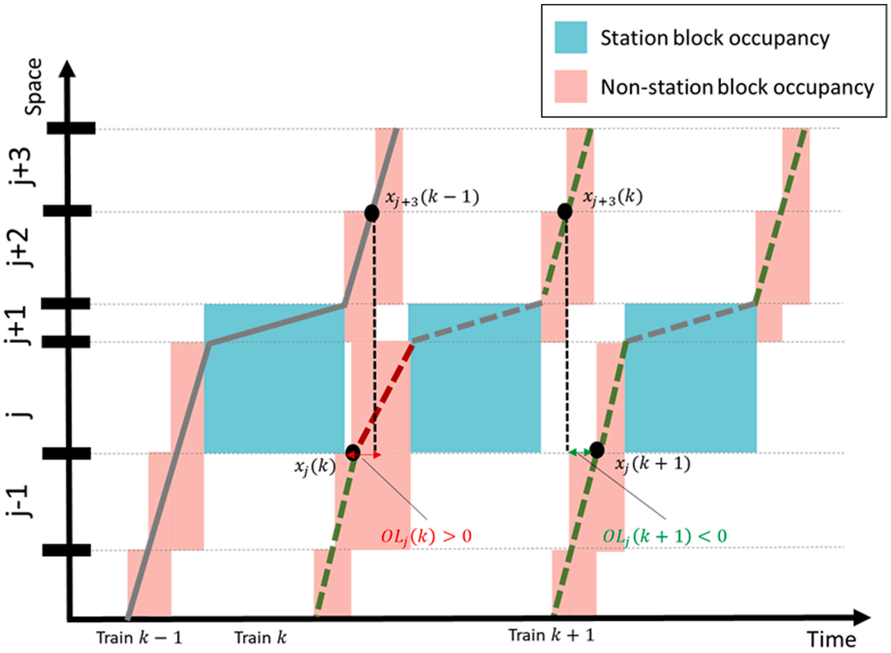


Fig. 2. Schematic of train interaction zone to determine interaction overlap time.

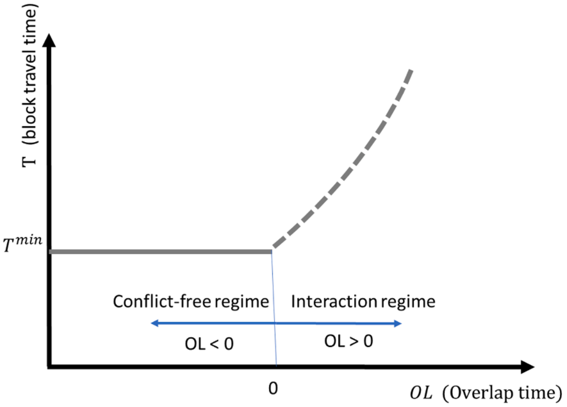


Fig. 3. Train travel time function for a hypothetical block. OL > 0 indicates closer spacing between consecutive trains.

block j under the train interaction regime.

According to Fig. 3, the block travel time of a trailing train is a function of the location of the lead train using the overlap time as the reference point for interaction between consecutive trains. Equation (2) can be rewritten as:

$$T_j(k) = T_j^{\min} + \max[0, f(OL_j(k))] \tag{3}$$

The first component is the minimum travel time in the case of conflict-free operations (when the overlap time defined by equation (1) is negative), and the second component is the expected delay due to the proximity to the lead train (the overlap time defined by equation (1) is positive). Assuming a linearly increasing function $\mathrm{f}\left(OL_j(\mathrm{k})\right)$, equation (3) can be expressed as follows:

$$T_j(k) = T_j^{\min} + \alpha_j \max[0, OL_j(k)] \tag{4}$$

Where α_j is the slope of the line.

3.2. *Train trajectory formulation*

The Train Following Model introduced in the previous section provides the required building block to develop train trajectories. Trains move on non-station and station blocks. Travel time on non-station blocks includes conflict free (free flow) travel time and possible delays due to train interactions. According to Fig. 1, assuming presence of a lead train on Δ_j downstream blocks (j to j + Δ_j-1, interaction zone of block j) affects the speed of the following train entering block j, the time train k-1 activates block j+Δ_j (entering block j + Δ_j) can be represented as:

$$\mathrm{X}_{j+\Delta_j}(k-1) = \mathrm{X}_j(k-1) + T_j^{min} + T_j^d(k-1) + S_{j+1}(k-1) + \cdots + T_{j+\Delta_j-1}^{min} + T_{j+\Delta_j-1}^d(k-1) \tag{5}$$

where:

T_j^{min} : Min occupancy time for block j.

T_j^d $(k-1)$: Delay for train k-1 approaching block j.

$\mathrm{S}_{j+1}(k-1)$: The time for train k-1 to traverse station block j+1(including dwell time, and the minimum run time on the station block).

$T_{j+\Delta_j-1}^d(k-1)$: Delay for train k-1 on block j + Δ_j −1 where the value of Δ_j depends on the signaling system and the specific track section characteristics.

Note that $\mathrm{S}_{j+1}(k-1)$ consists of a minimum travel time plus train k-1 dwell time on the station block S_{j+1}. Substituting (5) into (1) we obtain:

$$OL_j(k) = -H_j(k) + T_j^{\min} + T_j^d(k-1) + S_{j+1}(k-1) + \cdots + T_{j+\Delta_j-1}^{\min} + T_{j+\Delta_j-1}^d(k-1) \tag{6}$$

Where $H_j(k)$ is the preceding headway of train k entering block j: $H_j(k) = X_j(k) - X_j(k-1)$.

Thus, the delay of the next train (k) on block j shown in equation (3) can be written as

$$T_j^d(k) = \alpha_j \max\left[0, -H_j(k) + T_j^{\min} + T_j^d(k-1) + S_{j+1}(k-1) + \cdots + T_{j+\Delta_j-1}^{\min} + T_{j+\Delta_j-1}^d(k-1)\right] \tag{7}$$

And consequently equation (2) becomes:

$$T_j(k) = T_j^{\min} + \alpha_j \max[0, -H_j(k) + T_j^{\min} + T_j^d(k-1) + S_{j+1}(k-1) + \cdots + T_{j+\Delta_j-1}^{\min} + T_{j+\Delta_j-1}^d(k-1)] \tag{8}$$

Equation (8) above shows that the delay of train k on block j is a function of its headway, train (k-1) occupancy times on downstream blocks j + U where U={0,…,Δ_j −1} including station block occupancy (j + 1), and delay of the previous train on the same block (j). This recursive function shows that when $T_j^d(k) > 0$ which is more likely to happen when headways are short, the delay propagates to the following train(s). Besides short headways, a long delay of the lead train $T_j^d(k-1)$, or $T_{j+U}^d(k-1)$, or long dwell time of train k-1, $S_{j+1}(k-1)$, can also result in delay propagation. Thus, any initial disruption, or even a short interruption, could result in knock-on delays for several following trains.

3.3. Station occupancy time

A simple dwell time model that can capture the impact of increased numbers of passengers boarding and alighting due to higher demand or longer headways is employed. Station occupancy time includes a constant travel time (time to decelerate to a stop, door opening and closing time, time to accelerate when leaving a station, etc) and a variable dwell time. We assume that the station dwell time is a function of the headway and passenger demand (boarding and alighting). With a fixed demand rate per unit of time, increases in headway result in more passengers wait to board (and alight at destination stations) during one headway; resulting in increased dwell time. Although simple, this dwell time model can adequately represent observed dwell times and capture the impact of demand and headway variability. We define the station occupancy time of train k at station i as:

$$S_i(k) = \min\left[H_i(k)D_i\theta_i + S_i^{\min}, S_i^{\max}\right] \tag{9}$$

where:
D_i: Demand per unit of time or demand rate at station i.
θ_i: Boarding and alighting time per passenger at station i.
S_i^{min}: Station block i minimum occupancy time (run-time, acceleration and deceleration time, and minimum dwell time).
S_i^{max}: Maximum allowed dwell time at station i.

