IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 5, 2018, accepted November 19, 2018, date of publication November 28, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2883462

Spatial Analysis of Bikeshare Ridership With Smart Card and POI Data Using Geographically Weighted Regression Method

JIE BAO[1,2,3], XIAOMENG SHI[1,2,3], AND HAO ZHANG[1,2,3,4]

[1]Jiangsu Key Laboratory of Urban ITS, Nanjing 211189, China
[2]Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Nanjing 211189, China
[3]Southeast University, Nanjing 211189, China
[4]Faculty of Transportation Engineering, Huaiyin Institute of Technology, Huai'an 223001, China

Corresponding author: Hao Zhang (andyhao@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71874067, in part by the Social Science Foundation of Jiangsu Province under Grant 15EYC004, in part by the Science and Technology Project of Ministry of Housing and Urban-Rural Development under Grant 2016-R3-012, in part by the Philosophy and Social Science Foundation of Jiangsu Province Universities under Grant 2017SJB1641, and in part by the Scientific Research Foundation of Graduate School of Southeast University under Grant YBJJ1790.

ABSTRACT The primary objective of this paper is to explore the spatial analysis of bikeshare ridership with a consideration of the diversity across different station categories using smart card data and points of interests (POIs) data. The bikeshare trip records were obtained from the Citi Bike system of New York City. The POI data in the vicinity of each station were collected through the Google Places API. K-means clustering method was employed to classify the bikeshare stations into five categories. Then, the geographically weighted regression (GWR) method was applied to establish the relationship between bikeshare ridership and various kinds of influencing factors. To account for the diversity across different station categories, five separate GWR models for each station category were developed and compared with the joint model of all station categories. The results of likelihood ratio test confirmed the superiority and importance of building separate models for each bikeshare station category instead of a joint model. In addition, all the developed bikeshare ridership models were applied to predict the ridership of the newly opened stations in the next year. The results were indicated that the prediction performance of separate bikeshare ridership models was generally better than that of the joint model. The findings of this paper could help transportation agency to develop specific planning and management strategies for each station category of the entire bikesharing system.

INDEX TERMS Bikeshare, POI, K-means, geographically weighted regression, spatial analysis.

I. INTRODUCTION

Over the past few decades, bikesharing systems have grown rapidly in urban areas, providing individuals a healthy and green transport mode for short trips [1]–[9]. Currently, more than 800 cities have implemented their own bikesharing systems around the globe [9]. Bikeshare have been considered a promising way to resolving the last mile problem of public transit and reducing traffic congestion and air pollution in urban areas. However, a series of problems emerged recently due to the fluctuating spatial and temporal demand of shared bicycles, such as the high operating costs, low usage rates, and the inefficient bike repositioning [10]–[14].

To improve the operational efficiency of bikesharing systems and attract more people to use shared bicycles, considerable efforts have been devoted to examining the factors that affect the bikeshare ridership [1], [3], [10], [11], [13], [15]–[17]. Rixey [1] investigated the effects of demographic and built environment characteristics nearby stations on the bikeshare ridership in three different bikesharing systems. The results suggested that population density, job density and the income level surrounded by each station are all positively correlated with the bikeshare ridership for all the three bikesharing systems. Noland *et al.* [13] estimated the effects of bicycle infrastructure, employment population, land use mix and transit access on bikeshare trip generation by seasons of the year, weekday/weekend and user type. They found that bikeshare stations located near busy subway stations and bicycle infrastructure are usually related with higher usage frequency, and stations surrounded by greater population

and employment are usually related with greater usage. Faghih-Imani *et al.* [3] examined the influence of meteorological data, temporal characteristics, bicycle infrastructure, land use and built environment attributes on arrival and departure flows of the bikesharing system in Montreal. The results of multilevel regression models indicated that the bicycle flows are expected to decrease with the increased distance away from CBD, and increase with the increased number of restaurants, commercial enterprises and universities in the vicinity of a station.

Although these studies have provided important insights into the influencing factors of bikeshare ridership, two important issues have been greatly neglected. First, each category of stations have their unique characteristics in built environment, land use, travel patterns and trip purposes [15]. Neglecting the diversity across different station categories may hide some important findings associated with specific stations. Moreover, suggestions and strategies that aim at improving the operations of bikesharing systems should be developed separately for each station category, mainly due to the varied influencing of the contributing factors across different station categories. Second, the spatial autocorrelation problem should be considered in modelling the station-level ridership. Traditional ordinary least squares (OLS) multiple regression models usually have fixed coefficients of explanatory variables, which will fail to capture the spatial heterogeneity of influencing factors across stations.

Some researchers started to considering the bikeshare station clusters in the ridership modeling. For example, Hyland *et al.* [16] clustered the bikeshare stations based on the type of arriving trips using k-means or fuzzy c-means clustering techniques. Although the stations can be well clustered based on the number of arriving trips, this method may fail to allocate new stations to specific clusters due to the lack of recorded trip information. Recently, with the wide application of the location based services, many online social media such as Foursquare, Twitter, and Google Map can accurately recommend some places that may attract users to visit. In these social media applications, individuals can check in some point of interests (POIs) and share their activities, emotions, and experiences of these places. Traditionally, POI is defined as a specific point of location that someone may find useful or interesting. Most consumers use this term when referring to restaurants, hotels, park or any other categories applied in digital maps [17]. Some previous studies have suggested that the POIs surrounding each station have great potential to reveal the travel patterns and possible trip purposes [18], [19], and accordingly can be used to classify the category of bikeshare stations.

In addition, to address the spatial heterogeneity problem when modelling bikeshare ridership data, the geographically weighted regression (GWR) method was employed in this study. GWR method has been recently widely used to model the ridership of various transportation modes, such as taxi [20], public transit [21], and ride-sourcing [22]. Some recent studies also have started applying GWR to the spatial analysis of bikeshare-related problems [23]. For example, Ji *et al.* [23] applied geographically weighted Poisson regression (GWPR) model to explore the factors that influence the metro-bikeshare transfer. Compared with other spatial statistical methods such as spatial autoregressive models (SAR) and spatial error models (SEM), the GWR method can be specified easily, and the spatial distribution of the coefficients of variables can also be displayed in a more intuitive manner [24]. To the best knowledge of the authors, this paper is one of the first attempts to directly apply GWR to model the station-level bikeshare ridership with the consideration of diversity across different station categories. The key goal of this paper can be summarized as follows:

(a) To classify the bikeshare stations into different categories based on the POI data in the vicinity of each station.
(b) To explore the potential influencing factors of bikeshare ridership in New York City accounting for the diversity across different station categories.
(c) To show the justification of using separate models for each station category instead of a joint model.
(d) To investigate the prediction performance of developed bikeshare ridership models in new built stations.
(e) To provide insights and suggestions from models of each station category which lead to appropriate policy implications.

