The Journal of Transport and Land Use jtlu.org
Vol.17 No. 1 [2024] pp. 537–555

Effect of multiscale metro network-wide attributes on peak-hour station passenger and flow balancing

Haixiao Pan
College of Architecture and Urban Planning
Tongji University
hxpank@vip.126.com

Miao Hu (corresponding author)
Urban Mobility Institute
Tongji University
miaowhu@tongji.edu.cn

Xiyin Deng
College of Architecture and Urban Planning
Tongji University
471560175@qq.com

Ailing Liu
College of Architecture and Urban Planning
Tongji University
864053969@qq.com

Abstract: Analyzing the balance of station passenger and passenger flow is essential for understanding jobs-housing balance and built environment in station areas and network-wide range as well as for enhancing the efficiency of urban rail transit operations. Taking the Shanghai rail transit network as a case study, this paper defines the Multiscale Subnetwork (MSSN) based on a specific spatial scope. By extracting the network features and built-environment elements of the stations and the MSSN, this study analyzes the factors affecting the peak-hour station passenger and the imbalance of regional network passenger flow. The research suggests that the small MSSN analysis, within 6-8 km from a station, can provide valuable results from a network-wide perspective, rather than solely focusing on individual station areas or the entire network. The regional attributes of jobs-housing balance and the transportation conditions in the MSSN range have great impact on both station passengers and flow imbalance. This research provides theoretical insights for urban planners and policymakers to formulate effective strategies for urban rail transit networks.

Keywords: transportation planning, urban rail transit, passenger flow, geographically weighted regression, multiscale network

Article history:
Received: October 20, 2023
Received in revised form: March 3, 2024
Accepted: May 24, 2024
Available online: July 9, 2024

1 Introduction

With the rapid development of the economy and society, cities are continually expanding and spreading outward, leading to an increase in commuting between residential areas and workplaces. This has resulted in severe congestion during rush hours. Due to its large scale, high speed, punctuality, and safety, urban rail transit has become a crucial component of metropolitan public transport systems that support long-distance travel. In 2012, China had the longest domestic operating mileage of rail transit

https://doi.org/10.5198/jtlu.2024.2443
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

in the world. From 2012 to 2022, the average annual mileage of urban rail transit operating in China was 792.2 km. Among the top three cities in terms of operating mileage, Shanghai has seen an average increase of 50.9 km per year, with a growth rate of 8.1%. Beijing and Chengdu follow with increases of 46 km and 68 km per year and growth rates of 7.8% and 48.7%, respectively.

Cities in China are currently constructing multiple rail lines simultaneously, resulting in a rapid shift from individual lines to interconnected networks that cover the central city. However, the growth rate of passenger flow and the expansion of the network scale are unsynchronized. This is due to variations in the network characteristics of different stations, as well as differences in population and employment distribution. Some rail transit stations experience low peak-hour passenger flow, while others are facing high levels of congestion. Imbalanced passenger flow in the upstream and downstream directions of a line section can exacerbate congestion on regional networks and affect the comfort of rail transit passengers. Identifying and avoiding such imbalances, particularly the issue of low passenger flow during peak periods, is a topic worth studying. In Japan, Switzerland, and many other countries, researchers have already paid significant attention to this problem (Börjesson et al., 2021; Su, 2018). Extreme differences in passenger flow can have a negative impact on the sustainable development of urban rail transit and impose higher requirements on its efficient and safe operation (Zeng et al., 2018). While measures such as adjusting train operating schedules and implementing crowded fares can help alleviate peak passenger flow to some extent (Canca et al., 2016; Ceapa et al., 2012), it is also necessary to study the spatial structure and layout of the rail transit network, as well as its internal characteristics, from a network-wide perspective. In the context of the rapid expansion of rail transit networks, this study argues that the construction of rail transit and surrounding areas should be coordinated, and that research should be conducted at the network level rather than focusing solely on individual stations.

Relevant studies have demonstrated that the characteristics of land use around stations are among the most important factors contributing to peak-hour deviations (Li et al., 2020). Additionally, stations with similar surrounding land use commonly exhibit similar time-varying trends in passenger flow (Zhao et al., 2019). The location of a station also plays a role in peak-hour deviations. For instance, a station with high employment density and a status as a transfer node can result in an increase in the number of passengers (Cervero, 2007; Ewing & Cervero, 2001; Pan et al., 2017). In terms of urban planning elements, Ewing and Cervero conducted a meta-analysis to systematically determine the relationship between travel behavior and the built environment in station areas. They identified the “5D” of the built environment, which includes density (job density, population density, plot ratio, and residential density), diversity (land use diversity), and design (walkable environment). And the other two features added are distance to transit and destination accessibility, which have also been found to be crucial by both domestic and foreign scholars (Cervero, 2006; Pan et al., 2017; Peng et al., 2021). Furthermore, some scholars have suggested considering other factors outside the station area, such as network characteristics, and proposed the research concept of “5D+N” (Xia & Zhang, 2019).

Many studies on transit-oriented development (TOD) are based on stations or single line corridors (Liu et al., 2022). This makes the single station area the primary spatial scale unit when analyzing the relationship between TOD and passengers (Su et al., 2022). Furthermore, the characteristics of a station in the network are frequently regarded as external factors or simple assumptions, which can reduce the robustness of passenger analysis models (Andersson, 2021). In addition to the central urban areas of China, numerous enterprises gather near metro stations on the periphery of cities (Fang, 2021).

The commuting destinations of peak passenger flow are scattered across rail network nodes. Therefore, the peak passenger flow should be studied from a subnetwork-wide perspective, and the planning elements should be extended from stations to regional networks to fully coordinate and exploit the large-capacity advantages of rail transit systems. To this end, some scholars have proposed the concept of corridor-TOD (C-TOD) to capture the network interaction between TOD stations on the rail network, combining economic, social, and environmental indicators to comprehensively evaluate C-TOD in terms of facility density, accessibility of various services, and traffic emissions (Liu et al., 2020).

Nonetheless, with the increasing complexity of the urban rail transit network structure in China, and the widespread distribution of jobs on the network, the traditional approach of studying rail transit systems solely through the concept of corridors may no longer be suitable. In light of this, this study proposes a new subnetwork concept for analysis, building upon existing research on nodes and corridors. This concept expands the TOD area from individual stations or areas along the metro line to a network level. By examining the rail network at multiple scales, this study empirically investigates and discusses the reasons behind the peak passenger flow and imbalanced performance of the Shanghai rail transit network.

