Article

Exploring Spatiotemporal Variation in Hourly Metro Ridership at Station Level: The Influence of Built Environment and Topological Structure

Zhuangbin Shi [1] **, Ning Zhang** [1,*]**, Yang Liu** [2,3] **and Wei Xu** [4]

[1] Intelligent Transportation System Research Center, Southeast University, Southeast University Road 2, Nanjing 211189, China; shizhuangbin@gmail.com
[2] School of Transportation, Southeast University, Southeast University Road 2, Nanjing 211189, China
[3] Urban Planning Group, Eindhoven University of Technology, 5600 MB Eindhoven, the Netherlands; kmliuyang90@gmail.com
[4] School of Automation, Southeast University, Sipailou 2, Nanjing 210096, China; 220171541@seu.edu.com
[*] Correspondence: ningzhang1972@gmail.com or 101004970@seu.edu.cn; Tel.: +86-13814049777

check for updates

Abstract: Reliable and accurate estimates of metro demand can provide metro authorities with insightful information for the planning of route alignment and station locations. Many existing studies focus on metro demand from daily or annual ridership profiles, but only a few concern the variation in hourly ridership. In this paper, a geographically and temporally weighted regression (GTWR) model was used to examine the spatial and temporal variation in the relationship between hourly ridership and factors related to the built environment and topological structure. Taking Nanjing, China as a case study, an empirical study was conducted with automatic fare collection (AFC) data in three weeks. With an analysis of variance (ANOVA), it was found that the GTWR model produced the best fit for hourly ridership data compared with traditional regression models. Four built-environment factors, namely residence, commerce, scenery, and parking, and two topological-structure factors, namely degree centrality and closeness centrality, were proven to be significantly related to station-level ridership. The spatial distribution pattern and temporal nonstationarity of these six variables were further analyzed. The result of this study confirmed that the GTWR model can provide more realistic and useful information by capturing spatiotemporal heterogeneity.

Keywords: hourly station-level ridership; spatiotemporal variation; geographically and temporally weighted regression; built environment; topological structure

1. Introduction

Due to the competitive advantages in huge capacity, clean energy, and conservation of land, the metro has been widely perceived as a preferred public transport mode for major metropolitans [1,2]. With the emerging traffic congestion and overpopulation aggravated by rapid urbanization and motorization, a growing number of Chinese cities have taken steps to build well-connected metro networks. For those middle or long travel-distance individuals, the metro service is more attractive compared to other public transportation modes for travelers because of its high speed, reliable schedule and comfortable running smoothness. As a consequence, it can promote a significant modal shift to reduce the heavy use of cars [3,4]. Although the construction of metro lines has turned out to be an effective measure to alleviate transportation pressure and improve city image [5], some problems have occurred during the development and expansion of metro networks, such as uneven distribution of passengers [6], overcrowded carriages [7], and increased security risk [8]. A particular source of these problems is the unreasonable planning of route alignment and station location.

Ridership demand forecasting is a vital component for the analysis of project viability and sustainability in metro planning [9]. A traditional method, namely the four-step model, has been widely used for travel demand forecasting and transportation infrastructure planning because of its universal applicability [10]. Since the four-step model requires a large amount of activity surveys and complex modeling processes [11,12], many recent studies adopted an alternative approach by investigating the regression relationship between the ridership demands and the characteristics of metro stations [9,12–18]. Most of those studies focus on the ridership demand of metro stations in daily or annual profiles, but only a few concern the variation in hourly ridership demand. The hourly ridership demand, especially the ridership demand during peak periods that often results in heavy crowding in metro systems of many Chinese cities, can provide some considerable new insights for metro planning. This paper will attempt to address this gap with a case study from Nanjing, China.

In a recent study, Zhao et al. built an ordinary least squares (OLS) multiple regression model to investigate the influence of land utilization, external connectivity, intermodal connection, and station type on average weekday ridership at station level with data from Nanjing metro [15]. The results were compared with the outcomes of regression models developed for cities in the United States [13] and Seoul [14], and it was found that both the R^2 value and the significance of variables vary greatly across different cities. Another experiment in the Madrid Metro network demonstrated that the correlation between station-level transit ridership and variables of the characteristics have significant differences in space. Also, the geographically weighted regression (GWR), which has advantages in measuring spatial instability, manifested its superiority over the traditional OLS model [9]. A similar spatial autocorrelation across different metro stations was found in commercial property value in Wuhan [19] and metro-bikeshare transfer in Nanjing [20]. Besides, the short-term metro ridership was proven to be autocorrelative in temporal scales in many studies [17,21], which means that the hourly ridership in contiguous periods could present analogous correlations with variables of the characteristics.

The objective of this study is to investigate the spatial and temporal variations in hourly metro ridership demand at station level. The rest of this paper is organized as follows. The next section reviews existing literature of the influence factors on metro ridership and the direct forecasting models. Section 3 describes the methodology used in this paper. Section 4 is an introduction to the study area and the data set. Modelling results analysis and discussion are presented in Section 5 and our conclusions are presented in Section 6.

2. Literature Review

2.1. Characteristics Affecting Ridership

It is essential for ridership forecasting to explore what and how factors affect metro ridership at station level. In past decades, previous studies found that the station-level ridership demand in metro systems is significantly associated with the characteristics of the station environment [9,13,14,16,22–25]. Characteristics affecting station-level ridership mainly include built environment variables, socio-economic variables, connectivity with other metro stations and other traffic modes. Built environment is usually measured by building uses and building gross amount or density. The typical built uses can be categorized into: residence, service facilities, companies/offices, attractions, educations, hotels, bus stops, roads, and parking lots [26]. The built environment is thought of as the crucial driver of transit ridership [12]. Heavy land use means that there are more people living and/or working in the walking area around metro stations [27]. In general, the metro stations with a higher land use level in neighborhood areas, such as retail, hotel, education, and bus routes, should be more likely to aggregate crowded passenger flow; nevertheless it is not an incontestable truth. For instance, the station-level ridership will decrease with increasing storage area according to Zhang and Wang [25], and large-scale commercial building density has a negative correlation with station-level ridership in Seoul metro while taking 1 km as service boundary [24].

Many studies have found that the population and employment within walking area are important factors and have positive correlations with metro ridership [9,12,13,15]. According to the results of Kuby et al., lower income might lead to increasing metro use [13]. In terms of the Rio de Janeiro metro, metro ridership presents a stronger correlation with the number of jobs than the average income and population [16]. Apart from population, employment and income, some other socioeconomic characteristics, such as age group, gender, ethnicity, and unemployment rate are also considered to be correlated with transit ridership [28–30].

The influence of connectivity with other metro stations is usually quantified with the topological features of network structure. For example, the distance of the nearest station was found to be negatively associated with station-level ridership [16,31]. A more comprehensive research conducted by Sohn and Shim [14] introduced several external connectivity features: closeness centrality, betweenness centrality, straightness centrality, and average number of transfers calculated respectively for both the metro and highway networks. Taking the Seoul metro as a case study, they found that only the closeness centrality and average number of transfers have significant impacts on metro ridership at a station-level, and both the associations are negative. Regarding the impact of connectivity with other traffic modes, it can be counted by some indicators of built environment, such as feeder bus lines/stops, road length, and number of parking lots [13–15].

