The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2632-0487.htm

Exploring influencing factors on transit ridership from a local perspective

Yuxin He, Yang Zhao and Kwok Leung Tsui
School of Data Science, City University of Hong Kong, Kowloon, Hong Kong

Received 13 June 2019
Revised 27 August 2019
Accepted 6 September 2019

Abstract

Purpose – Exploring the influencing factors on urban rail transit (URT) ridership is vital for travel demand estimation and urban resources planning. Among various existing ridership modeling methods, direct demand model with ordinary least square (OLS) multiple regression as a representative has considerable advantages over the traditional four-step model. Nevertheless, OLS multiple regression neglects spatial instability and spatial heterogeneity from the magnitude of the coefficients across the urban area. This paper aims to focus on modeling and analyzing the factors influencing metro ridership at the station level.

Design/methodology/approach – This paper constructs two novel direct demand models based on geographically weighted regression (GWR) for modeling influencing factors on metro ridership from a local perspective. One is GWR with globally implemented LASSO for feature selection, and the other one is geographically weighted LASSO (GWL) model, which is GWR with locally implemented LASSO for feature selection.

Findings – The results of real-world case study of Shenzhen Metro show that the two local models presented perform better than the traditional global model (OLS) in terms of estimation error of ridership and goodness-of-fit. Additionally, the GWL model results in a better fit than GWR with global LASSO model, indicating that the locally implemented LASSO is more effective for the accurate estimation of Shenzhen metro ridership than global LASSO does. Moreover, the information provided by both two local models regarding the spatial varied elasticities demonstrates the strong spatial interpretability of models and potentials in transport planning.

Originality/value – The main contributions are threefold: the approach is based on spatial models considering spatial autocorrelation of variables, which outperform the traditional global regression model – OLS – in terms of model fitting and spatial explanatory power. GWR with global feature selection using LASSO and GWL is compared through a real-world case study on Shenzhen Metro, that is, the difference between global feature selection and local feature selection is discussed. Network structures as a type of factors are quantified with the measurements in the field of complex network.

Keywords Local, Influencing factors, LASSO, Geographically weighted regression, Metro ridership

Paper type Research paper

1. Introduction

Urban rail transit (URT) plays a critical role in maintaining effective passenger mobility nowadays. URT ridership at the station level is known to be influenced by interaction among multiple factors (e.g. land-use, socio-economics, intermodal traffic accessibility and

Smart and Resilient Transportation
Vol. 1 No. 1, 2019
pp. 2-16
Emerald Publishing Limited
e-ISSN: 2632-0495
p-ISSN: 2632-0487
DOI 10.1108/SRT-06-2019-0002

© Yuxin He, Yang Zhao and Kwok Leung Tsui. Published in *Smart and Resilient Transportation*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at: http://creativecommons.org/licences/by/4.0/legalcode

This work was supported by the National Natural Science Foundation of China (No. 71901188 and No. 11471275), and the Research Grants Council Theme-based Research Scheme (No. T32-101/15-R).

metro network structure, etc.). Exploring the influence of these factors is vital to accurately estimate travel demand and to effectively make design schemes of urban systems including the identification of which public infrastructures, services and resources need to be built and deployed. Modeling URT ridership at the station level can help to not only estimate and forecast ridership but also analyze the influencing factors on it.

Given the need to understand the effects of multiple factors on URT ridership, a growing number of recent studies have sought to model transit ridership. As one of the best-known models, the four-step (generation, distribution, mode choice and assignment) model has been widely used since the 1950s. However, its weaknesses are also obvious, such as low model accuracy, low data precision, insensitivity to land use, institutional barriers and high expense (Gutiérrez *et al.*, 2011). As an alternative to the four-step model, direct demand models have become popular in ridership estimation in recent decades. Direct demand models estimate ridership as a function of influencing factors within the pedestrian catchment areas (PCA) via regression analysis, which enable identifying factors that contribute to higher transit ridership (Gutiérrez *et al.*, 2011; Choi *et al.*, 2012; Cervero, 2006; Kuby *et al.*, 2004; Chu, 2004). In the models, a PCA is a geographic area for which a station attracts passengers. The size and shape of a catchment area depend on how accessible a station is and how far it is from alternative stations. One can use buffers to create circular catchment areas by a specific distance or use Thiessen polygons to illustrate the area most accessible to each station. Direct demand models have distinct advantages in travel analysis, such as simplicity of use, easy interpretation of results, immediate response, and low cost. As a kind of direct demand models, ordinary least square (OLS) multiple regression which assumes parametric stability is generally used (Gutiérrez *et al.*, 2011; Kuby *et al.*, 2004; He *et al.*, 2018; Sohn and Shim, 2010; Loo *et al.*, 2010; Sung and Oh, 2011; Thompson *et al.*, 2012; Zhao *et al.*, 2013; Chan and Miranda-Moreno, 2013; Singhal *et al.*, 2014; Liu *et al.*, 2014). In other words, OLS considers that the coefficients estimated do not have significant differences in space. With the development of spatial modeling, direct demand models could increase their spatial explanatory power by using geographically weighted regression (GWR), which is designed to model spatial parametric non-stationarity and variance heterogeneity. In recent years, Cardozo *et al.* (2012) compared the performance of OLS and GWR in modeling transit ridership and its influencing factors, and GWR showed better goodness-of-fit than OLS for forecasting station-level ridership. Furthermore, the study of GWR with penalized forms (e.g. ℓ_1 norm) can be found in Wheeler's (2009) studies. Wheeler (2009) introduced least absolute shrinkage and selection operator (LASSO) into the GWR framework, called geographically weighted LASSO (GWL) to simultaneously conduct coefficient regularization and local model selection, which has the capability to reduce prediction and estimation errors for estimating the response variable in GWR.