2 Station and multiscale subnetwork (MSSN)

2.1 Study site

As of December 30, 2021, Shanghai has 20 rail lines in operation (including one maglev line). The total operating mileage of the Shanghai rail transit has reached 831 km, ranking first in mainland China and worldwide. In 2021, the Shanghai rail transit network achieved an average daily passenger volume of 9.78 million. However, the passenger flow is still spatially imbalanced under the large-scale network operations. Based on the available data on passenger flow, population, jobs, road networks, and transportation conditions, this study considers a weekday in 2016 as the research object. The passenger flow data of 288 stations on 14 lines during the morning peak were provided and extracted from the metro card data.

2.2 Station passenger characteristics

The data of passenger was extracted from the IC card of ticket gates, in which only the starting and ending places of a trip were recorded; therefore, all of transferring behaviors were not included. The number of passengers at each station is the sum of inbound and outbound people during peak hours. Here, the term “outbound/inbound passenger” of a station refers to the number of people who tap in/out at a station. The outbound station is where a transit trip begins and the passengers here are leaving for another station, while the inbound station is the destination of the transit travel.

Table 1. Station passenger of Shanghai rail transit network in the morning peak

	Min	Pct10	Pct25	Median	Mean	Pct75	Pct90	Max
Passenger	476	2410	4126	7146	9028	10798	17869	43309

According to the statistics displayed in Table 1, the distribution of passengers during the morning peak among Shanghai rail transit network stations is clearly imbalanced. People’s Square Station on Lines 1, 2, and 8 had the highest number of passengers, which was 86 times that of Middle Huaxia Road Station, approximately 15 km away. Figure 1

illustrates the differences in passenger numbers between white stations (stations with fewer passengers than the 10% quantile) and dark red stations (stations with more passengers than the 90% quantile). Generally, stations with passengers falling below the 10% quantile had an average population of less than 1873 within their 500 m radius. The average rail travel distance between these stations and People's Square Station in the city center was 29.20 km. On the other hand, stations with passengers exceeding the 90% quantile had an average population of 7597 within their 500 m radius. Their average rail travel distance to People's Square Station was 9.56 km.

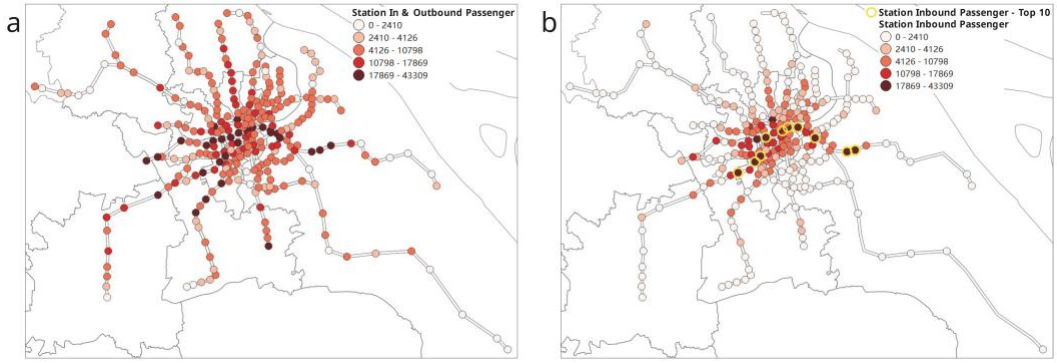


Figure 1. Spatial distribution of peak-hour station passenger of Shanghai rail transit network

Figure 1(b) shows the distribution of the peak-hour inbound passengers. Stations experience significant differences between inbound and outbound passengers during the morning peak because of the surrounding land use. Stations with a high volume of outbound passengers typically have more residential areas nearby, whereas stations with a large number of inbound passengers are likely to be located near workplaces. Regarding the inbound passengers, the city exhibits characteristics of a multicenter distribution of employment. Interestingly, not all of the top ten stations with the highest number of inbound passengers are situated in the downtown area, such as People's Square and Nanjing East Road. Some stations, like Caohejing Development Zone and Zhangjiang Hi-Tech, are located approximately 13 km away from the city center.

2.3 MSSN definition

The destinations of commuters are scattered over numerous rail transit stations. In addition to the central area, the periphery of the city also sees a significant number of inbound passengers during rush hours, indicating that job centers play a role in determining the direction of passenger flow. The imbalanced performance of passengers during the morning peak has similar distribution patterns in dense downtown areas around stations and along the metro lines on the outskirts.

Therefore, this study proposes a more comprehensive concept called the MSSN, in combination with the concepts of regional circular areas and rail corridors. The MSSN is a geographical range that is part of the entire urban rail network and is defined by a certain distance from a station. Each station has its own MSSN and the selection of the scale is explained in detail below.

The proposed MSSN is intended to better reflect the imbalanced performance of passenger flow at the regional level. To determine the geographical range of the MSSN, statistics on the travel distance of origin-destination (OD) passenger flow are used. Table 2 presents the distribution of passenger travel distances. The travel distance here was

obtained from the metro travel route planning service of the AutoNaviMap Web API. It is important to note that this distance is measured along the rail network, not in a straight line between stations.

Table 2. Rail travel distance between stations in the morning peak

	Min	Pct10	Pct25	Median	Mean	Pct75	Pct90	Max
Travel distance/km	0.6	5.0	8.4	13.5	15.3	20.0	27.5	115.4

For each pair of OD, the sum of passenger flows under different travel distances can be calculated to obtain Figure 2. This figure shows that the travel distance range with the highest passenger flow is 6–15 km, with 10 km having the highest passenger flow, followed by 12 km. Both of these distances are close to the median level (13.5 km) of travel distance. Based on these results, we select 6, 8, 10, 12, 14, and 16 km as the geographical range of the MSSN, and each station will have a corresponding MSSN under these spatial scales.

Taking 6 km-MSSN as an instance, the number and the distribution of stations within the range vary for three stations that in different locations (Figure 3). In the city center, the MSSN resembles a small regional rail network, while in the periphery, it takes the shape of a single-lane corridor (the colored area in Figure 3 is for illustrative purposes only).