Boarding and alighting time per passenger can be estimated using historical data for each station for different times of the day. In cases when denied boarding is common due to limited capacity or a strict dwell time policy is used, a maximum dwell time constraint is considered to avoid long dwell times under high demand scenarios. As such, we use the min function to ensure $S_i(k)$ does not exceed the

Table 1
Line Performance Model Algorithm.

```
Input: Conflict-free travel time on each block $T_i^{min}$, travel time function on each block i, $\Delta_i$ for each i, dwell time model parameters $S_i^{min}$ and $\theta_i$ , and analysis period δ.
    Initial train trajectory (time train enter each block $X_i(0)$)
    Trains departure at terminal $X_0(k)$
    Passenger demand rate at each station i during study period: $D_i$

Initialization
    Set (if any) disruptions (primary delay) for train k at block i, $T_i^d(k)$ for i $\in T$ (non-station block) or $S_i^d(k)$ for i$\in S$(station block).

Train Movement Model
    For each train k = 1, …, N
      For each block i = 1, …, I
        If i $\neq$ Station block Train Following Model
          Set $T_i^d(k) := \alpha_i \max[0, X_i(k) - X_i(k-1) + T_i^{min} + T_i^d(k-1) + \sum_{U=1}^{\Delta_i - 1}(T_{i+U}^{min} + T_{i+U}^d(k-1) + \sum_{U=1}^{\Delta_i - 1}(S_{i+U}^{min} + S_{i+U}^d(k-1))]$
          Set $X_{i+1}(k) := X_i(k) + T_i^{min} + T_i^d(k)$
        Else if i = Station block Dwell Time Model
          Set $S_i^d(k) = \min[(X_i(k) - X_i(k-1))D_i\theta_i , S_i^{\max}]$
          Set $X_{i+1}(k) := X_i(k) + S_i^d(k) + S_i^{min}$
        End if
    End for
  End for

  Output
  Travel time of individual trains
  Delay of individual trains (difference between predicted travel time and minimum possible travel time $T_i^{min}$)
  Train headways at each station (including bottleneck throughput)
  Knock-on delays (cumulative delay of the following N trains on non-station blocks)
  Total knock-on delay = $\sum_{k=1}^{N}\sum_{i=1}^{I} T_i^d(k)$
  Total delay = $\sum_{k=1}^{N}\sum_{i=1}^{I} T_i^d(k) + S_i^d(k)$
  Average delay per Train = $\frac{\text{Total Delay}}{N}$
  Line Throughput at i = $(N-1)/(X_i(N) - X_i(1))$
```

maximum dwell time at each station.

3.4. Line performance model

The Train Following Model described in the previous section, if applied to all track sections, can predict a train's progression (i.e. trajectory) along the line for a given initial train trajectory. The travel time on each block is expressed as the station occupancy time (equation (9)) which includes dwell time and minimum block occupancy time (for station blocks), or minimum travel time plus delay, as in equation (8), for running blocks. Assuming a set of dispatching headways, and the trajectory of the initial train, the line performance model can fairly represent delay propagation and recovery along the line, headways of trains at different stations, and the impact of disruptions and deviations from the operations plan. The travel time of train k on each block j is predicted based on equations (8) or (9) and sequentially applied for each block and each following train recursively. Thus, considering the trajectory of an initial train and subsequent dispatching headways from the terminal, the trajectories of the following trains can be predicted using the model.

Table 1 summarizes the algorithm driving the performance model using the Train Following Model described in the previous section. The process begins with the trajectory of an initial train ($k = 0$) as well as the following train departure time from the terminal (block $j = 1$) as inputs. The trajectory of the following train ($k = 1$) is then estimated based on the trajectory of the lead train $k = 0$. The process continues recursively to estimate the trajectories of the N following trains. The use of the travel time functions and the overlap described in Section 2.1 enables real-time prediction of train trajectories and delay propagation.

4. Empirical analysis and model validation

The proposed Train Following Model is calibrated and validated with data from the Red Line of the Massachusetts Bay Transportation Authority (MBTA). The data from the Operations Control System (OCS) includes the times each block is activated by a train over a period of three months. Automated passenger fare collection system (AFC) data provide passenger demand data at individual stations. Dwell times were inferred by combining the block activation times and the OD flow data obtained from the AFC transactions (Gordon et al., 2013).

This analysis focuses on the trunk section of the line. The MBTA Red Line (Fig. 4) has very different properties in terms of infrastructure and operational practices compared with many newer systems. The MBTA rail lines are driven by operators with a supervisory train control system activated for safety. Thus, there is variability in train speed and dwell times which affects operations. In addition, the MBTA Red Line has bottlenecks caused by two closely spaced stations in the downtown area, Park St. and Downtown Crossing stations, with only one short block between them. Both stations have heavy boardings and alightings as they are transfer stations to MBTA Green Line and Orange Line, as well as being major generators of commuting and recreation trips in downtown Boston. The causes and impact of this bottleneck on the operations will be further discussed later in the paper.

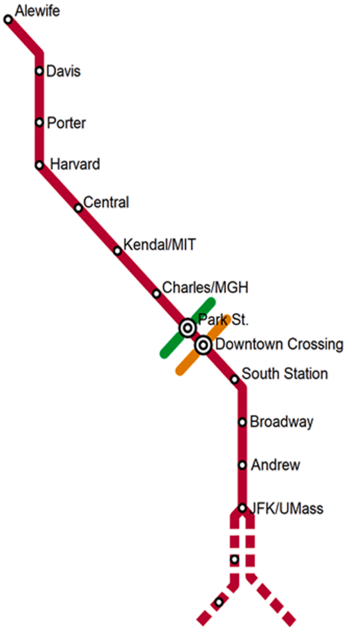


Fig. 4. Trunk section of the MBTA Red Line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Block occupancy analysis

We apply block occupancy analysis to understand the operations bottleneck, sources of primary delays, and delay propagation. A block occupancy diagram better illustrates the difference between normal and delayed train movements, and their variability, than a train trajectory graph. To illustrate the variability in blocking times, we present an example based on historical track data from the MBTA Red Line. The analysis illustrates that, even with homogenous train types, blocking times and minimum headways depend on train interactions.

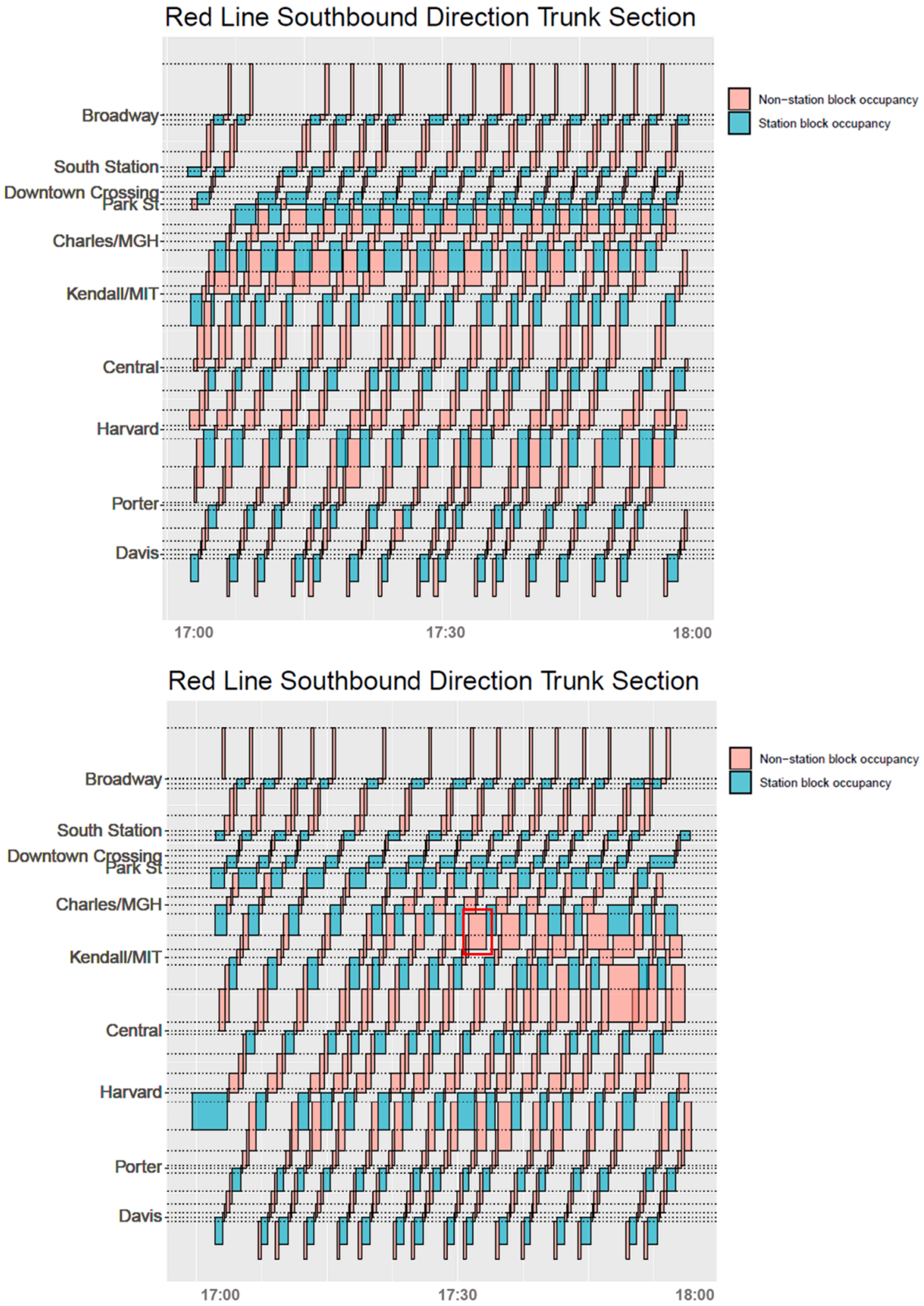
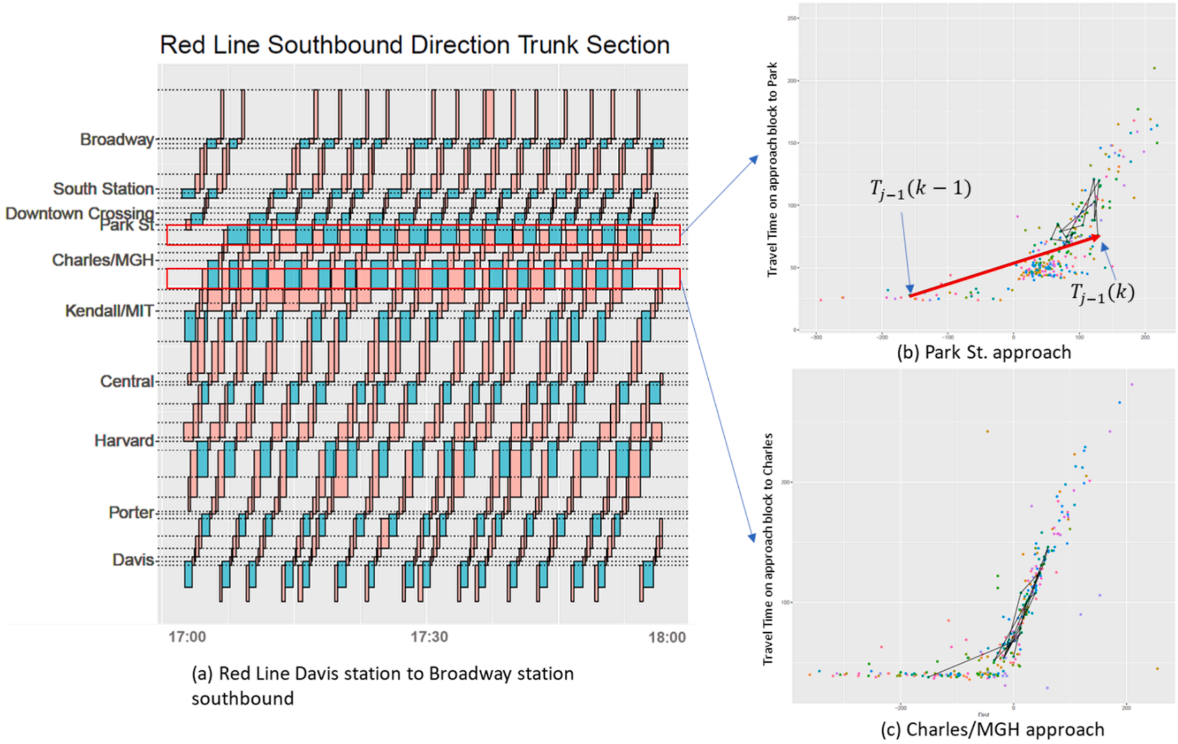