In light of deficiencies of popularly used global direct demand models (OLS multiple regression), considering the advantage of spatial models including GWR and GWL for modeling potentially spatially varying relationships, we applied two local direct demand models based on GWR with global implemented LASSO and GWL into modeling influencing factors on metro ridership. For the former, we select features by implementing LASSO globally for all calibration locations before the process of GWR, and for GWL, we can select features for each station by implementing LASSO locally. The ridership and its potential influencing factor data of Shenzhen Metro in the year of 2013 are used to elaborate the two models. Besides, we conduct a relevant comparison analysis of results generated from those two models.

Our main contributions are threefold:

- The approach taken is based on spatial models considering spatial autocorrelation of variables, which outperform the traditional global regression model OLS in terms of model fitting and spatial explanatory power.
- GWR with global feature selection using LASSO and GWL are compared through a real-world case study on Shenzhen metro, that is, the difference between global feature selection and local feature selection is discussed.
- Network structures as a type of factors are quantified with the measurements in the field of complex network.

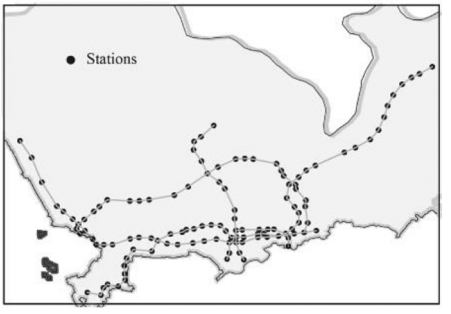
The remainder of this paper is organized as follows. In Section 2, we outline the profiles of study area and the data description. In Section 3, we provide a description of the methodology we used. Section 4 conducts results analysis and the comparison between GWR with global LASSO and GWL models. Finally, Section 5 contains concluding remarks.

2. Area of study and data

Our study focuses on Shenzhen Metro network, which consists of five lines and 118 stations in the year of 2013 (Figure 1)[1]. Figure 1 shows the spatial distribution of those stations. Shenzhen Metro ridership data at the station level were aggregated by using the data collected through AFC system of Shenzhen Metro Corporation in China. The data set includes the total information about entry–exit smart card records. The data used in the research cover a time span of seven days from October 14 (Monday) to 20 (Sunday) in 2013. We summed boarding and alighting ridership amounts and then calculated the average daily ridership of the whole week. The explanatory variables represent factors hypothesized to influence station ridership.

2.1 Response variable

This paper aims to identify and analyze multiple factors influencing station ridership. We conduct preliminary statistical analysis using metro AFC data on October 14. Figure 2(a) shows the spatial distribution of AFC data records in one day. It presents that the records are most densely distributed at Grand Theatre station and Laojie station, closely followed by Huaqiang Road station and Luohu station, and the records of other stations have relatively sparse distribution.



Note: Spatial distribution of Shenzhen Metro stations

Figure 1.
Shenzhen Metro map of 2013

Figure 2(b) is about temporal distribution of AFC data records, which shows that the spatial distribution of records has a peak value at both 8:00 and 18:00 on both weekdays and weekends. Additionally, the characteristics of temporal distribution of records on weekdays and weekends is quite similar, which suggests that there are similar metro travel patterns, with morning and evening peaks on weekdays and weekends in Shenzhen.

Therefore, the models with average daily ridership of the whole week (the operation times of Shenzhen metro is 6:30-23:00) as the response variable will be built intending to find the factors influencing the station-level ridership.

2.2 Explanatory variables

The explanatory variables represent factors hypothesized to influence station ridership (Table I). The variables can be classified into four categories:

(1) land use;

(2) social economics;

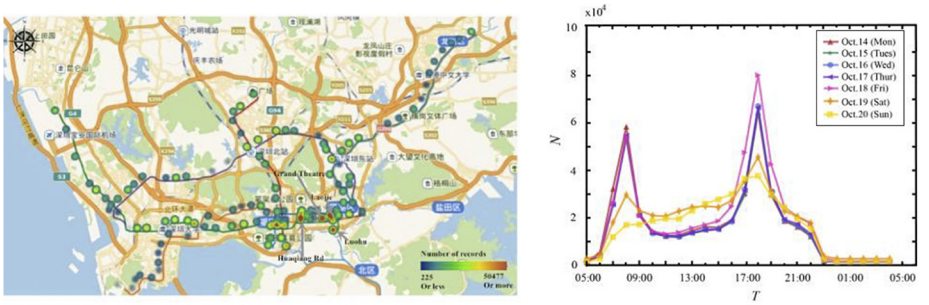
(3) intermodal traffic access variables; and

(4) network structure.

As the average friendly walking distance is generally assumed to be 500 m in large- and middle-sized cities according to Dovey *et al.* (2017), we also define the distance of PCA of each Shenzhen Metro station as 500 m. In our work, we use a buffer to create circular PCA by 500 m. Based on the buffer with a radius of 500 m determined, population, all of the land use-related data and the number of bus stations were collected subsequently.

2.2.1 Land use variables. All of the land use–related data within a PCA were collected from Baidu Map with the assistance of API, and land use variables consist of the residences, entertainment, services, business, education and offices closer to the station. Specifically, the information covers the numbers of residence, restaurants, schools, working buildings, hospitals, banks, shopping places and hotels within 500-m PCA.