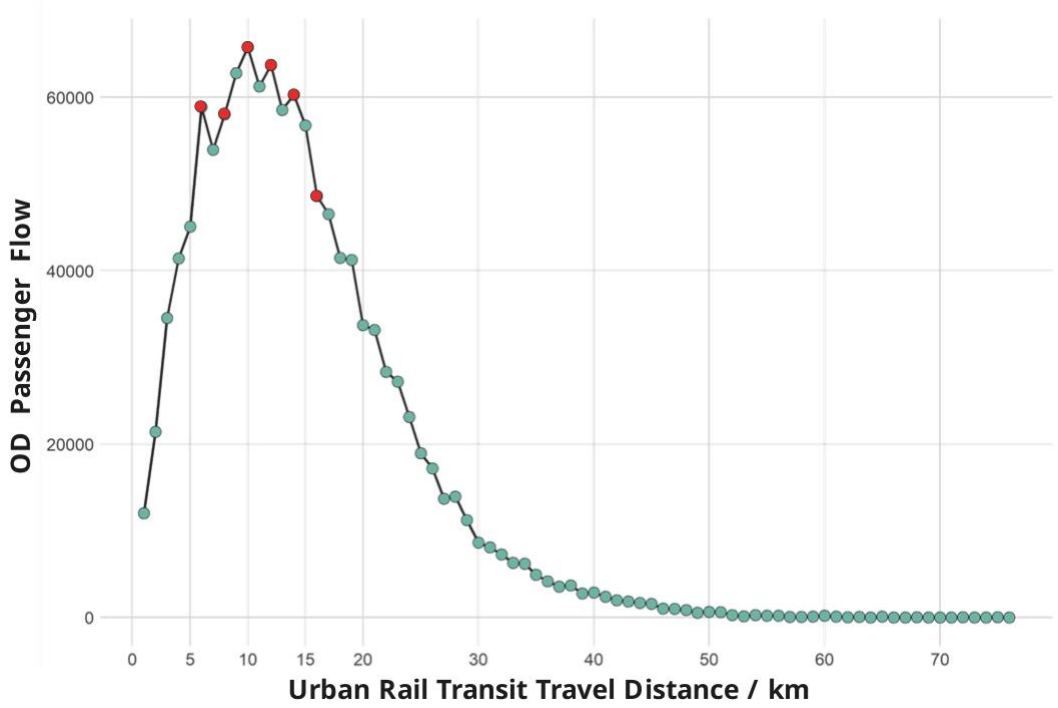


Figure 2. Distribution of OD passengers and rail transit travel distance

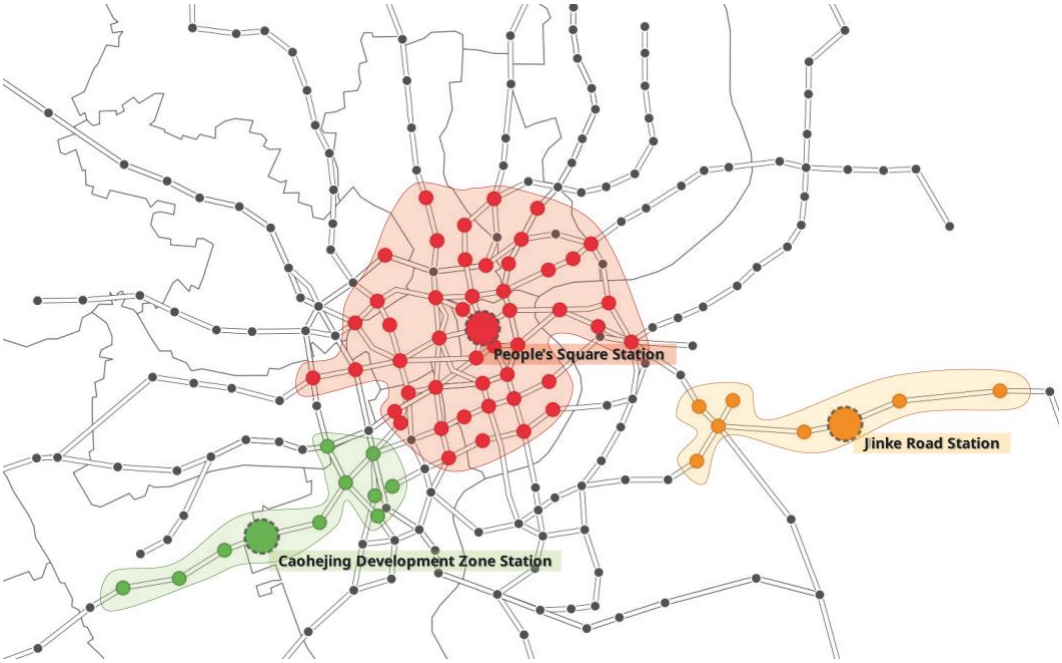


Figure 3. Example of 6 km-MSSN range for downtown and peripheral stations

3 Variables description

During peak hours, there is a noticeable spatial imbalance among passengers at train stations, often seen in the form of corridors or regional networks. In order to better understand the reasons for this imbalance, apart from the operation of the station itself, this study focuses on indicators related to the built environmental elements and network characteristics. These indicators will be used later in analysis models.

The traditional "5D" framework includes five categories of built environmental elements: density, diversity, design, distance to transit, and destination accessibility. For this study, the station is considered the primary spatial statistics unit, and then extended to the MSSN level to calculate corresponding coverage characteristics.

Additionally, network characteristics that contain the spatial information of the station and MSSN are also calculated. As previously mentioned, the location of a station within a city becomes increasingly important as the rail transit network expands and metro lines spread. The network characteristics can provide insights into the station's network structure, reflecting its stability and centrality, as well as its real-world location. Moreover, the station's opening year is included as a variable to measure its operational maturity. Table 3 summarizes all the variables used in this study. The columns "Station" and "MSSN" indicate whether the variables that can be statistically calculated around stations or within the MSSN ranges. These variables are also selected based on their accessibility during the research process. Their detailed explanations are as follows:

Table 3. Variables

"5D+N"	Variable description	Variable name	Station	MSSN
Network	Number of interchangeable metro lines	NumofLines	√	
	Betweenness centrality	Between	√	√
	Closeness centrality	Close	√	√
	Time to city center by transit	CTime	√	√
	Distance to city center by transit	CDis	√	√
Density	Population size	Ppl	√	√
	Number of jobs	Job	√	√
Diversity	Jobs-housing ratio	Jhr	√	√
Design	Walkable road network density	Road	√	√
Distance to transit	Number of bus stops	Bus	√	√
Destination accessibility	Time accessibility of populations	acs_PplMTime	√	√
	Time accessibility of workplaces	acs_JobMTime	√	√
Other characteristics	Opening years	Year	√	

3.1 Network

"Network" typically includes both the inherent network topology centrality and the travel time and distance to the city center, which correspond to the geographical location.

