Chin. Geogra. Sci. 2019 Vol. 29 No. 6 pp. 1065–1077
https://doi.org/10.1007/s11769-019-1065-8

Springer Science Press
www.springerlink.com/content/1002-0063

Analysis of Metro Station Ridership Considering Spatial Heterogeneity

GAN Zuoxian[1], FENG Tao[2], YANG Min[1], Harry TIMMERMANS[2], LUO Jinyu[1]

(1. *Jiangsu Key Laboratory of Urban ITS*, *Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies*, *School of Transportation*, *Southeast University*, *Nanjing* 211189, *China*; 2. *Urban Planning Group*, *Eindhoven University of Technology*, *Eindhoven* 5600*MB*, *The Netherlands*)

Abstract: This study aims to explore the role of spatial heterogeneity in ridership analysis and examine the relationship between built environment, station attributes and urban rapid transit ridership at the station level. Although spatial heterogeneity has been widely acknowledged in spatial data analysis, it has been rarely considered in travel behavior studies. Four models (three global models-ordinary least squares (OLS), spatial lag model (SLM), spatial error model (SEM) and one local model-geographically weighted regression (GWR) model) are estimated separately to explore the relationship between various independent variables and station ridership, and identify the influence of spatial heterogeneity. Using the data of built environment and station characteristics, the results of diagnostic identify evidence the existence of spatial heterogeneity in station ridership for the metro network in Nanjing, China. Results of comparing the various goodness-of-fit indicators show that, the GWR model yields the best fit of the data, performance followed by the SEM, SLM and OLS model. The results also demonstrate that population, number of lines, number of feeder buses, number of exits, road density and proportion residential area have a significant impact on station ridership. Moreover, the study pays special attention to the spatial variation in the coefficients of the independent variables and their statistical significance. It underlines the importance of taking spatial heterogeneity into account in the station ridership analysis and the decision-making in urban planning.
Keywords: spatial heterogeneity; rapid transit ridership; built environment; station level; spatial models

Citation: GAN Zuoxian, FENG Tao, YANG Min, Harry TIMMERMANS, LUO Jinyu, 2019. Analysis of Metro Station Ridership Considering Spatial Heterogeneity. *Chinese Geographical Science*, 29(6): 1065–1077. https: //doi.org/10.1007/s11769-019-1065-8

1 Introduction

Urban rapid transit systems have expanded worldwide, and are regarded as an effective and sustainable option to solve traffic congestion and cut down transport energy consumption. The historical experience of urban transportation development processes in western countries showed that car-based transportation structures have resulted in urban sprawl, land waste and air pollution. ‘New Urbanism’ including the Smart Growth and Neighborhood Conservation initiatives have been launched and diffused rapidly from developed countries to developing countries. Transit Oriented Development (TOD) caters to this development ideology and has become popular in urban design practice (Cervero et al., 2002). Squeezed by high congestion and technology costs, it is necessary for emerging markets, especially in some fast-growing developing countries to prioritize public transit (PT) over automobiles. For example, in recent years, the Chinese government has emphasized the ‘bus priority’ policy and built many urban rapid transit systems, including metro, light-rail and bus, in many major cities. Currently, 34 Chinese cities have their own urban rail transit system by the end of 2017.

Received date: 2018-09-09; accepted date: 2019-01-05
Foundation item: Under the auspices of National Natural Science Foundation of China (No. 71771049), the Six Talent Peaks Project in Jiangsu Province (No. 2016-JY-003), China Scholarship Council (No. 201606090149)
Corresponding author: YANG Min. E-mail: yangmin@seu.edu.cn

To justify the investments and subsidies, it is of paramount importance to increase the use of the PT system, which may imply planning a new transit route or station (Zhang and Wang, 2014; Kepaptsoglou et al., 2017). To make informed decisions in that context, it is important to understand the influence of the built environment and station attributes on transit ridership. Many transit demand forecasting models have been used to explore the relationship between these independent variables and station ridership. The traditional approach that is widely used in practice is the trip-based four-step model (McNally, 2007). However, this model has many deficiencies and does not adequately consider the built environment because its forecasts are based on relatively large traffic zones (Cardozo et al., 2012; Zhao et al., 2014; Kepaptsoglou et al., 2017). Another widely used approach is the activity-based models ((Rasouli and Timmermans, 2014). However, a large number of personal trip surveys demand a lot of manpower and financial resources.

Nowadays, direct demand models based on OLS regression are generally adopted to evaluate the relationship between built environment and ridership (Kuby et al., 2004; Ryan and Frank, 2009; Loo et al., 2010; Sung and Oh, 2011; Chan and Miranda-Moreno, 2013; Zhao et al., 2014; Durning and Townsend, 2015). Although rapid transit ridership involves count data, the preferable negative binomial regression and Poisson regression have been rarely used for station-level ridership analysis (Choi et al., 2012). The main reason is that count data models are perceived not to have any advantages for ridership analysis (Choi et al., 2012; Dill et al., 2013; Kim et al., 2016). Another violation of the assumptions underlying OLS regression analysis is that there is no correlation between the independent variables and the disturbance items. However, correlation always exists in reality and will cause simultaneity bias (also called endogeneity bias). Some studies attempted to apply two stage least squares (2SLS) to explore the relationship between independent variables and ridership (Estupiñán and Rodríguez, 2008). This approach divides endogenous variables into two parts: one part is linked to disturbance items and the other part has no relationship with the disturbance items. By looking for suitable 'instrumental variable(s)', 2SLS can obtain the consistent estimator.

Using the traditional regression models, population or population density have been found to have a significant positive effect on ridership (Kuby et al., 2004; Sung and Oh, 2011; Chan and Miranda-Moreno, 2013; Zhao et al., 2014; Jun et al., 2015). Land use mix was also tested for its effect on rapid transit ridership. However, reported effects are mixed (Sung and Oh, 2011; Durning and Townsend, 2015; Jun et al., 2015). Some previous studies showed that the areas of residential, commercial, business and other types of land use are considered to contribute to ridership (Sung and Oh, 2011; Durning and Townsend, 2015), but the correlation seems weak (Chan and Miranda-Moreno, 2013; Zhao et al., 2014). Estupiñán and Rodríguez (2008) interpreted road density as a connectivity factor in the BRT ridership analysis in Bogotá and concluded the relationship was positive but not significant. Similar studies conducted by Zhao et al. (2013; 2014) found that road density was positively and significantly related to metro station ridership in Nanjing, China. Many studies have examined the relationship between the number of metro lines, number of feeder buses and rapid transit ridership, and found that both factors were associated with increased ridership (Kuby et al., 2004; Zhao et al., 2013; 2014; Durning and Townsend, 2015).

