sustainability

MDPI

Article

Service-Oriented Load Balancing Approach to Alleviating Peak-Hour Congestion in a Metro Network Based on Multi-Path Accessibility

Zhiyuan Huang [1], Ruihua Xu [1,*], Wei (David) Fan [1,2], Feng Zhou [1,*] and Wei Liu [3]

[1] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, College of Transportation Engineering, Tongji University, Shanghai 201804, China; 9huangzhiyuan@tongji.edu.cn

[2] USDOT Center for Advanced Multimodal Mobility Solutions and Education, Department of Civil and Environmental Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223, USA; wfan7@uncc.edu

[3] Technology Center of Shanghai Shentong Metro Group Co. Ltd., Shanghai 201103, China; lw26star@sina.com

* Correspondence: rhxu@tongji.edu.cn (R.X.); zhoufeng24@tongji.edu.cn (F.Z.); Tel.: +86-021-69585730 (R.X.)

Received: 13 January 2019; Accepted: 26 February 2019; Published: 1 March 2019

Abstract: To further improve the service quality and reduce safety risks in current congested metro systems during peak hours, this paper presents a load balancing (LB) approach so that available capacity can be utilized more effectively in order to alleviate peak hour congestion. A set of under-utilized yet effective alternative routes were searched using a deletion algorithm (DA) in order to share the passenger loads on overcrowded metro line segments. An optimization model was constructed based on an improved route generalized time utility function considering the penalties of both in-vehicle congestion and transfers. A detailed load balancing solution was generated based on the proposed algorithm. A real-world example of three overloaded metro line segments in the Shanghai metro network were selected and used to verify the feasibility and validity of the developed load balancing method. The results show that the load balancing method can effectively reduce the overcrowding situation to a great extent. Finally, two prospective inducing schemes are discussed to help implement the load balancing solution in the actual metro system in an efficient and effective manner.

Keywords: metro system; peak hour congestion; overloaded segment; effective alternative routes; load balancing (LB)

1. Introduction

The metro system has been built as the backbone of urban public transportation in many large and medium-sized cities around the world due to its advantages of speed, capacity, punctuality, safety, lower-emission, etc. This is particularly the case in China, where there were 31 metro systems running in Mainland China with a total combined length of 3881.8 kilometers at the end of 2017. Today, China operates the world's longest, second, and fourth longest metro systems in its cities. In addition, four of the top ten busiest metro systems in the world are in China. In those busiest systems, many metro planning and operational problems arise accordingly. The most remarkable one is perhaps the significant recurrent congestion that many travelers have to experience during peak hours on a daily basis.

Congestion not only results in significant time delays and impose high safety risks, but also makes passengers perceive some other types of disutility which include, but are not limited to seat unavailability, undesirable but frequent physical contact with people, inability to read or use a smart phones when necessary, and/or legitimate concerns about sexual harassment [1]. Peak-period extreme crowding occurs on some metro line segments and at some stations when travel demands exceed the existing metro system capacity. For commuters, they prefer to use the route(s) with the shortest time and/or involving the least number of transfers. It is common that certain segments of a given transportation facility experience extreme congestion conditions during certain hours of the day while on other segments and during other hours the same facility may have very low rates of utilization [2]. Under extreme conditions, however, even if the headways of some lines are shortened to be 2 minutes during rush hours, the physical capacity may still not be able to meet the great peak hour travel demands correspondingly. As such, in the encouragement of "reasonable supply and controlled demand", active demand management (ADM) has been attracting more and more attention among many transportation researchers and practitioners. Along that line, the crowding could be effectively alleviated by managing passenger flow distribution in the network. In a complex network like the Shanghai metro system, trips between each of these large number of origin-destination (OD) pairs can be made by using two or more alternative routes. There may be a feasible way to divert passengers from the overcrowded metro line segments to other less heavily utilized routes. In this paper, the overcrowded line segment is defined as an *overloaded segment*, which can be quantified using a variable: *mean occupancy* (i.e., the ratio of the passenger loads to train capacity in peak period). By definition the rate can vary between 0 (trains travelling empty) and 1 (trains travelling fully loaded during the period). An overloaded segment can be declared when the occupancy of trains along it reaches a threshold of 85%. Note that, the threshold can be different from city to city. It is 70% on the London Underground while on the Santiago Metro the value is 85%, according to Reference [3,4]. Here, 85% is selected and deemed suitable for use in the metro systems in China. As one can imagine, when the occupancy is greater than or equal to 85%, the level of comfort can be greatly reduced due to the high passenger density. In addition, there is a chance that users may not even be able to board the first train which they would like to and have to wait for the next train. In this paper, it was assumed that the OD pairs and relevant ridership (that fundamentally causes the presence of overloaded segments and their corresponding congestion during the peak period) are known. The purpose of this research is to develop an efficient and effective method to find a set of under-utilized alternative routes for use so that the heavy passenger loads on those overloaded segments can be shared and balanced. As a by-product, the ridership being diverted will be made known as well.

The rest of this paper is organized as follows. In Section 2, previous studies on the methods to manage peak hour congestion are reviewed. Section 3 describes the problem to be solved in this paper. Section 4 constructs an optimization model based on an improved route generalized time utility function and develops a load balancing algorithm to calculate the OD ridership for diverting. In Section 5, the Shanghai metro system is used as a real case study to test the methodologies presented in this paper. In Section 6, some feasible inducing schemes that can be applied to actual active demand management in metro systems are discussed. Finally, conclusions are made and some future research directions are also presented in Section 7.

2. Literature Review

Restricted by the ability of infrastructure, equipment, and right-of-way, train capacity in metro system is reaching its limits. More and more transportation researchers and decision makers have begun to consider the optimal control of passenger flows from the "demand side" to reduce the peak hour congestion. In general, there are two strategies that can be used to manage the peak hour congestion: inflow control at stations and route guidance. Inflow control at stations is one of the most effective control means to ensure operational safety at stations. The theory of inflow control has been developed from inflow control at a single station, combined inflow control at multiple stations, to coordinated inflow control in the network. Zhao et al. [5] and Yao et al. [6] established a coordinated passenger inflow control model at the network level based on mathematical programming to minimize the number of delayed passengers and to provide quantitative basis for selecting control stations and determining control time and inflow rates. To reduce overloaded passengers on platforms and improve the service quality, Li et al. [7] and Shi et al. [8] integrated an optimal train timetable with passenger flow control strategy on an oversaturated metro line. Xu et al. [9] studied the passenger flow organization problem at subway stations under uncertain demand and developed a unified simulation-based algorithm to solve it. Jiang et al. [10] developed a new reinforcement learning-based method to optimize the inflow volume during a certain period of time at crowded metro stations. In principle, the studies mentioned above aimed to relieve the passenger congestion at stations. However, through our field investigation, inflow control at metro stations was not effective in mitigating the overcrowding situations in packed trains. In order to relieve the in-vehicle congestion, route guidance for passengers is a potentially very effective measure that can be used based on dynamic path planning and effective information release to induce passengers to choose a more reasonable route. Relevant research efforts are often made to reassign the passenger flows based on the degree of congestion on such efficient routes. For example, Si et al. [11] presented an urban transit assignment model based on augmented network with both the in-vehicle congestion and transfer congestion and developed an improved shortest path-based algorithm to solve it. Zhu et al. [12] presented a modified stochastic user-equilibrium assignment algorithm to calculate the passenger flow distribution for network operations. Though the impacts of in-vehicle congestion on passengers began to be considered, how to mitigate in-vehicle congestion is not the focus of these studies. Abadi et al. [13] developed a coordinated multimodal dynamic freight load balancing (MDFLB) system to balance freight loads across both the rail and road networks. The research idea is very important since load balancing improves the distribution of loads in the network and aims to optimize the use of resource and avoid overload in partial network.