Due to the socioeconomic, cultural and political specificities of Nanjing (e.g., large amount of floating population), accurate socio-economic variables within the walking area of metro stations are difficult to access. On the other hand, despite the fact that population and employment are not synonymous with the density or type of built environment, the built environment variables can be used as indicators of socioeconomic status. For example, high residential building density can be attributed to more residents around stations. Considering the accessibility of data, we mainly explored the correlation between station-level ridership with the variables of built environment and topological structure in this paper.

2.2. Direct Ridership Models

The history of demand modeling has been dominated by the four-step model [10]. The four-step model is developed by formulating the process as sequential and interrelated models (trip generation, distribution, mode choice and route choice). It has experienced two stages: activity-based and trip-based. Either of the two versions requires an enormous database and is designed for regional scales. Thus, the four-step model is less effective for forecasting traffic demand on station-level scales [9]. Regression-based direct ridership models, as a complementary method of the four-step model, are able to formula the relationship between ridership and station-level characteristics [32]. This method is calibrated with data that are easily accessible, such as smart card data and land use density. It makes direct ridership models relatively concise and inexpensive compared with the four-step model.

Linear regression is a statistical analysis method used to determine the linear quantitative relationship between two or more variables. It can be categorized into two general types: global and local regression models [9]. OLS regression is the most representative global model for revealing the influence of various factors on metro ridership [13,22,31]. It is based on a hypothesis that the prediction errors for all samples are independent. However, based on Tobler's First Law of Geography [33], the spatial data normally present similar patterns with short distances. The metro ridership has been proven to be autocorrelated in spatial scales according to the results of Cardozo et al. [9] and Jun et al. [18], who introduced the GWR model to station-level ridership forecasting.

Instead of the constant parameters of the global regression model, the parameters of GWR will change with the samples' position to capture spatial variations [34]. Taking the version of the OLS model, the GWR model can be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, i = 1, 2, \cdots n \tag{1}$$

where the spatial coordinates of the metro station i are denoted as (u_i, v_i); y_i, x_{ik} and ε_i are the ridership, the kth explanation variable and the error term for the metro station i, respectively; $\beta_0(u_i, v_i)$ and $\beta_k(u_i, v_i)$ represent the regression parameters at metro station i, which are allowed to vary across space.

As summarized by Cardozo et al. [9], the GWR model has some important advantages, which mainly refer to greater detail and accuracy, stronger explanatory power, smaller estimation errors, place-based decision support and measuring the degree of spatial similarity. Although the GWR model presents a significant superior capability in determining the spatial dependencies of station-level ridership, it is not adequate to model spatiotemporal data because it needs to aggregate or average the time-scale data by a certain period [26], such as average weekday boardings investigated in most previous studies on station-level ridership. A novel approach proposed by Huang et al. [35], the geographically and temporally weighted regression (GTWR), can offer a better improved fit from a new perspective by considering the spatial and temporal heterogeneity compared with the traditional OLS or GWR models. It has been applied to some certain topics, for example, housing price [35] and environment [36,37]. In a notable research, Ma et al. explored the influence of built environment on transit ridership using GTWR model [26]. However, his work focused on bus ridership in the region, in contrast to our study of metro stations.

3. Methodology

The GTWR model, which is an extend version of the GWR model, can take consideration of both spatial and temporal nonstationarity in real data. Thus, the GTWR model can presented as:

$$y_i = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i)x_{ik} + \varepsilon_i, i = 1, 2, \cdots n \tag{2}$$

where t_i is the index of observing time interval of metro ridership.

The estimates of regression parameters can be shown as follows:

$$\hat{\boldsymbol{\beta}}(u_i, v_i, t_i) = \left(\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i, t_i)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i, t_i)\boldsymbol{y} \tag{3}$$

where $\boldsymbol{W}(u_i, v_i, t_i)$ is an $n \times n$ matrix, $\boldsymbol{W}(u_i, v_i, t_i) = diag(w_{i1}, w_{i2}, \cdots, w_{in})$, whose diagonal elements denote weights based on the definition of the space–time distance to observation i, and its off-diagonal elements are zeros.

Considering that location and time usually have different scaling effects, we transfer the temporal distance into an additional spatial-scale distance with a time-space scale factor. If the Euclidean distance and absolute time difference are adopted as spatial distance and temporal distance, the spatiotemporal distance D^{ST} between observation i and observation j can be formulated as a linear combination of the spatial distance D^S and the temporal distance D^T.

$$D_{ij}^{ST} = D_{ij}^S + \tau D_{ij}^T = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} + \tau|t_i - t_j| \tag{4}$$

where the value of time-space scale factor τ means that the decay of weight for the increase in temporal distance by a unit (hours) equates to that for the increase in spatial distance by τ unit (kilometers). Actually, the GWR model is a special case of the GTWR model with the parameter $\tau = 0$.

The diagonal elements of weighting matrix $\boldsymbol{W}(u_i, v_i, t_i)$ are calculated according to a weighting function formulated with the spatiotemporal distance and a kernel function. An eligible kernel function should ensure that neighboring observation points from the spatiotemporal data are allocated relatively larger weighting value. The most common weighting kernel is fixed Gaussian-based function, which can be expressed as:

$$w_{ij} = exp\left(-\left(\frac{D_{ij}^{ST}}{h}\right)^2\right) \tag{5}$$

where *h* denotes a fixed parameter of spatio-temporal bandwidth.

A cross-validation (CV) minimization procedure is usually conducted to select the suitable parameters of time-space scale factor and bandwidth [26,34,37].

$$\text{Minimize} \; CV(\tau, h) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(\tau, h) \right)^2 \tag{6}$$

where $\hat{y}_{\neq i}(\tau, h)$ is the estimate of observation *i* from the GTWR model established with the training data after eliminating observation *i* in the situation of time-space scale factor τ and bandwidth *h*.

The over-fitting that resulted from the extraordinary large weight for observation *i* can be effectively avoided with the CV procedure. To improve the efficiency of parameter optimizing procedures, we use a hybrid model by the incorporation of grid search and particle swarm optimization (PSO) to search an approximation optimal solution of Equation (6).

4. Study Area and Data

As the capital of Jiangsu Province, Nanjing is the second largest city in eastern China with the Yangzi River flowing through the city. Like many other metropolitans in China, Nanjing has been experiencing rapid urbanization and motorization after the reform and opening-up in 1978. By the year 2016, the city contained 11 districts with an administrative area of 6600 km^2 (2500 mi^2) and a total population of 8.27 million, 82% of which have urban registration [38]. The consistently growing motor vehicle retention reached 2.28 million by the end of 2017.