However, one drawback of the direct forecasting models based on traditional regression models is that by assuming the relations between variables in the model are homogeneous, they ignore local properties (Páez, 2006). According to the First Law of Geography: 'Everything is related to everything else, but near things are more related than distant things', the spatial data are not static and usually show spatial autocorrelation (Tobler, 1970). Thereto, despite many studies analyzed the relationship between various independent variables and ridership at the station level, spatial effects were rarely considered. Spatial effects, including spatial dependence and spatial heterogeneity, result in the uneven distribution of many ridership-related factors across space. Ridership analysis using the OLS model may lead to biased forecasts because the spatial effects can play a very important role in the final results (Pavlyuk, 2016). Using geographically weighted regression (GWR) can enhance the explanatory power of the approach (Cardozo et al., 2012; Blainey and Preston, 2013).

Overall, in previous studies, direct ridership models have been widely used because of its relatively low cost and quick response (Jun et al., 2015), but spatial heterogeneity was often ignored. Hereinto, we proposed a framework to understand the influences of built envi-

ronment factors and station attributes on urban transit ridership considering spatial heterogeneity. Based on one-month's smart card data records from Nanjing metro system. Three local models ordinary least squares, spatial lag model and spatial error model, and one local model geographically weighted regression model were utilized in the present paper. The latter three spatial-econometrics-models permit the explicit inclusion of urban rail transit station heterogeneity into model specifications. Moreover, we explored the variation of each significant variable across space based on local analysis to provide great explanatory power. The systematic analyses of urban rail transit ridership presented in this study is expected to help better understanding the relationships between built environment, station attributes and station ridership. The outcomes of the paper provide the basis for urban and transportation planning with an objective to promote transit ridership and sustainable urban development.

2 Data and Methods

2.1 Data

Nanjing, the capital city of Jiangsu Province, is located in the eastern China. It consists of 11 districts (Fig. 1), covering an area of 6587 km^2 with a total population of 8.27 million people in 2016. Nanjing is one of the cities in China where the metro is widely used after the opening

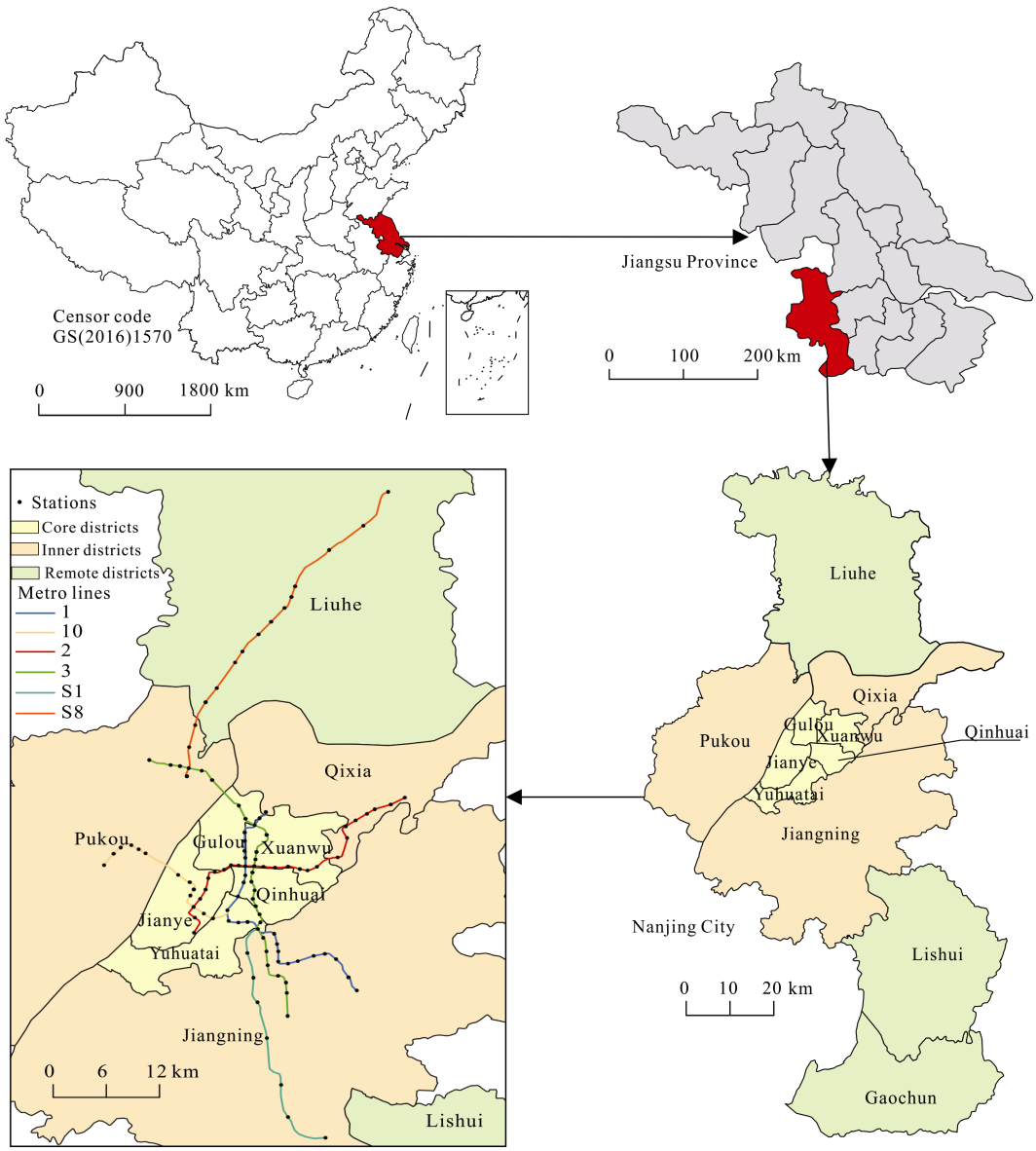

Fig. 1 Location of case study area (Jiangsu Province, Nanjing City, and Nanjing metro network in 2015)

of the first metro line in September 2005. To cope with the growing urban sprawl and huge travel demand, the local government started planning the urban rail transit network in 2002. Fig. 1 shows the Nanjing metro system in April 2015. It covered 6 lines and 112 stations, which makes it the fifth largest rail transit system in China (not including Hong Kong, Macao, Taiwan) (the top four cities are Shanghai, Beijing, Guangzhou and Shenzhen). The metro system carries 717 million passengers annually and its share in 2015 was about 34.8% of the passenger volume of public transport (Zhao et al., 2013; Gan et al., 2018).

