An empirical study on fundamental diagram of urban rail transit: The case of Boston's subway data

Jiahua ZHANG[1], Kentaro WADA[2] and Takashi OGUCHI[3]

[1]Non-member of JSCE, Graduate Student, Dept. of Civil Eng., the University of Tokyo
(4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan)
E-mail:zhangjh@iis.u-tokyo.ac.jp
[2]Member of JSCE, Associate Professor, Faculty of Engineering, Information and Systems, University of Tsukuba
(1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan)
E-mail:wadaken@sk.tsukuba.ac.jp
[3]Fellow of JSCE, Professor, Inst., Industrial Science, the University of Tokyo
(4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan)
E-mail:takog@iis.u-tokyo.ac.jp

Seo et al.[1)] proposed a fundamental diagram (FD) of urban rail transit to describe the interaction between passenger demand and train operation, in a simple manner. This paper investigates their proposed FD and its variants by using empirical data from the Boston subway system. Specifically, three FD models, which are based on different assumptions of the train dwelling time, are calibrated and evaluated using subway operation and passenger arrival data. The results show that the free-flow regime of the FD models can explain the empirical data well. In addition, the dwelling time monotonically increasing with the number of boarding passengers' might be enough to describe passengers' influence on transit operation.

Key Words : *Rail transit, passenger congestion, traffic flow theory, fundamental diagram*

1. Introduction

Urban rail transit generally serves as the primary solution for commuters' travel demand in metropolises, owing to its high capacity and punctuality[2)]. However, passengers typically suffer from severe congestion and frequent delays, especially during the rush hours. For example, during the morning rush hours in the Tokyo metropolitan area, more than 40 rail transit lines observed congestion rates over 150% (the number of passengers divided by designed car capacity > 150%) and 29 lines reported delays of more than 5 minutes, occurring on over 50% of the weekdays in a month[3), 4)]. Moreover, studies have shown that 96 of 311 metro stations adopted ordinary entry restriction in Beijing, China[5)] and only 58.1% of weekday trains arrived on time in New York City[6)]. To relieve congestion during rush hours and to prevent the occurrence of delays, engineering efforts have been attempted, such as physically increasing capacity, or enhancing the reliability of equipment. These approaches are undoubtedly effective, but limitations widely exist. On the other hand, severe congestion and most delays appearing in urban rail transit are essentially caused by surging demand and the improper behavior of passengers[3)]. Therefore, it is crucially important to understand congestion and the delay mechanisms that are caused by passenger influence.

In general, congestion and delay can easily develop into a vicious circle during the rush hours. This effect is comprehensively reviewed by Tirachini et al.[7)]. On the subject of rail transit, due to growing demand during rush hours, more passengers accumulate on the platform. Then, the dwelling time of trains is extended because of both in-vehicle and on-platform congestion. Next, the longer dwelling time leads to a delay in the following trains (also known as "knock-on delay", see Carey and Kwieciński[8)], especially in a high-frequency rail transit system. Once the delay occurs, it propagates such that travel time and headways increase, which finally causes further accumulation of passengers on the platform.

To describe the interaction between passenger demand and train operation in a high-frequency rail transit system in a simple manner, Seo et al.[1)] proposed a fundamental diagram (FD) of urban railway transit that expresses the train flow as a function of the train density and passenger arrival flow. The FD is analytically derived from the basic operation principles of trains. Seo et al.[1)] also discussed the applicability of the FD to a macroscopic simulation of rail transit operation dynamics through comparison with a microscopic simulation model.

This study aims to investigate their proposed FD and its variants using empirical data. Specifically, the three FD models, which are based on different assumptions of the train dwelling time, are calibrated and evaluated using Boston subway operation and passenger arrival data, provided by the Massachusetts Bay Transportation Authority (MBTA). The results show that the free-flow regime of the FD models can explain the empirical data well. In addition, the dwelling time monotonically increasing with the number of boarding passengers' might be enough to describe passengers' influence on transit operation.

This paper is structured as follows: the second section introduces the Boston subway data and depicts the relation among the train flow, density, and passenger arrival flow, based on the extracted data. The third section formulates the three FD models and provides numerical examples for each FD. Finally, the fourth section calibrates and evaluates the models using the empirical data, and a brief discussion on future work is provided.