The concept of centrality in network analysis represents the degree to which a node is at the core of a network (Xia & Zhang, 2019). When applied to rail transit networks, centrality can measure the topological characteristics and importance of stations within the network. There are three main types of centralities: degree, betweenness, and closeness. Degree centrality refers to the number of connecting edges of a station to other stations, that is, the number of adjacent stations. This centrality is similar to the number of interchangeable metro lines at a station. Thus, this study only selects the latter, as it holds a more realistic meaning.

(1) Station level:

The number of interchangeable lines at a station is the number of lines on which the station is located within the rail transit network.

Betweenness centrality is the number of times a station is "passed" in each shortest path within the network, meaning it measures the importance of the station as an intermediary when connecting different stations. In terms of the network structure,

$$C_B(N_i) = \sum_{s \neq i \neq t} \frac{d_{st}(i)}{d_{st}}, \tag{1}$$

where d_{st} is the number of shortest paths from station s to station t, and $d_{st}()$ is the number of paths that pass through station i in the shortest path from station s to station t.

Closeness centrality is calculated as the reciprocal of the sum of the shortest paths from one station to all other stations. A high closeness centrality indicates that a station is an effective transportation link:

$$C_C(N_i) = \frac{n-1}{\sum_{i \neq j} d_{ij}}, \tag{2}$$

where d_{ij} is the shortest path distance between station i and station j.

Travel time and distance to the city center: As displayed in Figures 4(d) and 4(e), the population and job distribution follows a decreasing trend from the city center to the periphery. Considering this strong pattern, the geographical location of a station can also affects passenger flow. People's Square Station is defined as the city center here, which has the largest number of station passengers, The actual travel time and distance to People's Square Station by transit were obtained from the AutoNaviMAp WebAPI service.

(2) MSSN level:

The betweenness and closeness centralities are only calculated for stations within the MSSN itself and not for the entire network. Since these two types of centralities can already reflect the network connectivity within the corresponding spatial range to some extent, the MSSN level does not require interchangeable metro line variables.

The travel time and distance to the city center are consistent with the station level, meaning that the MSSN central station is used to represent the MSSN level for this variable.

3.2 Density

"Density" typically encompasses various factors such as population, jobs, floor area ratio, residential and workplace density, and commercial retail density. In this study, the population and number of jobs are chosen as the density variables.

Currently, the definition of the scope of the rail transit station area is mainly based on the walking distance, which is closely tied to the function of the rail transit system (Liu, 2017). It is assumed that urban residents are not willing to walk much longer distances (500 m) to reach the nearest station, and this distance can also be affected by weather conditions, road patterns, and other factors (Dittmar & Ohland, 2004). In practice, the standards for the catchment area tend to vary between cities. For example, in the US, the walking distance for the station area is mostly between 400-800 m, while in Shenzhen, China, it is between 400-700 m based on a survey (Hennigan, 2016; Wang et al., 2011). In Seoul, scholars have proposed a reasonable walking distance of 500 m for the impact area of a rail transit station (Sung et al., 2014; Sung & Oh, 2011). Similarly, Pan et al. (2017) found that a 500 m walking distance was significant in their study of passenger volume in Shanghai, as employment opportunities and other factors within the 500 m buffer zone were statistically significant. In terms of management policies, the Ministry of Housing and Urban-Rural Development of China recommends that the core area of a rail station should be within 500 m, and the distance between stations and the main commercial center and transport hub should be less than 500 m. In Shanghai, the planning and management of the rail station area typically follows a 500 m scale. Therefore, this study also takes a 500 m range to determine the different attributes around stations.

(1) Station level: The number of residents and jobs within 500 m of a station.

(2) MSSN level: The total number of residents and jobs, which is the sum of the corresponding numbers for all stations within 500 m of the MSSN.

3.3 Diversity

"Diversity" typically includes land use diversity (the proportion of residential, commercial, and industrial lands) and jobs-housing relationships. Considering that commuting is primarily related to residences and workplaces, this study selects the jobs-housing ratio (JHR) as a measure of diversity.

(1) Station level: The JHR is a ratio index of jobs to the population within a 500 m radius of the station.

$$JHR(S_i) = \frac{J_i}{\sum_{j=1}^{n} J_j} / \frac{P_i}{\sum_{j=1}^{n} P_j}, \quad (3)$$

where P_i is the number of residents living within 500 m around the station i, and the J_i is the number of available jobs within the same radius.

(2) MSSN level: The JHR is a ratio index of jobs to population within the MSSN range.

$$JHR(CSN_i) = \frac{J_{csni}}{\sum_{j=1}^{n} J_j} / \frac{P_{csni}}{\sum_{j=1}^{n} P_j}, \quad (4)$$

where P_{csni} is the sum of residents living within 500 m around all stations in the MSSN (with the ith station as MSSN center) range, and the J_{csni} is the sum of available jobs within the same range.

3.4 Design

"Design" typically includes a walkable environment, intersection density, total road length, density of public spaces, average building size, station parking availability, and main and road mileage. In this study, the density of a walkable road network is chosen as the design variable. The road network data was obtained from the online geographic database OpenStreetMap.

(1) Station level: The density of the walkable road network within 500 m and 1 km of a station.

(2) MSSN level: The density of the walkable road networks around all stations within the MSSN range.

3.5 Distance to transit

"Distance to transit" typically refers to the distance to the transit station and the number of bus stops or lines. This study chooses the number of bus stops as the variable. This data was collected using the AutoNaviMap Web API service.

(1) Station level: The number of bus stops within 500 m and 1 km of the station.

(2) MSSN level: The total number of bus stops around all stations within the MSSN range.

3.6 Destination accessibility

"Destination accessibility" typically includes accessibility to jobs, points of interest, educational institutions, and public safety facilities. Since commuting is primarily linked to residential and workplace locations, this study focuses on the time accessibility of the residential population and the time accessibility of jobs as the key variables for destination accessibility. The time-related data below was also obtained using the AutoNaviMap Web API service.