The primary dataset contained metro ridership data covering all Nanjing metro stations in April 2015 from Nanjing Metro Corporation (NMC). It consisted of a full month of usage information and the total number of trip records was more than 43 million. The land use CAD DWG file of the Nanjing region from the Nanjing Urban Planning Bureau (NUPB) was used to generate the characteristics of the built environment. Furthermore, we downloaded the Nanjing road network from Open Street Map (OSM) to calculate road density. All variables were gathered and generated using ArcGIS Version 10.3. Thus, the following ten candidate variables were used:

Station ridership refers to the total number of passengers at each metro station in a month (April 2015). Moreover, boardings and alightings were examined. The two numbers are nearly the same (Fig. 2), indicating that the number of boardings at each station is highly consistent with the number of alightings. One possible reason for this is that people who go to a destination return by metro to the origin. Based on this finding (We also examined the daily boardings and alightings of each stations and found that there was not significant difference between them), the dependent variable used the boarding data only, instead of both boarding and alighting data.

Population (person/km^2): The total population per squared kilometer within a station's catchment area, defined by the 800 m buffer, based on Thiessen polygons.

Number of lines: Number of metro lines at each station.

Number of feeder buses.

Number of exists: The total number of exits for each station, based on field observations.

Road density (km/km^2): Road density within the station catchment area.

S_Sdistance (m): Distance to the nearest metro station, calculated by metro distance rather than linear distance.

Distance center (m): Distance to the city center, calculated as metro distance.

Business area (%): Percentage of business/commercial land use within the station catchment area.

Residential area (%): Percentage of residential land use within the station catchment area.

Industrial/manufacturing area (%): Percentage of industrial/manufacturing land use within the station catchment area.

The descriptive statistics of the selected variables are presented in Table 1.

2.2 Catchment area

First of all, a critical operational decision underlying station ridership analysis is how to define the catchment area. This decision influences catchment area variables, such as population, road density, number of feeder buses, and land use mix. The catchment areas used in previous studies are normally determined using an empirically value, ranging from 300 m to 900 m. Researchers usually accept that a 800 m (or half a mile) Euclidean distance or network distance that people are willing to walk to a station (Kuby et al., 2004; Cardozo et al., 2012; Guerra et al., 2012; Zhao et al., 2014). Zhao used a radius of 800 m as the catchment area in his study about Nanjing metro stations and obtained good results (Zhao et al., 2013; 2014). Moreover, the Ministry of Housing and Urban-Rural Development of People's Republic of China (MOHURD) launched a design guidance of urban rail in November, 2015: 'Guidelines

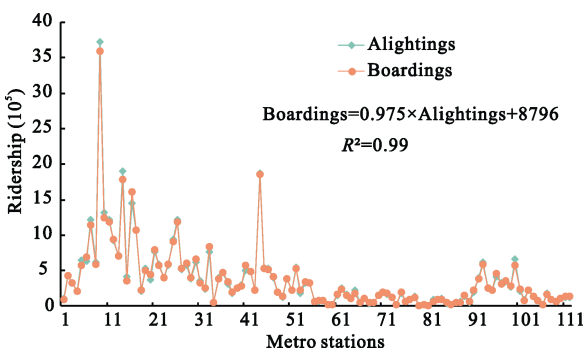

Fig. 2 Comparison between boardings and alightings of all Nanjing metro stations in April 2015

Table 1 Summary statistics for candidate explanatory variables

Variable	Mean	Median	SD	Min.	Max.
Station ridership	386307	231638	478855	10996	3599863
Population (person/km^2)	10214	7144	9511	574	46668
Number of lines	1.07	1	0.29	1	3
Number of feeder buses	9.74	8	7.16	0	44
Number of exits	3.28	3	2.29	1	22
Road density (km/km^2)	9.12	8.97	2.81	2.31	16.87
S_Sdistance (m)	1530	1270	930	745	5182
Distance_center (m)	16172	14377	12470	0	60379
Business area (%)	9	7	9	0	62
Residential area (%)	28	29	18	0	66
Industrial/manufacturing area (%)	10	4	12	0	54

for Planning and Design of Urban Rail', where a radius of 800 m was defined to determine the influence area of rail stations (Ministry of Housing and Urban-Rural Development, 2012). Thus, a fixed boundary buffer of 800 m (based on the longitude and latitude of central point of each station) was used as the catchment areas in this study. In order to split the overlapping parts of adjacent station buffers, Thiessen polygons were constructed using ArcGIS (Version 10.3).

2.3 Regression models

A regression model based on stepwise OLS was first applied to explore the link between independent variables and rapid transit ridership. The stepwise OLS model will help identifying the significant independent variables. The linear OLS regression model can be expressed as:

$$y = X\beta + \varepsilon \tag{1}$$

where y is an $n \times 1$ vector of ridership for n rapid transit stations, X is an $n \times k$ vector of ridership for n rapid transit stations and k explanatory variables, β is a $k \times 1$ vector of unknown coefficients for k explanatory variables, ε is an $n \times 1$ vector of disturbance items of n rapid transit stations.

The classical OLS model has the following requirements: one is the strict exogenous hypothesis $E(\varepsilon_i|X)=0$; the other is the spherical disturbance term $Var(\varepsilon_i|X)=\sigma^2 I_n$, which means that the random disturbance items should be identically and independently distributed. However, for spatial data, the observed values of a variable at adjacent places tend to be autocorrelated. Some diagnostic test of spatial autocorrelation such as Moran's statistic, Lagrange multiplier or Robust Lagrange multiplier can be used before any spatial model is recommended to analyze spatial heterogeneity. Because spatial dependence can be demonstrated by spatial lag or spatial error, two global spatial models (Spatial Lag Model, SLM and Spatial Error Model, SEM) will be utilized in this study. SLM, also referred to as SAR (Spatial Autoregression) can be presented as:

$$y = \lambda W y + \varepsilon \tag{2}$$

where W is the set of spatial weighting (w_{ij}) between stations i and j, λ is an unknown coefficient, also called spatial autoregressive parameter and used to measure the effect of spatial lag (Wy) on y.

More generally, if explanatory variables are added to the basic SEM, the Equation can be expressed as:

$$y = \lambda W y + X\beta + \varepsilon \tag{3}$$

In SLM, if λ=0, the equation transforms into a classical regression model (Equation (1))

The spatial dependence (disturbance items ε) can be defined as follows:

$$\varepsilon = \rho W \varepsilon + \mu \tag{4}$$

where ρ is an unknown coefficient and used to describe spatial heterogeneity, and $\mu \sim N(0, \sigma^2 I_n)$. Equations (1) and (4) together formed SEM. If ρ=0, SEM would be transformed into a classical regression model (Equation (1)).