To address the growing problems of urban expansion and traffic congestion, Nanjing began to construct its urban rail transit (URT) system from 2000. Since the first line formally started to operate in September 2005, Nanjing Metro had had 7 lines and 128 stations in operation as to November 2017 (Figure 1). By the year 2030, the projected metro network will contain 25 lines and cover 80% of suburban area with 800 m as service distance for metro stations [39]. Under such a circumstance, it will provide a more useful decision support for line alignment and station location to investigate the correlation between metro ridership at station level and influence factors of topological structure and built environment based on the insight into spatiotemporal variation.
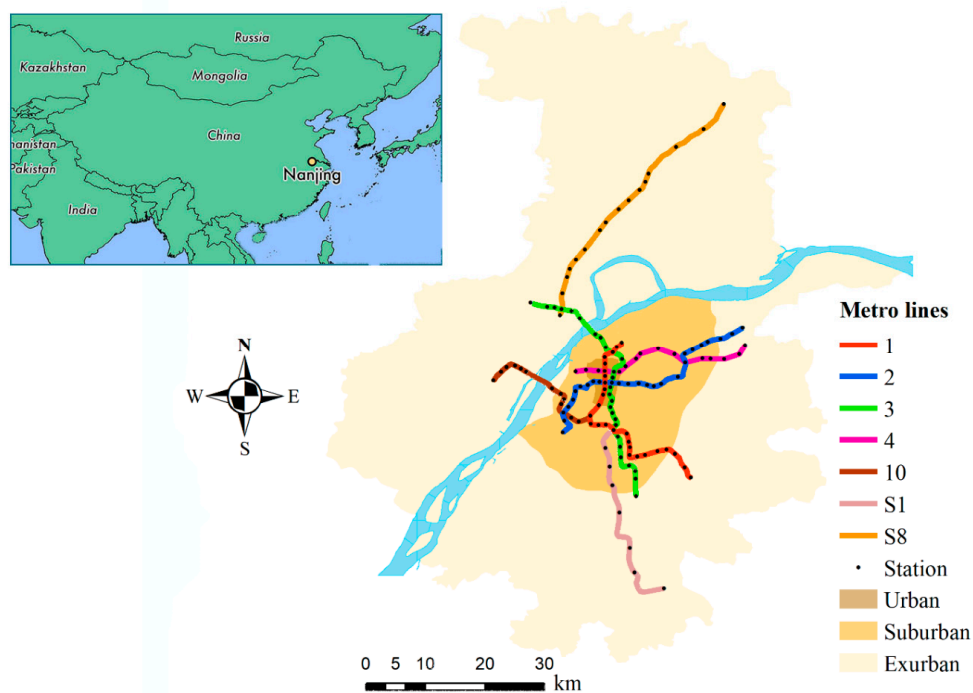

Figure 1. A map of the study area.

4.1. Metro Ridership Data

The automatic fare collection (AFC) system records the information of boarding station identification (ID), tap-in time, alighting station ID, and tap-out time for almost every passenger. In this study, the AFC data were collected from three weeks in October 2017 (9 to 29 October 2017) without festival. The raw transaction data were aggregated into hourly time interval from 5 a.m. to 12 p.m., 19 h totally with non-service time omitted, for all the 128 stations. The patterns of weekday travel and weekend travel are distinctly different from each other. Because the weekday travel demand always attracts more concern, the station ridership in this study only refers to the weekday boarding number. The average hourly ridership was derived from the same period during weekdays (from Monday to Friday) as the dependent variable. As shown in Figure 2, two obvious peaks can be observed for most stations, where the morning peaks mainly appear during 7:00–9:00 and the afternoon peaks mainly appear during 17:00–19:00.
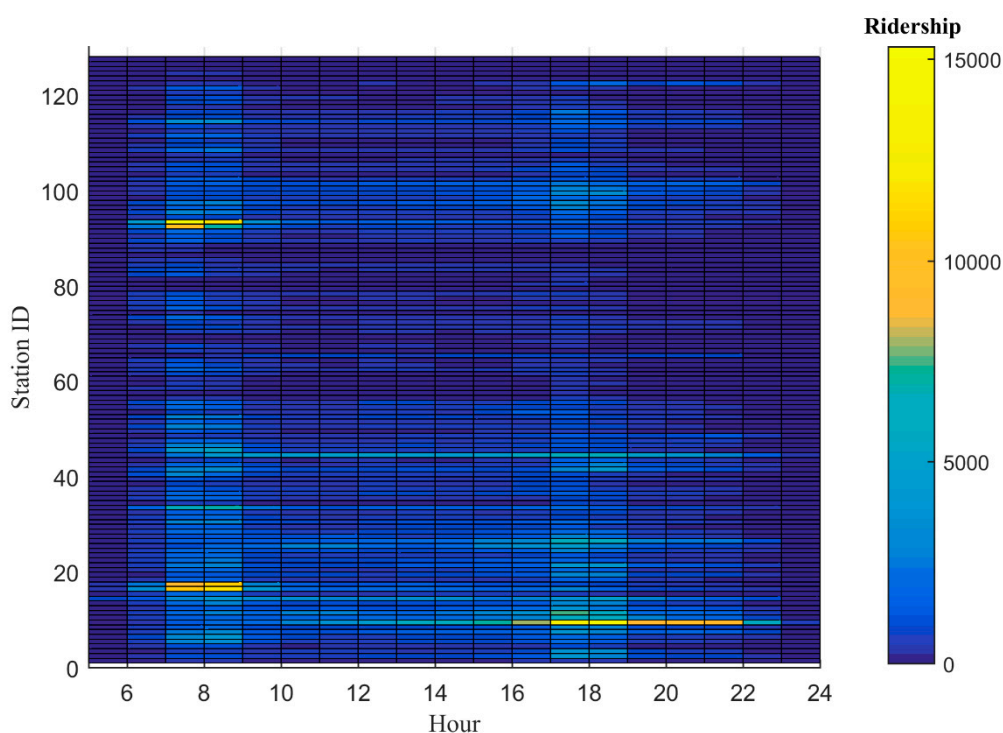

Figure 2. Average hourly ridership for metro stations.

4.2. Influence Factor Measurement

Considering the findings of existing studies and the accessibility of data, two categories of the factors that may influence metro demand at station level, namely built environment and topological structure, were investigated in this paper. The measured factors were regarded as the potential independent variables in our GTWR model. A summary of the potential independent variables for station-level ridership is shown in Table 1.

The service area for the metro station is usually taken as an area within a threshold walking distance of 800 m (0.5 miles) in many previous studies [9,13,15,16]. The built-environment related factors were calculated within this area with the assistance of a geographic information system (GIS) and point of interest (POI) data collected from Baidu Map. Among the built-environment related factors, the type of land use surrounding the metro station was categorized into eight types: residence, office, commerce, service, education, hotel, scenery, and parking, and the intensity of land use was described by the number within walking area of stations, while the five remaining built-environment related factors were used to indicate the convenience of intermodal connectivity with other traffic modes.

Table 1. Potential factors and their descriptions.