The rest of the paper is organized as follows. Section 2 discusses the procedures for gathering various types of data from multiple data sources. Section 3 describes the methodology of this paper. The results of data analysis are presented in Section 4. Section 5 is a summarization of this study.

II. DATA SOURCES

We collected data from the New York City to illustrate the procedure for the spatial analysis of bikeshare ridership. In the present study, five types of data were collected, including bikesharing trip data, POI data in the vicinity of each station, bicycle infrastructure data, weather data, and the socio-demographic characteristics. The data were collected from different data sources. More specifically, the bikesharing trip data were collected from the Citi Bike website [25]. Each trip record contains the following information: start time and date, stop time and date, station name and geo-location, user type and gender. In the present study, we selected the bikesharing trip records of the whole 2015 year, which can represent the variation of bikeshare usage across the whole year. Moreover, trip records with trip duration exceed 120 minutes or less than 2 minutes are further removed, mainly because some users might forget to return the bicycles to the dock correctly or only try to check their new smart cards [15]. Finally, a total of 9,787,566 bikesharing trip records were selected in the study period.

The POI data in the vicinity of each bikeshare station were collected through the Google Places application programming interface (API). For each query, the latitude, longitude and searching radius information of each station are

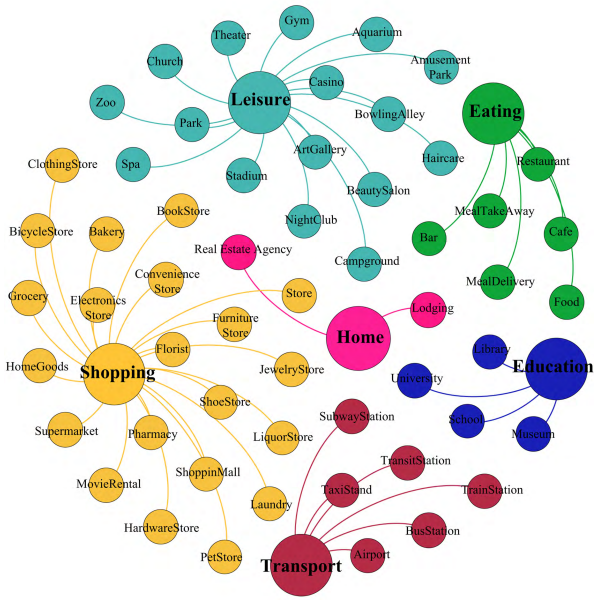


FIGURE 1. Classification of extracted POI data.

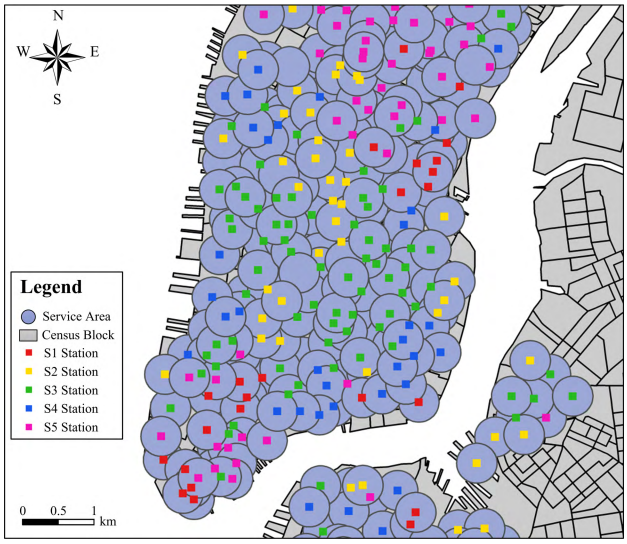


FIGURE 2. Spatial distribution of selected stations and service areas.

sent to the Google Places API. Then, feedback messages are returned, including nearby buildings, POIs with geo-location, and customer ratings. A web crawler was developed by Python in this study to extract the needed POI data from the feedback messages efficiently and to remove the redundant POIs automatically. As suggested by previous studies [26], we set the searching radius of each station as 250 meters regarding the dense urban form of New York City and the average distances between two stations. Finally, a total of 32,476 POIs within 301 bikeshare stations were extracted. To better understand the categories of bikeshare stations, the extracted POIs were further classified into six categories according to the associated trip purposes (See Figure 1).

The bicycle infrastructure, which includes the free sidewalk bicycle parking racks and all the bicycle routes, were collected from the New York City Department of Transportation (NYCDOT). Although the bicycle parking racks may not represent the actual parking locations of shared bicycles, we expect that areas with more bicycle racks will be involved with more cycling activities and thus may generate more bikeshare usage. The weather data were collected from the National Climate Data Center (NCDC) website which provide the monthly weather information across the United States. The socio-demographic characteristics were collected from the U.S. Census Bureau and the American Community Survey (ACS). The information obtained from the social-demographic data include the number of population segregated by ethnicity and age cohorts, the number of college enrollment, median household income, and the number of employment population.

Note that all the collected socio-demographic data are at the census-block-group level. To distribute the census-block-group data to each bikeshare station, we defined the service area for each station, and then distributed census-block-group based data to these service areas weighting by the share of each census-block-group's land area within the station's service area. In this study, we set the radius of service area for each station as 250 meters, which was applied in previous studies [26]. Figure 2 illustrates the spatial distribution of selected bikeshare stations and their service areas. Finally, the number of bicycle parking racks, the length of bike lanes as well as the socio-demographic data were aggregated into corresponding station service areas with the ArcGIS software. The descriptive statistics of all the variables are summarized in Table 1.

III. METHODOLOGY

In the present study, the K-means clustering method was firstly applied to classify bikeshare stations into several categories based on the spatial distribution of different POIs in the vicinity of each station. Then, for each station category, the geographically weighted regression (GWR) method was developed to establish the relationship between the station-level ridership and a variety of candidate explanatory variables listed in Table 1. The methods used in the present study are briefly discussed in this section.

A. K-MEANS CLUSTERING ANALYSIS

K-means clustering analysis is a classic unsupervised machine learning method, which aims to divide M objects in N dimensions into K clusters such that the within-group distances are minimized and between-group distances are maximized [27]. This algorithm firstly divides objects into k clusters randomly, then calculates the centroids of each cluster and assigns each object to the cluster with the nearest centroids. The algorithm runs iteratively until the centroids of each cluster do not change or the pre-defined number of iterations has reached.

TABLE 1. Descriptive statistics of variables.