(1) Station level: Time accessibility of the population is determined by considering the number of residents living around the station and the travel time between stations. This variable reflects the attractiveness of a station for outbound travel.

$$A(N_i) = \sum_{i \neq j} \frac{P_j}{t_{ij}}, \quad (5)$$

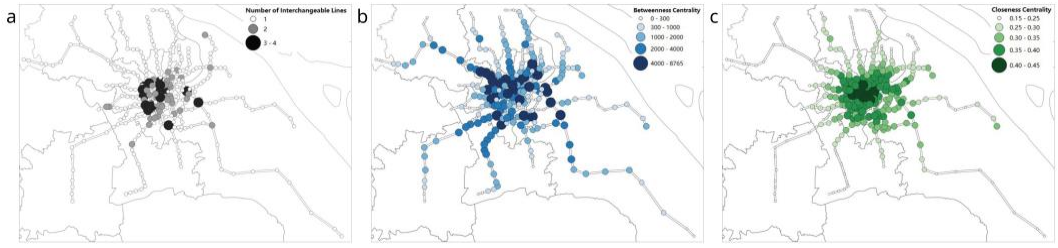
where P_i is the number of residents living within 500 m around the station i, and the t_{ij} is the travel time from station i to station j.

Job time accessibility is calculated based on the number of jobs located around a station and the travel time between stations. This variable reflects the attractiveness of a station for inbound travel.

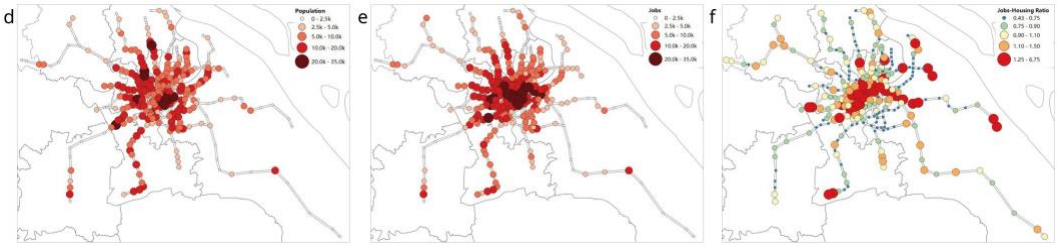
$$A(N_i) = \sum_{i \neq j} \frac{J_j}{t_{ij}}, \quad (6)$$

where the J_j is the number of available jobs within 500 m around the station j.

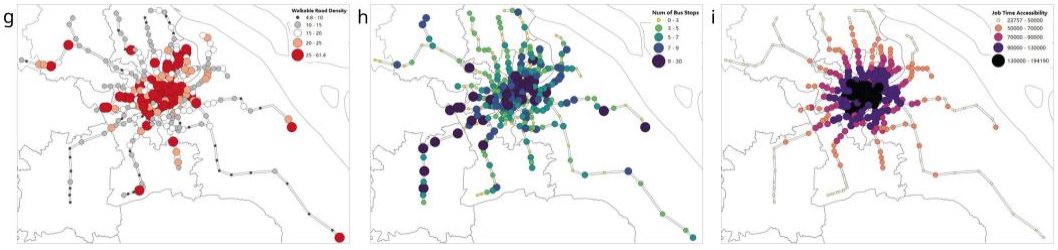
(2) MSSN level: The time accessibility values of population and jobs are calculated based on all stations within the MSSN and corresponding job centers of different stations, rather than on a network-wide basis. The entire process is described in Section 4.2.1.



Network characteristics: (a) Number of interchangeable metro lines; (b) Betweenness centrality; (c) Closeness centrality



Built environment characteristics: (d) Population; (e) Jobs; (f) Jobs-housing ratio



Built environment characteristics: (g) Density of walkable road; (h) Number of bus stops; (i) Job time accessibility

Figure 4. Example visualization of station level variables

3.7 Other

"Other" variable represent the age at which the station was opened.

(1) Station level: Opened years is the number of years since the station was put into operation reckoned from 2016.

(2) MSSN level: The stations vary greatly in terms of their opening year, and the average opened year has no strong explanatory meaning. Therefore, this variable is not taken into consideration in the MSSN.

4 Analysis of station passenger and flow imbalance

4.1 Station passenger analysis

Traditional linear regression, which has moderate predictive accuracy, is frequently used in research on station passengers. However, this model assumes a relatively stable relationship between passengers and other factors, treating variables as global variables (Gan, 2019). The transit stations investigated in this study are distributed in different geographical areas, but many of them are close together and have similar built environment, network characteristics, and even passenger flows. This suggests that the variables of stations between adjacent regions may have an impact on station passengers, indicating spatial heterogeneity. The ordinary least squares (OLS) regression does not consider the distance to the study object. Therefore, this study utilizes geographically weighted regression (GWR) to determine the correlation between variables and station passengers, and to analyze whether the selection of each variable can reflect the operating conditions of the station and at what spatial scale these variables influence station passengers. The GWR algorithm considers the influence of adjacent station variables when performing regression. GWR has already been widely used in passenger flow research and has shown better explanatory performance compared to traditional regression models (Gan, 2019; Qi & Hu, 2021). In GWR, the bandwidth is an important parameter for determining the weight of variables. To determine the optimal bandwidth for different stations, the adaptive bandwidth under the Akaike information criterion is used in this study.

Considering that transportation hub stations may have strong traffic attributes that can affect passenger flow, the Shanghai Railway Station, Hongqiao Terminal 1, Hongqiao Terminal 2, Hongqiao Stations, Hongqiao Railway Station, Shanghai South Railway Station, and Pudong International Airport were excluded from the analysis. In addition, stations with passengers below the 10% quantile were also excluded here.

The station and MSSN have many variables that may have similar impacts on passenger flow. Therefore, principal component analysis was conducted before performing GWR, which can replace some independent variables with several unrelated comprehensive variables. Taking the station-level variables as an example, six component variables were sufficient to explain 90.61% of the information, and the important variables in these components were used for regression analysis (Table 4).