Variable	Description	Mean	Std. Dev.	Min	Max
Built Environment					
Residence	Number of residence communities	17.40	19.17	0	92
Office	Number of companies and government agencies	79.02	140.88	0	866
Commerce	Number of shopping malls, restaurants, retail stores, and entertainment centers	261.93	403.23	0	2489
Service	Number of communications, financial buildings, hospitals	49.84	68.37	0	504
Education	Number of education agencies	26.16	37.63	0	178
Hotel	Number of hotels	14.00	20.89	0	119
Scenery	Number of famous sceneries	9.79	22.58	0	161
Parking	Number of parking lots	10.31	13.46	0	82
Connected Road	Total length of road (km)	29.64	12.59	1.56	55.85
Connected Bus	Number of bus stops	23.55	21.17	0	82
Connected Bike	Number of public-bike stations	3.53	4.68	0	18
Nearest Bus	The distance from the nearest bus stop (m)	5315.78	8435.76	16.21	42,693.46
Nearest Bike	The distance from the nearest public-bike station (m)	405.63	657.33	1.98	3695.94
Topological Structure					
Degree Centrality	Degree centrality of metro station	2.05	0.63	1	5
Closeness Centrality	Closeness centrality calculated based on Metro network	0.07	0.02	0.04	0.10
Betweenness Centrality	Betweenness centrality calculated based on Metro network	906.36	755.51	0	3570

Regarding the topological-structure-related factors, three common centrality metrics, namely degree centrality, closeness centrality, and betweenness centrality [40], were adopted in this paper, for the undirected metro network with edge weights as the line lengths between adjacent stations.

Degree centrality of an observation station is measured as the total number of adjacent edges of this station. Degree centrality is the simplest index in the notions of centrality, which can indicate the types of station (i.e., transfer station, terminal station, and intermediate station) from the perspective of network connectively.

However, a well-connected transfer station may be located in a relative remote region. Closeness centrality provides complementary mechanisms for the definition of centrality. The aim of closeness centrality is to describe the accessibility determined by how easily passengers get to all the stations from an observation station. It can be calculated as the reciprocal of the average shortest-path distances from observation station i to all other stations in the metro network, as given by:

$$C_i^C = \frac{R-1}{\sum_{\forall r} d_{ir}} \tag{7}$$

where d_{ir} defines the shortest topological distance between stations i and r; R is the total number of metro stations.

Betweenness centrality is a different extension of centrality, which is based on how often a special node acts as a bridge on the shortest path between any other two nodes. It can be introduced as a measure for quantifying the importance of a station on the connectivity between other stations in a metro network. The betweenness centrality of observation station i can be compactly defined as:

$$C_i^B = \sum_{i \neq k \neq m} \frac{\phi_{km}(i)}{\phi_{km}} \tag{8}$$

where ϕ_{km} is the total number of shortest paths from station k to station m, and $\phi_{km}(i)$ the number of these paths that pass through station i.

5. Modeling and Discussion

5.1. Variables Selection for GTWR

Although the GTWR model has a remarkable ability to capture spatial and temporal heterogeneity, it is difficult to implement the selection of candidate variables because of the diverse test results of parameters at different spatiotemporal observation points. Thus, a multiple OLS regression, based on average daily ridership data of 128 stations in Nanjing, was conducted at first. A stepwise procedure was executed to identify the significant ridership-related variables, where the selection of variables is a bidirectional elimination based on the criterion that variables will be included with a confidence level above 85% or excluded with a confidence level below 70%. Table 2 shows the summary for the final OLS model with the parameter estimates and parametric diagnosis.

Four built-environment related factors and two topological-structure related factors were included in the final OLS model. All of these factors show a highly significant correlation with daily ridership demand, and most of them have slight collinearity with other variables (VIF < 4). Only the variables of Commerce and Parking are multicollinear variables, but they will be retained in the OLS model because the Variance Inflation Factor (VIF) values are markedly smaller than 10 (the threshold indicating significant collinearity) [41]. Table 2 shows that the variables of Commerce, Parking, Degree Centrality, and Closeness Centrality have a positive impact on daily ridership demand, which demonstrates that an area with prosperous commerce, friendly parking systems, well-connected metro lines, and accessibility to other metro stations will attract more passengers. However, the negative coefficients of Residence and Scenery are contrary to general cognition and existing findings [13–15].

Table 2. Summary and diagnosis of ordinary least squares (OLS) model coefficients.

Variable	Coefficient	Std. Dev.	$t-$Statistic	Probability	VIF
Intercept	−17661.46	5042.78	−3.50	0.00 [a]	
Residence	−201.53	86.73	−2.32	0.02 [a]	3.48
Commerce	18.86	5.37	3.51	0.00 [a]	5.92 [b]
Scenery	−158.47	57.07	−2.78	0.01 [a]	2.09
Parking	333.73	166.97	2.00	0.05 [a]	6.36 [b]
Degree Centrality	5126.02	1594.41	3.22	0.00 [a]	1.28
Closeness Centrality	24,7791.70	80,625.68	3.07	0.00 [a]	2.41
Number of observations	128	R^2	0.57		
RSS (Residual Sum of Squares)	15,969,253,868.32	MS (Mean Square)	100,142,322.48		

[a] The correlation is significant at the 0.05 level. [b] VIF (Variance Inflation Factor) is greater than 4.

5.2. Modeling

Before building the GTWR model, we need to make a spatial and temporal nonstationarity test for the sample data. If the spatial and/or temporal nonstationarity is significant, it means that the GTWR model can provide significant better fits for the sample data than the OLS model. An effective means can be employed to assess the spatial and temporal nonstationarity with an analysis of variance (ANOVA) [35]. The results of ANOVA based the average hourly ridership data in Nanjing Metro are presented in Table 3. In this table, the nonstationarity was diagnosed from spatial, temporal, both spatial and temporal perspectives by comparing the residual mean squares (MS) for the global model (OLS) and different local models, namely GWR, temporally weighted regression (TWR) and GTWR. As shown in the last two columns, the statistics of the F-test demonstrate that there is both spatial and temporal heterogeneity in the correlation between ridership and influence factors in the Nanjing Metro. In addition, it was found that the GTWR model can describe the data significantly better than considering the nonstationarity from a single scale (GWR or TWR). According to the values of R^2 in Table 3, 93% of variation in station-level ridership is explained in GTWR, which is 24% in OLS, 52% in TWR, 42% in GWR. Notably, the TWR model achieved a better goodness-of-fit compared with the GWR model in terms of R^2, which indicates that nonstationarity is more prominent in temporal scale than that in spatial scale.

Table 3. Analysis of variance (ANOVA) comparison between OLS and Local Regression Models.

Source of Variation	RSS	DF	MS	R^2	Pseudo-F Statistic	*p*-Value
OLS residuals	2,647,094,052.17	2425.00	1,091,585.18	0.24		
TWR residuals	1,571,998,861.05	2333.35	673,710.14	0.52		
GWR residuals	2,007,357,668.66	2304.84	870,932.18	0.42		
GTWR residuals	218,308,621.64	960.02	227,399.57	0.93		
TWR/OLS improvement	1,075,095,191.12	91.65	1,172,9932.15		17.41	0.00
GWR/OLS improvement	639,736,383.51	120.16	5,323,953.40		6.11	0.00
GTWR/OLS improvement	2,428,785,430.53	1464.98	1,657,899.07		7.29	0.00
GTWR/TWR improvement	1,353,690,239.41	1373.32	985,703.58		4.33	0.00
GTWR/GWR improvement	1,789,049,047.02	1344.82	1,330,330.03		5.85	0.00

To realize the optimal selection of time-space scale factor τ and bandwidth h simultaneously, we converted the spatiotemporal distance into spatial scale (km) and combined grid search and PSO. We conducted the parameter optimizing process with the following steps:

Firstly, we generated two sequences for the parameters τ and h and built GTWR models with each pairwise set. The minimum CV appeared at the set of (3, 3) as shown in Figure 3.
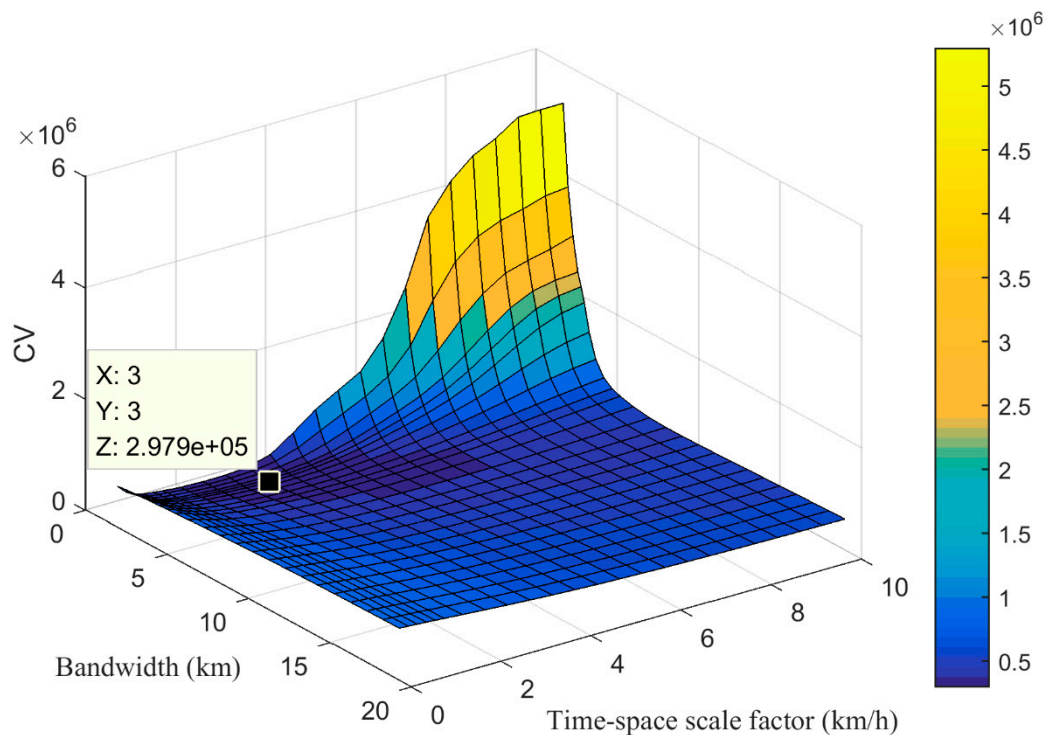
Figure 3. The parameters τ and h against cross-validation (CV).

We then initialized the population of particles in PSO as 50 and each particle with a random position on two dimensions. For each particle, its fitness was calculated as the CV value with the set of (τ, h) equal to its two-dimension coordinates.

Then, we gave a search range surrounding the set of (3, 3) and further minimized the CV value with iterations of PSO. The final optimal CV was attained with the parameter set of ($\tau = 3.13$, $h = 3.10$).

As stated before, the value of τ means that the decay of weight for the increase in temporal distance by one hours equate to that for the increase in spatial distance by 3.13 km. Since the average distance of adjacent metro stations is about 2 km for Nanjing metro, it means that the weights for adjacent metro stations within the same period approach twice those for adjacent time periods for the same station. In other word, the spatial data could provide more information for ridership forecasting than the temporal data.

5.3. Results and Discussion

The GTWR model was estimated with the optimal parameters based on average hourly data, and the results are presented in Table 4. Seven statistics were selected to describe the distributions of the estimated coefficients. Specifically, the lower quartile (LQ) and the upper quartile (UQ) were used to indicate interquartile range. The coefficient mathematically implicates that the metro ridership during the observed period will increase by the value of this coefficient due to unit change in the corresponding variable. Thereby, the positive mean coefficient for the Residence variable suggests that the metro station with more residual communities within its service area will generally attract more ridership, which is opposite to the result of OLS in Section 5.1 but identical with previous studies [14,15]. The sign of the mean Scenery is negative as same as that in OLS. A possible reason is that famous sceneries are usually located in underpopulated regions. Another counter-example is the variable of Closeness Centrality, where the signs of both mean and medium Closeness Centrality are negative. It is difficult to explain this abnormal phenomenon based on the information in Table 4. The positive mean and medium values of the variables of Parking and Degree Centrality arrive at a similar inference as the OLS model.

Table 4. Geographically and temporally weighted regression (GTWR) parameter estimate summaries.

Variable	Mean	Min	Max	LQ	Medium	UQ	Std. Dev.
Intercept	141.18	−52,864.92	78,513.04	−1234.82	2.43	1233.43	5079.40
Residence	28.13	−664.54	1070.38	−9.45	1.97	32.59	100.26
Commerce	−0.37	−197.84	77.47	−0.22	0.26	1.31	10.73
Scenery	−92.13	−7012.30	7548.52	−28.94	−5.65	4.98	510.39
Parking	34.40	−1576.47	4487.69	−4.53	12.78	45.46	232.10
Degree Centrality	313.77	−3168.00	6758.58	−31.92	110.47	700.74	694.48
Closeness Centrality	−3910.48	−1,191,278.07	901,756.46	−20,103.34	−1276.66	11,094.17	77,422.57

Since the parameters of the GTWR model were locally estimated and various across spatial and temporal scales, it could offer deep insight into the spatial and temporal heterogeneities of the influence of built environment and topological factors on metro ridership by visualizing the distribution patterns of GTWR-based estimators. Understanding how station demands are affected and determined, especially for the demand during peak period, is more of a concern in metro planning work. A comparison of the coefficient distributions during morning peak hours and evening peak hours is presented in Figures 4 and 5.

It is observed that all of the six variables have both positive and negative effects on metro ridership for morning and evening peaks, which is different from the fixed negative correlation of the OLS model in Table 2, as well as the result of Zhao et al. [15] that is also a case study of Nanjing. The coefficients for Residence are negative in the core urban area which is generally work-oriented and/or commercial-oriented and positive in the stations that are close to terminal stations which are generally residential-oriented. The negative coefficients in the core urban area could be reasonable, because there is higher car ownership and shorter commuting distance. During evening peak hours, the effect of Residence on metro ridership become negative for most stations in suburban areas, and the average absolute value of the negative coefficients is less than that in morning peak hours. It may be explained by that the number of residence communities primarily effect the trips departing from home regarding boarding ridership.

Figures 4 and 5 display the spatial distribution of the Commerce coefficients during peak hours. The number of shopping malls, restaurants, retail stores, and entertainment centers has a negative association with metro ridership in the core urban area during morning peak hours, and this association is converted to positive during evening peak hours. A possible explanation for this finding is that the trip purposes of boarding ridership relatively concentrate on commuting in the morning peak, whereas the ridership in evening peak may be mingled with large amounts of other trip purposes, such as dinner together, shopping, and entertainment. However, for the business districts in suburban areas, a contrary trend (i.e., from positive to negative) could be observed. It is not surprising that the suburban area is relative underdevelopment. Concerns of security and long travel distance could be possible reasons for the reduction in intensity of business activity in remote areas.