2. Boston Redline Operation Data

To properly investigate passenger influence on railway operation, the data should include both the movements of trains and the arrival of passengers. Fortunately, the 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[9]. Here we choose the busiest section of the Redline (from Alewife to JFK/Umass with 13 stations) as the analysis target. The flow and density of the railway system are calculated by employing Edie's definition[10] of traffic flow as shown in Eq. (1).

$$q(A) = \frac{\sum_{n \in N} d_n}{|A|}, \; k(A) = \frac{\sum_{n \in N} \tau_n}{|A|} \tag{1}$$

where A is the measurement time-space area and $|A| = L \times \Delta t$, and 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. 1** 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.

To obtain relatively steady state data, we filter out the unsteady data by judging the adjacent train flow change over 20%. Finally, the FD of the Boston Redline is depicted in **Fig. 2**. The color used represents the value of the passenger arrival rate, as illustrated in the color bar (pax/h).

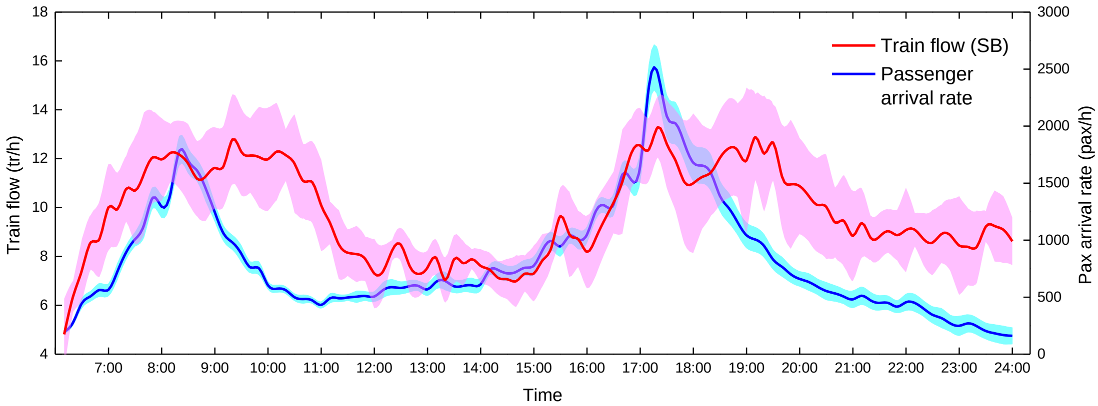


Fig. 1 Train and passenger flow transition during one day

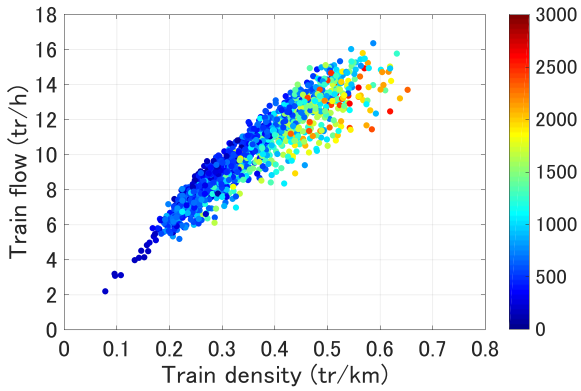


Fig. 2 Weekday FD of the Boston Redline (southbound)

3. Train Fundamental Diagram

In this section, we present three different train FD models. One is proposed by Seo et al.[1] and the others are slightly modified versions of the former, based on different train dwelling time assumptions.

(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 dwelling time t_b is determined by the number of boarding passengers N_p on the platform. Here, three assumptions for dwelling time are considered:

(a) t_b keeps constant at t_{bcon} regardless of N_p, which indicates that passenger congestion does not affect railway operation;
(b) t_b monotonically increases with N_p from a minimum value t_{b0} (buffer time), which is the same assumption made in Seo et al.[1];
(c) 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.[11].