Variable	Description	Min	Max	Mean	S.D.
Dependent variable					
Log (Bikeshare ridership)	The natural log of the number of monthly originating trips at a given station	0	9.72	5.26	1.58
Independent variables					
Population	Number of people per km^2 in each station's service area	62.18	12164	5031.2	2488.3
Non white	Percent of non-white population in each station's service area (%)	8.51	88.35	33.16	19.67
Employment	Number of employees per km^2 in each station's service area	37.65	7493.7	2998.1	1526.4
College enrollment	Number of enrolled college students per km^2 in each station's service area	4.34	3697.7	537.99	466.74
Bike racks	Number of bike racks in each station's service area	0	105	26.25	21.44
Bike length	The length of bike lanes in each station service area (km)	0	10.01	2.86	1.84
Station capacity	The capacity for each station	3	62	34.16	10.68
MIC	Median household income in each station's service area (10^3dollars)	29.33	435.67	149.26	72.11
Precipitation	The cumulative precipitation for each month (in)	1.86	5.23	3.41	1.24
Snowfall	The cumulative snowfall for each month (in)	0	18.6	4.09	7.16
Temperature	The average temperature for each month (□)	23.9	79	56.65	17.91
Station age	The number of months that station is operating until Dec, 2015	3	27	21.61	6.93
Month	Dummy variables for each month	-	-	-	-

In this study, we denoted each bikeshare station as a mixture of different POI category. For example, $s_1 = (p_1, p_2 \ldots, p_6)$, where $p_1, \ldots p_6$ represent the proportions of each POI category within the service area of station s_1. Then, all the bikeshare stations are labelled as POI feature vector, and input into the K-means clustering algorithms. The final output of the cluster centers could depict the characteristic and land use of each station category.

B. GEOGRAPHICALLY WEIGHTED REGRESSION

The spatial heterogeneity and spatial autocorrelation among explanatory variables are common issues when modelling geo-location data. To address these issues, some spatial statistical models were proposed, such as spatial autoregressive model [28], random-parameter model [29], and Bayesian spatial model [30]. Recently, the geographically weighted regression (GWR) method were proposed and widely applied in travel demand prediction [21], public transit usage forecasting [31], and traffic safety estimation [24]. The GWR method is considered as an extension of traditional linear regression framework, and accordingly easy for specification. Unlike the complex mechanism of Bayesian spatial model, the GWR method is easier for traffic engineers to understand and widely used in practical application. In particular, in the GWR models the coefficients of variables can be visualized in an easily identifiable manner, which could provide insightful suggestions for city planners and bikeshare system operators [24].

The GWR model is different from the traditional linear regression by allowing the coefficients of explanatory variables to vary over space [32]. In this manner, the spatial heterogeneity issue in the bikeshare ridership data could be addressed. The GWR model could be specified as follows:

$$y_{it} = \beta_0(u_{it}, v_{it}) + \sum\nolimits_k \beta_{kt}(u_{it}, v_{it}) x_{kit} + \varepsilon_{it} \tag{1}$$

where (u_{it}, v_{it}) represents the location of centroid of i^{th} bikeshare station during month t in this study. x_{kit} represents the k^{th} explanatory variable with varying coefficients at the i^{th} bikeshare station during month t. $\beta_0(u_{it}, v_{it})$ and $\beta_{kt}(u_{it}, v_{it})$ represents the intercept term and the coefficient of the k^{th} explanatory variable at the i^{th} bikeshare station during month t, respectively.

The expression of the local coefficients $\boldsymbol{\beta}$ suggests that the GWR method could address the spatial heterogeneity issue by allowing the estimated coefficients to vary across different bikeshare stations. Accordingly, the local coefficients $\boldsymbol{\beta}$ can be represented by the following matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \ldots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \ldots & \beta_k(u_2, v_2) \\ \ldots & \ldots & \ldots & \ldots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \ldots & \beta_k(u_n, v_n) \end{bmatrix} \tag{2}$$

where each row denotes the coefficients for each bikeshare station. The coefficients for each station are estimated as follows [32]:

$$\hat{\boldsymbol{\beta}}(i) = (\mathbf{X}^T \mathbf{W}(i) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(i) \mathbf{Y} \tag{3}$$

where $\mathbf{W}(i) = \text{diag}[w_{i1}, w_{i2}, \ldots, w_{in}]$, is a diagonal matrix. w_{ij} denotes the allocated weights for neighboring bikeshare station j in the estimation of the model for bikeshare station i. Several kernel functions were designed to define the weighting scheme, such as Gaussian, exponential and bi-square function [32]. The Gaussian weight function was selected

TABLE 2. The clustering performance of three clustering methods.

Clustering methods	Silhouette values
K-means	0.58
Agglomerative hierarchical clustering	0.33
DBSCAN	0.52

in this study to reflect the distance decay in the weighting scheme, which can be expressed in the following form:

$$w_{ij} = \exp[-\frac{1}{2} \times (d_{ij}/b)^2] \quad (4)$$

where d_{ij} denotes the distance between the centroids of two bikeshare stations, and the parameter b refers to the selected bandwidth. Considering that when the data become scarce the fixed kernel weighting scheme will produce inaccurate estimation results of coefficients, the Gaussian function with adaptive kernel was employed in this study [32]. For areas with more data points the bandwidth of the kernel will be lower, while for areas with few data points the bandwidth of the kernel will be larger. In the GWR method, the Corrected Akaike Information Criterion (AICc) value is a commonly used metric to determine the model specification. Finally, the model with the lowest AICc value will be considered as the best model.

IV. RESULTS OF DATA ANALYSIS

A. RESULTS OF K-MEANS CLUSTERING ANALYSIS

In this study, different clustering methods was employed to classify the bikeshare stations into several different categories based on the spatial distribution of POI data in the vicinity of each bikeshare station [27], [33]. Table 2 illustrates the clustering performance of K-means and other two traditional methods: hierarchical agglomerative clustering and density-based clustering of applications with noise (DBSCAN). The silhouette values of K-means clustering results were higher than the other two methods. In addition, the clustering centers output from K-means analysis could depict the general features in each cluster. Thus, the results of K-means clustering was used in this study.

In this study, we successively conducted the K-means clustering analysis by setting the number of clusters from 3 to 8 to obtain the optimal number of clusters. During this process, when the number of clusters was set as 5, the performance of K-means clustering reached the best. Accordingly, in this study the bikeshare stations were finally clustered into five categories. Table 3 gives the clustering centers for the five station categories, and Figure 3 illustrates the spatial distribution of the five station categories and their surrounding POIs.

The results of clustering analysis reveals that different station categories are surrounded by different land use and have their unique characteristics. More specifically, the clustering center of S1 stations (see Table 3) indicates that most of the POIs within service area are related with transportation facilities, such as transit stations, bus stops and subway stations. It can be inferred that this type of bikeshare station

TABLE 3. The centers of different bikeshare station categories in K-means clustering.