The Closeness Centrality provides a negative effect on station demand in the central area of the metro network and tends to present a general positive correlation with metro ridership when shifting away from the central area. The medium value of the coefficients for this variable during evening peak hours is much larger than that during morning peak hours (5403.02 versus 0.34). It may be a result from the resident lifestyle that many people live in rural areas and work in urban areas.

An obvious tidal traffic phenomenon can be observed from Figure 6. It can be seen that the stations at the junctures of suburban areas and exurban areas experience high densities of boarding passengers during morning peak hours; nevertheless the vast majority of crowded boarding ridership always occurs in the core urban stations during evening peak hours. This unbalanced movement can also partially explain the temporal variation between morning peak and evening peak.
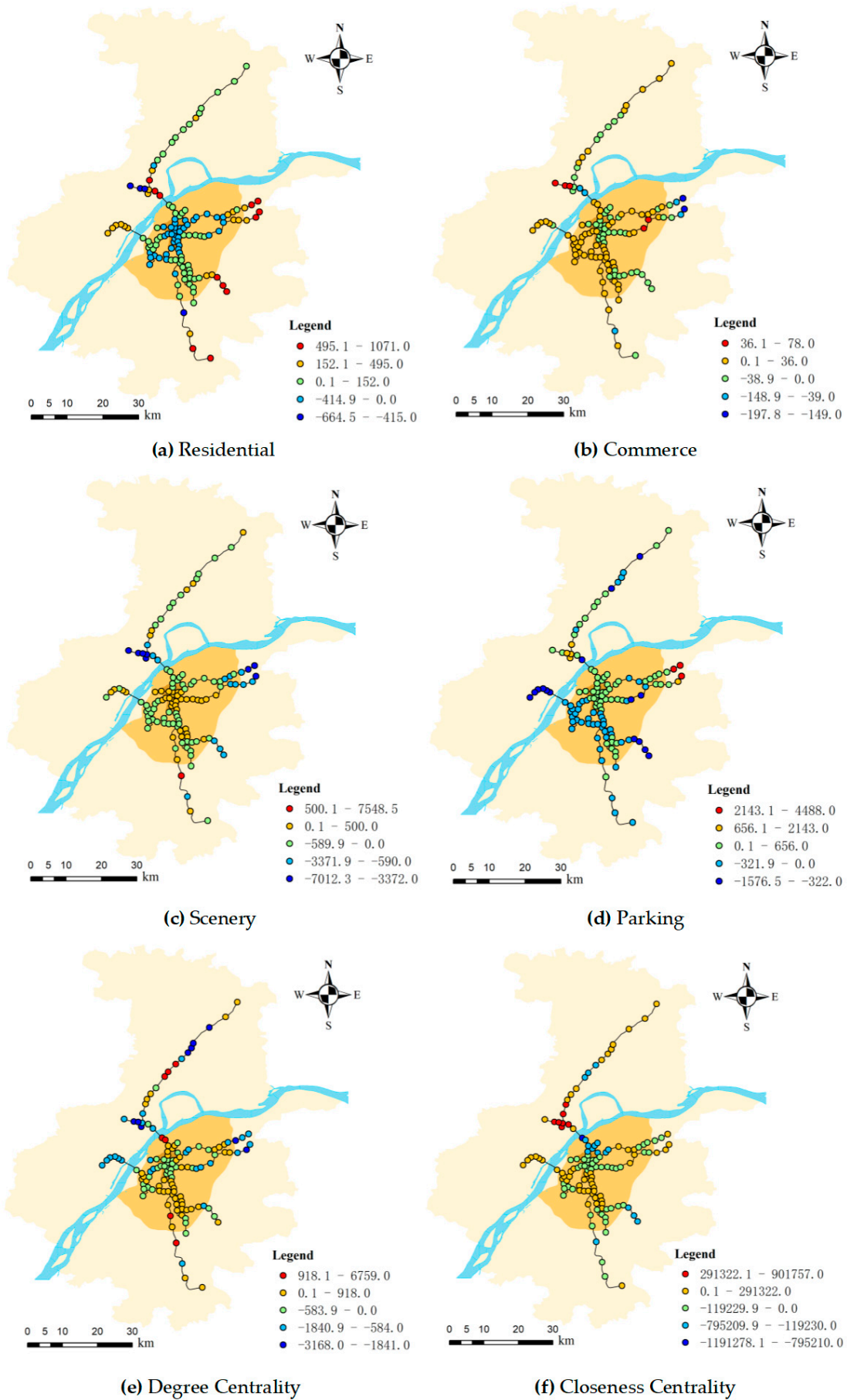
Figure 4. Spatial distribution of the coefficient during morning peak 7:00–8:00.
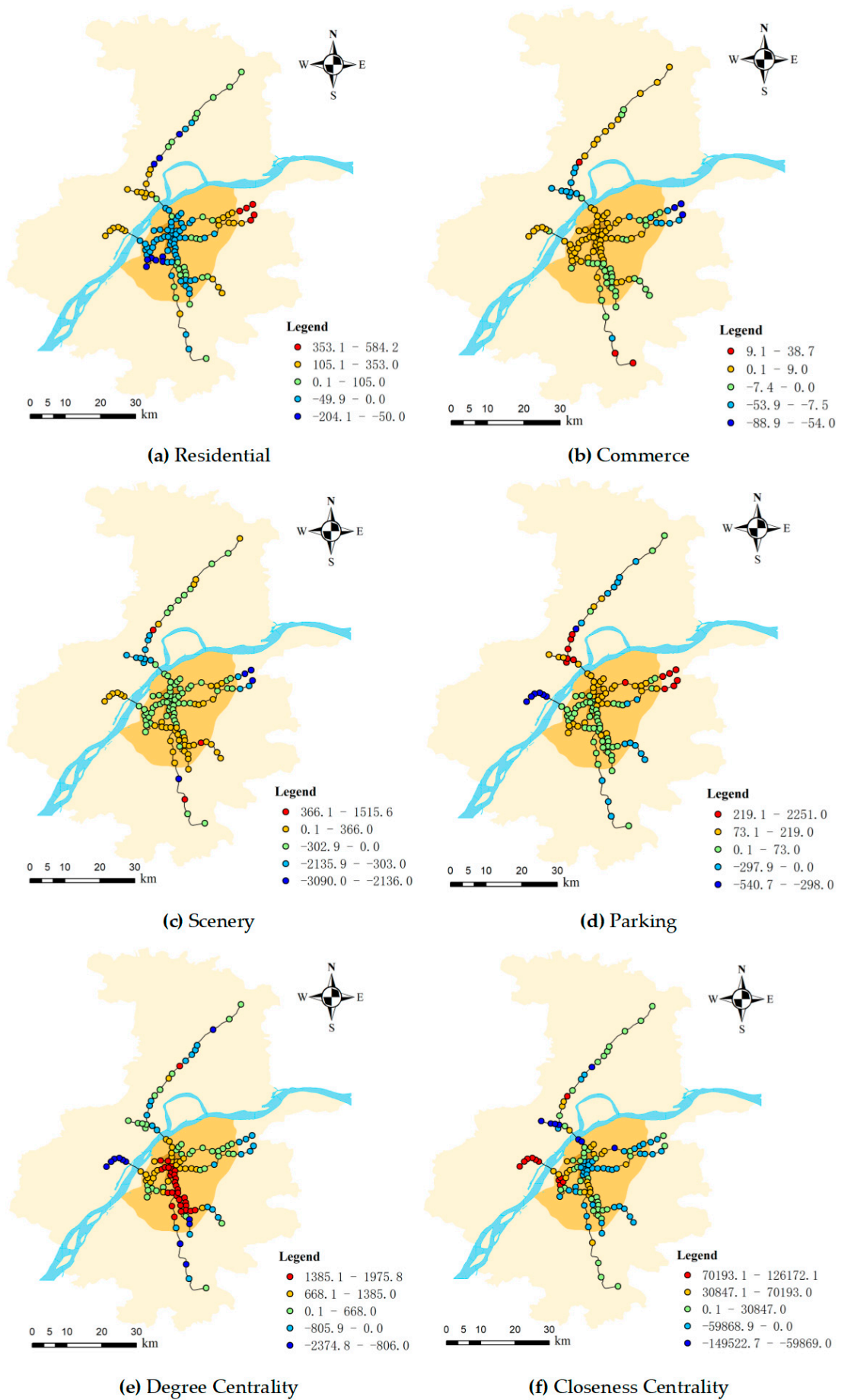
(a) Residential