In summary, to effectively manage the increasingly high travel demands in peak hours, passenger flow control methods are becoming more and more important and valued both in theory and practice. Although the negative effects in metro systems caused by in-vehicle congestion seems as serious as crowding on platforms, research on in-vehicle congestion-mitigation are not perfect and systematic. Based on recent studies, the main purpose of this paper is how to alleviate peak-hour in-vehicle congestion in metro systems in a proactive, targeted, and humanized way.

3. Problem Description

A metro system is an extremely complicated double-load network that is composed of three layers: static physical topology with stations and connections, dynamic network with running trains, and trip network with passengers (see Figure 1). Due to the drastic increase of travel demands, the current passenger flow loading in the network is becoming more and more unbalanced, particularly during peak hours, which can result in several adverse effects as mentioned above.

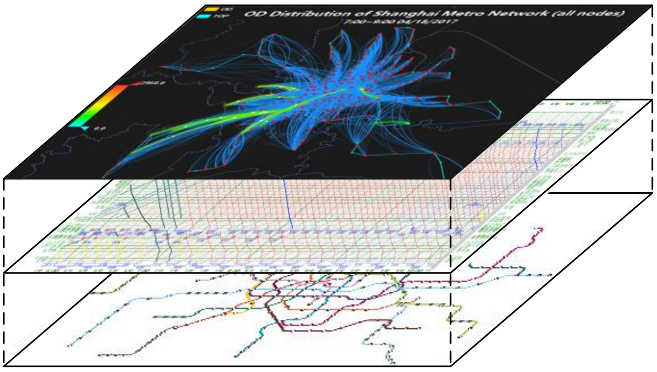


Figure 1. Double-load network illustration of the Shanghai metro system.

The travel demands of passengers often have a specific requirement on the use of certain routes particularly during peak hours. However, restricted by the ability of the infrastructure, equipment and right-of-way, train capacity cannot adequately match the peak hour excessive travel demands along those routes or segments because increasing the capacity is simply not always a feasible solution. Fortunately, with the development of the metro system, the network accessibility has been greatly enhanced. Passengers may have several routes to select from for each OD pair so as to complete their trips. As such, this paper intends to alleviate the peak hour congestion from the demand side using a service-oriented load balancing approach. By optimizing the passenger flow distribution in the metro network, the oversaturated situations may be gradually reduced. For example, Figure 2 shows a simple metro network. Route 1 is the shortest route in travel time from station O to D, which will be chosen by most passengers travelling in the peak hours. However, with the heavy passenger loads on it, Route 1 may be extremely crowded on some segments (which are shown as overloaded segments in Figure 2), and therefore at some stations, passengers may find it extremely hard to board the first train. The overcrowding will have an enormous impact on passengers under such scenarios. Some studies have been investigated that crowding affects the route choice of metro passengers and people will reroute or even have a willingness to pay (WTP) to avoid the in-vehicle crowding [1,14,15]. All the studies that have been reviewed indicate that crowding will increase the value of generalized travel time. Based on this inner motivation of passengers, to mitigate the severe congestion on overloaded segments of Route 1, an alternative Route 2 with surplus capacity can be effectively leveraged to share the large passenger loads on Route 1.

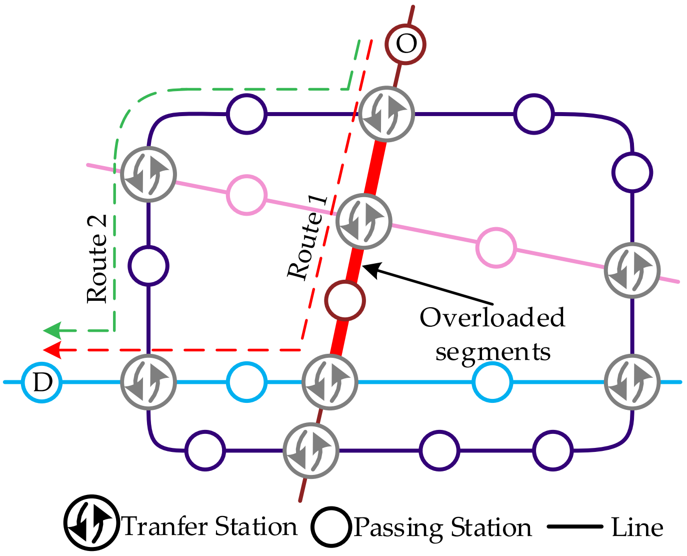


Figure 2. Overview of the load balancing approach.

The load balancing approach proposed in this paper aims at providing passengers a comfortable trip and reducing or preventing relevant safety risks caused by congestion. More specifically, this paper intends to develop an efficient method to find effective alternative routes in order to share the heavy passenger loads on overloaded segment(s) and decide how many passengers to divert is reasonable without new overloaded segment(s) appearing. The implementation of the load balancing method based on multi-path accessibility can be evaluated using both the occupancy of trains along related segments and load balancing indicators. Some assumptions need to be made before the presentation of the methodologies: (1) No delays and disruptions during the daily operations; (2) The average interval between departures in the timetable is fixed; (3) The slight difference between planned train timetables and actual train operation is ignored.

4. Methodologies

4.1. Notations

The following notations are defined and used to formulate the optimization model. Relevant sets, parameters, and variables are listed in Table 1.

Table 1. Relevant sets, parameters, and variables.

Symbols	Definition
V^{sec}	Set of OD ridership loaded on overloaded segment(s);
V_i	Ridership of OD pair i, $V_i \in V^{\text{sec}}$;
R_i	Set of efficient routes between OD pair i indexed by k;
$r_{i,k}$	Efficient route between OD pair i, $r_{i,k} \in R_i$; If $k = 0$, $r_{i,0}$ is the route of V_i which includes overloaded segment(s), otherwise, $r_{i,k}$ is the effective alternative route of $r_{i,0}$;
$[T_0, T_1]$	Study peak period;
Q	Set of involved passengers indexed by q;
$U_{k,q}$	Generalized time utility of route k chosen by passenger q;
$V_{k,a}$	Ridership loading on segment a on route k;
$t_{k,a}^{\text{in-vehicle}}$	In-vehicle-movement time along segment a on route k;
$t_{k,s}^{\text{dwell}}$	Dwell time at passing station s on route k;
$t_{k,l}^{\text{transfer}}$	Transfer time at transfer station l on route k, $t_{k,l} = t_{k,l}^{\text{walk}} + t_{k,l}^{\text{wait}}$;
$t_{k,l}^{\text{walk}}$	Walking time at transfer station l on route k;
$t_{k,l}^{\text{wait}}$	Waiting time at transfer station l on route k;
C_k	Train capacity of route k;
H_k	Headway between trains on route k;
P_k	Rated passenger capacity of per train car on route k;
m_k	Number of trains marshalling on route k;
$\beta_{k,l}$	Penalty coefficient of transfer crowding and the undesirable walking at transfer station l on route k;
$\varphi_{k,a}$	Occupancy of trains along segment a on route k;
$\alpha_{k,a}$	Penalty coefficient of in-vehicle crowding on segment a on route k;
ε	Binary variable; If route k is the route with overloaded segment(s), then $\varepsilon = -1$; otherwise, $\varepsilon = 1$.