2.2.2 Social – economics variables. Social–economic variables consist of the population distribution of Shenzhen in 2013 and operation days since the metro stations opened. The information of days since the metro lines and stations opened was collected from a website



(a) (b)

Figure 2. Spatial and temporal distribution of AFC data records

Notes: (a) Spatial distribution of records in one day; (b) temporal distribution of records per hour in one day

Categories	Explanatory variables	Acronym of variables	Source
Land Use	No. of residential units	*Residence*	Baidu map
	No. of restaurants	*Restaurant*	Baidu map
	No. retailers/shopping	*Shopping*	Baidu map
	No. of schools	*School*	Baidu map
	No. of offices	*Offices*	Baidu map
	No. of banks	*Bank*	Baidu map
	No. of hospitals	*Hospital*	Baidu map
	No. of hotels	*Hotel*	Baidu map
Network Structure	Distance to the city center	*Dis_to_cent*	Calculated
	Degree centrality	*Degree*	Calculated
	Betweenness centrality	*Between*	Calculated
Social Economics	Population	*Pop*	Worldpop
	Days since opened	*Days_open*	Baidu baike
Intermodal Traffic Access	No. of bus stations	*Bus*	Baidu map

Table I.
Summary of explanatory variables

named "UrbanRail"[2]. The higher residential population is hypothesized to be positively associated with ridership. Here, we obtained information about population distribution in the whole city of Shenzhen in 2013 from the website of Worldpop[3]. During data pre-processing, the population within each buffer can be obtained by summing up the value of the grid falls into the metro station buffer by using ArcGIS 10.2. Figure 3 shows the population distribution of the whole city of Shenzhen in 2013 and 500-m buffers of metro stations.

Through the preliminary visualization in Figure 3, it is noted that population is densely distributed near the metro region. The influence of population density within each station buffer on ridership is pending for analysis in the model.

2.2.3 Intermodal traffic access variables. As for intermodal traffic access, here we considered the feeder bus system. The number of bus stations near a metro station was hypothesized to be positively related to station ridership, which was also collected from the Baidu Map.

2.2.4 Network structure variables. In this paper, network structure variables comprise the degree centrality and betweenness centrality of the metro network nodes and the distance to the city center. In the field of complex networks, as the degree is a simple

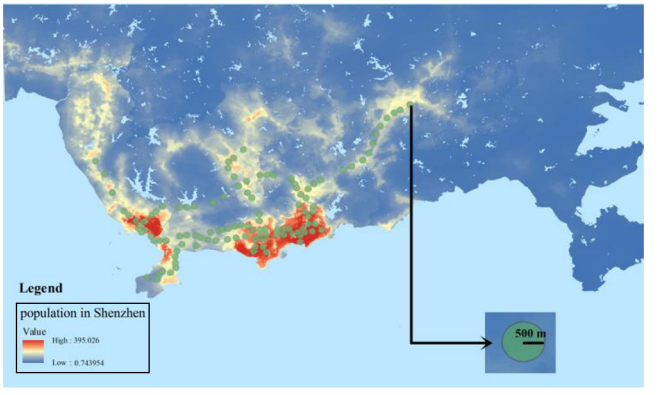


Figure 3.
Population distribution and 500-m buffers of metro stations

centrality measure that counts how many neighbors a node has, and the betweenness centrality for each node refers to the number of shortest paths that pass through the node (Erciyes, 2014); thus, they are correlated to the information for transfer stations or terminal stations, and the importance of stations in the aspect of their controlling overflows passing between others of metro networks. As for the distance Dist_i of each station to the city center, which is Shenzhen Municipal People's Government, located in Futian District, we calculate it by the following equation (1) considering the effect of the radius of the earth:

$$\mathrm{Dist}_i = \mathrm{R} \cdot \arccos\begin{pmatrix} \cos(Lat_0) \cdot \cos(Lat_i) \cdot \cos(Lon_0 - Lon_i) \\ +\sin(Lat_i) \cdot \sin(Lat_0) \end{pmatrix} \cdot \frac{\pi}{180} \tag{1}$$

Where, R is the radius of the earth, and (Lat_0, Lon_0) and (Lat_i, Lon_i) are the latitude and longitude of the city center and station i, respectively. The related geographical data were collected from Google Maps.

3. Methodology

3.1 GWR with global LASSO

The first method is to implement LASSO for all stations' variables first to perform variable selection, and after that feed the selected explanatory variables into the GWR model to understand the spatially varied effects of those selected factors on metro station ridership.

3.1.1 Geographically weighted regression. In this study, we use geographically weighted regression (GWR) models to estimate station-level ridership. GWR model is an extension of ordinary least squares (OLS) or linear least squares, which is shown as follows:

$$y_i = \beta + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \tag{2}$$

Geographical location factors are introduced into regression parameters to allow local parameter estimation, and the extended GWR model is as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{3}$$

Where y_i and $x_{i1}, x_{i2},\ldots,x_{ip}$ are observed values of the response variable y and explanatory variables $x_1, x_2,\ldots,x_p$ at the location of (u_i, v_i), which is geospatial coordinates of the observation point $i = (1,2,\ldots,n)$, and ε_i is the normally distributed error term (with the expected value 0 and constant variance). $\beta_k(u_i, v_i)(k = 1,2,\ldots,p)$ refers to p unknown functions associated with the spatial position. The geographic location of each observation point (u_i, v_i) is weighted by GWR model, and the weight generally is a kind of the distance decay function (Fotheringham and O'Kelly, 1989). In the model, the determination of bandwidth will directly affect the weight function and also the precision of the model, thus the determination of bandwidth is crucial.