(b) Commerce

(c) Scenery

(d) Parking

(e) Degree Centrality

(f) Closeness Centrality

Figure 5. Spatial distribution of the coefficient during evening peak 17:00–18:00.
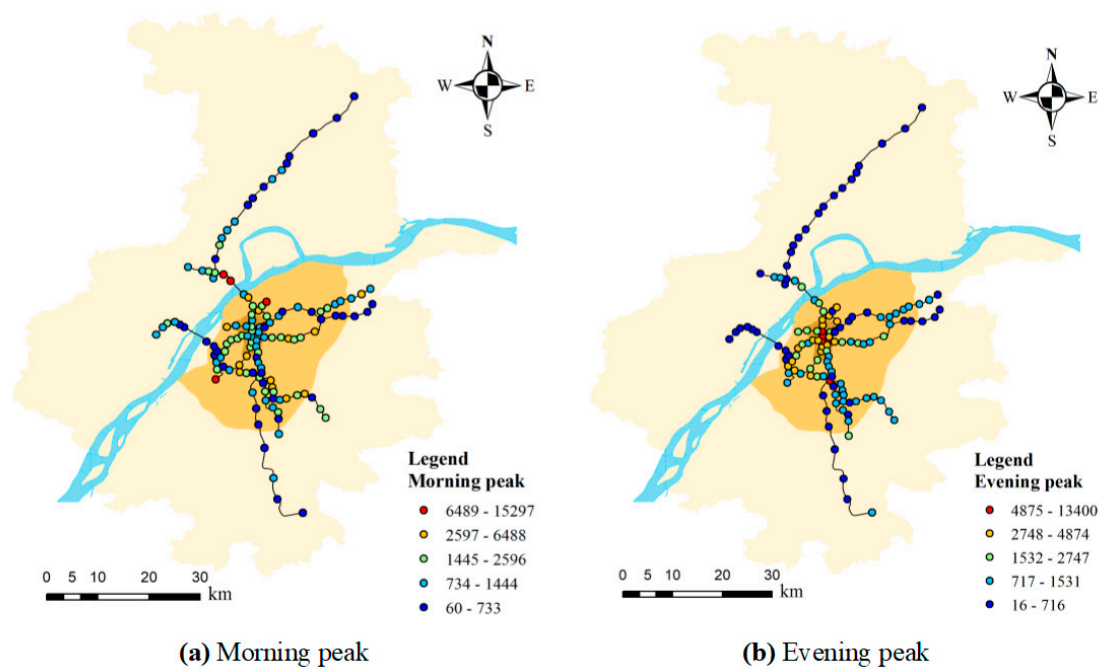
(a) Morning peak

(b) Evening peak

Figure 6. Spatial distribution of hourly ridership during peak hours.

6. Conclusions

This paper contributes by exploring the spatial and temporal distributions of the association between hourly boarding ridership at station level and the factors related to built environment and topological structure in Nanjing city. Two statistical techniques (OLS and GTWR) were combined to investigate this relationship. Firstly, the global OLS was conducted to auto-select the potential variables by a bidirectional stepwise procedure. The GTWR model, an extended version of the GWR model, can capture the spatial and temporal variation between hourly ridership and the selected variables. In the GTWR model, we transferred the temporal distance into spatial scales, which not only could reduce the number of parameters but also endow the transfer factor with a practical meaning. To improve optimized efficiency, we used the incorporation of grid search and PSO to release solution optimal selection of weighting parameters.

The results of OLS based on daily ridership data suggested that six variables including four built-environment factors (i.e., Residence, Commerce, Scenery, and Parking) and two topological-structure factors (i.e., Degree Centrality and Closeness Centrality) had a significant effect on station-level ridership. The diagnosis of ANOVA demonstrated that there is significant spatial and temporal nonstationarity in the relationship between ridership and influence factors. The GTWR model offered a significantly better goodness-of fit for hourly ridership data than the traditional OLS, GWR and TWR models by producing a more complete picture of ridership data analysis from the perspective of both spatial and temporal scales. The results of the GTWR model suggested that residence communities primarily effect the trips departing from home regarding boarding ridership. The commercial buildings in the central city mainly attract metro ridership during evening peak hours, whereas in suburban areas they serve more metro ridership during morning peak hours. Nevertheless, the findings of GTWR are not completely different from those of OLS; for example, adequate parking lots and transfer stations could attract more ridership. Consequently, the findings of the GTWR model can not only provide more reliable and accurate estimates for metro demand, but also help planners to better understand the spatial and temporal variation of the correlation between ridership and influencing factors.

Future work can enrich the factor data that relate to metro ridership, such as the frequency of buses, the number of sharing bikes, and traffic conditions. The POIs used in this study were counted

in the same service area without considering their diverse attractions and the decay influences with increasing distance. Besides, advanced weighting kernel function and the GTWR model that can estimate both local and global parameters deserve further investigation.