Fig. 5. Block occupancy time–space diagram for MBTA Red Line southbound on a) April 20, 2018 (top) and b) April 6, 2018 (bottom) from 17:00 to 18:00. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 5 shows a time–space diagram for the Red Line (southbound) operations for the trunk section during a one-hour period between 17:00 and 18:00 on two different weekdays. Gaps between consecutive trains show available slack times to absorb schedule deviations. The signal block just before Park St. station, has the lowest slack times. As seen in Fig. 5-a, this section of the line is a bottleneck. With the existing train specifications and dwell times, there is a maximum train throughput which can be calculated at this bottleneck. No matter how many trains are dispatched, the bottleneck block, or "weakest link" according to the Transit Capacity and Quality of Service Manual (Kittelson, et al., 2013), will control train throughput. Headways lower than the minimum headway at the bottleneck will lead to an upstream queue of trains and associated train delays. Fig. 5-b, on the other hand, shows a different scenario where operations seems to be smooth until the 9th train at Charles/MGH. A primary delay because of two closely spaced trains at Charles/MGH station at around 17:30 (as shown in Fig. 5b) results in a propagation of delays first for trains that arrive at Charles/MGH and later at the Kendal/MIT station. As shown in the two scenarios in Fig. 5a and 5b, the propagation of delays depends on the initial conditions (i.e. dispatch headway, occurrence of a primary delay and its time and location). In the scenario shown in Fig. 5a, delays initiated and then propagated to trains at the same stations (i.e. Park St. and Charles/MGH). In contrast, in the scenario shown in Fig. 5b, delays propagated to trains at upstream stations (i.e. from Charles/MGH to Kendal/MIT).

The block occupancy analysis can help detect bottlenecks in operations and determine the bottleneck capacity section(s) of the line. In addition, the block occupancy analysis clearly illustrates the times and sections of the line that are primary sources of delay in operations and how primary delays result in secondary delays for following trains. As such, the analysis will be useful in developing the Train Following Model and analyzing line performance.

Fig. 6 illustrates the travel times experienced by consecutive trains at the approach block before Park St. and Charles/MGH stations – the main operational bottleneck on the MBTA Red Line, during the 5–6 pm period on April 20, 2018. Each dot in Fig. 6(b) and 6(c) represents train travel time observations. The color of each dot represents a different day during the peak hour. The transition patterns of travel time on the preceding block to Park St. and on the preceding block to Charles/MGH are annotated with the black solid lines in Fig. 6(b) and Fig. 6(c) during the one-hour operation shown in Fig. 6(a). The travel time observation for train k-1 on the block is connected to the travel time observation for the following train k. During this one-hour operation, the transition of travel time for the Park St. approach block remains in the interaction regime while the transition of travel time for the Charles/MGH approach block seems to have higher variability due to a mix of conflict-free and interaction regimes. This could also be observed by the size of the occupancy blocks (in red) in Fig. 6(a). It can be observed that train operations were in the uncongested regime at the beginning of the time period upstream of the bottleneck. All consecutive trains are in the congested regime near Park St. station while the amount of delay varies. This analysis helps detect delay propagation to the following trains at upstream stations.



(a) Red Line Davis station to Broadway station southbound

(b) Park St. approach

(c) Charles/MGH approach

Fig. 6. A) block occupancy time–space diagram for MBTA Red Line southbound, afternoon peak b) Time series of train travel times on the approach block before Park St station c) Time series of train travel on the approach block before Charles/MGH station. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Train following model validation

The travel time function proposed in Section 2.1 is validated using the Red Line OCS data. Fig. 7 shows the empirical analysis of the travel time function using the definitions in Section 2.1 for several blocks on the (southbound) MBTA Red Line. We have included all key blocks: blocks approaching the Charles/MGH and Park St. stations; blocks approaching the Harvard, Kendal/MIT, and Central square stations; a block after the Charles/MGH station. The empirical analysis indicates that linear travel time functions within the