OLS, SLM and SEM mentioned above are all global regression models, and they are based on the assumption of a global nature of the production function (Pavlyuk, 2016). They yield global coefficients for each inde-

pendent variable, whereas GWR is a local model that presents local coefficients for each independent variable. The GWR can be stated as (Fotheringham et al., 2003):

$$y = X\beta(i) + \varepsilon \quad (5)$$

where $\beta(i)$ is an $n \times k$ vector of coefficients for n rapid transit stations and k explanatory variables, presenting the local production function for each station. The main advantage of GWR model is that not only the coefficients are estimated separately in GWR but also the model provides specific diagnostics for every sample, such as goodness-of-fit measures (R^2) and t-values. Thus, the GWR model is very helpful to observe how the relationships between independent variables and rapid transit ridership vary across space, which enables better understanding of the specific causes in every station or in different parts of a region (Fotheringham et al., 2003; Lloyd and Shuttleworth, 2005).

In addition, to check the multi-collinearity among explanatory variables, the variance inflation factor (VIF) is used in this study. The global R^2, adjusted-R^2 and Akaike Information Criterion (AIC) are used to comprise the goodness-of-fit of different models (OLS, SLM, SEM, GWR).

3 Results

The estimation results of the stepwise OLS regression analysis are presented in Table 2. Among the ten candidate explanatory variables, six were found significant: population, number of lines, number of feeder buses, number of exits, road density and residential area. These significant variables were thus kept in the final model.

Table 2 Summarized results of the stepwise OLS model

Variables	Coef.	SE	Z	*P*-value	VIF
Intercept	–572614	124292.9	–4.61	0.000	–
Population	18.7002	3.3855	5.52	0.000	2.09
Number of lines	245012.3	103918.6	2.36	0.020	1.85
Number of feeder buses	10565.6	3614.959	2.92	0.004	1.35
Number of exits	94176.4	13897.67	6.78	0.000	2.05
Road density	20521.1	10330.76	1.99	0.050	1.70
Residential area	–334297.1	146448.5	–2.99	0.024	1.35
R^2	0.7731				
Adjusted R^2	0.7602				
AIC	3094.42				

Note that the overall impact of built environment variables is relatively small. As expected, ridership is closely related to population and corresponding variables. The explained variance of the model with 6 significant independent variables is 0.7731, while the adjusted-R^2 is 0.7602. All independent variables are significant at the 0.05 level, and the values of VIF are less than 10, which means there is no significant multicollinearity (Mason et al., 1989).

Spatial effects were tested using Moran's statistic, the Lagrange multiplier and the Robust Lagrange multiplier in GeoDa. Table 3 shows the results. For the spatial error test, the Moran's I statistic is 1.980 and the P-value is 0.047, which indicates that SEM would be appropriate for considering spatial error. The statistic and P-value of Lagrange multiplier are 2.140 and 0.143, and the corresponding values for the Robust Lagrange multiplier are 5.139 and 0.023. The latter tends to be more accurate in model selection (Anselin et al, 1996; Macdonald-Wallis et al., 2011). For the spatial lag test, both the p-value of the Lagrange multiplier and Robust Lagrange multiplier are higher than 0.05, suggesting that SEM should be considered due to spatial autocorrelation.

Table 4 presents the estimation results for both SLM and SEM. According to the P-value of the spatial autoregressive parameter rho of SLM and the spatial autoregressive parameter of the error lambda of SEM, only the former is significant at the 0.05 level, indicating that spatial effect exists mostly in the disturbance terms. The values of R^2 and the AIC test suggest there are barely any improvements after using SLM (R^2=0.7741, AIC= 3096.03) instead of OLS (R^2=0.7731, AIC= 3094.42), whereas SEM (R^2=0.7820, AIC= 3091.48) may be more appropriate.

After OLS, SLM and SEM were used in the study

Table 3 Diagnostic tests for spatial effects in OLS regression

	Test	Value	MI/DF	*P*-value
Spatial error	Moran's I	1.980	0.081	0.047
	Lagrange multiplier	2.140	1	0.143
	Robust Lagrange multiplier	5.139	1	0.023
Spatial lag	Lagrange multiplier	0.198	1	0.656
	Robust Lagrange multiplier	3.198	1	0.073

Table 4 Estimation results of SLM and SEM

Variables	SLM			SEM		
	Coef.	Z	*P*-value	Coef.	Z	*P*-value
Intercept	–564448	–4.69	0.000	–613006	–4.96	0.000
Population	19.9539	5.39	0.000	18.5853	5.22	0.000
Number of lines	244814	2.44	0.015	236913	2.41	0.016
Number of feeder buses	11426.8	3.08	0.002	10507.3	2.95	0.003
Number of exits	93369.6	6.95	0.000	97444.9	7.35	0.000
Road density	20860.9	2.08	0.038	24369.3	2.39	0.016
Residential area	–336935	–2.38	0.017	–301792	–2.10	0.035
rho (lambda)	–0.074	–0.680	0.496	0.2872	2.06	0.039
R^2	0.7741			0.7820		
AIC	3096.03			3091.48		

based on a global perspective, the GWR model was used to provide the coefficients and test values of each station from a local perspective. The local statistics can report more details and differences between stations and identify non-regularity across space (Cardozo et al., 2012; Pavlyuk, 2016). Table 5 details the estimation results of the GWR model relative to OLS, SLM and SEM results. The R^2 and adjusted R^2 are 0.8418 and 0.8001, indicating the GWR model improved the goodness-of-fit by four percent compared to the OLS model. The AIC, sigma-square and Residual sum of squares of GWR model were all the lowest among the four models, which suggests that the GWR model is the model with the best performance.

The spatial distribution of the residuals from OLS, SLM, SEM and GWR models are presented in Fig. 3. The difference between the residuals in the OLS and SLM models is not obvious, whereas those of the SEM model are smaller. Fig. 3 also shows that the residuals of the GWR model are the smallest among all four models. Both the diagnostic test in Table 5 and spatial distribution of residuals in Fig. 3 show that the local model-GWR has the best model fit, followed by the global models-SEM, SLM and OLS.