4.2. Set of Effective Alternative Routes

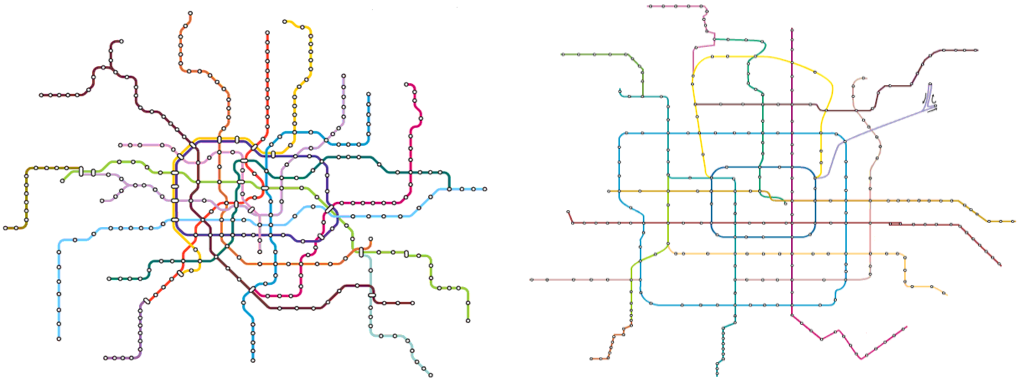
Metro systems in many large cities in China are in the process of being extended to a multilayer dynamic complex network from a single static network, with continuous narrowing headways to better serve growing users. The detailed information of the two busiest metro systems in China is presented in Table 2. As is evidently shown in Table 2, even if the minimum headway is shortened to be nearly 2 minutes in peak hours, the physical capacity may still not be able to meet the great peak hour travel demands. Therefore, increasing the capacity to satisfy booming demands is a straightforward but not always feasible in practice due to potential long construction time periods, budget limitations, right-of-way constraints, etc. [10].

Table 2. An overview of two metro systems in China.

System	Shanghai Metro	Beijing Subway
Lines	16	22
Stations	389	370
Transfers	52	53
Length (km)	666	608
Maximum daily ridership (million)	12.355	10.52
Minimum headway in peak hours	75s	70s

Note: As of 3/16/2018.

The physical metro topology presents the characteristics of both small world and the power-law distribution of distance between stations [16]. Dynamic accessibility is greatly enhanced in a complex network in which more routes are generally available for passengers to select, see Figure 3. Route choice modelling is normally based on time variables such as the in-vehicle time, transfer time and dwell time. However, passengers' perceptions of available alternative routes are such that they do not always choose what the modeler would consider as the "lowest cost" option [13]. Passengers are becoming more and more focused on the in-vehicle and on-platform crowding instead of focusing on such traditional time variables only. The exceedingly overcrowding not only brings a high degree of discomfort to users, but also impinges directly on the performance of the metro system, leading to significant train delays and a high likelihood of passengers being left behind. In comparison, the demand elasticities in association with route choice should be larger if the trip maker is presented with alternative routes that are similar in terms of comfort, punctuality and other travel attributes [17]. Assume that the set of OD ridership loaded on overloaded segment(s) during peak period $[T_0, T_1]$ is $V^{sec}= \{V_1, V_2, \ldots, V_i, \ldots, V_I\}$. And I is the total number of V_i. For each V_i, a set of effective routes are defined as $R_i= \{r_{i,0}, r_{i,1}, r_{i,2}, \ldots, r_{i,k}, \ldots, r_{i,K}\}$. $r_{i,0}$ is defined as the route of V_i with overloaded segment(s). $r_{i,k}(k = 1, 2, \ldots, K)$ is the effective alternative route of $r_{i,0}$. Commonly, both the K-shortest algorithm and Dial algorithm can be used to find routes for all OD pairs based on the route generalized cost function given by Bai, et al. [18]. However, the K-shortest route approach only adapts well to the condition of K $\leq$ 3 due to the repeated iterations, and Dial algorithm would skip over some routes when the metro system has a loop [11,18,19]. Hence, a deletion algorithm (DA) based on the depth-first traversal (DFT) method is used to find the set of effective routes for any given OD pair, as applied in [12,19]. The core feature of DA is to find the next alternative shortest route by deleting a link on the existing shortest route in the directed graph and searching for a replacement link. The DA is actually implemented by adding additional nodes and corresponding links in the directed graph.



(**a**) Shanghai metro (**b**) Beijing subway

Figure 3. Physical topology of Shanghai metro (**a**) and Beijing subway (**b**).

However, not all the routes between an OD pair can be used as an efficient alternative to $r_{i,0}$. The effective alternative routes must satisfy two requirements as follows:

(1) The difference of generalized time utility cannot exceed a threshold between the alternatives and the shortest route. The threshold ΔU can be calculated by using Equatioin (1)

$$\Delta U = \min\{\theta U_{\min}, T\} \tag{1}$$

where $U_{\min}$ is the generalized time utility of the shortest route for an OD pair. θ denotes a coefficient of proportionality. T represents the maximum time that passengers can tolerate beyond $U_{\min}$. θ and T are 0.6, 10 min respectively which are calibrated based on the results of travel surveys [12].

(2) Furthermore, all routes not only are accessible in the physical topology, but also need to meet the constraint of lines' service time. A route is invalid if some connections occur beyond the service time and thus will be excluded from the candidate set.

4.3. Mathematical Model

The load balancing (LB) approach developed in this paper aims at providing passengers a comfortable trip by relieving the in-vehicle congestion and achieving the best match between transportation capacity supply and travel demands during the peak period $[T_0, T_1]$. During the loading process, the loading will change the distribution states of passenger flow. The generalized time utility $U_{k,q}$ of a route k which an involved passenger q chooses will be continually changing. The approach presented in this paper is based on the generation of the segment ridership which is then used to calculate the utilities of the routes in the optimization procedure. The best load balancing solution will be the output when all of the involved passengers have the minimum total travel time utility with no new overloaded segment(s) emerging. Therefore, an optimization model for the load balancing problem during peak hours can be formulated as follows.

$$\min U(n) = \sum_{k\in K}\sum_{q\in Q} U_{k,q} \tag{2}$$

subject to

$$U_{k,q} = \sum_{a\in A_k}(1+\alpha_{k,a})t_{k,a}^{in-vehicle} + \sum_{s\in S_k} t_{k,s}^{dwell} + \sum_{l\in L_k}\beta_{k,l}t_{k,l}^{transfer} + \delta_k \tag{3}$$

$$t_{k,l}^{\text{transfer}} = t_{k,l}^{\text{walk}} + t_{k,l}^{\text{wait}} \tag{4}$$

$$t_{k,l}^{\text{wait}} = H_k/\pi \tag{5}$$

$$\beta_{k,l} \geq 1 \tag{6}$$

$$\alpha_{k,a} = \begin{cases} 0 & \varphi_{k,a} \leq 0.85 \\ \gamma(\varphi_{k,a} - 0.85) & 0.85 < \varphi_{k,a} < 1 \\ \infty & \varphi_{k,a} \geq 1 \end{cases} \tag{7}$$