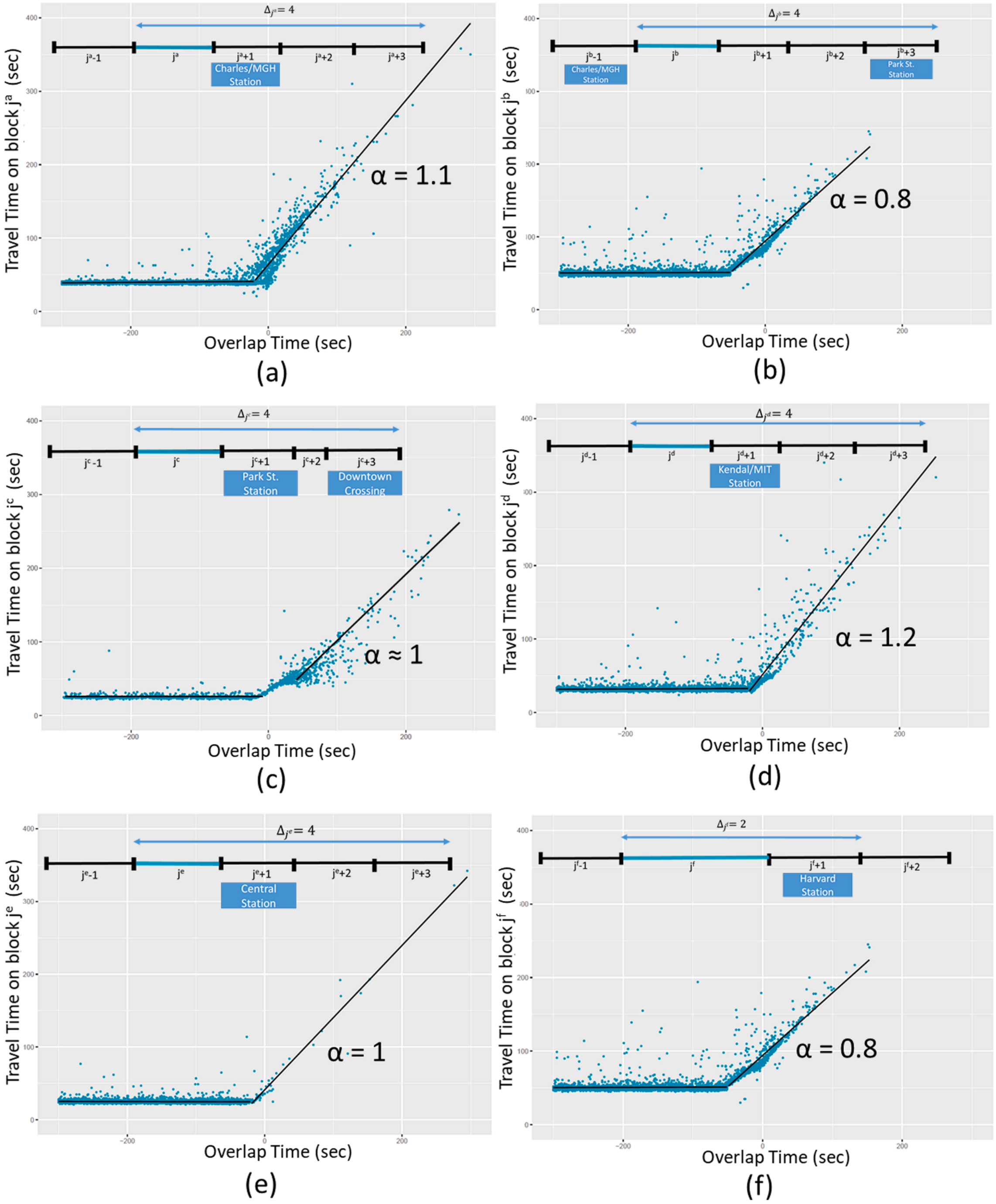


Fig. 7. Travel time function for different blocks of the MBTA Red Line based on OCS data. a) block before Charles/MGH Station b) block after Charles/MGH Station c) block before Park St. Station d) block before Kendal/MIT Station e) block before Central Station f) block before Harvard Station. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interaction zone, as assumed in equation (4), are reasonable. The estimated parameter α for each block is also shown. Similar to the hypothetical delay curve, travel time for the following train within the conflict-free (uncongested) zone is the minimum possible travel time for the block, or within the conflict (congested) zone is an increasing function of the overlap time.

Fig. 7 indicates that the block run-time of the following train can be approximated by a linearly increasing function of the overlap time as defined in Section 2.1. According to the sample block travel time plots, this is a reasonable assumption for a mesoscopic model. It should be noted that train travel times may not necessarily be equal for the same overlap times and same train types. There are some variability due to differences in operator behavior, train configuration, and actual demand and train load. The slope of the function in the congested section, estimated by ordinary least squares linear regressions, differs slightly across train blocks, ranging from 0.8 to 1.2. The difference in the slope depends on the location, and geometry of the signal blocks, and associated speed codes. Temporary speed restrictions (for reasons such as safety, construction, etc.) also affect the slope. For instance, on the block before Kendal/MIT (Fig. 7(d)), the slope is 1.2 while on the block before Harvard (Fig. 7(f)), the slope is around 0.8. The higher the slope, the higher the sensitivity to the overlap time and the higher the delay due to train interactions. However, it should be noted that the travel time function for the sequence of blocks governs the overall throughput and ability to absorb delays on a line segment. The block before Park St. (Fig. 7(c)) exhibits high variability in travel time especially at small overlap times. One reason for this may be the presence of two closely spaced stations (Park St and Downtown Crossing) with only one block separating the two. The travel time function does not have a similar pattern to the station blocks and is not a function of the train ahead, rather it depends on the dwell time (number of boarding and alighting passengers).

The empirical analysis also reveals that the train interaction zone for each block may contain different numbers of blocks as dictated by the speed codes of the signaling system. Thus, choosing the right reference point to determine the overlap time for each block requires either access to the speed codes or analysis of the empirical data. The reference point should be chosen as the earliest possible block where the presence of the lead train does not affect the speed of the following train entering block j, a conflict-free operation. It should be noted that the available MBTA OCS data only included circuit track activation times and not deactivation times. As such, occupancy times in the analysis are slightly underestimated by excluding the duration between the train entering the current block and the time clearing the preceding block. This is the reason for a slight lag between the 'zero overlap threshold' in Fig. 7 and the start of the increasing slope of the train travel time function on the interaction regime.

As indicated, travel times on station blocks are mainly governed by the dwell time and passenger demand, and not by the Train Following Model. The dwell time model coefficients, such as boarding and alighting times per passenger and the intercept, at each station during the peak hour period are obtained from Wolofsky et al. (2019). In that study, separate dwell time model coefficients were obtained based on the dominant passenger flow direction (predominantly boardings, predominantly alightings, and mixed flow) for different stations and times of day. In this study, we used model coefficients related to the more dominant passenger flow directions for each station. Thus, coefficients defined in equation (9) were calibrated using historical station level data.

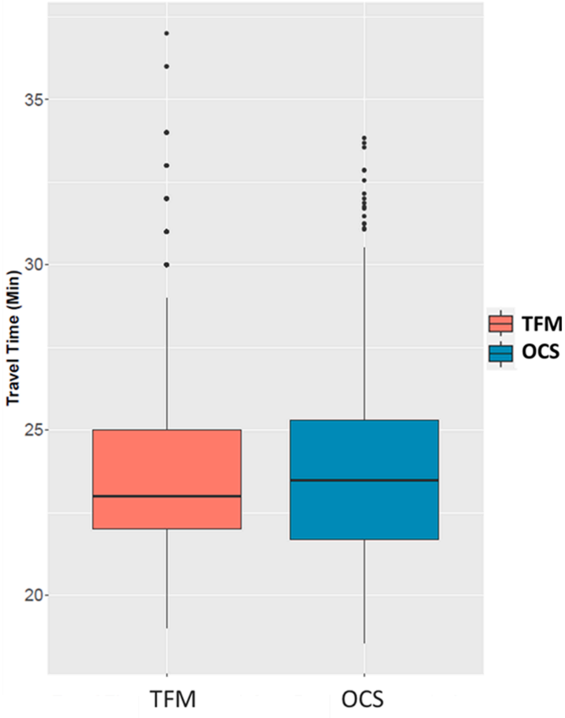


Fig. 8. Run time distribution comparison for OCS and the TFM (Davis to South Station).

4.3. Line performance model validation

The proposed model is validated using the OCS data from the Red Line which include the times a block is activated by a train. For the model, we assumed normal operation of the first train during each time period and dispatched headways at the terminal following the same mean and standard deviation as observed in the OCS data during that time period. We compare line performance measures including headways at key stations, run times, and a scenario-based comparison of train trajectories.

We first compare run times predicted by the model with actual run times from the OCS database. Fig. 8 presents the run time comparison between Davis and South stations during the PM peak period. The predicted run times exhibit a distribution close to the observed. The mean run times according to the actual data and the model results are 23.9 and 23.6 min respectively (1.2 % difference). The variability in run times was also captured by the Train Following Model with standard deviations of 2.78 min compared to 3.02 min in the actual data (8.6 % difference). The Wilcoxon signed-rank test was performed to compare the mean of the two samples. The null hypothesis that the two samples have the same mean cannot be rejected at the 95 % level of confidence (p-value = 0.057). Similarly, the F test was conducted to compare the difference between the two sample variances. With p-value of 0.13, the hypothesis that the sample variances are equal cannot be rejected at the 95 % level of confidence.

We also compare the headway distributions at key stations at the bottleneck and upstream of the bottleneck. As Fig. 9 illustrates, the median headway based on model output is consistent with the OCS data. Headway variability is also well captured. Headway variability consistently declines approaching the bottleneck station as headways are regulated at the bottleneck due to the bottleneck capacity constraint. We will discuss how the bottleneck headway can be analytically obtained from the proposed model in Section 4.1.

The ability of the proposed model to represent system performance under interruptions or disruptions is also important. We use OCS data from 04/18/2018 during the 4–5 pm peak period when the headway of>10 min is experienced at Davis station due to delay in train dispatch departing Alewife terminal at 4:10 pm. The inputs to the model are the trajectory of the train departing just before the incident and the departure headways at Davis station. Fig. 10 compares the trajectories of the trains based on the OCS data and model output. The Train Following Model is able to capture the trajectories of the trains after the long delay reasonably well. Run-time differences between OCS and the model are mainly due to dwell time differences. Dwell times are impacted by various external factors (e.g. demand spikes, operator variability, events such as a door jam, etc). For instance, we observe a discrepancy between the modelled dwell time at Harvard station of the fifth train and the actual dwell time. The actual dwell time is longer than that predicted by the model. The longer dwell time may have been caused by any external factor including another interruption. This may also be the source of observed discrepancies between the fifth and sixth trains compared to the actual data. The overall results illustrate the value of the model to predict train trajectories and system performance under a variety of operating conditions.