We also assume that all of the waiting passengers can always board the next approaching train, which means $N_p = a_p \cdot H$, where a_p is the passenger arrival rate at stations and H is the time headway of successive trains. Now, the three assumptions of t_b can be expressed as Eq. (2) - (4), respectively. γ here can be interpreted as the dwelling time growth rate with increase in the number of boarding passenger.

$$t_b = t_{bcon} \tag{2}$$

$$t_b = t_{b0} + \gamma a_p H \tag{3}$$

$$t_b = \begin{cases} t_{b0} & \text{if } a_p H \le N_0 \\ t_{b0} + \gamma\left(a_p H - N_0\right) & \text{if } a_p H > N_0 \end{cases} \tag{4}$$

On the other hand, the cruising behavior of a train is modeled using Newell's simplified car-following model[12]. More specifically, the position $x_m(t)$ of train m at time t is expressed as:

$$x_m(t) = \min\left\{x_m(t-\tau) + v_f \tau, x_{m-1}(t-\tau) - \delta\right\} \tag{5}$$

Where m-1 indicates the preceding train, τ is the minimum time headway of successive 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 (a), (b) and (c), as model A, B, and C, respectively.

(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}, γ, a_p, N_0, τ, δ are time-independent
- headway H and desired cruising speed v_f are also time-independent

Also, for simplicity, we assume a homogeneous railway system, 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 as l

Now, the train FD as expressed in Eq. (6) can be separately derived in free-flow and congested regime by combining the above-mentioned assumptions.

$$q = Q(k, a_p) = k\bar{v}, \tag{6}$$

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. (7).

$$\bar{v} = \frac{l}{t_b + l/v}, \tag{7}$$

where v is the average cruising speed of a train. In the free-flow regime, $v = v_f$ so that the explicit expression of q for model A, B and C can be easily derived by substituting Eq. (2) - (4) and Eq. (7) into Eq. (6).

In the congested regime, the headway H should satisfy:

$$H \geq t_b + \frac{\delta + v\tau}{v}. \tag{8}$$

By taking the equal boundary condition of Eq. (8) and employing $q = 1/H$, Eq. (2) - (4) can be substituted into Eq. (8) so that q can be described as a function of v and a_p:

$$q = f_1(v, a_p). \tag{9}$$

Then, by inserting Eq. (7) and Eq. (9) into Eq. (6), we can also obtain k as a function of v and a_p:

$$k = f_2(v, a_p). \tag{10}$$

By using Eq. (9) and Eq. (10), the slope of the 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^* :

$$q^* = \frac{1}{t_b + \delta/v_f + \tau}, \tag{11}$$

as a boundary condition, the FDs of model A, B and C can be formulated in Eq. (12) (see also, Seo et al[1], for the details of the derivation of the FD).

When $a_p/N_0 < q_1^{*c}$,

$$Q(k, a_p) = \begin{cases} \dfrac{kl - \gamma a_p}{t_{b0} + l/v_f - \gamma N_0} & \text{if} \quad \gamma a_p/l \leq k < k_1^c \\ \dfrac{kl}{t_{b0} + l/v_f} & \text{if} \quad k_1^c \leq k < k_1^{*c} \end{cases} \tag{12a}$$

$$Q(k, a_p) = \begin{cases} -\dfrac{\delta l}{(l-\delta)t_{b0} + \tau l}\left(k - k_1^{*c}\right) + q_1^{*c} & \text{if} \quad k_1^{*c} \leq k < k_2^c \\ -\dfrac{\delta l}{(l-\delta)(t_{b0} - \gamma N_0) + \tau l}\left(k - k_2^{*c}\right) + q_2^{*c} & \text{if} \quad k \geq k_2^c \end{cases} \tag{12b}$$

When $a_p/N_0 \geq q_1^{*c}$

$$Q(k, a_p) = \begin{cases} \dfrac{kl - \gamma a_p}{t_{b0} + l/v_f - \gamma N_0} & \text{if} \quad \gamma a_p/l \leq k < k_2^{*c} \\ -\dfrac{\delta l}{(l-\delta)(t_{b0} - \gamma N_0) + \tau l}\left(k - k_2^{*c}\right) + q_2^{*c} & \text{if} \quad k \geq k_2^{*c} \end{cases} \tag{12c}$$

Where

$$q_1^{*c} = \frac{1}{t_{b0} + \delta/v_f + \tau}, \quad k_1^{*c} = \frac{t_{b0} + l/v_f}{\left(t_{b0} + \delta/v_f + \tau\right)l}$$

$$q_2^{*c} = \frac{1 - \gamma a_p}{t_{b0} + \delta/v_f + \tau - \gamma N_0}, \quad k_2^{*c} = \frac{\left(1 - \gamma a_p\right)\left(t_{b0} + l/v_f - \gamma N_0\right)}{\left(t_{b0} + \delta/v_f + \tau - \gamma N_0\right)l} + \frac{\gamma a_p}{l}$$

$$k_1^c = \frac{a_p}{N_0 l}\left(t_{b0} + l/v_f\right), \quad k_2^c = \left(q_1^{*c} + \frac{\delta l k_1^{*c}}{(l-\delta)t_{b0} + \tau l} - \frac{a_p}{N_0}\right)\cdot\frac{(l-\delta)t_{b0} + \tau l}{\delta l}$$