	S1	S2	S3	S4	S5
Home (%)	3.1	4.1	4.37	5.76	14.99
Eating (%)	11.36	22.34	43.01	19.2	23.4
Leisure (%)	6.63	10.03	10.18	15.02	7.02
Shopping (%)	11.68	40.14	22.45	16.58	17.87
Transport (%)	53.21	16.98	13.27	20.46	31.41
Education (%)	14.03	6.4	6.71	22.97	5.31

mainly serves as an important transferring tool for other public transit systems, particularly during commuting time. The dominant POIs surrounding S2 stations are related with shopping stores, and most of this type of stations are located nearby the famous shopping malls of Manhattan like Times Square and Herald Square. In contrast with S2 station, S3 stations are mainly located in the central business district (CBD) area of Manhattan, and have a relatively high proportion of POIs associated with eating places within their service areas. Accordingly, it can be inferred that people are more likely to ride to S2 stations for shopping activities and to S3 stations for eating activities. As is shown in Table 3, the clustering center of S4 stations is mainly related with educational POIs and also surrounded with eating, shopping and transportation related POIs. The characteristics of this station category exhibits a typical school living land use pattern, indicating a main usage for college students. Compared with the other four stations categories, S5 stations have a relatively high proportion of POIs related with residential places, and are mainly located around the Central Park (see Figure 3). Previous studies have revealed that people living around this station type are more likely to ride for commuting on weekdays and ride for outdoor sports on weekends [34], [35].

In general, the clustering analysis results of bikeshare stations indicate that different categories of stations are associated with different dominant individual activities and thus may show different usage patterns and trip purposes. Accordingly, it is highly desirable to conduct the spatial analysis of bikeshare ridership regarding the diversity across different station categories.

B. SPECIFICATION OF GWR MODELS

In this study, six different types of GWR models were developed and compared: one joint model and five separate models for each station category. The GWR models were specified using the software of GWR 4.0 [36]. The specification results of the GWR models are given in Table 4.

Similar with the specification of general linear regression, the explanatory variables in the GWR model were selected through a stepwise procedure. The variables were input into GWR models one by one by checking the significance and the AICc values of the model. However, when checking the significance of explanatory variables in GWR model,

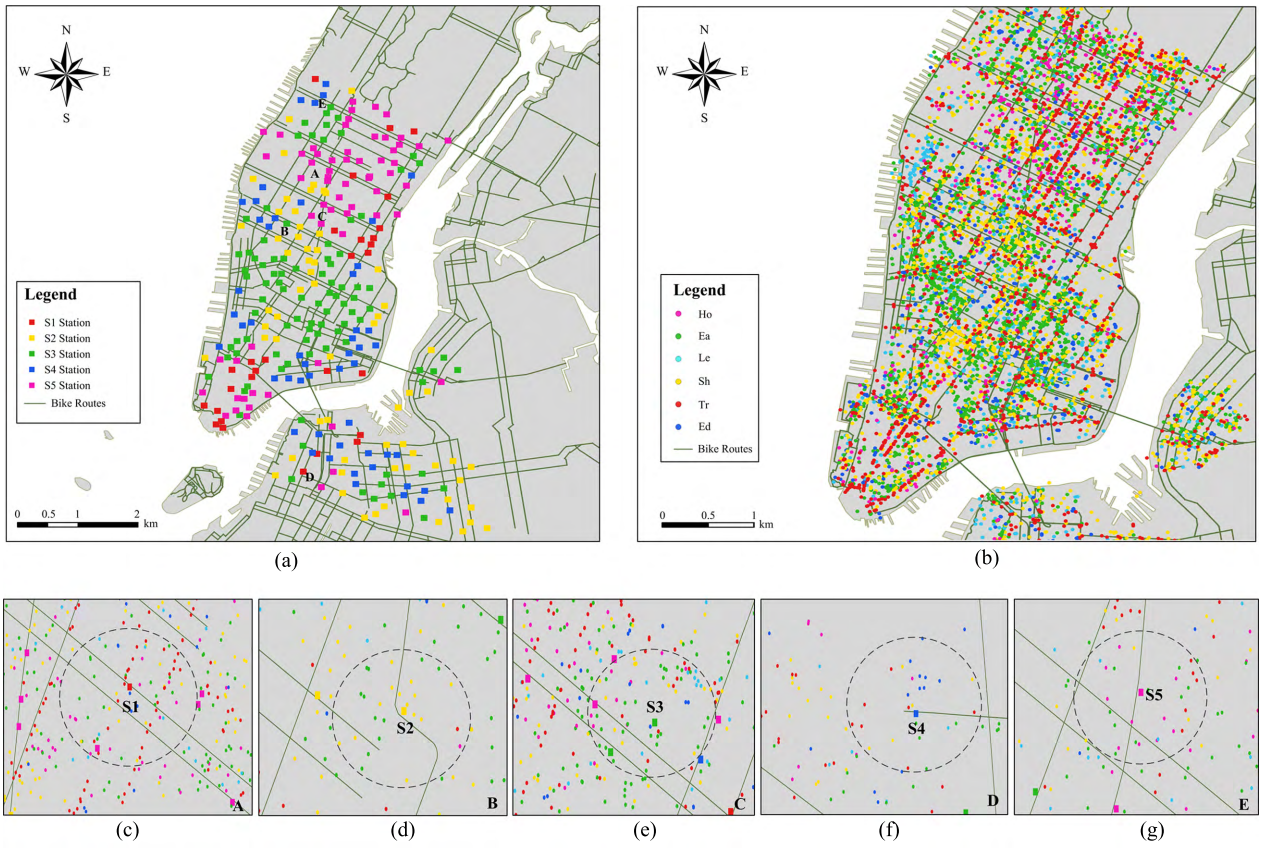


FIGURE 3. Spatial distributions for the five categories of bikeshare station and their surrounding POIs. (a) Spatial distribution of bikeshare stations. (b) Spatial distribution of POIs. (c) Zoom in on A area. (d) Zoom in on B area. (e) Zoom in on c area. (f) Zoom in on D area. (g) Zoom in on E area.

a potential problem is that one variable may be significant in some certain bikeshare stations while insignificant in other bikeshare stations. To account for this problem, in this study the GWR models selected the variables that are significant in over 80% of bikeshare stations with at least 90% level of confidence during the modeling procedure [37]. The variable selection procedure was repeated until the AICc value of the models reached the minimum. In the end, the GWR model which have the lowest AICc value, were considered as the best models. In addition, to verify the potential multicollinearity between selected variables, the variance inflation factors (VIF) values of the explanatory variables were calculated following the ordinary least square analysis. As shown in Table 4, the VIF values for all the selected explanatory variables were less than 5, indicating that the problem of multicollinearity did not exist.

Furthermore, the spatial correlation in the residuals of predictions in the GWR model and the global model for all the stations were computed and compared. The global model assumes that effects of explanatory variables are fixed over space. As is shown in Table 5, the spatial correlation in the residuals of GWR model is not significant at 90% confidence level. However, in the global model, a significant spatial correlation is found in the residuals of predictions. The tests of Moran's I statistics indicate that by accounting for the spatial heterogeneity in the station-level bikeshare ridership data, the residuals in the GWR model are less spatially correlated than that of global model.