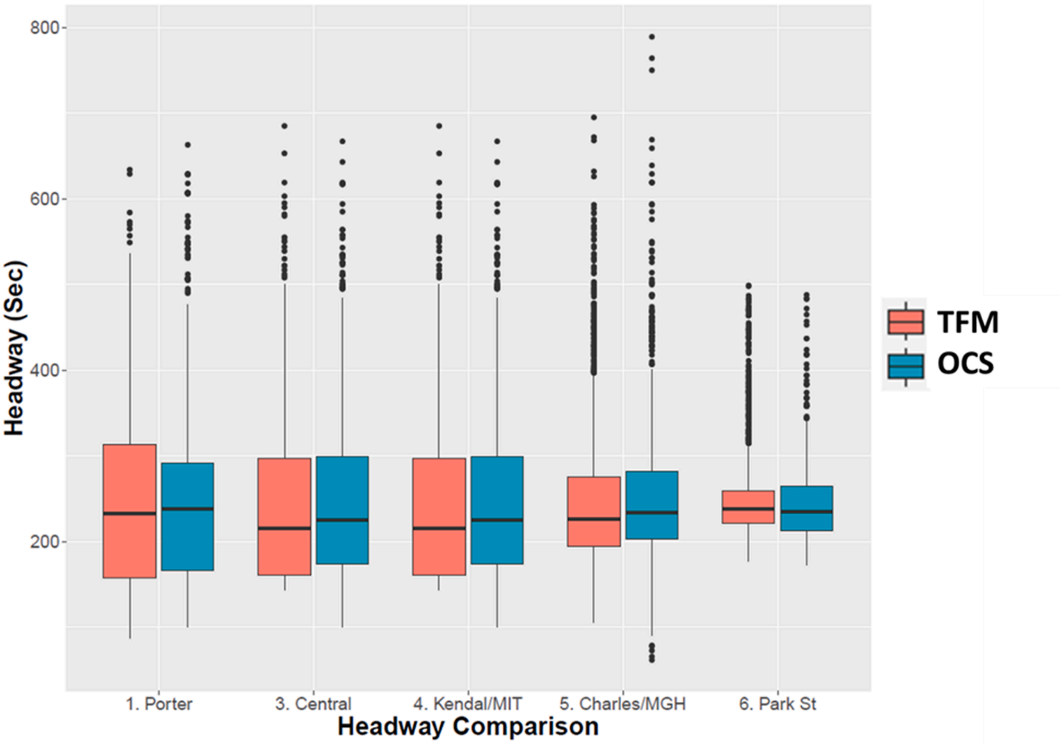


Fig. 9. Headway distribution comparison between OCS and the TFM at key stations.

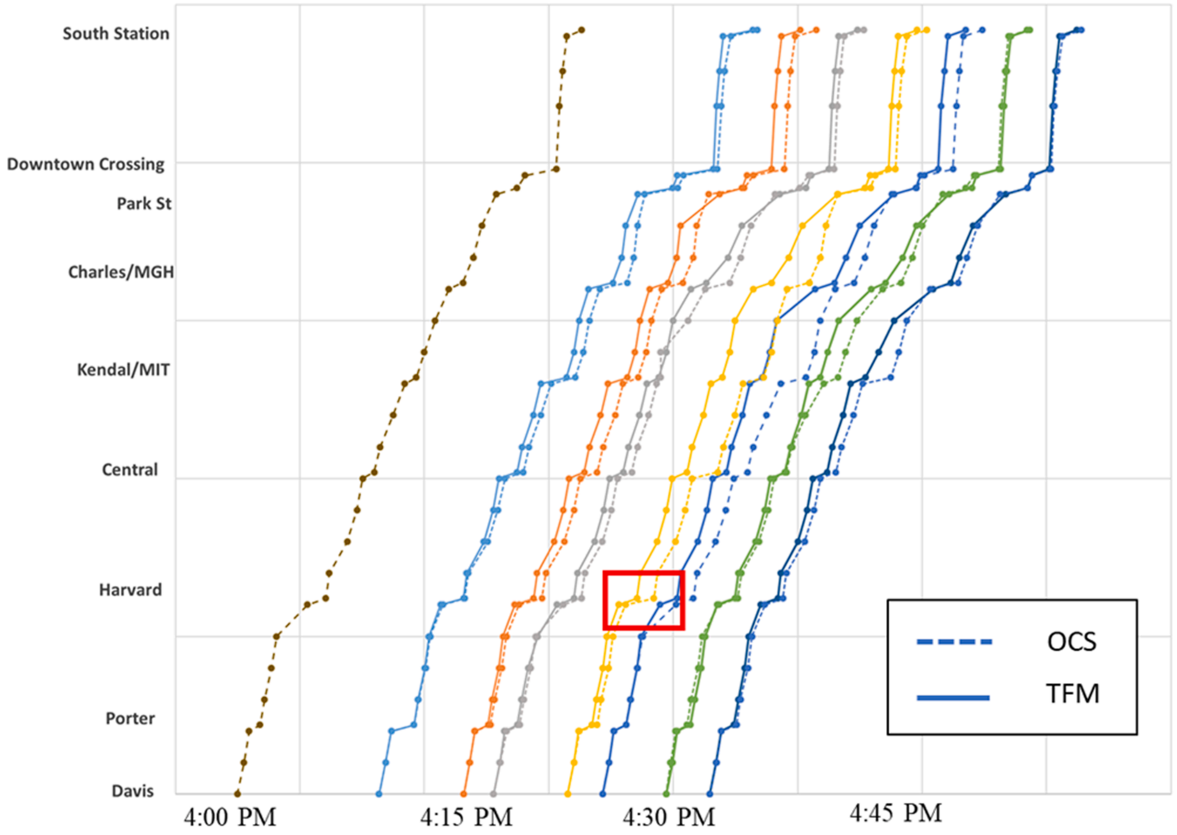


Fig. 10. Train trajectory comparison between OCS and the TFM during a disruption on 2018–04-18.

5. Applications and discussion

In this section we present some applications of the Train Following Model. We first discuss how the model can be used to determine bottleneck capacity. Then the model is used to evaluate performance under various service frequency assumptions. We also examine a disruption scenario focusing on knock-on delays. Lastly, we present the impacts of dispatching headway variation and passenger demand levels on overall line performance.

5.1. Assessment of bottleneck capacity

The proposed model can be used to assess capacity at bottlenecks. Park St. and Downtown Crossing stations on the MBTA Red Line are close together (with only one short block between them) as shown in Fig. 5. The travel time function on the approach block to Park St. depends on the next three downstream blocks (including the Park St. and Downtown Crossing station blocks). This means that, unlike other stations where we observe conflict-free train travel time on the block after the station, the travel time on the blocks after Park St.(including the dwell time at Downtown Crossing) also includes delays and affects the travel time for the approach block to Park St. This may be the reason that empirical travel time for this block has a different shape and higher variability.

Denoting by j the approach block to Park St.; j + 1 the Park St. station block; j + 2 the short block between Park St. and Downtown Crossing; and j + 3 the Downtown Crossing station block, the travel time formulation for the approach block to Park St. is as follows:

$$\mathrm{T}_j(k) = T_j^{min} + T_j^d(k) \tag{10}$$

where:

$$T_j^d(k) = \alpha_j \max[0, -H_j(k) + T_j^{\min} + T_j^d(k-1) + S_{j+1}(k-1) + T_{j+2}(k-1) + S_{j+3}(k-1)] \tag{11}$$

According to the travel time model calibration, $\alpha \simeq 1$ for block j (Park St. Station). Since the interstation block j + 2 is very short and just after Park St., the travel time on block j + 2 , $\mathrm{T}_{j+2}(k-1)$, is constant and independent of train interactions. Thus, assuming that the trajectory of train k from entering block j to exiting j + 2 is based on equation (11), the headway of train k when entering block j + 2 (leaving the Park St. station block) is given by:

$$H_{j+2}(k) = \begin{cases} \mathrm{H}_j(k) + \mathrm{S}_{j+1}(k) - S_{j+1}(k-1) - T_j^d(k-1) & \text{if } T_j^d(k) \ = 0 \qquad (i.e.\ long\ headway) \\ T_j^{min} + \mathrm{T}_{j+2}(k-1) + \mathrm{S}_{j+3}(k-1) + \mathrm{S}_{j+1}(k-1) & \text{if } T_j^d(k) \ > 0 \qquad (i.e.\ short\ headway) \end{cases} \tag{12}$$

Equation (12) shows that departing Park St, when the initial headway for train k before block j is short enough - resulting in a delay on the approach block j - the headway departing Park St. (block j + 1), is independent of the headway of train k, and dominated by the dwell time of train k-1 at Park St. (block j + 1) and Downtown Crossing (block j + 3) as well as the constant travel time on the short interstation block between Park St. and Downtown Crossing. Thus, for trains arriving during the congested regime of the travel time function ($T_j^d > 0$):

$$H_{j+2}(k) = T_j^{min} + T_{j+2} + S_{j+3}(k-1) + S_{j+1}(k-1) \tag{13}$$

This observation is consistent with the bottleneck effect shown in Fig. 9. Both OCS data and the model results show that headways are regulated there with considerable reduction in headway variability. Reduced headway variability at Park St. station indicates that the headways at Park St. are a function of factors other than dispatched headway variability; most likely factors such as typical dwell times at Park St. and Downtown Crossing as indicated by equation (13). Using the minimum travel times for blocks j and j + 2 and the median dwell times at Park St. (S_{j+1}) and Downtown Crossing (S_{j+3}) (based on OCS data) in equation (13), we obtain a headway value around 230 *sec* for the existing operations. This is consistent with the actual headways based on the OCS data shown in Fig. 9.