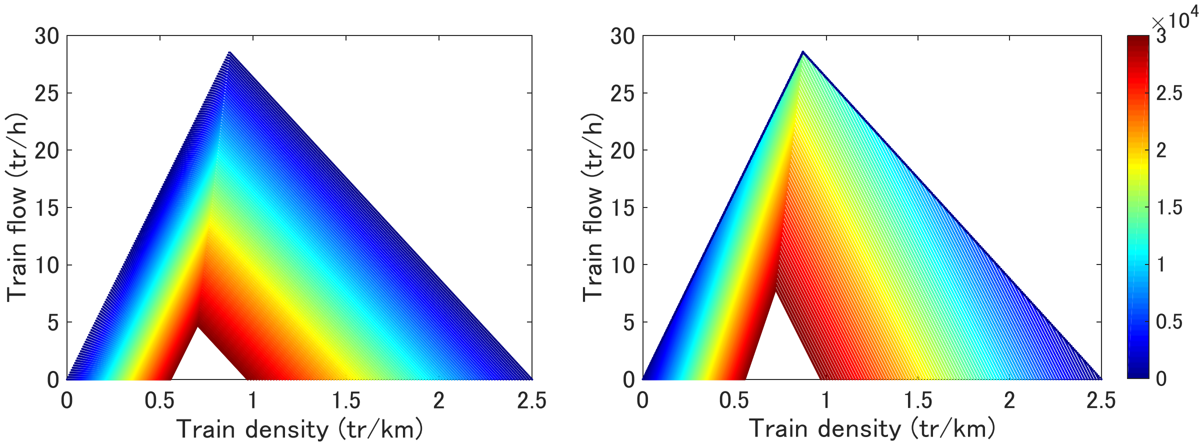
Model A is expressed by Eq. (12c) by taking $\gamma = 0$, $N_0 = 0$ and $t_{b0} = t_{bcon}$. Model B is also expressed by Eq. (12c) by taking $N_0 = 0$. For model C, the equation has to be separately written, depending on the relationship between a_p/N_0 and q_1^{*c}.

Eqs. (12a) and (12b) actually describe the situation when a_p is not large enough to force a condition in which the dwelling time is always larger than t_{b0}. More specifically, when $k_1^c \leq k \leq k_2^c$, the dwelling time $t_b = t_{b0}$, which implies that operation under this condition can guarantee the dwelling time is not extended due to passenger influence. While out of this range, dwelling time would be extended either because trains in operation are insufficient or abundant (train bunching). On the contrary, Eq. (12c) describes the situation when a_p is large enough so that dwelling time is always larger than t_{b0}.

For a better understanding of passenger influence on train flow, we present two numerical examples of the FDs for model B and C, as shown in **Fig. 3**, based on the parameters in **Table 1**. From the comparison of **Fig. 3(a)** and **Fig. 3(b)**, it can be observed that under the same a_p, model C can achieve higher train flow due to a relatively short dwelling time.

Table 1 Parameters used in the numerical example

Parameter	Value
t_{b0}, N_0, γ	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



(a) (b)

Fig.3 Numerical examples of the FDs, (a) model B, and (b) model C

4. Model Calibration and Evaluation

In this section, we calibrate and evaluate the proposed models employing data from the Boston Redline.

(1) Calibration based on enumeration method

The three models are calibrated using the enumeration method. Specifically, we begin by preselecting the variation range for each parameter, and we then build a parameter set by accounting for all possible combinations. The parameter set is then sorted based on root mean square error (RMSE), which is calculated by Eq. (13):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_i^m - q_i^e\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_i^m - Q\left(k_i^m, a_{p,i}^m\right)\right)^2} \tag{13}$$

where q_i^m, k_i^m, $a_{p,i}^m$ are, respectively, train flow, density, and passenger arrival rate, measured from the Boston Redline FD (southbound), while q_i^e is the train flow estimated from the three models (calculated by Eq. (12)).