In this study, the likelihood ratio tests (LRT) were conducted to test whether separate model estimation for each station category is superior to a joint model for all the station categories. The null hypothesis for the LRT test is that the joint model for all the five station categories does not have a significantly lower log-likelihood value than the separate models for each station category together. This null hypothesis also indicates that there are no significant differences between separate models and joint model. The test statistic can be calculated as follows:

$$LR = -2[LL\left(\beta_{joint}\right) - LL\left(\beta_{s1}\right) - LL\left(\beta_{s2}\right) - LL\left(\beta_{s3}\right) - LL\left(\beta_{s4}\right) - LL\left(\beta_{s5}\right)] \quad (5)$$

This statistic follows a chi-square distribution. The degrees of freedom (n) equals to the difference between the number of estimated variables in the joint model and the number of all the five separate models together. If the test statistics value is more than the chi-square value with n degrees of freedom at a certain confidence level, then the null hypothesis should be rejected.

In this study, two groups of likelihood ratio tests were conducted. The first test compared the joint model without

TABLE 4. Results of gwr models for one joint model and five separate models.

	Joint model (GWR_J)		S1 model (GWR_{S1})		S2 model (GWR_{S2})		S3 model (GWR_{S3})		S4 model (GWR_{S4})		S5 model (GWR_{S5})	
	Coeff (mean)	VIF	Coeff (mean)	VIF	Coeff (mean)	VIF	Coeff (mean)	VIF	Coeff (mean)	VIF	Coeff (mean)	VIF
Intercept	-5.28	-	-16.88	-	-15.1	-	-7.33	-	-3.44	-	3.411	-
Socio-economic and demographic variables												
Population	-	-	-	-	-	-	0.938	1.6	-	-	0.128	1.3
Non-white	-	-	-	-	-0.32	2.3	-	-	-0.45	2.5	-	-
Employment	0.299	1.6	0.532	2.1	-	-	-	-	-	-	-	-
College enrollment	-	-	-	-	-	-	-	-	0.552	3.6	-	-
MIC	0.582	1.1	0.519	1.3	1.283	1.2	0.486	1.2	-	-	-	-
Bikeshare infrastructure related variables												
Bike racks	0.126	1.5	-	-	-	-	-	-	0.207	1.8	0.089	1.3
Bike lane length	0.035	1.2	0.052	1.4	-	-	0.071	1.5	0.163	1.6	-	-
Station capacity	0.617	1.1	1.416	1.2	1.123	1.3	0.748	1.1	1.772	1.3	0.794	1.5
Station age	0.015	1.1	0.017	1.3	0.011	1.2	0.011	1.1	0.016	1.2	0.022	1.4
Weather related variables												
Precipitation	-0.232	2.1	-0.368	2.4	-0.188	2.0	-0.156	2.1	-0.177	2.2	-0.341	2.6
Snowfall	-0.638	1.8	-0.483	1.9	-0.521	2.1	-0.374	1.6	-0.481	1.5	-0.732	2.2
Temperature	0.11	1.2	0.088	1.3	0.133	1.3	0.097	1.2	0.183	1.6	0.264	1.5
Month variables												
January	-	-	-	-	-	-	-	-	-	-	-	-
February	-	-	0.197	1.1	-	-	-	-	0.222	1.8	-	-
March	0.231	1.1	0.228	1.2	0.221	1.3	0.231	1.2	0.265	1.5	0.466	1.3
April	0.267	1.2	0.274	1.4	0.32	1.5	0.458	1.6	0.426	1.3	0.641	2.1
May	0.31	1.5	0.324	1.6	0.468	1.6	0.476	1.8	0.489	1.8	0.774	2.5
June	0.396	1.6	0.416	1.8	0.472	1.9	0.411	1.5	0.576	2.3	0.732	2.6
July	0.472	2.1	0.46	2.2	0.521	2.3	0.683	2.8	0.764	3.1	0.821	3.3
August	0.338	1.3	0.369	1.3	0.442	1.2	0.555	2.6	0.701	2.8	0.96	3.1
September	0.351	1.6	0.376	1.5	0.387	1.7	0.432	1.3	0.683	2.2	1.24	2.5
October	0.39	2.2	0.401	2.0	0.413	2.3	0.39	1.8	0.44	1.8	0.931	2.4
November	0.181	2.1	0.222	1.8	0.221	1.9	0.216	1.7	0.338	1.6	0.712	1.8
December	-0.122	1.6	-0.089	1.5	-	-	-0.165	1.4	-0.201	1.6	-	-
Goodness of fit												
R^2	0.623		0.79		0.697		0.596		0.764		0.46	
AICc	477.338		48.633		121.523		166.377		66.271		189.536	

TABLE 5. Moran's I statistics for residuals of predictions in the GWR and global model.

Model	Moran's index	Expected index	z-score	p-value
GWR model	0.012	-0.003	0.551	0.583
Global model	0.203	-0.003	4.168	0.000

station category variables and separate models, while the second test compared the joint model with station category variables and separate models. As is shown in Table 6, the results of likelihood ratio rest indicate that the separate models for each station category are significantly better than the joint model at the 95% confidence level. The test confirms the necessary of considering the diversity across different station categories in spatial analysis of bikeshare ridership. In addition, the results in Table 6 also suggest that the separate models for each station category are significantly better than the joint model with station category variables at the 95% confidence level. This finding indicates that building separate models for each station category is better than incorporating station category as dummy variables in a joint model.

Moreover, in this study we further compared the performance of GWR model and spatial autoregressive model, which is commonly used for addressing the issue of spatial correlation in many previous studies [28], [29]. Table 7 gives the results of goodness-of-fit between the two methods for each station category. It can be found that in general the GWR models exhibit better performances than SAR models for each station category. The comparison analysis results suggest that the GWR is a superior technique for the modelling of station-level bikeshare ridership data than SAR method.

TABLE 6. Likelihood ratio test for the separate models and joint models.

	Description	Log likelihood values
$LL(\beta_{joint1})$	Log likelihood value of joint model without station category variables	-3819.706
$LL(\beta_{joint2})$	Log likelihood value of joint model with station category variables	-3605.227
$LL(\beta_{S1})$	Log likelihood value of S1 station model	-584.268
$LL(\beta_{S2})$	Log likelihood value of S2 station model	-774.324
$LL(\beta_{S3})$	Log likelihood value of S3 station model	-823.33
$LL(\beta_{S4})$	Log likelihood value of S4 station model	-641.607
$LL(\beta_{S5})$	Log likelihood value of S5 station model	-690.05
$LR1$=-2[$LL(\beta_{joint1})$-$LL(\beta_{S1})$-$LL(\beta_{S2})$-$LL(\beta_{S3})$-$LL(\beta_{S4})$-$LL(\beta_{S5})$] = **612.254** (Critical χ^2 = **168.61** with 95% confidence level)		
$LR2$=-2[$LL(\beta_{joint2})$-$LL(\beta_{S1})$-$LL(\beta_{S2})$-$LL(\beta_{S3})$-$LL(\beta_{S4})$-$LL(\beta_{S5})$] = **183.296** (Critical χ^2 = **159.81** with 95% confidence level)		

TABLE 7. Measures of goodness of fit for the GWR and SAR.