5.2. Performance evaluation with respect to service frequency

Several performance metrics, including line capacity and train delay, can be evaluated using the Train Following Model. Capacity or throughput of the line is defined as the maximum number of trains that can pass the critical section of the line during a time period (Martin, 2014). In the case of the MBTA Red Line, as explained in sections 2.1 and 4.1, the capacity of the line is determined by the bottleneck section at the downtown stations. We compare the throughput after the bottleneck (number of trains per hour that pass through the bottleneck section) as a function of the scheduled frequency of service. As expected, the throughput is bounded by the bottleneck capacity so that any increase in frequency beyond the bottleneck capacity increases delays due to train congestion upstream of the bottleneck.

Fig. 11 shows the throughput and delay as a function of scheduled frequency. As mentioned in Section 2.3, delay is defined as the travel time difference between observed/predicted train travel time and the travel time under conflict-free operations of trains along the line. The results are consistent with the capacity determination presented in Martin (2014). The model can be used to determine the trade-offs between serving more passengers and train delays as a measure of efficiency.

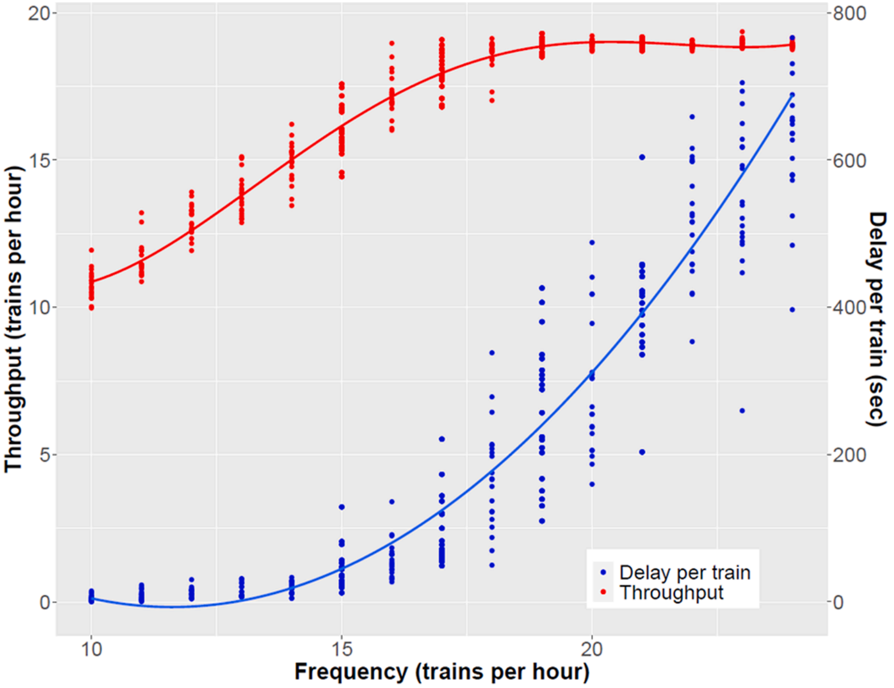


Fig. 11. Frequency, throughput, and delay relationships using the Train Following Model (train frequencies at the MBTA Red Line during peak hour periods during spring 2018 were between 13 and 16 trains per hour). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this analysis typical variability is assumed in both the dispatching headways and dwell times at downtown stations. The results in Fig. 11 indicate that there is little benefit in terms of throughput from increasing frequency above 17 trains per hour (tph) and no gains after 20 tph. At the same time, delays increase rapidly for scheduled frequencies over 15 tph. As shown in Fig. 11, frequencies of 17 and 20 tph result in actual throughput of around 16 tph and 17 tph respectively, with average delays per train of 140 s and 300 s respectively. Note that the delays reported are for the southbound direction between Davis and South Station. Extreme dispatching scenarios (frequencies over 20 tph) represent situation when additional train dispatches do not improve line throughput (around 18 tph as shown in Fig. 11) while average delay to each train increases exponentially with each additional hourly train dispatch. As such, going over certain level of train dispatch will negatively affect the line performance and can lead to infeasible timetables.

Considering average dwell time at stations and no primary delay, the minimum travel time (when there are no train interactions) from Davis to South Station is about 19.5 min at a frequency of 15 tph, based on the model output. The estimated delay for higher frequencies is consistent with the results reported in the literature. Heimburger et al. (1999) report average delays around 3, 6.5, and 10.5 min for frequencies of 18, 20, and 22 tph, respectively compared with the baseline minimum travel time. The corresponding values from our model are 4, 7.5, and 12 min respectively. Zhou el al. (2020) also report around 7 min of additional travel time on average when the frequency increases from 14 tph to 20 tph (assuming that the coefficient of variation for dispatched headways is 0.30). According to the results in Fig. 11 the corresponding increase in travel time is around 6.5 min.

5.3. Stability and knock-on delay propagation

The proposed model can also be used for initial assessment of alternative schedules in terms of operations stability. Stability is the ability to return to normal operations after a primary delay or interruption in service. Primary delays are common in operations and are caused by incidents which result in longer than scheduled train occupancy times. Secondary delays refer to delays experienced by a following train due to a conflict with the lead train through the signaling system. Knock-on delays are defined as cumulative secondary delays by a set of following trains caused by a primary delay of a train. (Bešinović and Goverde 2019, Goverde and Hansen 2013).

As a performance measure for stability, we use the knock-on delays and number of affected trains after a primary delay until the effect is absorbed and the system returns to normal operations. Primary delays or interruptions are unavoidable in practice. For this analysis, the primary delay is an input to the model in order to study its impact on performance as measured by the resulting knock-on delays.

Fig. 12 compares the performance in terms of knock-on delays assuming primary delays for the lead train of one and five minutes respectively at the bottleneck station (Park St.). The knock-on delays are estimated for sixteen following trains (which was the maximum frequency of service during the peak hour of operations on the MBTA Red Line in Spring 2018). Knock-on delays are reported between Davis and South Station.

As observed in Section 4.2, regular operations may also result in knock-on delays especially at higher service frequencies. At lower scheduled frequencies the primary delay can be absorbed; however, at higher frequencies, delays propagate upstream to following

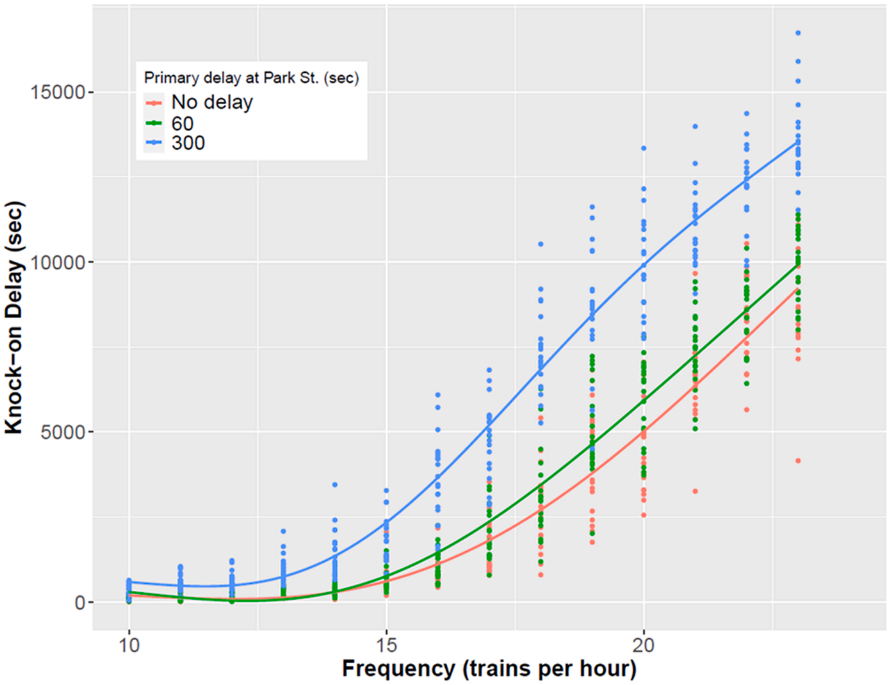


Fig. 12. Effect of initial disruption on knock-on delays as a function of scheduled frequency.

trains. For instance, at a frequency of 15 tph, the initial 60-*sec* service interruption (i.e. primary delay) at Park St. results in an additional 230 *sec* of knock-on delays to following trains on average. However, at a frequency of 20 tph, the same train interruption results in an additional 670 *sec* of knock-on delays on average.