Finally, the parameter vector with the smallest RMSE will be regarded as the best fitting one, denoted by $\hat{\boldsymbol{\beta}}$. Since we found that almost all of the datapoints from the Boston Redline lay in the free-flow regime of the FD (as shown in **Fig. 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}, γ, v_f), and four (t_{b0}, N_0, γ, 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**

2. Δ is the increment.

Table 2 Variation range of the calibrated parameters

Model A: $116 \times 71 = 8236$	**Model B:** $56 \times 100 \times 71 = 397{,}600$	**Model C:** $56 \times 50 \times 100 \times 71 = 19{,}880{,}000$
t_{bcon}: $[5,\ 120,\ \Delta = 1]\ s$	t_{b0}: $[5,\ 60,\ \Delta = 1]\ s$	t_{b0}: $[5,\ 60,\ \Delta = 1]\ s$
v_f: $[10,\ 80,\ \Delta = 1]\ km/h$	γ: $[0.01,\ 1,\ \Delta = 0.01]\ s/pax$	N_0: $[5,\ 250,\ \Delta = 5]\ pax$
	v_f: $[10,\ 80,\ \Delta = 1]\ km/h$	γ: $[0.01,\ 1,\ \Delta = 0.01]\ s/pax$
		v_f: $[10,\ 80,\ \Delta = 1]\ km/h$

As can be seen from **Table 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**. From **Table 3** we can see that model A has a larger RMSE than both model B and model C. Also, in model A, the values of $\hat{t}_{bcon}$ and $\hat{v}_f$ 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 $\hat{\gamma}$, and similar values of $\hat{v}_f$ and 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 Calibration results

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

(2) Sensitivity analysis and Evaluation using the Akaike information criterion (AIC)

To show the sensitivity of the parameters, we draw the contour maps of the RMSE with respect to the variation of parameters for the three models in **Fig. 4 - Fig. 6**. On the contour maps, color is used to represent the value of the RMSE.

In **Fig. 4**, low RMSE values appear as a blue strip. Because t_{bcon} and l/v_f in Eq. (12c) 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 relationship between γ - t_{b0} and $l/v_f\ -\ \gamma$ minimize in the enclosed area (blue ellipses in **Fig. 5(b) - (c)**, **Fig. 6(e) - (f)**), which yields realistic estimates for the parameters. Here, other parameters in **Fig. 5** and **Fig. 6** take the corresponding values of $\hat{\boldsymbol{\beta}}$ in **Table 3**. From **Fig. 6(b)**, we can conclude that the RMSE is not sensitive to the change of N_0, since l/v_f dominates the denominator of Eq. (12a) when $t_{b0} = \hat{t}_{b0}$ and $\gamma = \hat{\gamma}$. In addition, the RMSE in the blank area of **Fig. 6(c)** is not available because the product of γ and N_0 approaches $\hat{t}_{b0} + l/\hat{v}_f$, causing the denominator of Eq. (12a) to become zero.

To compare the performance of the three models, we adopt the AIC 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. (14):

$$AIC = -2\ln L + 2p \tag{14}$$

where L 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, L can be derived by Eq. (15) - (17).

$$L(\hat{\boldsymbol{\beta}}) = \prod_{i=1}^{n} f\left(q_i^m \middle| \hat{\boldsymbol{\beta}}\right) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2\right) \tag{15}$$

$$\hat{q}_i^e = Q\left(k_i^m, a_{p,i}^m \middle| \hat{\boldsymbol{\beta}}\right) \tag{16}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2 \tag{17}$$

Finally, submitting Eq. (15) - (17) into Eq. (14), the AIC can be obtained by Eq. (18). The calculation results for the AIC of the three models are listed in **Table 4**.

$$AIC = nln(2\pi) + nln\left(\sum_{i=1}^{n}(q_i^m - \hat{q}_i^e)^2 / n\right) + n + 2p \tag{18}$$

Table 4 Calculation results

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

The calculation results show that model B produces the smallest AIC, although it is quite similar to model C. The AICs of both models are much smaller than that of model A. Therefore, it can be concluded that models that conider the influence of passenger congestion on railway operation perform much better.