Table 4. Matrix of principal components of the station level variables

Variables	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
NumofLines	0.244	0.510	0.116	0.070	0.272	0.099
Year	0.175	0.218	0.709	-0.258	-0.562	0.037
Ppl500m	0.182	-0.197	-0.065	0.594	-0.328	0.290
Job500m	0.329	-0.199	0.199	0.510	-0.003	-0.021
Jhr500m	0.084	0.018	0.201	-0.107	0.247	-0.223
Close	0.404	0.105	-0.363	-0.198	-0.001	-0.061
Between	0.165	0.648	-0.045	0.307	0.170	-0.049
CDis	-0.239	0.045	0.203	0.213	0.226	-0.013
CTime	-0.333	0.059	0.234	0.260	0.159	0.025
acs_PplMTime	0.362	-0.035	-0.140	-0.047	-0.089	-0.008
acs_JobMTime	0.366	-0.039	-0.075	-0.053	-0.044	-0.042
Road500m	0.201	-0.214	0.252	-0.162	0.397	0.443
Bus500m	0.115	-0.158	0.172	0.053	0.143	-0.527
Road1km	0.214	-0.225	0.168	-0.124	0.360	0.329
Bus1km	0.196	-0.232	0.151	0.089	0.136	-0.514
Eigenvalue	0.299	0.054	0.043	0.035	0.030	0.019

When using the OLS method, the variables that have significant impacts on station passengers are the number of bus stops, number of interchangeable lines, centrality, and number of years opened. In contrast, the GWR method allows for the value of a variable to vary based on geographical location, resulting in that the variables of population, jobs, and road density to play an explanatory role in the regression model.

Table 5. Regression results and used variables at the station and different MSSN level

Regression	500m	6km	8km	10km	12km	14km	16km
OLS-R^2	0.617	0.681	0.668	0.634	0.670	0.663	0.657
GWR-R^2	0.752	0.793	0.797	0.791	0.784	0.790	0.798
Station Variables included	NumofLines, Year, Ppl500m, Job500m, Close, Between, Bus500m						
MSSN Variables included	-	Ppl6km, Close6km	CTime, acs_JobMTime8km, Close8km	Ppl10km, Road10km, Between10km	Job12km, Cose12km, Between12km	Job14km, Close14km, Between14km	Job16km, Close16km, Between16km

As shown in Table 5, compared with the traditional least squares regression, the GWR model better reflects the station passengers. After adding MSSN-level variables, particularly those related to jobs, population, and centrality, the interpretation is further strengthened. Moreover, the R^2 values of the 8 km and 16 km spatial scales in the GWR regression are relatively high. This highlights the importance of considering the characteristics of an 8 km-MSSN surrounding the station in addressing the imbalance of the rail transit network.

4.2 MSSN imbalance performance

4.2.1 Data preparation

Taking the MSSN defined in Section 2.3 as the analysis object, the balance of peak-hour passenger flow from a region-wide perspective was evaluated and the correlation between the balanced performance and variables was further investigated. In terms of passenger spatial balance, existing studies usually consider a station, a line section, or a single line as the object and use parameters such as mean, variance, standard deviation, or the Gini index to quantify the passenger difference between stations or upstream and downstream. Considering that an MSSN can be spatially viewed as multiple line sections, this study proposes the following MSSN balance index inspired by the concepts of upstream and downstream passenger flows. This index measures the difference between centripetal and centrifugal passenger flows in regional areas with job centers as the core. The closer the index value to 1, the more balanced the passenger flow interactions between stations within the MSSN and outside; conversely, a lower value indicates that the flow interactions are uneven:

$$B_{(i,scope)} = \frac{InFlow_{(i,scope)}}{OutFlow_{(i,scope)}}, \tag{7}$$

where $InFlow_{i,scope}$ is the passenger flow from stations within MSSN (with the ith station as the center and $scope$ as the geographical range) to each stations' corresponding job center; whereas $OutFlow_{i,scope}$ is the passenger flow from each stations' corresponding job center to itself, and the stations here are also within MSSN (with the ith station as the center and $scope$ as the geographical range).

According to the number of inbound passengers during peak hours and the geographical locations, Lujiazui (urban central business district), Xujiahui (urban subcenter), and Jinke Road Station (suburban employment center) were selected to represent the employment centers in Shanghai, as illustrated in Figure 5. These three job centers were chosen to gain a better understanding of the imbalance between different regions (Figure 6) rather than calculating the actual job center for each station. The balance index was calculated for each MSSN to these three job centers, and the one with the most imbalanced performance (the absolute value of the difference from 1 is the largest) was considered the final job center for the corresponding MSSN.

Moreover, the "destination accessibility" variable, which includes the time accessibility of populations and workplaces, should be calculated based on the job center. Taking the time accessibility of population as an example:

$$A(N_{(i,\,scope)}) = \sum_{(i \in CSN)} (P_{JobCenter\ (i,scope)} / t_{JobCenter\ (i,scope)}), \tag{8}$$

where $P_{Job\_Center(i,scope)}$ is the population within 500 m of each stations' corresponding $Job\_Center_{i,scope}$, and $t_{i,Job\_Center_{i,scope}}$ is the transit travel time from stations within the MSSN (with the ith station as the center and $scope$ as the geographical range) to the corresponding $Job\_Center_{i,scope}$.

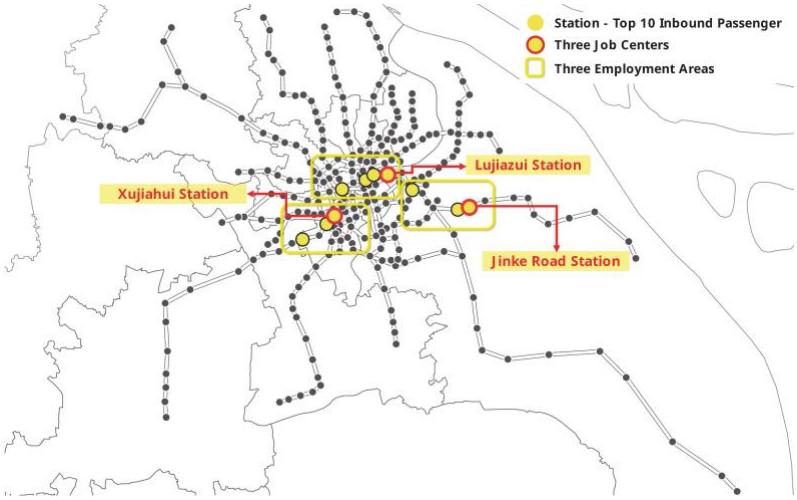


Figure 5. Distribution of the three selected job centers on Shanghai rail transit network