$$\varphi_{k,a} = \frac{V_{k,a}(n+1)}{C_k} \tag{8}$$

$$C_k = \frac{T_1 - T_0}{H_k}\cdot m_k \cdot P_k \tag{9}$$

$$V_{k,a}(n+1) = V_{k,a}(n) + \varepsilon f(V_i, \eta(n)) \tag{10}$$

$$\varepsilon = \begin{cases} 1 & \text{if route } k \text{ is the effective alternative route} \\ -1 & \text{if route } k \text{ is the congested route} \end{cases} \tag{11}$$

given

$$V^{\sec} = \{V_1, V_2, \cdots, V_i, \cdots, V_I\} \tag{12}$$

$$R_i = \{r_{i,0}, r_{i,1}, r_{i,2}, \cdots, r_{i,k}, \cdots, r_{i,K}\} \tag{13}$$

where the objective function (2) aims at minimizing the total utility of all involved passengers. Equation (3) defines $U_{k,q}$, the improved generalized time utility of route k which an involved passenger q chooses, and relates it to all the time elements with penalties levied on in-vehicle congestion and transfer(s) that a traveler spends for a trip in metro systems. $U_{k,q}$ is calculated using $t_{k,a}^{\text{in-vehicle}}$, $t_{k,s}^{\text{dwell}}$, $t_{k,l}^{\text{transfer}}$ and other parameters, where $t_{k,a}^{\text{in-vehicle}}$ is the rail-ride time along segment a, $a \in A_k$; $t_{k,s}^{\text{dwell}}$ is the dwell time at passing station s, $s \in S_k$; and $t_{k,l}^{\text{transfer}}$ denotes the transfer time at transfer station l, $l \in L_k$. δ_k is a constant term introduced to represent omitted factors other than $t_{k,a}^{\text{in-vehicle}}$, $t_{k,s}^{\text{dwell}}$ and

$t_{k,l}^{\text{transfer}}$, if any. Equations (4)–(5) indicate the transfer time at the transfer station l of route k which contains two parts: Average walking time for transferring $t_{k,l}^{\text{walk}}$ and average waiting time $t_{k,l}^{\text{wait}}$ at transfer station. $\beta_{k,l}$ in Constraint (6) is a penalty coefficient of transfer crowding and the undesirable walking at transfer stations which can be estimated from the travel surveys. Constraints (7)–(11) denote different load states on each segment, corresponding to the occupancy $\varphi_{k,a}$ of trains along segments which are determined by $V_{k,a}(n+1)$ and C_k. The threshold values of 0.85 and 1 in Constraint (7) are set according to both the General Technical Specifications for Metro Vehicles (GB/T 7928-2003) [20] and the literature [3,4]. The ridership of segment(s) $V_{k,a}(n+1)$ is determined by constraint (10), in which $f(V_i, \eta(n))$ is a nonlinear function of the segment ridership on the route with overloaded segment(s) denoted by $\eta(n)$. The specific form of function f is unknown. Two approaches can be used to determine the function f according to Abadi et al. [13]. One is to identify the non-linear terms of function f in a closed form which is very complicated and lack of effective identification schemes. Another is using simulation to estimate the segment ridership by processing available historical/real time ridership data which is computationally feasible and provide a much better accuracy. $\eta(n)$ is a setting proportion for every loading and n is the number of iterations. Constraint (11) shows how the ridership should change between the route with overloaded segment(s) and effective alternative routes. Generally, $t_{k,a}^{\text{in-vehicle}}$, $t_{k,s}^{\text{dwell}}$, H_k, m_k, and P_k are given by train timetables. $t_{k,l}^{\text{walk}}$ is decided by the layout of the transfer station. If the arriving of passengers is uniform and the interval of train departure is fixed, $t_{k,l}^{\text{wait}}$ often takes half the headway H_k, i.e., π value for 2 [21]. V^{sec} in Equation (12) is a known quantity given by Huang et al. [22], and R_i in Equation (13) is obtained using the method as mentioned above.

In order to evaluate the performance of the LB method, two indicators are defined to reflect the matching between train capacity and travel demands over all involved segments, including the penalty for crowding z_1 and the penalty for wasted capacity z_2. During peak hours, (50%, 85%) is the reasonable range of the mean occupancy. When the mean occupancy exceeds the threshold of 85%, the train will become overcrowded, reducing the level of service. In addition, if the mean occupancy is less than 50%, it means that transportation capacity is surplus. To evaluate the train load balancing level, the two penalty terms can be calculated as follows.

$$z_1 = \omega_1 \sum_{k \in K} \sum_{a \in A} \max\{0, V_{k,a} - C_k \times 85\%\} \tag{14}$$

$$z_2 = \omega_2 \sum_{k \in K} \sum_{a \in A} \max\{0, C_k \times 50\% - V_{k,a}\} \tag{15}$$

where ω_1 is the penalty coefficient for crowding, and ω_2 means the penalty coefficient for the wasted capacity.

4.4. Algorithm of Load Balancing

In this section, a load balancing algorithm is presented to utilize available capacity more effectively to alleviate peak hour congestion in the metro network. The essence of the method is to search under-utilized alternative routes to share passenger loads on overloaded segment(s) with no new overloaded segment(s) emerging. The load balancing algorithm procedure is presented as follows.

Input data: (1) Train timetable data: $t_{k,a}^{\text{in-vehicle}}$, $t_{k,s}^{\text{dwell}}$, H_k, m_k, P_k; (2) Given data: T_0, T_1, V^{sec}, R_i, π; (3) Estimated data: $t_{k,l}^{\text{walk}}$, $\beta_{k,l}$.

Output: load balancing solution E, the total travel utility U and two evaluation indicators z_1, z_2.

STEP 1: Input data, initialize $i = 0$;

STEP 2: $i = i + 1$, if $i \leq I$, initialize $k = 0$, go to Step 3; Else, end;

STEP 3: $k = k + 1$, if $k \leq K$, initialize $n = 0$, go to Step 4; Else, go to Step 2;

STEP 4: $n = n + 1$;

STEP 5: V_i will be shifted from the route $r_{i,0}$ with overloaded segment(s) to an effective alternative route $r_{i,k}$ iteratively.

STEP 6: Since the loading will change the states of the network, the occupancy of trains along each relevant segment and the total travel time utility of involved passengers will be recalculated and updated after every loading.

STEP 7: During the loading process, if new overloaded segment(s) appears on $r_{i,k}$, then the algorithm ends and goes back to Step 3; Else if no overloaded segment(s) appears on $r_{i,k}$ and the total travel time utility $U(n+1) < U(n)$, then the algorithm goes back to Step 4; Till up to $U(n+1) \geq U(n)$, and if the overloaded segment(s) on $r_{i,0}$ disappears, then the algorithm ends and two evaluation indicators will be computed. Return with the load balancing solution E, U, z_1, and z_2.

STEP 8: If one loading finished, no new overloaded segment(s) arises and the overloaded segment(s) on $r_{i,0}$ still exists, once V_i is loaded completely, the algorithm goes back to Step 2; else, the algorithm will go back to Step 3.

Figure 4 illustrates the process of the load balancing.