With the GWR model, it is possible to analyze the spatial variability of estimated coefficients of each explanatory variable. Fig. 4 presents the spatial distribution of the estimated coefficients and *t*-values. The population is significant almost over the whole area. The mean of the population coefficient is 17.5808, which suggests that ridership would increase by 17.5808 for one more person per square kilometer within the metro

Table 5 Estimation results and diagnostic statistics of the models

Variables	OLS	SLM	SEM	GWR		
				Mean	Min.	Max.
(Intercept)	−572614	−564448	−613006	−650046	−891213	−322134
Population	18.7002	19.9539	18.5853	17.5808	7.7731	24.8272
Number of lines	245012	244814	236913	191826	−133729	408285
Number of feeder buses	10565.6	11426.8	10507.3	9883.9	−2640.9	19900.7
Number of exits	94176.4	93369.6	97444.9	106197.7	58947.46	136729.3
Road density	20521.1	20860.9	24369.3	28608.6	6440.4	42091.8
Residential area	−334297.1	−336935	−301792	−290156.2	−598940.3	273818.6
R^2	0.7731	0.7741	0.7820		0.8418	
Adjusted R^2	0.7602	–	–		0.8001	
AIC	3094.42	3096.03	3091.48		3089.566	
Sigma-square	5.499e+10	5.133e+10	4.953e+10		4.583e+10	
Residual sum of squares	5.774e+12	5.749e+12	5.547e+12		4.026e+12	

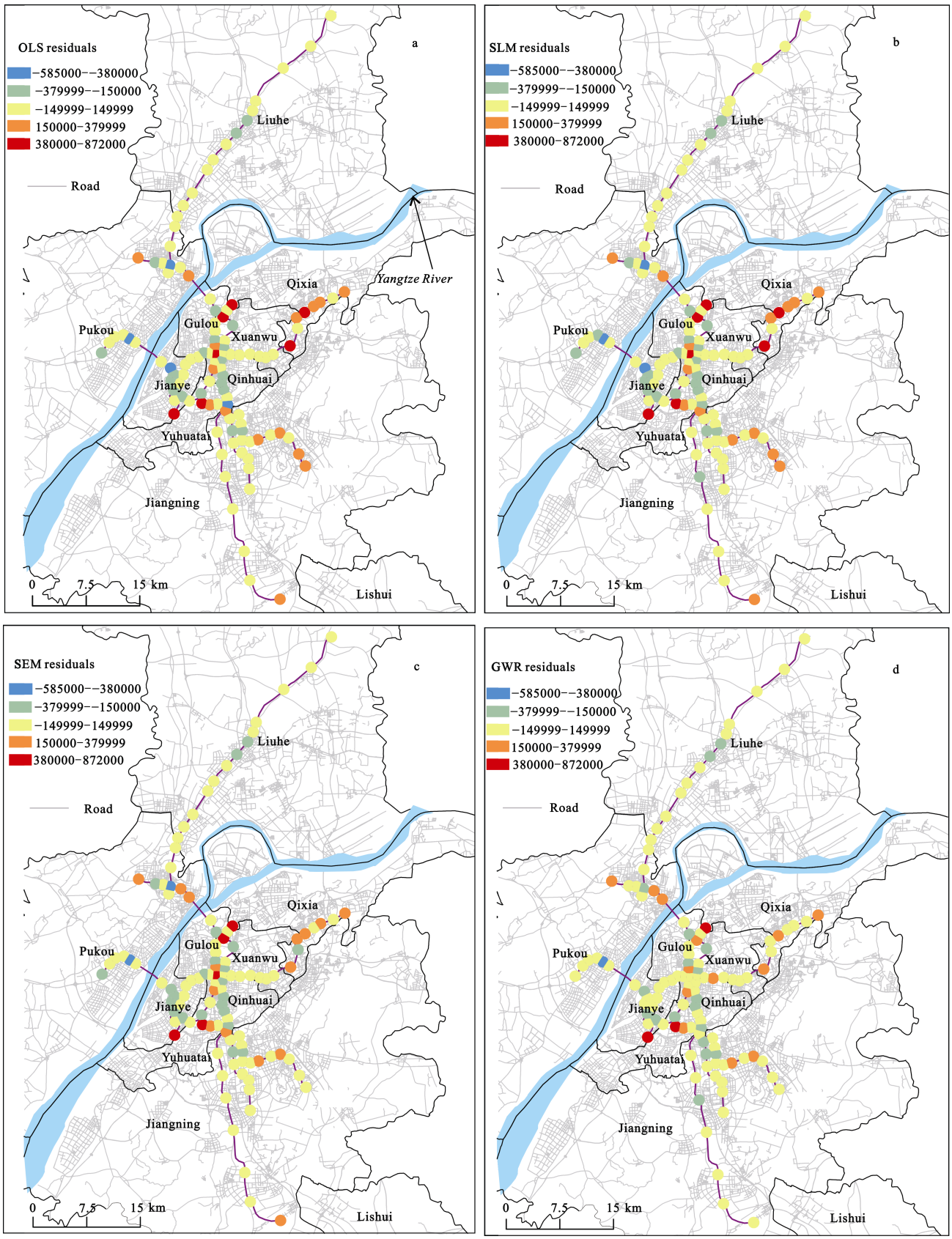

Fig. 3 Spatial distribution of residuals

station catchment area. However, the coefficients were higher (more trips per person) in the northern, western and southern parts than in the central and eastern parts of Nanjing. The number of metro lines was found most significant in the central and southern parts, while the Number of feeder buses in the catchment area is most significant in the north-central and western parts. The central part is the city center of the Nanjing region with Nanjing Railway Station located in the north-central part, and the northern and western parts are separated from the main city by the Yangtze River, where the trip demand is higher than that in other parts of Nanjing. Fig.4 also shows that the number of exits is significant at a 0.05 level (the absolute value of *t*-value >1.96) all over the metro catchment areas. Road density had a positive relationship with ridership (Table 5). The relationship is significant in the center, western and southern parts, while it is not significant in the northern and eastern parts which indicates the spatial variation of Road density in these areas should be interpreted