Measures of goodness of fit	R^2	AICc
S1 GWR model	**0.79**	**48.633**
S1 SAR model	0.724	72.533
S2 GWR model	**0.697**	**121.523**
S2 SAR model	0.628	160.32
S3 GWR model	**0.596**	**166.377**
S3 SAR model	0.536	194.33
S4 GWR model	**0.764**	**66.271**
S4 SAR model	0.701	80.988
S5 GWR model	**0.46**	**189.536**
S5 SAR model	0.457	210.307

C. DISCUSSION OF MODEL SPECIFICATION RESULTS

As is shown in Table 4, for the socio-economic and demographic variables, it can be found that the number of employed people in each station's service area is positively correlated with the bikeshare ridership in the joint model. This finding is expected and consistent with many previous studies [3], [11]. However, by further inspecting this coefficient in separate models, we found that the employment variable only has significant impact on the ridership of S1 stations. Considering that the S1 stations are surrounded by many public transit stations and mainly serve for commuting trips, the bikeshare ridership of S1 station are more easily to be affected by the employment population. Similarly, the number of college enrollment in each station's service area is found to be only positively correlated with the bikeshare ridership of S4 stations. The S4 station represents a typically bikeshare usage for school living. Accordingly, bikeshare of this station type mainly serve for college students and are more easily affected by the number of college enrollment.

In addition, the population in each station's service area is usually found to be positively correlated with the bikeshare ridership in many previous studies [1], [3], [11], [15], [38]. In this study, from the results of separate models we further found that the population variable only has significant impact on the bikeshare ridership of S3 and S5 stations. This finding is intuitive because stations of this two types are usually located nearby residential areas and are more easily to be affected by the surrounding population. The percent of non-white population in each station's service area is found to be negatively correlated with the ridership for some certain station categories, which is consistent with many previous studies [39]. Moreover, the median household income in each station's service area seems to have a positive effect on the bikeshare ridership in most station categories. This finding is also consistent with that of other bikesharing systems in Minneapolis, Denver, and Washington [1].

For the bikeshare infrastructure related variables, it can be found that the number of bike racks in each station's service area is positively correlated with the bikeshare ridership in the developed joint models. This finding reveals that areas with more bike racks will be involved with more bicycle activities and accordingly generate more potential trips [13]. Similarly, the bike lane length in each station's service area is also found to be positively correlated with the bikeshare ridership in the joint model. Figure 4 further illustrates the spatial pattern of the coefficients of the two variables. It can be found that the coefficient of bike racks reaches the highest value in the stations around the Lower Manhattan area, while the coefficient of bike lane length reaches the highest value in the stations around the Midtown area. This finding can guide transportation planners where to add some bike racks or new bike lanes to improve the bikeshare ridership more effectively, particularly with the budget limitations. By further inspecting the results of separate models, we found that the bike lane length variable only have significant effect on the ridership of S1, S3 and S4 stations. These types of stations are mainly located within areas that have a relatively high proportion of bike lane, such as the school-living area or near the public transit stations. Thus, the higher density of bike lanes in the service area of these station types will generate more bikeshare trips. For the weather related variables, the monthly accumulative precipitation and snowfall are found to be negatively correlated with the bikeshare ridership, while the average temperature is found to be positively correlated with the bikeshare ridership for each month. These findings are intuitive because good weather and warm temperature will attract more people to take shared bicycle [16], [38].

In general, the findings from the spatial analysis of bikeshare ridership in New York City are consistent with many previous studies. By building the separate models for each bikeshare station category, the fitness of developed GWR models are greatly improved, and many interesting findings associated with specific station categories can be revealed. Accordingly, when designing planning scheme and operation strategies, policy makers should be cautious about the varied influencing of the contributing factors across different station categories.

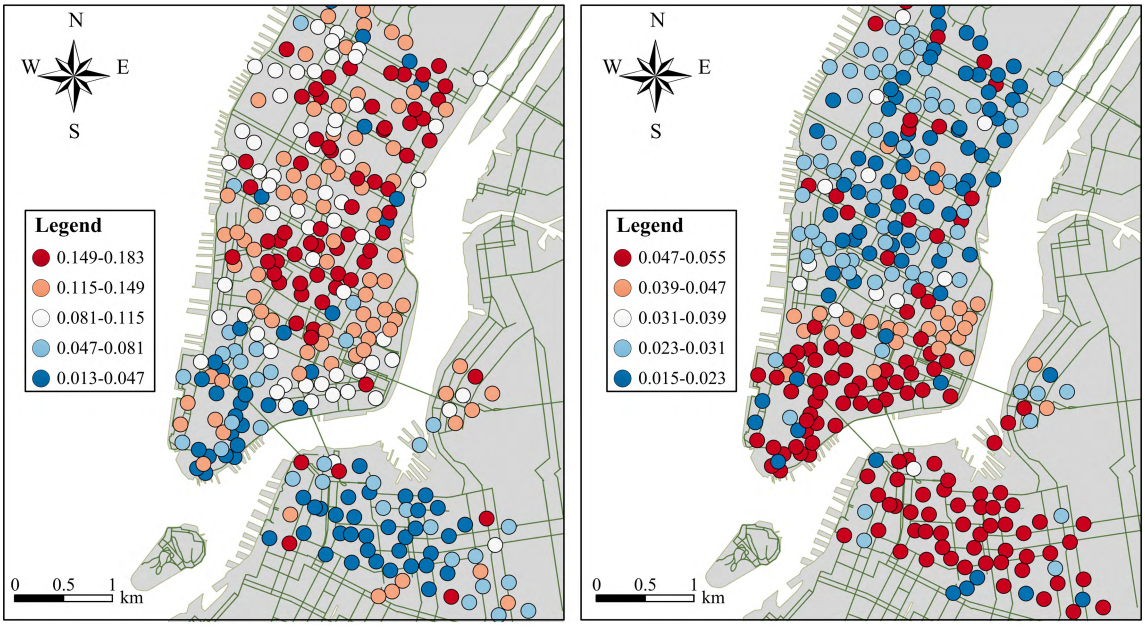


FIGURE 4. Spatial pattern of coefficients of bike racks and bike lane length in GWR models. (a) Spatial pattern of coefficients of bike racks in GWR models. (b) Spatial pattern of coefficients of bike lane length in GWR models.