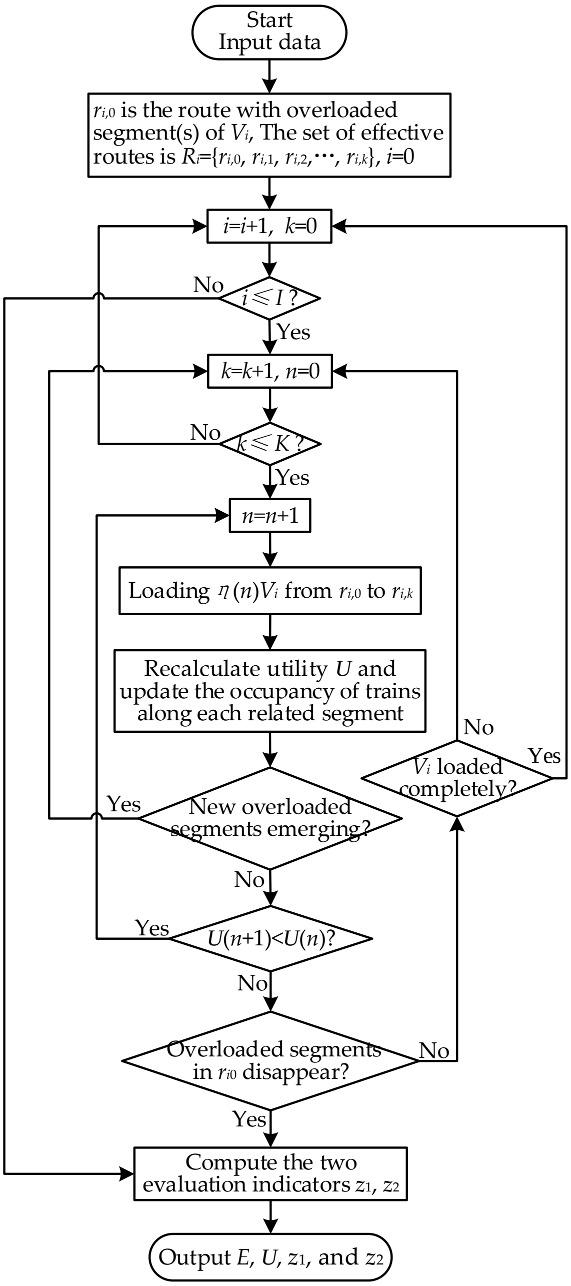

Start
Input data
$r_{i,0}$ is the route with overloaded segment(s) of V_i, The set of effective routes is R_i={$r_{i,0}$, $r_{i,1}$, $r_{i,2}$,⋯, $r_{i,k}$}, i=0
i=i+1, k=0
i≤I ?
No
Yes
k=k+1, n=0
k≤K ?
No
Yes
n=n+1
Loading $\eta(n)V_i$ from $r_{i,0}$ to $r_{i,k}$
Recalculate utility U and update the occupancy of trains along each related segment
New overloaded segments emerging?
Yes
No
V_i loaded completely?
No
Yes
$U(n+1)<U(n)$?
Yes
No
Overloaded segments in r_{i0} disappear?
No
Yes
Compute the two evaluation indicators z_1, z_2
Output E, U, z_1, and z_2

Figure 4. The algorithmic procedure to generate the load balancing solution.

5. Case Study and Discussions

The LB approach as presented in this paper was evaluated and illustrated using a real-world case based on the Shanghai metro system. At the end of 2017, the Shanghai metro system consisted of 16 lines with 389 stations (see Table 2 and Figure 6). Due to huge travel demands of commuters, the Shanghai metro system had been exhibiting high occupancies during peak hours over many line segments of the metro network, as shown in Figure 5.

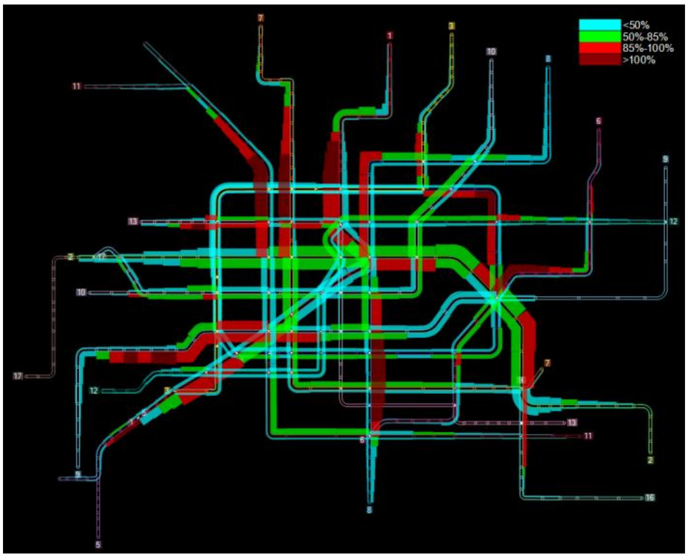


Figure 5. Occupancy of trains on all segments in Shanghai Metro system in morning peak.

On 3 March 2016, during the morning peak period of 08:30–09:00, the mean occupancies of trains along segments Caoyang Rd.–Longde Rd.–Jiangsu Rd. (in the direction from Caoyang Rd. to Jiangsu Rd.) and Xujiahui–Yishan Rd. (from Xujiahui to Yishan Rd.) are 95.63%, 94.92%, and 97.64% respectively. Such occupancies far exceed the threshold of 85%. Hence, Caoyang Rd.–Longde Rd.–Jiangsu Rd., and Xujiahui–Yishan Rd. can be declared as overloaded segments. In Figure 6, all of the involved lines are colored, and all of the related stations are presented by numbers and letters (e.g., 11a). The overloaded segments 11a–11b–11c on Line 11 and 11e–3/4e on Line 9 are marked in red shadow boxes. The OD pairs and ridership causing the congestion on overloaded segments 11a–11b–11c in peak period of 08:30–9:00 have been acquired [22]. Among them, three clusters of OD pairs (OD_1, OD_2, OD_3) and 41.07% of total ridership, have effective alternative routes and are shown in Figure 6, where red dotted lines represent the routes with overloaded segments and green dotted lines denote under-utilized alternative routes. The OD pairs, ridership and the set of effective alternative routes for testing the performance of load balancing method are provided in Table 3. Note that, the effective alternative route $r_{2,1}$ had an overloaded segment Baoshan Rd. (3/4h)–Hailun Rd. (10a) with an occupancy of 97.37%, so $r_{1,1}$ and $r_{3,1}$ are the preferential effective alternative routes used in the loading process.