3.1.2 Least absolute shrinkage and selection operator. The structured data has 14 explanatory variables (shown in Table I) with a limited amount of observations, which may cause multicollinearity and overfitting. Redundant variables should be removed to make the process of modeling more efficient. Therefore, before fitting the regression model, it is necessary to select features from the original variables candidates. As a kind of shrinkage

methods, LASSO tends to not only reduce the variability of the estimates, thus improving the model's stability, but also set some of the coefficients to zero, enabling variable selection. LASSO makes use of the ℓ_1 norm. ℓ_1 penalties are convex and the assumed sparsity can lead to significant computational advantages. LASSO is defined as follows:

$$\hat{\beta}^R = arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 \tag{4}$$

Subject to:

$$\sum_{k=1}^{p} \left| \beta_k \right| \leq s \tag{5}$$

where s is a parameter that controls the degree of coefficient shrinkage. Tibshirani (1996) proved that LASSO constraint $\sum_k |\beta_k| \leq s$ is equivalent to adding the penalty term $\lambda \sum_k |\beta_k|$ to the residual sum of squares (RSS). Thus a direct relationship between s and $\lambda \geq 0$ which is a complexity parameter that controls the degree of shrinkage of coefficients. Hence, coefficients of LASSO can also be expressed as:

$$\hat{\beta}^R = arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} \left| \beta_k \right| \right\} \tag{6}$$

The generally used methods for solving LASSO are standard convex optimizer (Gauraha, 2018) and least angle regression (LARS) (Efron *et al.*, 2004). In this study, we adopted LARS to solve LASSO.

3.2 Geographically weighted LASSO model

The second method is based on the GWL framework developed by Wheeler (2009). This method performs the local model selection by implementing LASSO for each station, so that one can understand what factors influence which stations and how strong the influencing effects are.

The algorithm to estimate the GWL solutions is shown as following:

Step 1: estimate the local scaling GWL parameters (shrinkage parameter s_i at each location i and bandwidth b) by minimizing leave-one-out-cross-validation (LOOCV) root mean square error (RMSE). Here we choose the bandwidth b in the binary search for the minimum RMSPE.

- Calculate the $n \times n$ inter-point distance matrix $\mathbf{D}$ with the coordinates (u_i, v_i) of station i.
- Calculate the $n \times n$ weights matrix $\boldsymbol{W}$ using the distance matrix $\mathbf{D}$ and the initial b according to $w_j(i) = \exp\left[-\left(\frac{d_{ij}}{b} \right)^2 \right]$. The diagonal elements in the weights matrix are defined as $\boldsymbol{W}(i) = diag[w_1(i),\ldots,w_n(i)]$.
- For each station i, $i = 1,\ldots,n$:

- Set the square root of $\mathbf{W}(i)$ as $\mathbf{W}^{1/2}(i)$ and $\mathbf{W}^{1/2}(i)_{ii} = 0$, i.e., set the (i,i) element of the square root of the diagonal weights matrix to 0 to delete the observation point i.
- Calculate $X_w = \mathbf{W}^{1/2}(i)\mathbf{X}$ and $y_w = \mathbf{W}^{1/2}(i)\mathbf{y}$ with $\mathbf{W}^{1/2}(i)$at station i.
- Call **lars** (X_w,y_w), seek the lasso solution that minimizes the error for y_i, and save it.
- Stop when there is a slight change in the estimated b. Save the estimated b, parameter s_i, the matrix of regression coefficients $\hat{\boldsymbol{\beta}}(i) = \left(\hat{\beta}_{i0}, \hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip}\right)^T$ and indicator vector $\mathbf{z}$ of which variable coefficients are shrunken to zero.

Step 2: estimate the final local scaling GWL solutions using the shrinkage parameter s_i and kernel bandwidth parameter b estimated in *Step 1*.

(1) Calculate the weights matrix $\boldsymbol{W}$ using the distance matrix $\mathbf{D}$ and the b estimated in *Step 1*.

(2) For each location i, $i = 1,\ldots,n$:
- Set the square root of $\mathbf{W}(i)$ as $\mathbf{W}^{1/2}(i)$.
- Calculate $X_w = \mathbf{W}^{1/2}(i)\mathbf{X}$ and $y_w = \mathbf{W}^{1/2}(i)\mathbf{y}$ using $\mathbf{W}^{1/2}(i)$ at station i.
- Call **lars** (X_w, y_w) and save the series of lasso solutions.
- Choose the lasso solution that matches the LOOCV solution on the basis of the shrinkage parameter s_i and the indicator vector $\mathbf{z}$.

4. Results and discussion

4.1 Spatial autocorrelation test to variables

Before building GWR and GWL models, the analysis is performed to determine if the candidate variables are spatially autocorrelated. The test of spatial autocorrelation can detect how strong spatial correlation of variables is, which will provide a theoretical basis for the feasibility of applying a GWR model. Moran's I is a measure of spatial autocorrelation developed by Moran (1950). Moran scatter plot can reflect the spatial autocorrelation intuitively. The scatter plot has four quadrants. If the observed value falls to the first and third quadrants, it indicates that there is a strong positive spatial correlation. If it falls to the second and fourth quadrants, it indicates there is a strong negative spatial correlation. Figure 4 shows several variables' Moran scatter plot.

According to Moran scatter plots (see Figure 4), it can be seen that Moran's I values of all variables above aren't equal to 0, which indicates these variables are not randomly distributed in space, and mostly falls to the first and the third quadrants. It shows that each variable is positively spatially correlated more or less, especially three explanatory variables, namely, population, distance to the city center and days since opening, have strong spatial correlations as the values of Moran's I are all greater than 0.3 (Cressie, 1992). The result also lays the foundation for the feasibility of the follow-up study.