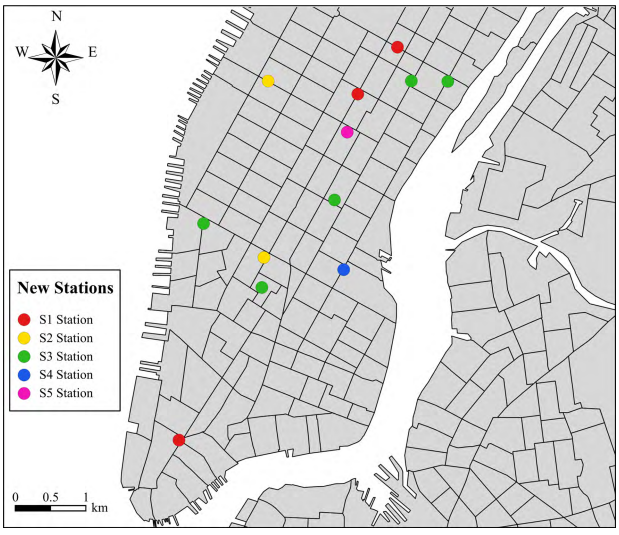


FIGURE 5. The selected new stations opened in 2016 year.

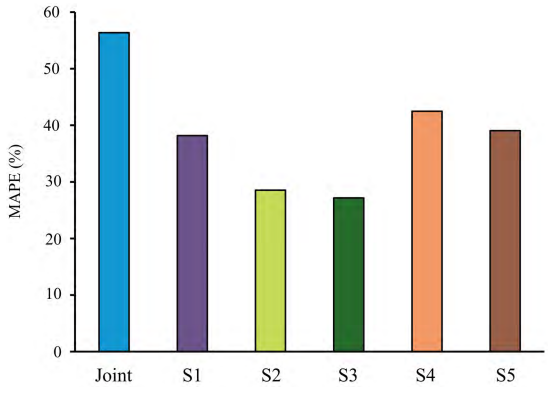


FIGURE 6. Model prediction performance for the selected new stations.

D. MODEL PREDICTION AND VALIDATION

The scale of the bikesharing system in New York City has expanded year by year. Although the developed GWR models have exhibited decent performance for the study data, the applicability for forecasting ridership at new stations needs to be further investigated. In the present study, 12 newly opened stations in downtown Manhattan during 2016 year were selected to validate the prediction performance of developed models. Figure 5 illustrates the locations of selected new stations and their related station categories.

More specifically, we applied the developed GWR models of 2015 year to predict the bikeshare ridership of 12 new stations opened in 2016 year. Then, we calculated the actual ridership of the selected stations from the trip records of 2016 year. The mean absolute percentage error (MAPE) was calculated to assess the prediction performance for each station category (See Figure 6).

As is shown in Figure 6, the MAPE values of the separate models were generally lower than that of the joint model. The findings again confirmed the fact that building bikeshare ridership models for each station type can greatly improve the model prediction performance. In addition, the prediction quality of the S1, S4 and S5 station category models seem not to be very satisfactory as most MAPE values were higher than 30%. The reasons are twofold. First, the weather related variables are aggregated monthly, which are not fine enough to reflect the weather variation in predicting bikeshare ridership for new stations. Second, there was a significant growth

of the bikesharing user base in New York City during the two years. It was reported that the registered bikesharing users of New York City have increased by nearly 150% during these two years.

V. CONCLUSIONS AND DISCUSSIONS

The present study conducted the spatial analysis of bikeshare ridership for different station categories using smart card data and online POI data. We collected bikeshare trip records from the Citi Bike system in New York City to illustrate the procedure. The POIs in the vicinity of each bikeshare station were obtained from the Google Places API. The K-means clustering analysis was firstly conducted to classify the bikeshare stations into five categories based on the spatial distribution of their surrounding POIs. Then, the GWR models were developed to establish the relationship between the bikeshare ridership and various contributing factors, such as the bicycle infrastructures, station capacity, weather variables, and socioeconomic and demographic variables in each station's service area. A joint model and five separate models for each station category were built, respectively. The results of likelihood ratio test confirmed the superiority and necessary of building separate models for each bikeshare station category instead of a joint model. Moreover, all the developed GWR models were applied to predict the bikeshare ridership of the newly opened stations in the next year. The results suggested by building separate bikeshare ridership models for each station category, the model prediction accuracy are greatly improved, particularly for station category S1, S2, and S3.

The results of the clustering analysis revealed that different bikeshare station categories are associated with different dominant activities and may exhibit different usage patterns. By capturing the diversity across different station categories, the fitness of developed GWR models and the model prediction performance at new stations were both greatly improved. Moreover, the varied influencing of the contributing factors across different station categories can be identified, and some interesting findings associated with specific station category can be revealed. Accordingly, when designing planning scheme and operation strategies, policy makers could provide insightful suggestions for each station category that aim at improving the operations of the whole bikesharing systems.

Even though the developed models can provide some guidance and effective strategies to improve the operations of bikesharing systems, several limitations are still needed to be addressed before the results of this study are used in practical engineering:

(a) Attaching the potential trip purpose to each collected POI data may suffer from some uncertainty. For example, a commercial building including restaurants, bars, and fitness room may have different types of trip purposes. In this study, we checked the collected POI data and found that POI of this kind only constitute a relatively small proportion (7.23%) of the whole dataset. Thus, in general the POI category in this study can be accurately classified by the type of associated trip purposes.

(b) Clustering the station category based on surround POIs could be appropriate in a suburban context where land uses are typically separate without mixed-uses. However, in a dense, urban, and predominantly mixed-use environment such as Manhattan, the collected POI may not well reflect the land use function.

(c) Some specific category of POIs are more likely be labeled than the other category of POIs in the Google Map Application such as recreational and leisure places, leading to potential biases of the station clustering results.

(d) The difference between weekday and weekend trips and the difference between member and non-member trips were not investigated in this study. The further segment of trip types may help to better reveal the influencing factors to ridership.

The limitation of POI data is a prevalent problem in many previous transportation studies which also applied the POI data to cluster the bus stations [19] and subway stations [18]. Some of these limitations may go away since more and more human activity information can be collected in urban areas, such as twitter check-in data and mobile phone data. In the future, these data sources could be combined with POI data for better clustering the station category and understanding the spatial variation of bikeshare ridership. The authors recommend that future studies may focus on these issues.

REFERENCES

[1] R. Rixey, "Station-level forecasting of bikesharing ridership station network effects in three U.S. systems," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2387, pp. 46–55, Dec. 2013.

[2] N. Borecki *et al.* (2012). *Virginia Tech Capital Bikeshare Study: A Closer Look at Casual Users and Operations*. [Online]. Available: https://ralphbu.files.wordpress.com/2012/01/vt-bike-share-study-final3.pdf

[3] A. Faghih-Imani, N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq, "How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in Montreal," *J. Transp. Geogr.*, vol. 41, pp. 306–314, Dec. 2014.

[4] S. Susan, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia: Past, present, and future," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2143, pp. 159–167, Oct. 2011.