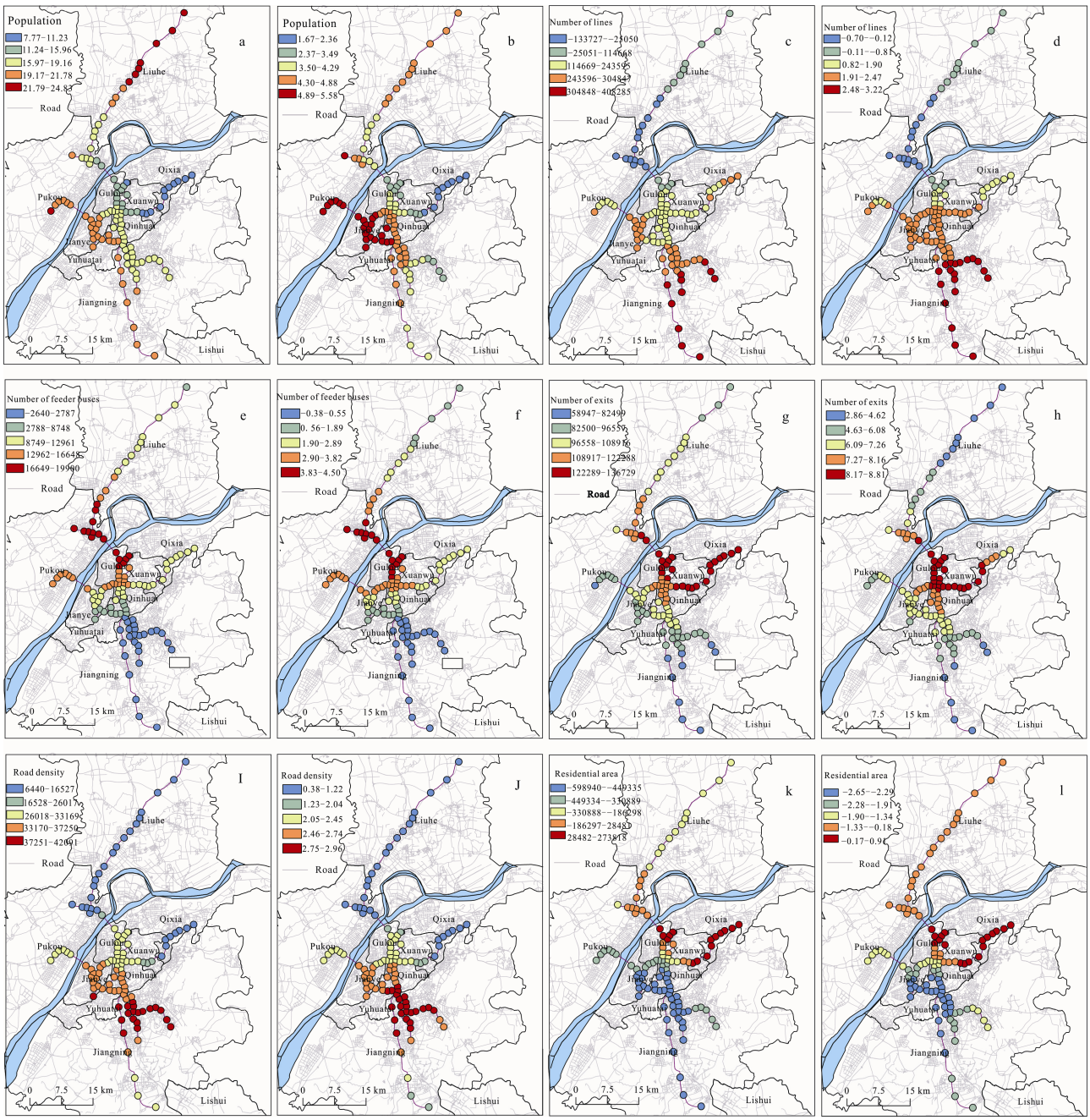

Fig. 4 Spatial distribution of estimated coefficients (a, c, e, g, i, k) and corresponding t-values (b, d, f, h, j, l) of the GWR model

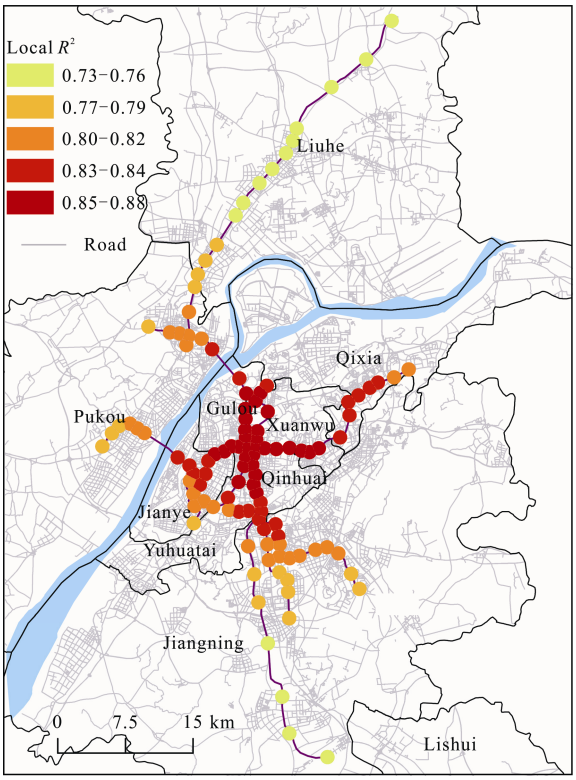

Fig. 5 Spatial distribution of local R^2 of the GWR model

carefully. By contrast, the spatial distribution of t-values of residential area suggests that the spatial variation in the center, western and southern parts should be interpreted carefully, because the corresponding coefficients were not significant at the 0.05 level.

In addition, it is possible to analyze the spatial variability of the GWR model's explanatory power according to the spatial distribution of local R^2. Fig. 5 presents the spatial distribution of local R^2. The model has a higher explanatory power in the central parts than in outskirt areas with a low metro station density. The same finding was also discovered in Madrid (Cardozo et al., 2012). A probable reason is that the distribution of metro station in the central area is more intensive than that in outskirt areas so that more samples (stations) make the goodness of fit superior. Nevertheless, all local R^2 values, mapped in Fig. 5, are higher than 0.73.

4 Discussion

4.1 Spatial heterogeneity

The purpose of this study is two-folded: one is discovering the spatial heterogeneity in station ridership, the other is examining the relationship between independent variables and station ridership, especially from a local perspective. The diagnostic tests for spatial effects (Table 3) showed the existence of spatial heterogeneity in the ridership analysis, which is ignored in the OLS model and usually results in biased estimation results. To cope with the spatial effect issue, SLM and SEM are applied and the estimated coefficients in the two models are all statistically significant. According to the value of spatial autoregressive parameters rho and lambda, the positive spatial autocorrelation is addressed in the disturbance items and the spatial dependence (spatial lag) of station ridership is not very obvious. The values of R^2, AIC, Sigma-square and residual sum of squares also suggest that SEM outperforms SLM in the ridership analysis. The spatial heterogeneity played an important and necessary role in station ridership. The coefficients of each variable vary across space, which also verifies the existence of spatial heterogeneity. It is possible to use the GWR model to better reflect the spatial heterogeneity or dependence from a local perspective.