4.2 Results of Model 1 (GWR with global LASSO)

Since strong spatial correlation has been found for the variables in the research, it is reasonable to build GWR models to analyze influencing factors on station ridership of Shenzhen Metro. Through variables' selection based on LASSO, the explanatory variables selected from the candidate variables listed in Table I are *pop, Between, Days_open, Shopping, Dis_to_cent*. It indicates that population, betweenness centrality, days since stations opened, numbers of shopping places within PCA and distance of stations to the city

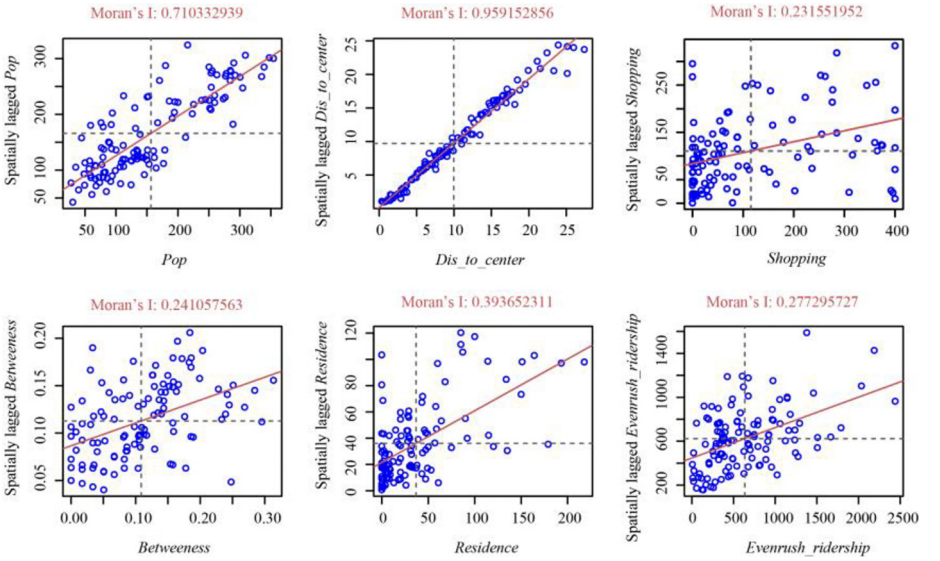


Figure 4.
Moran scatterplot of variables

center are important features for influencing metro station ridership for most of metro stations in Shenzhen.

Next up, GWR for modeling average daily ridership of the whole week and its influencing factors is built. The results of GWR model compared with those of OLS model that also includes the same variables selected by implementing LASSO are presented in Table II.

First, according to Table II, AICc value of GWR model is less than that of their corresponding global regression (OLS) model. According to the evaluation criterion of Brunsdon *et al.* (1996), if the AICc value of GWR model is at least 3 less than that of OLS model, we can consider that GWR model fits better than OLS model even considering the complexity of GWR model. What is more, the adjusted R^2 value of GWR model is obviously greater than that of the corresponding OLS model, which shows that GWR model has strong explanatory power even under consideration for model complexity. Likewise, the parameter value (Sigma) indicating the model error of GWR model is also lower, and the residual sum of square from the GWR model is smaller than that from the OLS model. Generally speaking, the results show that the goodness-of-fit indicators of GWR model perform better than those of OLS model. Additionally, ANOVA tests shown in Table II are carried out to find out if the global (OLS) regression model and the GWR model have the same statistical performance (the same size of error variance). The results of ANOVA test suggest that there is a significant improvement when GWR is adopted.

GWR model for average daily ridership of a whole week regression performs pretty well in terms of the value of R^2, which means that we only need to know the information of population distribution, betweenness centrality, days since stations opened, number of shopping places within PCA and distance of stations to the city center; we can use GWR model to explain 81 per cent of the response variable and average daily ridership, and meanwhile, the data related to these explanatory variables are quite easy to collect.

According to Voronoi algorithm (Fu *et al.*, 2006), the Shenzhen Metro coverage area can be divided into several Thiessen polygons according to the locations of stations. In this context, the spatial distribution of local coefficients is visualized by Thiessen polygons. Through

Varaiables	Global (OLS) Estimate	Standard error	*t*(Est/SE)	Local (GWR) Minimum	Maximum	Mean	STD
Intercept	148902.14	8373.14	17.78	−220846.06	1259330.97	230997.13	231395.86
Pop	29682.08	9949.31	2.98	−87929.62	174341.19	24227.35	57929.46
Betweenness	25971.88	9096.29	2.86	−75396.67	122215.16	30440.91	53705.86
Days_open	36322.47	8796.81	4.13	−449460.34	868771.29	72348.20	158603.43
Shopping	15242.75	9217.93	1.65	−68601.68	58895.47	1474.80	29877.34
Dis_to_cent	−10074.73	9248.92	−1.09	−207415.70	755777.31	43828.83	157781.05