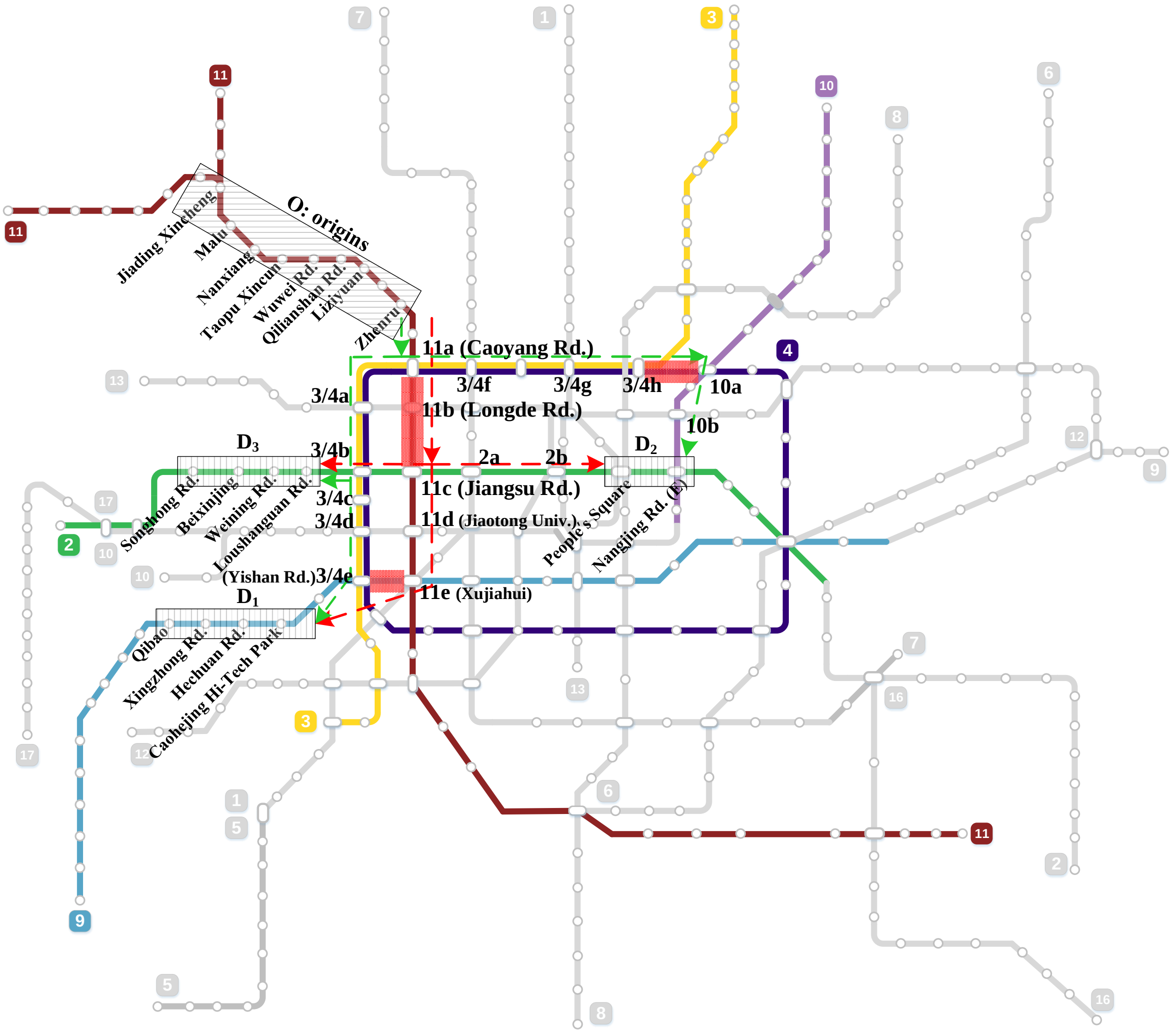


Figure 6. Origin-Destination (OD) pairs and effective routes in network topology of the Shanghai metro system.

Table 3. Input data for load balancing (LB).

OD Pair(s)	Ridership	Route with Overloaded Segments	Effective Alternative Route
OD$_1$	V_1= 3169	$r_{1,0}$: O→11a→11c→11e→D$_1$	$r_{1,1}$: O→11a→3/4a→3/4e→D$_1$
OD$_2$	V_2= 2874	$r_{2,0}$: O→11a→11c→D$_2$	$r_{2,1}$: O→11a→3/4h→10a→D$_2$
OD$_3$	V_3= 1500	$r_{3,0}$: O→11a→11c→3/4b→D$_3$	$r_{3,1}$: O→11a→3/4a→3/4b→D$_3$

Basic raw data about train operation are obtained from the train scheduled timetables used on 3 March 2016 which are provided by Operation Center of Shanghai Metro. The passenger travel data of the records of rail journeys are acquired from the Automatic Fare Collection Cleaning Centre (ACC) of Shanghai Metro. π value for 2 [21]. The waking time $t_{k,l}^{\text{walk}}$ for transferring and $\beta_{k,l}$ adopted in the given model should be estimated in advance though the results of travel surveys conducted in the Shanghai metro system in peak periods on weekdays. Tables 4 and 5 presents the values of parameters used in the model solution process, and values of the penalty coefficients for compute z_1 and z_2 are ω_1 = 2 and ω_2 = 1.

Table 4. Parameters of train operation during the morning peak period of 08:30–09:00.

Line	Number of Trains Marshalling m_k	Rated Capacity P_k (Passengers/Car)	Headway H_k (s)	Correction Factor π
Line 2	8	310	180	2
Line 3–4	6	310	150	2
Line 4	6	310	300	2
Line 9	6	310	225	2
Line 10	6	310	200	2
Line 11	6	310	180	2

Table 5. Parameters of involved transfer stations.

Involved Transfers	Transfer Direction	Walking Time for Transferring $t_{k,l}^{walk}$ (s)	Penalty Coefficient $\beta_{k,l}$
Caoyang Rd. (11a)	From Line 11 to Line 3/4	206	1.67
Jiangsu Rd. (11c)	From Line 11 to Line 2	220	2.95
Xujiahui (11e)	From Line 11 to Line 9	220	1.86
Zhongshan Park (3/4b)	From Line 3/4 to Line 2	271	2.13
Yishan Rd. (3/4d)	From Line 3/4 to Line 9	354	2.19
Hailun Rd. (10a)	From Line 4 to Line 10	150	1.50

By applying the proposed LB approach, a load balancing block is developed in the prototype system using VB.net language to calculate the process of passenger flow load balancing. The load balancing block has to balance the relevant utilization by shifting passenger loads from oversaturated segments to less utilized alternatives. Once the stopping criterion is satisfied, the final load balancing solution is returned. A portion of the numerical calculation of simulation is shown in Table 6.

Table 6. Numerical calculation of the LB process.