References

1. Zhang, J.; Xu, X.; Hong, L.; Wang, S.; Fei, Q. Networked analysis of the Shanghai subway network, in China. *Phys. A Stat. Mech. Its Appl.* **2011**, *390*, 4562–4570. [CrossRef]
2. Kim, C.; Kim, S.; Kang, H.; Song, S.-M. What Makes Urban Transportation Efficient? Evidence from Subway Transfer Stations in Korea. *Sustainability* **2017**, *9*, 2054. [CrossRef]
3. Mu, R.; de Jong, M.; Yu, B.; Yang, Z. The future of the modal split in China's greenest city: Assessing options for integrating Dalian's fragmented public transport system. *Policy Soc.* **2012**, *31*, 51–71. [CrossRef]
4. Li, M.; Dong, L.; Shen, Z.; Lang, W.; Ye, X. Examining the Interaction of Taxi and Subway Ridership for Sustainable Urbanization. *Sustainability* **2017**, *9*, 242. [CrossRef]
5. Anderson, M.L. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *Am. Econ. Rev.* **2014**, *104*, 2763–2796. [CrossRef]
6. Kim, H.; Kwon, S.; Wu, S.K.; Sohn, K. Why do passengers choose a specific car of a metro train during the morning peak hours? *Transp. Res. Part A Policy Pract.* **2014**, *61*, 249–258. [CrossRef]
7. Shi, Z.; Zhang, N.; Zhang, Y. Hazard-based model for estimation of congestion duration in urban rail transit considering loss minimization. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2595*, 78–87. [CrossRef]
8. Zhang, L.; Liu, M.; Wu, X.; AbouRizk, S.M. Simulation-based route planning for pedestrian evacuation in metro stations: A case study. *Autom. Construct.* **2016**, *71*, 430–442. [CrossRef]
9. Cardozo, O.D.; García-Palomares, J.C.; Gutiérrez, J. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Appl. Geogr.* **2012**, *34*, 548–558. [CrossRef]
10. Mcnally, M.G. The four step model. In *Handbook of Transport Modelling*; Hensher, D.A., Button, J.K., Eds.; Emerald Group Publishing Limite: Bingley, UK, 2007; pp. 35–52.
11. Ceccato, V.; Oberwittler, D. Comparing spatial patterns of robbery: Evidence from a Western and an Eastern European city. *Cities* **2008**, *25*, 185–196. [CrossRef]
12. Gutiérrez, J.; Cardozo, O.D.; García-Palomares, J.C. Transit ridership forecasting at station level: An approach based on distance-decay weighted regression. *J. Transp. Geogr.* **2011**, *19*, 1081–1092. [CrossRef]
13. Kuby, M.; Barranda, A.; Upchurch, C. Factors influencing light-rail station boardings in the United States. *Transp. Res. Part A Policy Pract.* **2004**, *38*, 223–247. [CrossRef]
14. Sohn, K.; Shim, H. Factors generating boardings at Metro stations in the Seoul metropolitan area. *Cities* **2010**, *27*, 358–368. [CrossRef]
15. Zhao, J.; Deng, W.; Song, Y.; Zhu, Y. What influences Metro station ridership in China? Insights from Nanjing. *Cities* **2013**, *35*, 114–124. [CrossRef]
16. De Andrade, G.T.; Gonçalves, J.A.M.; da Silva Portugal, L. Analysis of Explanatory Variables of Rail Ridership: The Situation of Rio de Janeiro. *Procedia Soc. Behav. Sci.* **2014**, *162*, 449–458. [CrossRef]
17. Ding, C.; Wang, D.; Ma, X.; Li, H. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability* **2016**, *8*, 1100. [CrossRef]
18. Jun, M.-J.; Choi, K.; Jeong, J.-E.; Kwon, K.-H.; Kim, H.-J. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *J. Transp. Geogr.* **2015**, *48*, 30–40. [CrossRef]
19. Tao, X.; Ming, Z.; Aditjandra, P.T. The impact of urban rail transit on commercial property value: New evidence from Wuhan, China. *Transp. Res. Part A Policy Pract.* **2016**, *91*, 223–235.

20. Ji, Y.; Ma, X.; Yang, M.; Jin, Y.; Gao, L. Exploring Spatially Varying Influences on Metro-Bikeshare Transfer: A Geographically Weighted Poisson Regression Approach. *Sustainability* **2018**, *10*, 1526. [CrossRef]

21. Wang, X.; Zhang, N.; Zhang, Y.; Shi, Z. Forecasting of Short-Term Metro Ridership with Support Vector Machine Online Model. *J. Adv. Transp.* **2018**, *2018*, 3189238. [CrossRef]

22. Chan, S.; Miranda-Moreno, L. A station-level ridership model for the metro network in Montreal, Quebec. *Can. J. Civil Eng.* **2013**, *40*, 254–262. [CrossRef]

23. Choi, J.; Lee, Y.J.; Kim, T.; Sohn, K. An analysis of Metro ridership at the station-to-station level in Seoul. *Transportation* **2011**, *39*, 705–722. [CrossRef]

24. Sung, H.; Choi, K.; Lee, S.; Cheon, S. Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership. *J. Transp. Geogr.* **2014**, *36*, 134–140. [CrossRef]

25. Zhang, D.; Wang, X. Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC. *J. Transp. Geogr.* **2014**, *41*, 107–115. [CrossRef]

26. Ma, X.; Zhang, J.; Ding, C.; Wang, Y. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Comput. Environ. Urban Syst.* **2018**, *70*, 113–124. [CrossRef]

27. Murray, A.T.; Davis, R.; Stimson, R.J.; Ferreira, L. Public Transportation Access. *Transp. Res. Part D Transp. Environ.* **1998**, *3*, 319–328. [CrossRef]

28. Zhang, Q.; Han, B.; Li, D. Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. *Transp. Res. Part C Emerg. Technol.* **2008**, *16*, 635–649. [CrossRef]

29. Brown, J.; Thompson, G.; Bhattacharya, T.; Jaroszynski, M. Understanding transit ridership demand for the multidestination, multimodal transit network in Atlanta, Georgia: Lessons for increasing rail transit choice ridership while maintaining transit dependent bus ridership. *Urban Stud.* **2014**, *51*, 938–958. [CrossRef]

30. Giuliano, G. Travel, location and race/ethnicity. *Transp. Res. Part A Policy Pract.* **2003**, *37*, 351–372. [CrossRef]

31. Zhao, J.; Deng, W.; Song, Y.; Zhu, Y. Analysis of Metro ridership at station level and station-to-station level in Nanjing: An approach based on direct demand models. *Transportation* **2014**, *41*, 133–155. [CrossRef]

32. Cervero, R. Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *J. Am. Plan. Assoc.* **2006**, *72*, 285–295. [CrossRef]

33. Tobler, W.R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46* (Suppl. 1), 234–240. [CrossRef]

34. Fotheringham, A.S.; Brunsdon, C.; Charlton, M. *Geographically Weighted Regression*; John Wiley and Sons, Limited West Atrium: Chichester, UK, 2003.

35. Huang, B.; Wu, B.; Barry, M. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 383–401. [CrossRef]

36. Chu, H.-J.; Huang, B.; Lin, C.-Y. Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. *Atmos. Environ.* **2015**, *102*, 176–182. [CrossRef]

37. Du, Z.; Wu, S.; Zhang, F.; Liu, R.; Zhou, Y. Extending geographically and temporally weighted regression to account for both spatiotemporal heterogeneity and seasonal variations in coastal seas. *Ecol. Inform.* **2018**, *43*, 185–199. [CrossRef]

38. NPB. Nanjing Statistic Yearbook 2016. Available online: http://221.226.86.104/file/nj2004/2017/zonghe/index.htm (accessed on 15 June 2018).

39. Survey of Nanjing Metro Network Planning by 2030. Available online: http://www.njmetro.com.cn/build_030.aspx (accessed on 15 June 2018).

40. Newman, M.E. Analysis of weighted networks. *Phys. Rev. E* **2004**, *70*, 056131. [CrossRef] [PubMed]

41. Kutner, M.H.; Nachtsheim, C.; Neter, J. *Applied Linear Regression Models*; McGraw-Hill/Irwin: New York, NY, USA, 2004.