4.2 Relationship between independent variables and station ridership

We estimated the global parameters of the OLS, SLM and SEM models, and the local parameters of the GWR model. Estimation results between different models were generally consistent. The coefficients of population vary from 17.5808 to 19.9539 (Table 5). Similar findings can also be found in other studies (Kuby et al., 2004; Zhao et al., 2014; Durning and Townsend, 2015; Jun et al., 2015). A positive relationship was also found between number of lines and station ridership. Station ridership increases from 191826 to 245012 if there was one more metro line passing through the station. The number of feeder buses in the catchment area is positively associated with ridership, which indicates that the metro-feeder bus intermodality within metro station areas have synergistic impacts on increasing metro riders. This finding is consistent with most previous studies, such as Kuby et al., 2004; Sohn and Shim, 2010; Gutiérrez et al., 2011; Zhao et al., 2014; Jun et al., 2015. The coefficients (9883 to 11427 passengers a month) were smaller than those in the previous studies in Nanjing (503 passengers a weekday) and Seoul (1382 passengers a weekday) (Sohn and Shim, 2010; Zhao et al., 2014). This may because of the longer distance of remote and suburban metro stations that were built in

Nanjing.

The number of exits has a significant positive effect on metro station ridership (Table 5). The coefficient for the number of exits ranges from 93369.6 to 106197.7, indicating it has a strong effect on station ridership after controlling for other variables. An interpretation may be the number of exits (entrances) improves the accessibility of the station and gives the perception of shorter waking distance. For example, Xinjiekou Station has 19 exits in April 2015 and the linear distance between the two farthest exits (No. 14 and No. 20 exits) is 450 m, while the catchment area was set to 800 m in the current paper. There is no doubt that having more exits may extend the actual catchment area through increasing the station's external accessibility and connectivity, which would attract more people to use metro.

As expected, road density was positively related to station ridership. This finding corresponds with Zhao et al. (2014) results, who analyzed Nanjing metro system and stated that increasing 10 m road length in the catchment area would add six passengers. Durning and Townsend (2015), on the other hand, found road density is negatively associated with ridership and recognized this difference was likely a product of using network-based buffers (Durning and Townsend, 2015). According to the GWR model the spatial distribution of the coefficients has hot spots in the center and central-southern parts of Nanjing, where road density played a more significant role (Fig. 4).

Residential area has a negative relationship with ridership, indicating that a larger percentage of residential areas within the station catchment area tends to decrease the number of passengers. However, results of prior research are rather mixed. Pulugurtha and Agurla (2012) found that the share of residential land use within the buffer of a bus stop has a negative association with ridership, while Sung and Oh (2011), and Durning and Townsend (2015) found a positive correlation with transit station ridership. The results of other previous studies on Nanjing metro system showed a weak negative association between residential areas with transit station ridership (Zhao et al., 2013; 2014). This result may in part be mitigated by the fact that business, office and commercial area may attract more people than residential area in general. The GWR model suggests that a higher residential area ratio decreases metro ridership in the center and central-southern parts of Nanjing.

5 Conclusions

This study aimed to identify spatial heterogeneity in the Nanjing Metro System and examine the influence of built environment and transit attributes on metro ridership. Traditional stepwise OLS model, global spatial models- SLM and SEM, and local spatial model-GWR were applied to that end. These models extend classical rapid transit ridership analysis by considering spatial heterogeneity.

The major findings of this study can be summarized as follows. First, based on ten candidate independent variables, six variables are significantly related with station ridership: population, number of lines, number of feeder buses, number of exits, road density, and residential area. Especially, number of exits and residential area, which were not examined or turned out to be positively related with station ridership in existing studies, seems significantly and negatively associated with station boardings in Nanjing. Second, the evidence of spatial heterogeneity is found and it is most reflected in the error items. The spatial autocorrelation is decreased or eliminated and the estimation errors become smaller by using the SEM instead of the OLS model. Third, the GWR model has the best explanatory power, and it provides a local perspective to better understand the spatial heterogeneity. This finding provides additional insights into the analysis of urban rail transit ridership and support evidence-based urban and transportation planning policy development. To sum up, the main contribution of the present study is two-fold. One is that this study proved that it is necessary to take spatial heterogeneity into consideration when predicting urban rail transit ridership and the other is that the provided methodological framework of spatial analysis of urban rail transit ridership can be adopted for urban and transportation policy-makers.

The findings of this study have several implications. First, the results suggest that it is reasonable for Chinese government to give priority to public transit development and build urban rapid transit systems. Second, in general, the number of feeder buses affects all metro station ridership, indicating that metro station ridership would be significantly increased by an additional feeder bus line due to the synergistic impacts of metro-feeder bus intermodality. The evidence in the paper supports the importance of feeder buses to increase metro station

ridership, especially for the area separated by the Yangtze River in Nanjing, even though the coefficient of number of feeder bus lines is smaller than in previous studies in Nanjing or Seoul (Sohn and Shim, 2010; Zhao et al., 2013). Third, the number of exits on station ridership is positively associated with station ridership, indicating that multi-exits help metro stations attract more passengers, because multi-exits actually extend the passenger-attracting scope, especially for the city center and densely populated areas. It also suggests it is important to reserve several exits for new-built stations instead of building few exits at the beginning, considering that metro systems need to carry increasingly more passengers during its growth. Fourth, road density is found to be an important factor in promoting metro station ridership. One possible reason for this is that the small block with high-density road network provides a good environment of walking and bicycling for passengers to get access to metro stations. It implies that designing relatively small block with high-density road network in new metro station areas is expected to attract more metro users. Therefore, road density should be underlined to promote metro station ridership, especially in achieving transit-oriented development (TOD) purpose (Sung and Oh, 2011). Fifth, unlike what previous research concluded, the estimation results of the models indicate that the proportion of residential area does not have a positive effect on station boardings. In contrast, the proportion of residential area is negatively associated with station ridership, especially in the areas where large parts of the land are used for commerce and offices. Moreover, the results are not very consistent across various Nanjing regions and in some parts not significant. It suggests researchers should be hesitant to use proportions to explain absolute ridership. Policy-makers should be cautious to make their decisions on such possible spurious results about the impact of the built environment on ridership. Finally, the GWR model demonstrates local ridership effects of built environment and transit attributes by stations. It implies that it is necessary for urban and transportation policy-makers to refer to the GWR results to determine which factors should be improved in different station area in order to increase metro ridership at station-level.