<table>
<tr><th rowspan="3">Iteration</th><th colspan="5">Route with Overloaded Segments</th><th colspan="5">Effective Alternative Route</th></tr>
<tr><th colspan="10">The Occupancy of Trains Along Involved Segments (%)</th></tr>
<tr><th colspan="10">Load V_1 from $V_{1,0}$ to the under-utilized route $V_{1,1}$</th></tr>
<tr><td><i>n</i></td><td><b>11a–11b</b></td><td><b>11b–11c</b></td><td><b>11c–11d</b></td><td><b>11d–11e</b></td><td><b>11e–3/4e</b></td><td><b>11a–3/4a</b></td><td><b>3/4a–3/4b</b></td><td><b>3/4b–3/4c</b></td><td><b>3/4c–3/4d</b></td><td><b>3/4d–3/4e</b></td></tr>
<tr><td>0</td><td>95.63</td><td>94.91</td><td>54.72</td><td>48.34</td><td>97.46</td><td>60.70</td><td>65.36</td><td>56.45</td><td>48.06</td><td>44.17</td></tr>
<tr><td>1</td><td>94.55</td><td>93.84</td><td>53.64</td><td>47.26</td><td>96.12</td><td>61.60</td><td>66.25</td><td>57.34</td><td>48.97</td><td>45.06</td></tr>
<tr><td>2</td><td>91.33</td><td>90.61</td><td>50.41</td><td>44.04</td><td>92.08</td><td>64.29</td><td>68.94</td><td>60.03</td><td>51.65</td><td>47.75</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>20</td><td>89.18</td><td>88.46</td><td>48.26</td><td>41.89</td><td>89.40</td><td>66.08</td><td>70.73</td><td>61.82</td><td>53.45</td><td>49.54</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>40</td><td>88.10</td><td>87.39</td><td>47.19</td><td>40.81</td><td>88.05</td><td>66.98</td><td>71.63</td><td>62.72</td><td>54.34</td><td>50.44</td></tr>
<tr><td>41</td><td>87.83</td><td>87.12</td><td>46.92</td><td>40.54</td><td>87.72</td><td>67.20</td><td>71.85</td><td>62.94</td><td>54.57</td><td>50.66</td></tr>
<tr><td>42</td><td>87.63</td><td>86.92</td><td>46.72</td><td>40.34</td><td>87.47</td><td>67.37</td><td>72.02</td><td>63.11</td><td>54.73</td><td>50.83</td></tr>
<tr><td colspan="11">The total travel time utility reaches the minimum and $r_{1,0}$ only has one effective alternative route. Therefore, the loading of V_1 ends.</td></tr>
<tr><td colspan="11">Continue to load V_3 from $r_{3,0}$ to the under-utilized route $r_{3,1}$</td></tr>
<tr><td><i>n</i></td><td colspan="2"><b>11a–11b</b></td><td colspan="2"><b>11b–11c</b></td><td colspan="2"><b>11c–3/4b</b></td><td colspan="2"><b>11a–3/4a</b></td><td colspan="2"><b>3/4a–3/4b</b></td></tr>
<tr><td>0</td><td colspan="2">87.63</td><td colspan="2">86.92</td><td colspan="2">54.01</td><td colspan="2">67.37</td><td colspan="2">72.02</td></tr>
<tr><td>1</td><td colspan="2">87.10</td><td colspan="2">86.38</td><td colspan="2">53.60</td><td colspan="2">67.81</td><td colspan="2">72.46</td></tr>
<tr><td>2</td><td colspan="2">86.02</td><td colspan="2">85.31</td><td colspan="2">52.80</td><td colspan="2">68.71</td><td colspan="2">73.36</td></tr>
<tr><td>...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td></tr>
<tr><td>10</td><td colspan="2">85.48</td><td colspan="2">84.77</td><td colspan="2">52.40</td><td colspan="2">69.16</td><td colspan="2">73.81</td></tr>
<tr><td>...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td><td colspan="2">...</td></tr>
<tr><td>24</td><td colspan="2">85.35</td><td colspan="2">84.63</td><td colspan="2">52.29</td><td colspan="2">69.27</td><td colspan="2">73.92</td></tr>
<tr><td>25</td><td colspan="2">85.22</td><td colspan="2">84.50</td><td colspan="2">52.19</td><td colspan="2">69.38</td><td colspan="2">74.03</td></tr>
<tr><td>26</td><td colspan="2">85.01</td><td colspan="2">84.29</td><td colspan="2">52.04</td><td colspan="2">69.56</td><td colspan="2">74.21</td></tr>
<tr><td colspan="11">The total travel time utility reaches the minimum. The overloaded segments disappear without new overloaded segments emerging. As such, the load balancing solution is returned.</td></tr>
</table>

As one can see in Table 7, the load balancing solution is presented. Afterwards, to verify the validity of the methodologies, a comparative analysis before and after the LB was also conducted, where the values of the total travel generalized time, penalty for crowdedness, and penalty for wasted capacity are listed in Table 8. Additionally, the benefits of the solution obtained by optimization method were presented via the relative deviations of the optimal results from the no balancing results. The occupancy of trains along all involved segments are shown in Figure 7. As can be clearly seen, with load balancing, the total travel generalized time of involved passengers was reduced by 0.6%, which is a slight decrease. The penalty for crowdedness was reduced by 71.54%, so it can prove that the LB method presented in this paper can effectively alleviate the overcrowding situation. However, the penalty for wasted capacity increases by 8.79%, but this loss can easily be compensated by other benefits. For the overloaded segments marked in circles in Figure 7, their occupancies effectively drop to 85%. In other words, the overcrowding on segments between Caoyang Rd.–Longde Rd.–Jiangsu Rd. and Xuijiahui–Yishan Rd. were effectively alleviated to a great extent.

Table 7. Load balancing solution.

Ridership for Loading	Routes for Balancing
1487	From $r_{1,0}$ to $r_{1,1}$
489	From $r_{3,0}$ to $r_{3,1}$

Table 8. Performance comparison before and after LB.

Indictor	No LB	With LB	Variations
Total travel generalized time (h)	4070.8266	4046.5331	−0.60%
Penalty for crowdedness	14,836	4222	−71.54%
Penalty for wasted capacity	4139	4503	+8.79%

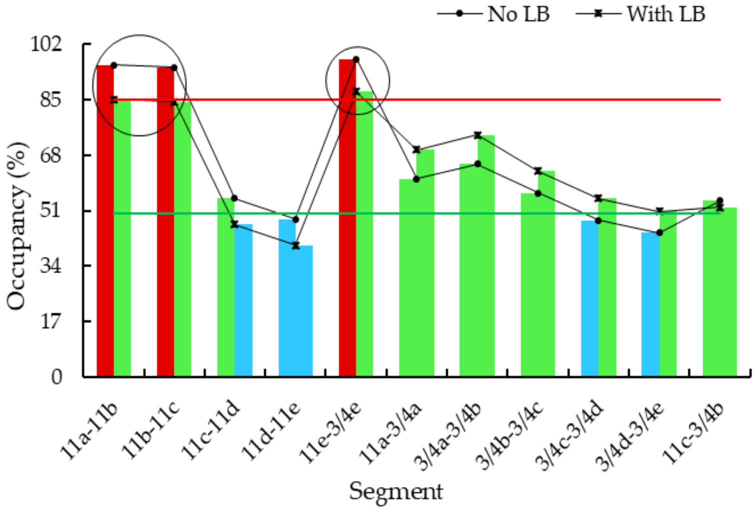


Figure 7. Changes of occupancy before and after LB.

6. Perspectives in the Smart Metro Systems

Nowadays, in the Shanghai metro system, Beijing subway or other metro systems in China, mobile internet technology, artificial intelligence (AI), and other advanced technologies have been widely used to help trips in the metro system to be made in a more and more intelligent way. Some metro lines are even automatic without drivers. One can pay online using a smartphone when passing the gates. Wi-Fi is available at every corner of the station, even when the trains are running underground. With tracks being obtained from the available data of cellphone signals and the Wi-Fi being used, every user's trip can be accurately estimated and the distribution of the demands in the network can also be reliably forecasted throughout the day.

From the methodologies developed and numerical examples presented in this paper, it is possible to disperse travel demands to reduce the peak hour congestion. For the final load balancing solution to be effectively used and applied to the actual active demand management in the real world, some proactive inducing schemes are discussed in this section.

- **An Application (APP) downloaded onto the smartphone will be helpful.** For passengers, when overcrowded stations and segments are made available and shown, it can greatly help inform users to take an under-utilized route among all effective ones and allow passengers to reserve their trips ahead of time to avoid congested segments. For metro operation and management related enterprises, it is a real-time information broadcasting platform which can be effectively leveraged and developed to directly deliver important messages to passengers.
- **A route pricing strategy can be another novel scheme to balance the travel demands in a metro system.** A stated preference survey has been conducted in morning peak hours in the Shanghai metro system. Preliminary analysis of the results shows that more than 40% of commuters would choose a slightly longer but less crowded route [23]. If passengers shift their travels to a less heavily utilized routes in peak hours, they should be incentivized and rewarded (e.g., free of charge). It should be noted that, the incentives will need to be dynamically adjusted to the level of congestion on the segments in the metro network.

Peak hour congestion is a collective issue created by individual behavior. Such proactive inducing schemes, other than coercive measures at the cost of users' satisfaction, can be positively leveraged to mitigate the congestion on those overloaded segments by influencing personal travel behavior decisions.