When the primary delay is more extreme (i.e. service disruptions), the knock-on delay is considerably higher. As shown in Fig. 12, a 5-minute primary delay at the bottleneck station will result in about 2,400 *sec* and 4,000 *sec* of knock-on delays at frequencies of 15 tph and 20 tph, respectively. Each dot in Fig. 12 represents an independent run of the model for the scheduled frequency and primary delay at the bottleneck station. Choosing the right trade-off between scheduled frequency and service stability has been explored by other researchers using microsimulation; however, applying such approaches to daily operating problems is usually not feasible. The model proposed here is promising both for planning and real-time decision support applications.

Fig. 13 shows the individual train secondary delays caused by 1- and 5-minute primary delays at Park St. station. We assume a scheduled headway of 240 *sec* (15 tph) and no variability in headways at the terminal (under these conditions, trains are expected to travel with no secondary delays along the line). We observe decreasing delays for the following trains between Davis and South Station. Under the 1-minute primary delay there are 5 affected trains that experience knock-on delays. However, the number of affected trains increases to 11 for a 5-minute primary delay at Park St. station.

5.4. *Impact of demand and headway regularity*

In this section, we explore the impacts of passenger demand and headway regularity on overall line performance. As expected, an increase in passenger demand will result in an increase in dwell times, especially at the busy downtown (also bottleneck) stations, and thus further impact train congestion, throughput, and delays. Higher headway variability results in degradation of effective capacity (i.e. reduced line throughput) and larger train delays. We consider different demand levels multiplying the base demand by a factor. For headway regularity, we consider the dispatched headway variability at terminal. We use the headway coefficient of variation (CV), a commonly used metric for headway regularity, defined as the standard deviation divided by the average headway.

Fig. 14 shows line performance in terms of delays and bottleneck throughput as a function of total passenger demand and headway CV. Line performance is assessed for a range of CV values (0 to 0.8) and demand factors (0.8 to 1.8). The scheduled headway is 240 *sec* (peak-hour period headway in Spring 2018). As the headway variation increases, delays increase significantly while throughput decreases. Increased demand also results in increased delay and decreases in throughput. As expected, the best performance happens under the lowest demand factor (0.8) and regular headways: The average delay is 50 *sec* per train while the throughput goes as high as 17.5 tph. Under high demand (factor 1.8) and headway variability (CV = 0.8) average delay becomes 450 *sec* per train and the effective line throughput decreases to 13.5 tph. On average, a 10 % increase in demand results in an 0.25 tph decrease in throughput, and a 10 % increase in CV results in an 0.20 tph decrease in throughput. A 10 % increase in passenger demand results in 35 s increase in delay per train while a 10 % increase in CV results in 28 s increase in delay per train on average. Thus, consistent dispatching headways can improve the performance of the line in terms of both throughput and train delays. To the best of our knowledge, this relationship is

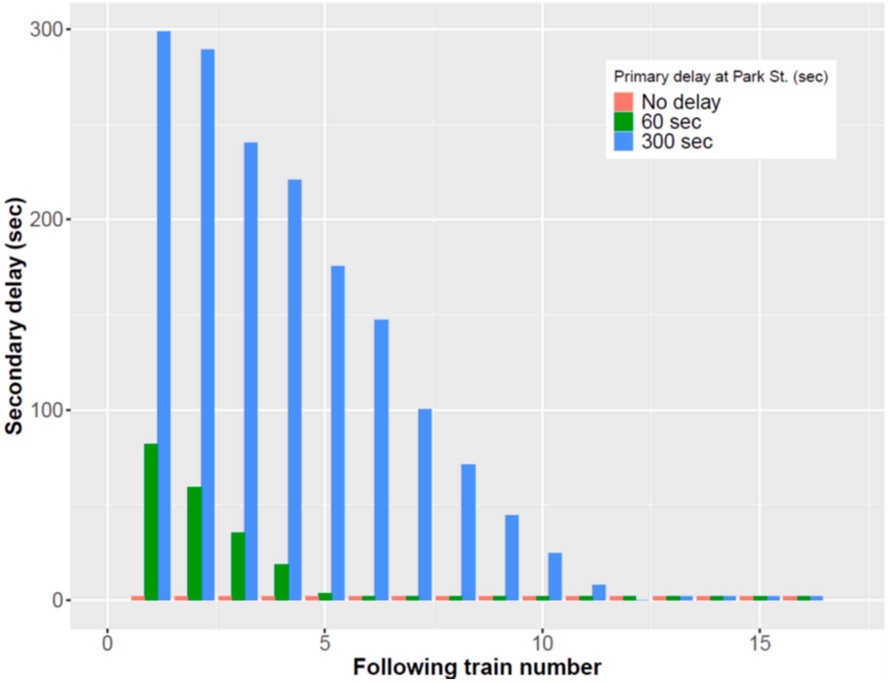


Fig. 13. Effect of a primary delay on the secondary delay for the following trains assuming no dispatched headway variability.

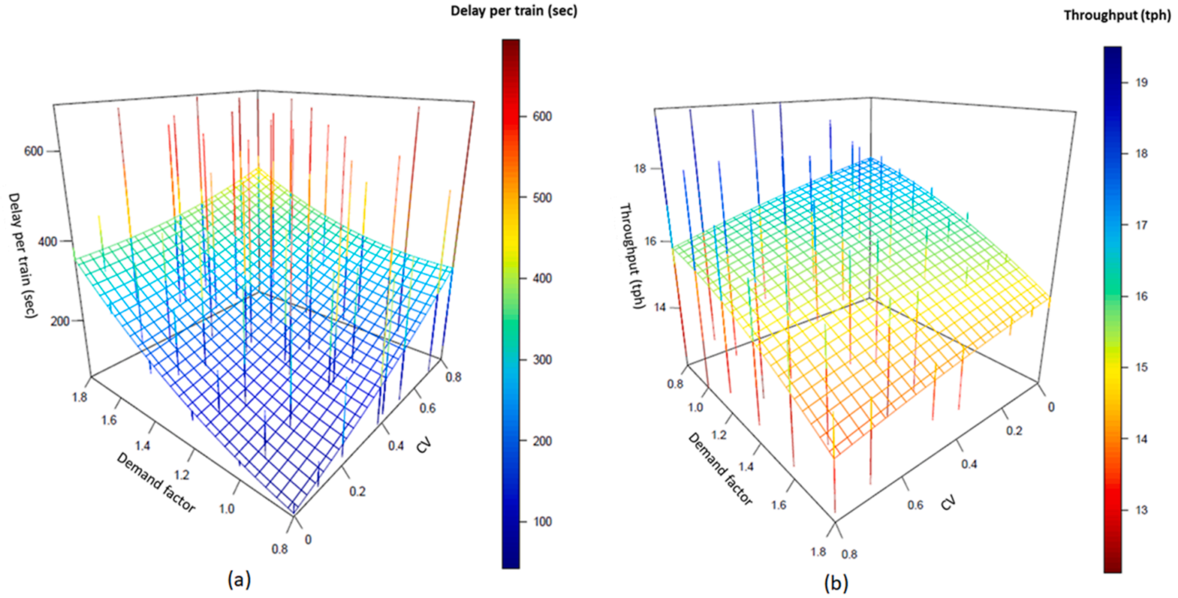


Fig. 14. A) delay per train and b) capacity as a function of departure headway coefficient of variation and passenger demand (scheduled headway of 240*sec*).

demonstrated here for the first time, and it should prove useful in developing fundamental relationships between demand, regularity, and line performance for rail transit operations.

6. Conclusion

In this paper, we proposed a Train Following Model which can efficiently capture the effects of train interactions. The model is based on a recursive estimation of the trajectory of a train based on the proximity to the lead train. By constructing a series of train travel time functions for each block, trajectories of trains can be estimated given initial conditions (i.e. the trajectory of an initial train and dispatching headway of following trains). The proposed model can explore relationships among train delay propagation, service frequency, line capacity, headway variation, and passenger demand, as well as assess the stability of a given schedule to disruptions. The model can serve as a quick response tool both in real-time and offline. Because of its recursive structure the model is implemented in a spreadsheet and is not as time- and resource- intensive as more detailed simulation models.

This model is sensitive to variability in dwell times, departure headways, and delays due to disruptions. The outputs are various performance measures such as train delays at the line or block level, headways along the line, and line throughput. A possible application of the model is its use as real-time disruption mitigation decision support tool. The proposed train following model was calibrated and validated using historical train tracking data for the MBTA Red Line. Several applications were also presented in the Red Line context. According to the model, increases in demand results in increases in delay and decreases in throughput. On average, a 10 % increase in demand results in an 0.25 tph decrease in throughput, and a 10 % increase in the dispatched headway coefficient of variation results in an 0.20 tph decrease in throughput. Similarly, a 10 % increase in passenger demand results in 35 s more delay per train, while a 10 % increase in headway coefficient of variation results in 28 s more delay per train on average. Stability analysis showed how much a primary delay can affect operations of several following trains. Impacts of service frequency on line performance were also explored.