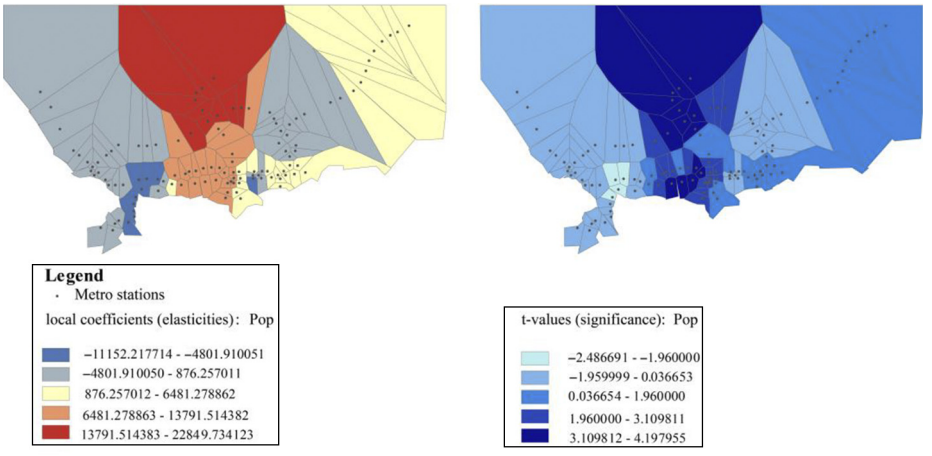
Diagnostic		
R^2	0.35	0.81
Adjusted R^2	0.32	0.65
Sigma	90925.69	64811.17
AICc	3038.33	3035.27
Residual sum of squares	925957952739.49	269172991697.74
Number of parameters	6	37.83

GWR ANOVA Table

Source	SS	DF	MS	*F*	*p*-Value
Global residuals	925957952739.493	6.000			
GWR improvement	656784961041.749	47.919	13706255804.666		
GWR residuals	269172991697.743	64.081	4200487280.123	3.263016	0.00000618

Table II.
Results of GWR with global LASSO model and OLS model with LASSO

understanding the spatial distribution of local coefficients (elasticities) and *t*-values (significance), it is possible to know how relations between the variables vary across space (estimated coefficients) and with what statistical significance. Take *pop* as an example. The mean of the coefficients from the population variable is 3284.89 [see Figure 5(a)]. It means that for each person within the station catchment area, the number of trips increases by



Notes: (a) Local coefficients distribution; (b) t-values distribution

Figure 5.
Spatial distribution of local coefficients for *pop* (elasticities) and *t*-values (significance)

3284.89 per day. However, these elasticities distribute unevenly in space. More trips per capita were expected in the center and mid-north, where commerce, administration and education are concentrated, while elasticity values were lower in the west and east. Moreover, the *t*-values map on the right shows that the effect of population is more significant in the middle area at a 0.05 level (the absolute value of *t*-values larger than 1.96) [Figure 5(b)].
In general, GWR has strong spatial explanatory power based on the local analysis of the variation of each coefficient across space (elasticities).

4.3 Results of Model 2 (GWL) and comparative analysis of two models

GWL model (GWR with locally implemented LASSO enabling simultaneous coefficient penalization and model selection) is conducted on the Shenzhen Metro data set. The comparison of results of GWL model and OLS model, OLS with LASSO for feature selection, GWR and GWR with global LASSO for feature selection for estimating average daily ridership of the whole week are shown in Figure 6.

The accuracy of the estimated responses is measured by calculating RMSE. The RMSE is the square root of the mean of the squared deviations of the estimates from the true values and should be small for accurate estimators. R^2 is a statistical measure that represents the proportion of the variance for a response variable that's explained by explanatory variables.

In general, the performance rank of five methods is "GWL>GWR with global LASSO>GWR>OLS>OLS with LASSO".

First, we can see the superiority of three local models for feature selection (GWL, GWR with global LASSO, GWR) over the global models (OLS, OLS with LASSO) in terms of the estimation error of response variable and goodness-of-fit. Second, it should be noted that the local models such as GWR with global LASSO and GWL perform better than the original version of GWR model, which proves the importance of feature selection. Third, GWL performs better than GWR with global LASSO, which indicates that the locally implemented LASSO for each station during the procedure of GWR performs better than the globally implemented LASSO for feature selection before GWR. Fourth, GWL for metro network performs substantially better than the other four models at estimating the response variable. Therefore, we can conclude that GWL model which incorporates locally implemented LASSO for the metro network is able to estimate Shenzhen Metro ridership more accurately.

To investigate which station a certain variable has the most impact on, the local regression coefficients' distribution for each variable of all stations in GWL model is plotted in bubble plots. The spatial distribution of local coefficients is shown in Figure 7.

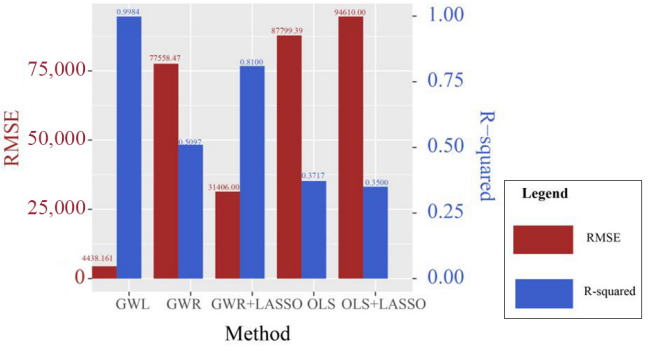


Figure 6.
Comparison of regression performance of models for ridership estimation