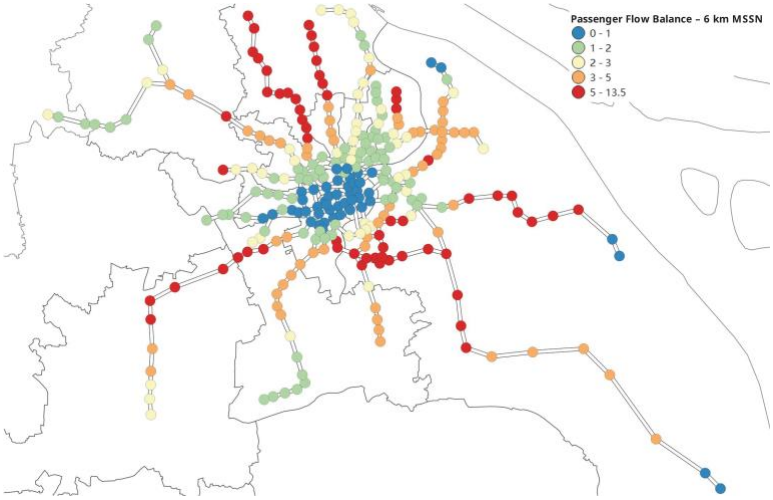


Figure 6. Imbalance performance of peak-hour passenger flow of 6 km-MSSN

4.2.2 Results

The MSSN is a specific region along a rail transit network, it is possible for subnetworks within this region to overlap in real-world situations. In such cases, variables such as network characteristics are calculated internally for the MSSN, which are reflective of the characteristics of the adjacent regions. As a result, traditional linear regression is utilized instead of GWR.

Take 6 km-MSSN for example, the findings presented in Table 6 indicate that factors such as population, jobs, jobs-housing ratio, closeness centrality, road density, travel time, and distance to the city center are most likely to influence passenger flow balance. As the spatial scale increases, the significant variables shift to primarily include the jobs-housing ratio, road density, and number of bus stops. This suggests that key factors that vary with geographical scale include the total number of population and jobs, the

convenience of traffic conditions, the number of transportation facilities, and the jobs-housing balance.

Table 6. Multiple linear regression results of 6 km-MSSN passenger flow balance

Variables	Coefficient	P.value	Conf.low	Conf.high
Ppl6k	-2.6E-05	4.13E-06	-3.7E-05	-1.5E-05
Job6k	2.16E-05	2.24E-06	1.28E-05	3.04E-05
Jhr6k	-8.09103	2.38E-18	-9.78733	-6.39472
Close6k	5.043495	0.000301	2.331355	7.755635
Between6k	0.002677	0.336324	-0.0028	0.00815
CDis	-0.05459	0.174383	-0.13352	0.024329
CTime	-0.06139	0.016124	-0.11131	-0.01147
acs_PplMTime6k	1.96E-05	0.636008	-6.2E-05	0.000101
acs_JobMTime6k	5.83E-06	0.48915	-1.1E-05	2.24E-05
Road6k	-9.6E-06	0.019887	-1.8E-05	-1.5E-06
Bus6k	0.003116	0.714823	-0.01366	0.019888
Intercept	15.44627	5.71E-32	13.18247	17.71008

The passenger flow balance cannot be reflected well by these variables within 8–12 km. The R^2 change rate, as displayed in Table 7, exhibits a significant downward trend from 6 km to 8 km, but bouncing back at 14 km, and the MSSNs at 6 km and 14 km have similar interpretation performance ($R^2 \approx 0.52$). While the R^2 value at 16 km is the highest, it is important to note that the average rail travel distance is only 15 km (Table 2). A 16km-MSSN can include hundreds of downtown stations in the MSSN, making it less useful for smaller-scale analysis. Hence, for a more effective interpretation of passenger flow interactions between different regions around rail transit networks, we recommend using a MSSN of 6 km and focusing on the important variables mentioned above at the regional level.

Table 7. Multiple linear regression results of MSSN passenger flow balance

MSSN	6km	8km	10km	12km	14km	16km
R2	0.521	0.473	0.459	0.476	0.522	0.560
R2 change	-	-9.21%	-2.96%	3.70%	9.66%	7.28%

5 Conclusion and future work

With the rapid construction and development of urban rail transit in China, the geographical scale of transit has rapidly produced networks instead of independent lines. However, the asynchronous increase in rail transit networks and ridership aggravates the imbalanced passenger flow during peak hours. This serious imbalance not only affects the passenger experience but also causes a wastes of construction resources. Based on the morning peak-hour station passenger data of Shanghai rail transit, this study established the Multiscale Subnetwork (MSSN) to analyze the factors of station passenger and regional flow balance. The following are the findings:

(1) The passenger flow distribution in this large-scale rail transit network is extremely uneven. Combined with the inbound passenger data, it was found that the commuting destinations exhibit a multi-center characteristic. This indicates that the studying of the

built environmental variables and network characteristics around a single station may not be sufficient to address the imbalance in ridership. Therefore, perspectives from different spatial scales need to be considered.

(2) The spatial scope of the MSSN was proposed and the MSSN-wide attributes were studied in two issues.

For station passenger analysis, the regional characteristics of an 8 km MSSN have a significant impact on station passengers. Because of the similarity of nearby stations, in addition to the centrality, number of transferable lines, time accessibility, and number of bus stops that will affect the station passengers under traditional linear regression, GWR can enhance the explanatory effects of population, jobs, and road density. The jobs-housing ratio in a multiscale spatial network also significantly influences station passengers.

For MSSN passenger flow imbalance analysis, the imbalance between upstream and downstream passenger flows during peak hours is tenfold. In addition, the study found that with the varying ranges of MSSNs, the impact of these attributes on passenger flow balance also changes. However, in general, an MSSN range of 6 km is suggested to guide the balance analysis for passenger flow in combination with urban elements within the range. The population, the number of jobs, and the location of MSSNs in the entire network are most likely to impact the balancing performance of passenger flow.