Interpreting our findings, it should be noted that our study has two limitations. First, because of the difficulty of collecting individual or household information within the station catchment areas, socio-economic variables were not considered in the study. Although some researchers stated there were no significant associations between these variables and ridership (e.g., Durning and Townsend, 2015), there is an overwhelming amount of evidence that activity participation varies with socio-demographics, while in addition car ownership in developing countries is strongly related to sociodemographic variables. Second, station ridership needs to be validated by using longitudinal or panel data. We plan to conduct such validation when we obtain the ridership data for the year 2015.

References

Anselin L, Bera A K, Florax R et al., 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1): 77–104. doi: 10.1016/0166-0462(95)02111-6

Blainey S P, Preston J M, 2013. Extending geographically weighted regression from points to flows: a rail-based case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(6): 724–734. doi: 10.1177/0954409713496987

Cardozo O D, García-Palomares J C, Gutiérrez J, 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34: 548–558. doi: 10.1016/j.apgeog.2012.01.005

Cervero R, Ferrell C, Murphy S, 2002. *Transit-Oriented Development and Joint Development in the United States: A Literature Review*. Washington, DC: Transportation Research Board.

Chan S, Miranda-Moreno L, 2013. A station-level ridership model for the metro network in Montreal, Quebec. *Canadian Journal of Civil Engineering*, 40(3): 254–262. doi: 10.1139/cjce-2011-0432

Choi J, Lee Y J, Kim T et al., 2012. An analysis of Metro ridership at the station-to-station level in Seoul. *Transportation*, 39(3): 705–722. doi: 10.1007/s11116-011-9368-3

Dill J, Schlossberg M, Ma L et al., 2013. *Predicting Transit Ridership at the Stop Level: the Role of Service and Urban Form*. Washington, DC: Transportation Research Board.

Durning M, Townsend C, 2015. Direct ridership model of rail rapid transit systems in Canada. *Transportation Research Record: Journal of the Transportation Research Board*, 2537(1): 96–102. doi: 10.3141/2537-11

Estupiñán N, Rodríguez D A, 2008. The relationship between urban form and station boardings for Bogota's BRT. *Transportation Research Part A: Policy and Practice*, 42(2): 296–306. doi: 10.1016/j.tra.2007.10.006

Fotheringham A S, Brunsdon C, Charlton M E, 2003. *Geographically Weighted Regression: the Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Gan Z, Yang M, Feng T et al., 2018. Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations. *Transportation*. doi: 10.1007/s11116-018-9885-4

Guerra E, Cervero R, Tischler D, 2012. Half-mile circle: does it best represent transit station catchments? *Transportation Research Record: Journal of the Transportation Research Board*, 2276(1): 101–109. doi: 10.3141/2276-12

Gutiérrez J, Cardozo O D, García-Palomares J C, 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6): 1081–1092. doi: 10.1016/j.jtrangeo.2011. 05.004

Jun M J, Choi K, Jeong J E et al., 2015. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *Journal of Transport Geography*, 48: 30–40. doi: 10.1016/j.jtrangeo.2015.08.002

Kepaptsoglou K, Stathopoulos A, Karlaftis M G, 2017. Ridership estimation of a new LRT system: direct demand model approach. *Journal of Transport Geography*, 58: 146–156. doi: 10.1016/j.jtrangeo.2016.12.004

Kim D, Ahn Y, Choi S et al., 2016. Sustainable mobility: longitudinal analysis of built environment on transit ridership. *Sustainability*, 8(10): 1016. doi: 10.3390/su8101016

Kuby M, Barranda A, Upchurch C, 2004. Factors influencing light-rail station boardings in the United States. *Transportation Research Part A: Policy and Practice*, 38(3): 223–247. doi: 10.1016/j.tra.2003.10.006

Lloyd C, Shuttleworth I, 2005. Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. *Environment and Planning A: Economy and Space*, 37(1): 81–103. doi: 10.1068/a36116

Loo B P Y, Chen C, Chan E T H, 2010. Rail-based transit-oriented development: lessons from New York City and Hong Kong. *Landscape and Urban Planning*, 97(3): 202–212. doi: 10.1016/j.landurbplan.2010.06.002

Macdonald-Wallis K, Jago R, Page A S et al., 2011. School-based friendship networks and children's physical activity: a spatial analytical approach. *Social Science & Medicine*, 73(1): 6–12. doi: 10.1016/j.socscimed.2011.04.018

Mason R L, Gunst R F, Hess J L, 1989. *Statistical Design and Analysis of Experiments: with Applications to Engineering and Science*. New York: Wiley.

McNally M G, 2007. The four-step model. In: Hensher D A, Button K J (eds). *Handbook of Transport Modelling*. Oxford: Pergamon, 35–53.

Ministry of Housing and Urban-Rural Development, 2012. *Guidelines for Planning and Design of Urban Rail*. Beijing: China Construction Industry Publishing House. (in Chinese)

Páez A, 2006. Exploring contextual variations in land use and transport analysis using a probit model with geographical weights. *Journal of Transport Geography*, 14(3): 167–176. doi: 10.1016/j.jtrangeo.2005.11.002

Pavlyuk D, 2016. Implication of spatial heterogeneity for airports' efficiency estimation. *Research in Transportation Economics*, 56: 15–24. doi: 10.1016/j.retrec.2016.07.002

Pulugurtha S S, Agurla M, 2012. Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, 15(1): 33–52. doi: 10.5038/2375-0901.15.1.3

Rasouli S, Timmermans H, 2014. Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, 18(1): 31–60. doi: 10.1080/12265934.2013.835118

Ryan S, Frank L F, 2009. Pedestrian environments and transit ridership. *Journal of Public Transportation*, 12(1): 39–57. doi: 10.5038/2375-0901.12.1.3

Sohn K, Shim H, 2010. Factors generating boardings at metro stations in the Seoul metropolitan area. *Cities*, 27(5): 358–368. doi: 10.1016/j.cities.2010.05.001

Sung H, Oh J T, 2011. Transit-oriented development in a high-density city: identifying its association with transit ridership in Seoul, Korea. *Cities*, 28(1): 70–82. doi: 10.1016/j.cities.2010.09.004

Tobler W R, 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(S1): 234–240.

Zhang D P, Wang X K, 2014. Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC. *Journal of Transport Geography*, 41: 107–115. doi: 10.1016/j.jtrangeo.2014.08.021

Zhao J B, Deng W, Song Y et al., 2013. What influences metro station ridership in China? Insights from Nanjing. *Cities*, 35: 114–124. doi: 10.1016/j.cities.2013.07.002

Zhao J B, Deng W, Song Y et al., 2014. Analysis of Metro ridership at station level and station-to-station level in Nanjing: an approach based on direct demand models. *Transportation*, 41(1): 133–155. doi: 10.1007/s11116-013-9492-3