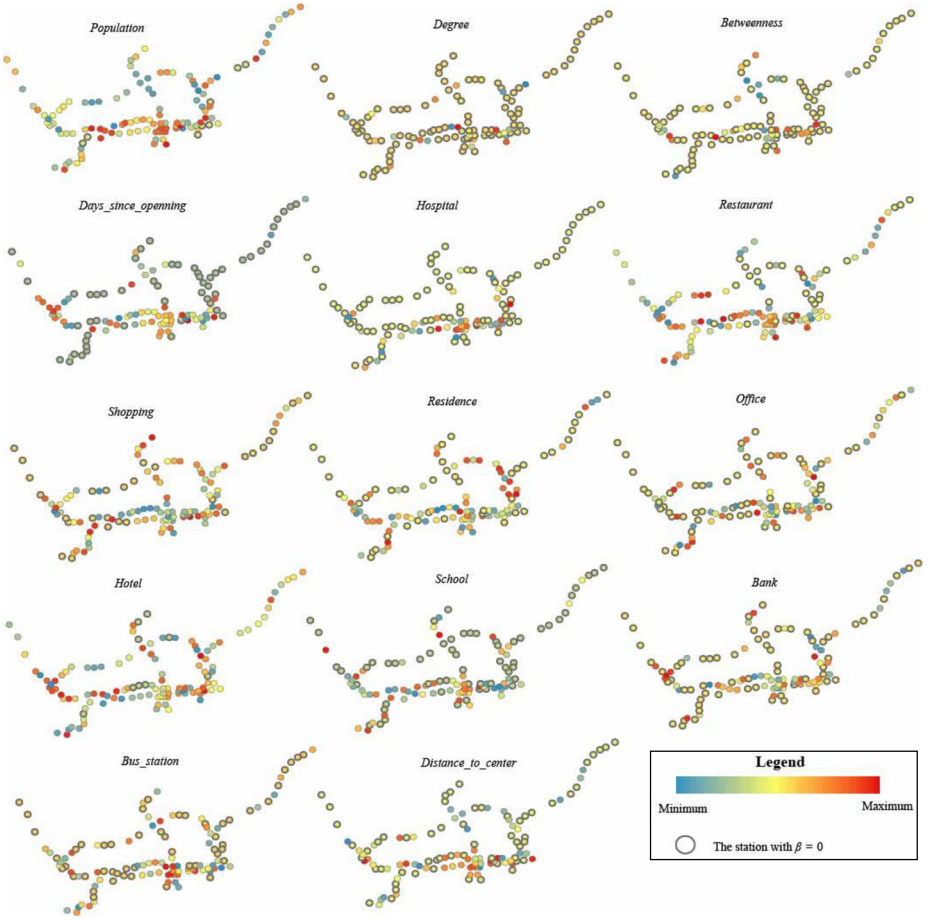


Figure 7. Spatial distribution of local coefficients (elasticities) of GWL model

Through understanding the spatial distribution of local coefficients (elasticities), the relations between the variables varying across space (estimated coefficients) and variables selection of all stations can be revealed. In Figure 7, the bubble with the bold outline demonstrates the coefficient for the variable in the station equals to 0, and other bubbles' colors identify the range of coefficients; the bubble with lighter colors means the coefficient is larger, and vice versa. Take the population factor as an example; stations with large positive coefficients are mainly distributed in the center, indicating that more trips per capita were expected in the central south area of the metro network, where commerce, administration and education are concentrated. Besides, we can note that for the factors of degree and hospital, there are numerous stations with zero coefficient, so these factors are not so important factors for influencing most of Shenzhen Metro stations' ridership.

Through comparing the interpretation of the coefficients of GWR with global LASSO and GWL models, we can find that the coefficients of both models are spatially varied, helping us gain local insights into analyzing the influencing factors of Shenzhen Metro ridership. In addition, for Model 1, the explanatory variables are selected for all stations

uniformly before conducting GWR, whereas for Model 2, the variables of each station are selected respectively during the procedure of GWR, and therefore, the difference between the coefficients of two models are: first, the coefficients of Model 2 include all potential candidate variables initially but Model 1 selects several important factors at the beginning. Second, for some stations, the coefficients for the certain variables of Model 2 may be shrunk to zero, but the coefficients for variables of Model 1 cannot be zero, which means that for Model 2, different stations may have different influencing factors and the degree of impact also can be varied, and for Model 1, we can only discuss the spatially varied impacts of those common important factors on the metro ridership of stations in different locations. Moreover, in Model 2, factors with coefficients of numerous stations being zero are in accordance with the factors which are not selected in Model 1, such as hospital and degree. In other words, GWL model paid more attention to the spatial difference of influence of factors on metro ridership at each station than GWR model with LASSO does. Generally, both models can provide us local perspectives more or less while interpreting coefficients.

5. Conclusion

In summary, this paper builds two spatial models to analyze the influencing factors of Shenzhen Metro ridership at the station level from a local perspective. One model is GWR model with global LASSO for variables selection, and the other one is GWL model, which implements LASSO for each calibration location during the procedure of GWR, i.e. GWR with local LASSO. We demonstrate the applicability of these two models through the spatial autocorrelation test and superiority of them over global models through a real-world case study of Shenzhen Metro systems, and meanwhile, we not only analyze the influencing factors of Shenzhen metro station-level ridership from a local perspective but also conduct a comparative analysis on these two models. Additionally, different from previous work, we borrow the conceptions, including degree centrality and betweenness centrality, from complex network theory to better quantify the network structure factors related to the practical significance of metro networks, which cover comprehensive information compared with dummy variables.

The results of the case study show that the local models including GWL model, GWR without feature selection and GWR model with global LASSO perform better than global models including OLS and OLS with LASSO in terms of estimation error and goodness-of-fit. Besides, the estimation error of GWL is lower than that of GWR with global LASSO, which indicates the locally implemented LASSO for each station during the procedure of GWR performs better than globally implemented LASSO for feature selection before GWR. With regards to the interpretation of coefficients of two models, the coefficients of GWL model include all potential candidate variables initially but GWR with global LASSO model select several important factors at the beginning. Additionally, for GWL model, different stations may have different influencing factors and the degree of impact also can be varied, and for GWR with global LASSO, we can only discuss the spatially varied impacts of those common important factors on the metro ridership of stations in different locations. To sum up, GWL model pays more attention to the spatial difference of influence of factors on metro ridership at each station than GWR model with global LASSO does.