(3) During the peak hours in Shanghai, there are still many stations with relatively few inbound and outbound passengers (Figure 1). These stations are usually located on the outskirts, and their connectivity with other stations is relatively poor. Although several stations near the city center have slightly higher centrality, the surrounding road network density and bus levels remain low. Stations with imbalanced passenger flow exhibit similar characteristics. To improve these stations and areas with uneven passenger flows, it is necessary to study the combination of networks and land construction around stations. At the network level, instead of expanding outward, the city planners can consider increasing downtown network density to alleviate the imbalanced flow. At the operation level, commuter express lines, hop-off stops during peak hours, and integrating operation services with multi-mode buses at the planning stage may be considered. Most importantly, these measures should be formulated from a multiscale area network perspective, rather than just individual stations.

In this study, there are also some limitations when exploring the imbalance of rail networks, such as the lack of an in-depth discussion of the imbalance within the MSSN and other imbalance definitions. In future work, we plan to expand the analysis by incorporating passenger data from a wider range of time periods and taking into account the mixed land-use conditions. Moreover, as the public transportation system in Shanghai continues to improve, rail transits may not fully reflect the behavioral characteristics of public transportation. With the complex networks of both bus and rail transits, the correlation between the built environment and ridership should be further explored. This could provide more comprehensive decision-making support for regional public transport and multimode transportation planning.

Acknowledgments

This research is part of the NSFC project about the optimization of livable transit-oriented development with a large rail transit network. The project is supported by the National Natural Science Foundation of China (No. 51778431), the Key Laboratory of Ecology and Energy-saving Study of Dense Habitat and Urban Mobility Institute of Tongji University.

Author contribution

Haixiao Pan: Conceptualization, supervision, project administration, funding acquisition, writing (review and editing). Miao Hu: Data curation, methodology, formal analysis, writing. Xiyin Deng: Conceptualization, investigation, writing. Ailing Liu: Conceptualization, investigation, writing.

All authors have read and approved the published version of the manuscript.

Data availability

The data that support the findings of this study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.25904512.v2.

References

Andersson, D. E., Shyr, O. F., & Yang, J. (2021). Neighborhood effects on station-level transit use: Evidence from the Taipei metro. *Journal of Transport Geography, 94,* 103127.

Börjesson, M., Rushid, A. R., & Liu, C. (2021). The impact of optimal rail access charges on frequencies and fares. *Economics of Transportation, 26–27,* 100217.

Canca, D., Barrena, E., Laporte, G., & Ortega, F. A. (2016). A short-turning policy for the management of demand disruptions in rapid transit systems. *Annals of Operations Research, 246*(1), 145–166.

Ceapa, I., Smith, C., & Capra, L. (2012). Avoiding the crowds: Understanding tube station congestion patterns from trip data. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 134–141.

Cervero, R. (2007). Transit-oriented development's ridership bonus: A product of self-selection and public policies. *Environment and Planning A: Economy and Space 39*(9), 2068–2085.

Cervero, R. (2006). Alternative approaches to modeling the travel-demand impacts of smart growth. *Journal of the American Planning Association, 72*(3), 285–295.

Dittmar, H., & Ohland, G. (2004). *The new transit town: Best practices in transit-oriented development.* Chicago: Bibliovault OAI Repository, University of Chicago Press.

Ewing, R., & Cervero, R. (2001). Travel and the built environment: A synthesis. *Transportation Research Record, 1780*(1), 87–114.

Fang, H. (2021). Application of big data in Beijing rail transit network planning. *Urban Rapid Rail Transit, 34*(1), 37–44.

Gan, Z. (2019). Impact of built environment on the passenger flows and transfer behavior of urban rail transit (PHD thesis), Southeast University, Nanjing, China.

Hennigan, M. F. (2006). Station area access within transit-oriented development: A typological analysis (Master's thesis), University of Texas at Austin, Austin, TX.

Li, S., Lyu, D., Liu, X., Tan, Z., Gao, F., Huang, G., & Wu, Z. (2020). The varying patterns of rail transit ridership and their relationships with fine-scale built environment factors: Big data analytics from Guangzhou. *Cities, 99,* 102580.

Liu, Y., Bao, Z., & Tian, W. (2022). Station-city integrate development path and practice in Guangzhou TOD: Multi-level spatial governance and collaborative planning. *Planners, 38*(2), 5–15.

Liu, L., Zhang, M., & Xu, T. (2020). A conceptual framework and implementation tool for land use planning for corridor transit oriented development. *Cities, 107,* 102939.

Liu, Q. (2017). Factors affecting circle structure of TOD concentric models. *Urban Planning International, 32*(5), 72–79.

Pan, H., Li, J., Shen, Q., & Shi, C. (2017). What determines rail transit passenger volume? Implications for transit oriented development planning. *Transportation Research Part D: Transport and Environment, 57,* 52–63.

Peng, S., Chen, S., Xu, Q., & Niu, J. (2021). Spatial characteristics of land use based on POI and urban rail transit passenger flow. *Acta Geographica Sinica, 76*(2), 459–470.

Qi, C., & Hu, H. (2021). Research on ridership forecast of urban rail transit station based on mixed geographic weighted regression. *Journal of Railway Science and Engineering, 18*(7), 1903–1909.

Su, X. (2018). Problems of Tokyo Metro delays and its countermeasures. *Modern Urban Rail Transit, 11,* 80–83.

Su, S., Zhao, C., Zhou, H., Li, B., & Kang, M. (2022). Unraveling the relative contribution of TOD structural factors to metro ridership: A novel localized modeling approach with implications on spatial planning. *Journal of Transport Geography, 100*(6), 103308.

Sung, H., Choi, K., Lee, S., & Cheon, S. (2014). Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership. *Journal of Transport Geography, 36*, 134–140.

Sung, H., & Oh, J. (2011). Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea. *Cities, 28*(1), 70–82.

Wang, J., Zheng, X., & Mo, Y. (2011). Establishment of density zoning and determination of floor area ratio along rail transit line based on TOD: A case study on rail transit line 3 in Shenzhen. *City Planning Review, 35*(4), 30–35.

Xia, Z., & Zhang, Y. (2019). From "5D" to "5D+N": Research published in English on the factors influencing TOD performance. *Urban Planning International, 34*(5), 109–116.

Zeng, L., Liu, J., Qin, Y., Wang, L., & Yang, J. (2018). A Passenger flow control method for subway network based on network controllability. *Discrete Dynamics in Nature and Society, 6*, 1–12.

Zhao, X., Wu, Y., Ren, G., Ji, K., & Qian, W. (2019). Clustering analysis of ridership patterns at subway stations: A case in Nanjing, China. *Journal of Urban Planning and Development, 145*(2), 04019005.