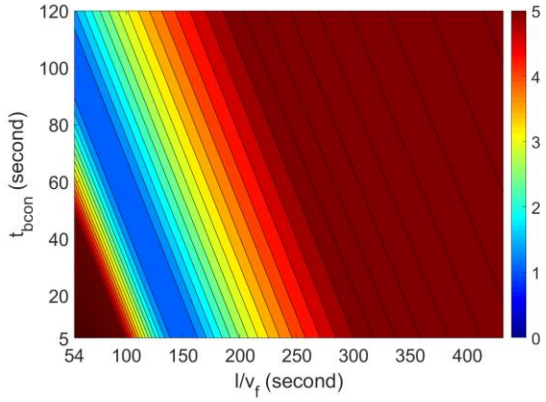


Fig. 4 Parameter sensitivity of model A

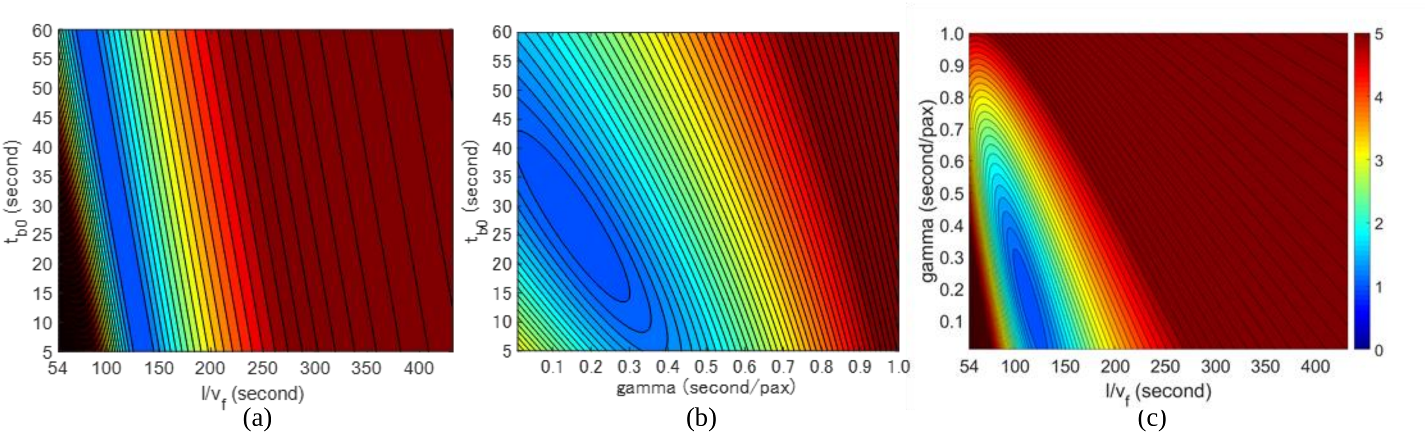


Fig. 5 Parameter sensitivity of model B

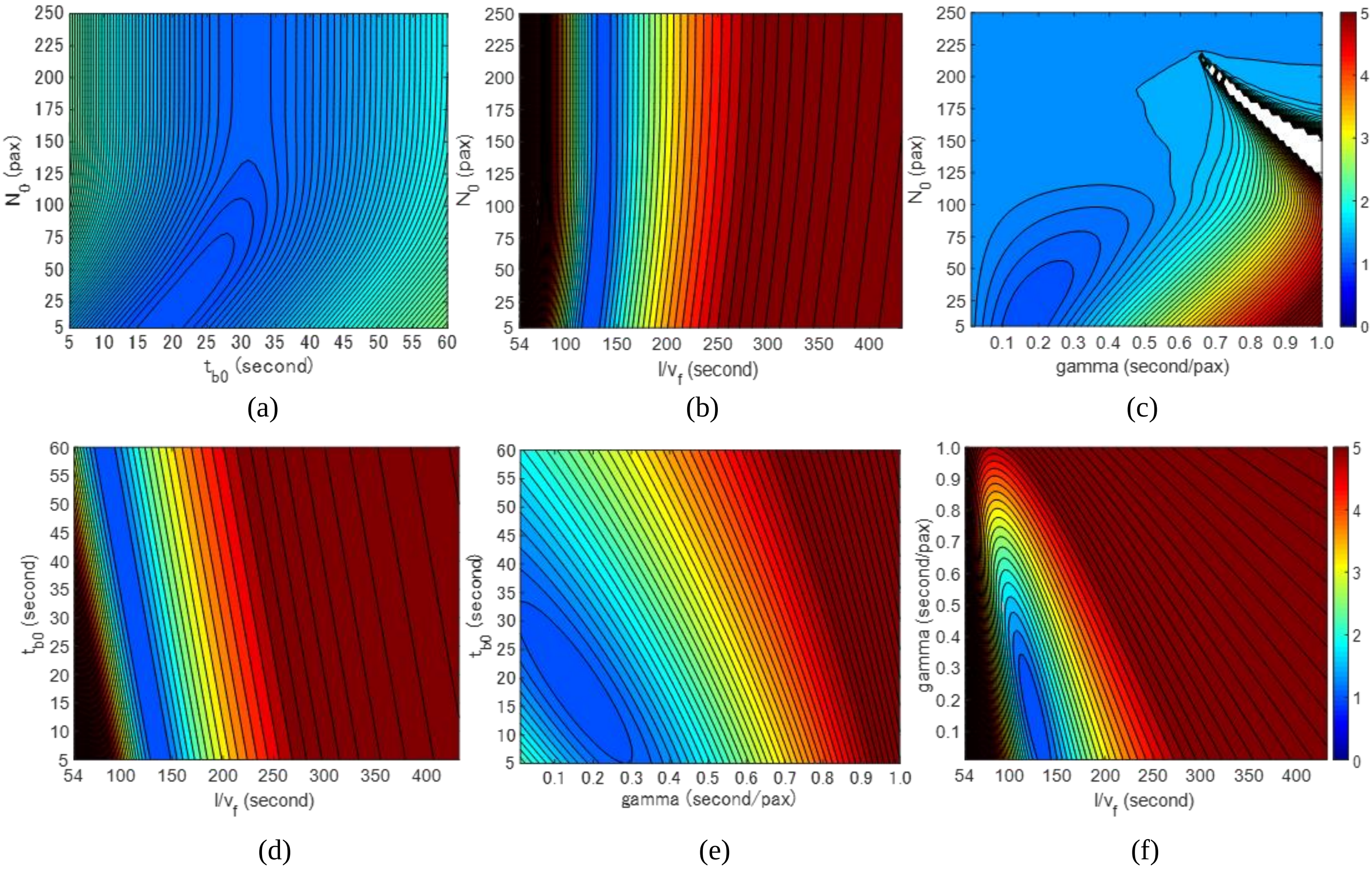


Fig. 6 Parameter sensitivity of model C

5. Discussion

This study investigated three models for train FDs, which are based on different assumptions of train dwelling time, by employing operation data from the Boston Redline. The conclusion indicates that passenger congestion influence on urban rail transit system operation was significant, and that train dwelling 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.

REFERENCES

1) Seo, T., Wada, K., and Fukuda, D. (2017). A macroscopic and dynamic model of urban rail transit with delay and congestion. *In Transportation Research Board 96th Annual Meeting*.
2) Vuchic, V. R. (2017). Urban transit: operations, planning, and economics. John Wiley & Sons.
3) Ministry of Land, Infrastructure, Transport and Tourism. (2017a). Railway related data: congestion rate. Retrieved from http://www.mlit.go.jp/common/001245932.pdf.
4) Ministry of Land, Infrastructure, Transport and Tourism. (2017b). Visualize delay in railway system. Retrieved from http://www.mlit.go.jp/common/001215328.pdf.
5) Sohu News. (2018, Jan. 8). Another 23 metro stations will adopt ordinary admission restriction (in Chinese). Retrieved

from https://www.sohu.com/a/215355638_780771.
6) Hu, W. (2018, Mar. 19). New York subway's on-time performance hits new low. Retrieved from https://www.nytimes.com/2018/03/19/nyregion/new-york-subways-on-time-performance-hits-new-low.html
7) 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, pp. 36-52.
8) 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), pp. 251-267.
9) Barry, M., and Card. B. (2014). Visualizing MBTA Data. [online] Github. Available at: http://mbtaviz.github.io/ [Accessed 7 Jan. 2019].
10) Edie, L. C. (1963). Discussion of traffic stream measurements and definitions. Port of New York Authority.
11) 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), pp. 92-98.
12) Newell, G. F. (2002). A simplified car-following theory: a lower order model. *Transportation Research Part B: Methodological*, 36(3), pp. 195-205.