In general, the two local models presented in this paper not only improve the performance of traditional OLS multiple regression on modeling metro ridership and its influencing factors in terms of goodness-of-fit and estimation error but also inspired metro planning, passenger flows management and periphery development from a local perspective.

Notes

1. Source: http://toursmaps.com/wp-content/uploads/2017/02/shenzhen_metro_map-1.gif
2. Source: www.urbanrail.net/as/cn/shen/shenzhen.htm
3. Source: www.worldpop.org.uk/data/get_data/

References

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E. (1996), "Geographically weighted regression: a method for exploring spatial nonstationarity", *Geographical Analysis*, Vol. 28 No. 4, pp. 281-298.

Cardozo, O.D., García-Palomares, J.C. and Gutiérrez, J. (2012), "Application of geographically weighted regression to the direct forecasting of transit ridership at station-level", *Applied Geography*, Elsevier Ltd, Vol. 34 No. 4, pp. 548-558.

Cervero, R. (2006), "Alternative approaches to modeling the travel-demand impacts of smart growth", *Journal of the American Planning Association*, Vol. 72 No. 3, pp. 285-295.

Chan, S. and Miranda-Moreno, L. (2013), "A station-level ridership model for the metro network in montreal, Quebec", *Canadian Journal of Civil Engineering*, Vol. 40 No. 3, pp. 254-262.

Choi, J., Lee, Y.J., Kim, T. and Sohn, K. (2012), "An analysis of metro ridership at the station-to-station level in Seoul", *Transportation*, Vol. 39 No. 3, pp. 705-722.

Chu, X. (2004), *Ridership Models at the Stop Level Final Report*, Fehrs and Peers Associates.

Cressie, N. (1992), "Statistics for spatial data", *Terra Nova*, Vol. 4 No. 5, pp. 613-617.

Dovey, K., Pafka, E. and Ristic, M. (Eds) (2017), *Mapping Urbanities: Morphologies, Flows, Possibilities*, Routledge.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least angel regression", *The Annals of Statistics*, Vol. 32 No. 2, pp. 407-499.

Erciyes, K. (2014), *Complex Networks: An Algorithmic Perspective*, CRC Press.

Fotheringham, A.S. and O'Kelly, M.E. (1989), *Spatial Interaction Models: Formulations and Applications*, Vol. 1, Kluwer Academic Publishers, Dordrecht, p. 989.

Fu, T.L., Yin, X.T. and Zhang, Y. (2006), "Voronoi algorithm model and the realization of its program", *Computer Simulation*, Vol. 23 No. 10, pp. 89-91.

Gauraha, N. (2018), "Introduction to the LASSO", *Resonance*, Vol. 23 No. 4, pp. 439-464.

Gutiérrez, J., Cardozo, O.D. and García-Palomares, J.C. (2011), "Transit ridership forecasting at station level: an approach based on distance-decay weighted regression", *Journal of Transport Geography*, Vol. 19 No. 6, pp. 1081-1092.

He, Y., Zhao, Y. and Tsui, K.L. (2018), "An analysis of factors influencing metro station ridership: Insights from Taipei metro", *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, IEEE*, Vol. 2018, pp. 1598-1603.

Kuby, M., Barranda, A. and Upchurch, C. (2004), "Factors influencing light-rail station boardings in the United States", *Transportation Research Part A: Policy and Practice*, Vol. 38 No. 3, pp. 223-247.

Liu, C., Erdogan, S. and Ducca, F.W. (2014), "How to increase rail ridership in Maryland? Direct ridership models (DRM) for policy guidance", *Journal of Urban Planning and Development*, Vol. 142 No. 4, pp. 1-10.

Loo, B.P.Y., Chen, C. and Chan, E.T.H. (2010), "Rail-based transit-oriented development: lessons from New York City and Hong Kong", *Landscape and Urban Planning*, Elsevier B.V., Vol. 97 No. 3, pp. 202-212.

Moran, P.A.P. (1950), "Notes on continuous stochastic phenomena", *Biometrika*, Vol. 37 Nos 1/2, pp. 17-23.

Singhal, A., Kamga, C. and Yazici, A. (2014), "Impact of weather on urban transit ridership", *Transportation Research Part A: Policy and Practice*, Elsevier Ltd, Vol. 69, pp. 379-391.

Sohn, K. and Shim, H. (2010), "Factors generating boardings at metro stations in the Seoul metropolitan area", *Cities*, Elsevier Ltd, Vol. 27 No. 5, pp. 358-368.

Sung, H. and Oh, J.T. (2011), "Transit-oriented development in a high-density city: identifying its association with transit ridership in Seoul, Korea", *Cities*, Elsevier Ltd, Vol. 28 No. 1, pp. 70-82.

Thompson, G., Brown, J. and Bhattacharya, T. (2012), "What really matters for increasing transit ridership: understanding the determinants of transit ridership demand in Broward county, Florida", *Urban Studies*, Vol. 49 No. 15, pp. 3327-3345.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society*, Vol. 58 No. 1, pp. 267-288.

Wheeler, D.C. (2009), "Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso", *Environment and Planning A: Economy and Space*, Vol. 41 No. 3, pp. 722-742.

Zhao, J., Deng, W., Song, Y. and Zhu, Y. (2013), "What influences metro station ridership in China? Insights from Nanjing", *Cities*, Elsevier Ltd, Vol. 35, pp. 114-124.

Corresponding author
Yang Zhao can be contacted at: yangzhao9-c@my.cityu.edu.hk

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**