7. Conclusions

Congestion in the metro systems is an issue of imbalance between the supply and demand in principle due to a large number of passengers demanding for a limited amount of capacity in peak hours. The passenger flow distribution in the metro network from the demand-side should be effectively managed. Aimed at improving service quality to passengers traveling in rush hours, this paper presents an LB method to shift loads from oversaturated segment(s) to under-utilized effective alternative routes. Based on the multi-path accessibility of a complex metro system, a set of effective alternative routes are searched using the deletion algorithm. Then, a load balancing solution is obtained when the iterative process satisfies the stopping criterion. A numerical experiment based on a real-world example of the Shanghai metro system is conducted to demonstrate the flexibility and validity of the proposed approach. The results are verified in that the LB method can effectively reduce penalty for crowdedness and the occupancy of trains along overloaded segment(s) in peak hours, leading to improved safety and comfort of the metro system. For the final load balancing solution to be effectively applied in practice, two practical inducing schemes are also discussed.

As the line of research matures, further research will focus on how to reduce congestion on the overloaded segment(s) without effective alternative routes. The method proposed cannot be applied to the overloaded segment(s) in which the passenger flow is mostly composed of traveling only on one line since there are no effective alternative routes. Therefore, to mitigate the network congestion completely, some supplemental methods and measures are necessary in future research. Furthermore, more accurate ridership data should be applied to verify the presented approach since the AFC data may have some drawbacks in estimating the ridership on segments.

Author Contributions: Conceptualization, R.X., Z.H. and F.Z.; Methodology, Z.H.; Software, F.Z.; Formal analysis, Z.H.; Investigation, Z.H.; Data curation, W.L.; Project administration, R.X.; Supervision, R.X.; Funding acquisition, R.X.; Writing—original draft preparation, Z.H.; Writing—review and editing, W.F.

Funding: This research was funded by the Research Project of Shanghai Science and Technology Commission, grant number 16DZ1203802; and International Exchange Program for Graduate Students, Tongji University, grant number 2018020027.

Acknowledgments: The acquisition of data in the paper is supported by the Shanghai Shentong Metro Group Co., Ltd. The authors are very grateful for this support. Special thanks to all anonymous reviewers for their helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, K.M.; Hong, S.-P.; Ko, S.J.; Kim, D. Does crowding affect the path choice of metro passengers? *Transp. Res. Part A: Policy Pract.* **2015**, *77*, 292–304. [CrossRef]
2. Li, S.M.; Wong, F.C.L. The effectiveness of differential pricing on route choice: the case of the mass transit railway of Hong Kong. *Transportation* **1994**, *21*, 307–324. [CrossRef]
3. Raveau, S.; Munoz, J.C.; Grange, L.D. A topological route choice model for metro. *Transp. Res. Part A: Policy Pract.* **2011**, *45*, 138–147. [CrossRef]
4. Raveau, S.; Guo, Z.; Muñoz, J.C.; Wilson, H.M.N. A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics. *Transport. Res. Part A: Policy Pract.* **2014**, *66*, 185–195. [CrossRef]
5. Zhao, P.; Yao, X.M.; Yu, D.D. Cooperative passenger inflow control of urban mass transit in peak hours. *J. Tongji Univ. (Nat. Sci.)* **2014**, *42*, 1340–1346.
6. Yao, X.M.; Zhao, P.; Qiao, K.; Yu, D.D. Modeling on coordinated passenger inflow control for urban rail transit network. *J. Cent. South Univ. (Sci. Technol.)* **2015**, *46*, 342–350.
7. Li, S.K.; Dessouky, M.M.; Yang, L.X.; Gao, Z.Y. Joint optimal train regulation and passenger flow control strategy for high-frequency metro lines. *Transp. Res. Part B: Method.* **2017**, *99*, 113–137. [CrossRef]
8. Shi, J.G.; Yang, L.X.; Yang, J.; Gao, Z.Y. Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach. *Transp. Res. Part B: Method.* **2018**, *110*, 26–59. [CrossRef]
9. Xu, X.Y.; Liu, J.; Li, H.Y.; Jiang, M. Capacity-oriented passenger flow control under uncertain demand: Algorithm development and real-world case study. *Transp. Res. Part E: L. Transp. Rev.* **2016**, *87*, 130–148. [CrossRef]
10. Jiang, Z.B.; Fan, W.; Liu, W.; Zhu, B.Q.; Gu, J.J. Reinforcement learning approach for coordinated passenger inflow control of urban rail transit in peak hours. *Transp. Res. Part C: Emerging Technol.* **2018**, *88*, 1–16. [CrossRef]
11. Si, B.F.; Mao, B.H.; Liu, Z.L. Passenger flow assignment model and algorithm for urban railway traffic network under the condition of seamless transfer. *J. Ch. Railw. Soc.* **2007**, *29*, 12–18.
12. Zhu, W.; Hu, H.; Xu, R.H.; Hong, L. Modified stochastic user-equilibrium assignment algorithm for urban rail transit under network operation. *J. Cent. South Univ. Engl. Ed.* **2013**, *20*, 2897–2904. [CrossRef]
13. Abadi, A.; Ioannou, P.A.; Dessouky, M.M. Multimodal dynamic freight load balancing. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 356–366. [CrossRef]
14. Whelan, G.; Jon, C. An investigation of the willingness to pay to reduce rail overcrowding. In Proceedings of the first International Conference on Choice Modelling, Harrogate, England, 30 March–1 April 2009.
15. Li, Z.; Hensher, D.A. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. *Transp. Policy* **2011**, *18*, 880–887. [CrossRef]
16. Feng, J.; Xu, Q.; Li, X.M.; Yang, Y.Z. Complex network study on urban rail transit systems. *J. Transp. Syst. Eng. Inf. Technol.* **2017**, *17*, 242–247.
17. Rosenbloom, S. Peak-period traffic congestion: A state-of-the-art analysis and evaluation of effective solutions. *Transportation* **1978**, *7*, 167–191. [CrossRef]
18. Bai, Y.; Liu, J.F.; Sun, Z.Z.; Mao, B.H. Analysis on route choice behavior in seamless transfer urban rail transit network. In Proceedings of the IEEE International Workshop on Modelling, Simulation and Optimization, Hong Kong, China, 27–28 December 2008; pp. 264–267.
19. Xu, R.H.; Luo, Q.; Gao, P. Passenger flow distribution model and algorithm for urban rail transit network based on multi-route choice. *J. Ch. Railw. Soc.* **2009**, *31*, 110–114.
20. National Standardization Administration of China. General technical specification for metro vehicles. 2003. Available online: http://www.gb688.cn/bzgk/gb/newGbInfo?hcno=657AECE6BFA031382A2645D61363896C (accessed on 5 December 2003).

21. Cea, J.D.; Fernández, E. Transit assignment for congested public transport systems: An equilibrium model. *Transp. Sci.* **1993**, *27*, 133–147. [CrossRef]
22. Huang, Z.Y.; Xu, R.H.; Zhou, F.; Xu, T.J. Estimation method of section passenger flow composition for metro network with constraints of time and route. *J. Tongji Univ. (Nat. Sci.)* **2018**, *46*, 920–925.
23. Xu, R.H.; Huang, Z.Y.; Zhou, F.; Li, C.F. Does differential pricing affect route choice of the peak hour metro passengers? A case study in Shanghai metro system. In Proceedings of the IEEE International Conference on Intelligent Rail Transportation, Singapore, 12–14 December 2018; pp. 1–5.