While the model cannot replace detailed micro-simulation models, it can efficiently address the drawbacks of macro-scale analytical models and complex discrete algebraic models. The model is general, requiring only a limited amount of historical train tracking data. In this paper, we implement the model using the OCS data from MBTA. Implementing the model for other transit systems may require adjustments depending on how the track circuit data is recorded. Further research should implement the model for other urban rail transit systems to evaluate the performance of the model for similar systems. A detailed analysis of train delay functions based on other rail transit systems would be an interesting future research direction. While this study focused on only homogenous train types, the model can be extended for heterogenous train types. Multiple train travel time functions should be constructed based on the sequence of train types and the right travel time function should be used for constructing each train trajectory. Adopting the model for moving block signaling systems would be another key future research direction. The insights from car following models can be a relevant and useful resource to develop a TFM for moving block signaling systems. A dwell time model that takes on-board passengers and denied boarding into account can be employed for rail transit systems that commonly deal with denied boarding passengers due to capacity constraints. While not within the scope of this study, issues such as delay propagation to the opposite direction can be captured by the proposed model. Another interesting future research direction is extending the model to allow real-time actions such as expressing or short-turning of trains in response to service disruptions.

CRediT authorship contribution statement

Saeid Saidi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Haris N. Koutsopoulos:** Supervision, Conceptualization, Methodology, Funding acquisition, Investigation, Writing – review & editing. **Nigel H.M. Wilson:** Supervision, Conceptualization, Methodology, Funding acquisition, Investigation, Resources, Writing – review & editing. **Jinhua Zhao:** Supervision, Conceptualization, Funding acquisition, Investigation, Resources, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) PDF-502602-2017; NSERC Discovery Grant RGPIN-2020-03944 on Urban Transit System Diagnosis, Monitoring, and Management Using Mobility Sensing Data; and the Massachusetts Bay Transportation Authority (MBTA). Authors would also like to thank John Attanucci and Fred. P. Salvucci for their contribution and advise on this project.

References

Bešinović, N., Goverde, R.M., 2019. Stable and robust train routing in station areas with balanced infrastructure capacity occupation. Public Transport 11 (2), 211–236.

Bešinović, N., Goverde, R., Quaglietta, E., Roberti, R., 2016. An integrated micro–macro approach to robust railway timetabling. Transp. Res. B Methodol. 87, 14–32.

Braker, J. G., 1993. Algorithms and applications in timed discrete event systems. Delft University of Technology.

Büker, T., Seybold, B., 2012. Stochastic modelling of delay propagation in large networks. J. Rail Transp. Plann. Manage. 2 (1–2), 34–50.

Carey, M., Kwieciński, A., 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. Transp. Res. B Methodol. 28 (4), 251–267.

Corman, F., Kecman, P., 2018. Stochastic prediction of train delays in real-time using Bayesian networks. Transportation Research Part C: Emerging Technologies 95, 599–615.

Corman, F., Trivella, A., Keyvan-Ekbatani, M., 2021. Stochastic process in railway traffic flow: models, methods and implications. Transportation Research Part C: Emerging Technologies 128, 103167.

Farhi, N., Van Phu, C.N., Haj-Salem, H., Lebacque, J.P., 2017. Traffic modeling and real-time control for metro lines. Part I-A Max-plus algebra model explaining the traffic phases of the train dynamics. In 2017 American Control Conference, IEEE. pp. 3834-3839.

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H., Attanucci, J.P., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Rec. 2343 (1), 17–24.

Goverde, R., 2007. Railway timetable stability analysis using max-plus system theory. Transp. Res. B Methodol. 41 (2), 179–201.

Goverde, R.M., Hansen, I.A., 2013. Performance indicators for railway timetables. In *2013 IEEE International Conference on intelligent rail transportation proceedings.* 301-306.

Goverde, R.M., Corman, F., D'Ariano, A., 2013. Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions. J. Rail Transp. Plann. Manage. 3 (3), 78–94.

Goverde, R.M., Bešinović, N., Binder, A., Cacchiani, V., Quaglietta, E., Roberti, R., Toth, P., 2016. A three-level framework for performance-based railway timetabling. Transportation Research Part C: Emerging Technologies 67, 62–83.

Hansen, A., Pachl, J., 2014. Railway Timetabling & Operations. Eurailpress, Hamburg.

Heimburger, D.E., Herzenberg, A.Y., Wilson, N.H., 1999. Using simple simulation models in operational analysis of rail transit lines: Case study of Boston's Red Line. Transp. Res. Rec. 1677 (1), 21–29.

Huang, P., Lessan, J., Wen, C., Peng, Q., Fu, L., Li, L., Xu, X., 2020. A Bayesian network model to predict the effects of interruptions on train operations. Transportation Research Part C: Emerging Technologies 114, 338–358.

International Union of Railways (UIC), 2013. Capacity (UIC Code 406) 2nd edition. Technical report.

Jensen, L.W., Landex, A., Nielsen, O.A., Kroon, L.G., Schmidt, M., 2017. Strategic assessment of capacity consumption in railway networks: Framework and model. Transportation Research Part C: Emerging Technologies 74, 126–149.

Ketphat, N., Whiteing, A., Liu, R., 2022. State movement for controlling trains operating under the virtual coupling system. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 236 (2), 172–182.

Kittelson & Assoc, Inc., Parsons Brinckerhoff, Inc., KFH Group, Inc., Texas A&M Transportation Institute, Arup, 2013. Transit Capacity and Quality of Service Manual. Third Edition. Transit Cooperative Highway Research Program (TCRP) Report 165, published by Transportation Research Board, Washington.

Koutsopoulos, H.N., Wang, Z., 2007. Simulation of Urban Rail Operations. Transportation Research Record: Journal of the Transportation Research Board 2006 (1), 84–91.

Landex, A., 2008. Methods to estimate railway capacity and passenger delays. Technical University of Denmark.

Lessan, J., Fu, L., Wen, C., 2019. A hybrid Bayesian network model for predicting delays in train operations. Comput. Ind. Eng. 127, 1214–1222.

Li, K., Gao, Z., 2013. An improved car-following model for railway traffic. J. Adv. Transp. 47 (4), 475–482.

Liu, Y., Zhou, Y., Su, S., Xun, J., Tang, T., 2021. An analytical optimal control approach for virtually coupled high-speed trains with local and string stability. Transportation Research Part C: Emerging Technologies 125, 102886.

Martin, U., 2014. Performance Evaluation. in *Railway timetabling & operations*, I. A. Hansen and J. Pachl, Eurailpress, Hamburg.

Nash, A., Huerlimann, D., 2004. Railroad simulation using OpenTrack. WIT Trans. Built Environ. 74.

Pachl, J., 2002. Railway operation and control. Mountlake Terrace, USA: VTD Rail Publishing.

Quaglietta, E., 2014. A simulation-based approach for the optimal design of signalling block layout in railway networks. Simul. Model. Pract. Theory 46, 4–24.

Schanzenbächer, F., Farhi, N., Leurent, F., Gabriel, G., 2020. Feedback control for metro lines with a junction. IEEE Trans. Intell. Transp. Syst. 22 (5), 2741–2750.

Schlechte, T., Borndörfer, R., Erol, B., Graffagnino, T., Swarat, E., 2011. Micro–macro transformation of railway networks. J. Rail Transp. Plann. Manage. 1 (1), 38–48.

Weik, N., Nießen, N., 2020. Quantifying the effects of running time variability on the capacity of rail corridors. J. Rail Transp. Plann. Manage. 15, 100203.

Wendler, E., 2007. The scheduled waiting time on railway lines. Transp. Res. B Methodol. 41 (2), 148–158.

Wolofsky, G., Saidi, S., Attanucci, J., Salvucci, F.P., 2019. Modelling Rail Transit Dwell Time Using Automatically Collected Passenger Data. *Transportation Research Board 98th Annual Meeting Transportation Research Board.* No. 19-03817.

Wolofsky. G. T., 2019. Towards 3-Minutes: Application of Holding and Crew Interventions to Improve Service Regularity on a High Frequency Rail Transit Line. Massachusetts Institute of Technology, Cambridge, MA.

Xun, J., Ning, B., Li, K.P., Zhang, W.B., 2013. The impact of end-to-end communication delay on railway traffic flow using cellular automata model. Transportation Research Part C: Emerging Technologies 35, 127–140.

Zhou, J., Koutsopoulos, H.N., Saidi, S., 2020. Evaluation of Subway Bottleneck Mitigation Strategies using Microscopic, Agent-Based Simulation. Transportation Research Record: Journal of the Transportation Research Board 2674 (5), 649–661.

Zhou, J., 2021. Urban Rail Simulation and Applications in Service Planning and Operations. Northeastern University.