[5] S. A. Shaheen, E. W. Martin, N. D. Chan, A. D. Cohen, and M. Pogodzinski, "Public bikesharing in North America during a period of rapid expansion: Understanding business models, industry trends and user impacts," MTI, San Jose, CA, USA, MTI Tech. Rep. 12-29, 2014. [Online]. Available: http://transweb.sjsu.edu/project/1131.html

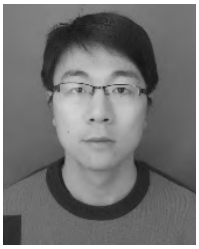
[6] G. Rebecca and L. Miskimins, "Exploring bicycle options for federal lands: Bike sharing, rentals and employee fleets," Federal Highway Admin., Vancouver, WA, USA, Tech. Rep. FHWA-WFL/TD-12-001, 2012. [Online]. Available: http://www.nps.gov/transportation/pdfs/FHWA_bicycle_options.pdf

[7] S. Shaheen, E. Martin, and A. Cohen, "Public bikesharing and modal shift behavior: A comparative study of early bikesharing systems in North America," *Int. J. Transp.*, vol. 1, no. 1, pp. 35–54, 2013.

[8] Institute for Transportation and Development Policy. (2013). *Riding the Bike-Share Boom: The Top Five Components of a Successful System*. [Online]. Available: www.itdp.org/riding-the-bike-share-boom-the-top-five-components-of-a-successful-system/

[9] O. O'Brien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," *J. Transp. Geogr.*, vol. 34, pp. 262–273, Jan. 2014.

[10] S. Shaheen and E. Martin, "Unraveling the modal impacts of bikesharing," *Access Mag.*, vol. 47, pp. 8–15, Jan. 2015.
[11] E. Fishman, S. Washington, N. Haworth, and A. Watson, "Factors influencing bike share membership: An analysis of Melbourne and Brisbane," *Transp. Res. A, Policy Pract.*, vol. 71, pp. 17–30, Jan. 2015.
[12] M. Elliot, A. Cohen, J. L. Botha, and S. Shaheen. (2016). *Bikesharing and Bicycle Safety*. [Online]. Available: http://transweb.sjsu.edu/PDFs/research/1204-bikesharing-and-bicycle-safety.pdf
[13] R. B. Noland, M. J. Smart, and Z. Guo, "Bikeshare trip generation in New York city," *Transp. Res. A, Policy Pract.*, vol. 94, pp. 164–181, Dec. 2016.
[14] N. Rahul, E. Miller-Hooks, R. C. Hampshire, and A. Bušić, "Large-scale vehicle sharing systems: Analysis of Vélib,"' *Int. J. Sustain. Transp.*, vol. 7, no. 1, pp. 85–106, 2013.
[15] J. Bao, C. Xu, P. Liu, and W. Wang, "Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests," *Netw. Spatial Econ.*, vol. 17, no. 4, pp. 1231–1253, 2017.
[16] M. Hyland, Z. Hong, H. K. R. de Farias Pinto, and Y. Chen, "Hybrid cluster-regression approach to model bikeshare station usage," *Transp. Res. A, Policy Pract.*, vol. 115, pp. 71–89, Sep. 2018.
[17] M. Pouke, J. Goncalves, D. Ferreira, and V. Kostakos, "Practical simulation of virtual crowds using points of interest," *Comput. Environ. Urban Syst.*, vol. 57, pp. 118–129, May 2016.
[18] J. Wang, X. Kong, A. Rahim, F. Xia, A. Tolba, and Z. Al-Makhadmeh, "IS2Fun: Identification of subway station functions using massive urban data," *IEEE Access*, vol. 5, pp. 27103–27113, 2017.
[19] S. Wang, L. Sun, J. Rong, and J. Ma, "Using point of interest data from electronic map to predict transit station ridership," in *Proc. TRB 93rd Annu. Meeting Compendium Papers*, 2014, pp. 1–15.
[20] X. Qian and S. V. Ukkusuri, "Spatial variation of the urban taxi ridership using GPS data," *Appl. Geogr.*, vol. 59, pp. 31–42, May 2015.
[21] O. D. Cardozo, J. C. García-Palomares, and J. Gutiérrez, "Application of geographically weighted regression to the direct forecasting of transit ridership at station-level," *Appl. Geogr.*, vol. 34, pp. 548–558, May 2012.
[22] J. Bao, P. Liu, P. Blythe, and J. Wu, "Exploring contributing factors to the usage of ridesourcing and regular taxi services with high-resolution GPS data set," in *Proc. TRB 97th Annu. Meeting Compendium Papers*, 2018, pp. 1–9.
[23] Y. Ji, X. Ma, M. Yang, Y. Jin, and L. Gao, "Exploring spatially varying influences on metro-bikeshare transfer: A geographically weighted poisson regression approach," *Sustainability*, vol. 10, no. 5, p. 1526, 2018.
[24] J. Bao, P. Liu, H. Yu, and C. Xu, "Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas," *Accident Anal. Prevention*, vol. 106, pp. 358–369, Sep. 2017.
[25] (2018). *Citi Bike Trip Histories*. [Online]. Available: http://www.citibikenyc.com/system-data
[26] A. Faghih-Imani and N. Eluru, "Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system," *J. Transp. Geogr.*, vol. 54, pp. 218–227, Jun. 2016.
[27] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
[28] A. D. Cliff and J. K. Ord, *Spatial Autocorrelation*. London, U.K.: Pion, 1973.
[29] L. Anselin, *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer, 1988.
[30] H. Huang, M. Abdel-Aty, and A. Darwiche, "County-level crash risk analysis in Florida: Bayesian spatial modeling," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2148, pp. 27–37, Aug. 2010.
[31] L.-F. Chow, F. Zhao, X. Liu, M.-T. Li, and I. Ubaka, "Transit ridership model based on geographically weighted regression," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1972, pp. 105–114, Jan. 2006.
[32] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, U.K.: Wiley, 2002.
[33] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, vol. 1, 1996, pp. 226–231.
[34] J. Bachand-Marleau, B. H. Y. Lee, and A. M. El-Geneidy, "Better understanding of factors influencing likelihood of using shared bicycle systems and frequency of use," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2314, pp. 66–71, Jan. 2012.
[35] J. Pfrommer, J. Warrington, G. Schildbach, and M. Morari, "Dynamic vehicle redistribution and online price incentives in shared mobility systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1567–1578, Aug. 2014.
[36] T. Nakaya, "GWR4 user manual," 2014. [Online]. Available: http://www.st-andrews.ac.uk/geoinformatics/wp-content/uploads/GWR4manual_201311.pdf
[37] J. Bao, P. Liu, X. Qin, and H. Zhou, "Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data," *Accident Anal. Prevention*, vol. 120, pp. 281–294, Nov. 2018.
[38] K. Gebhart and R. B. Noland, "The impact of weather conditions on bikeshare trips in Washington, DC," *Transportation*, vol. 41, no. 6, pp. 1205–1225, 2014.
[39] L. K. Maurer, "Feasibility study for a bicycle sharing program in Sacramento, California," in *Proc. TRB 91th Annu. Meeting Compendium Papers*, 2012, pp. 1–14.



JIE BAO was born in Taizhou, Jiangsu, China, in 1987. He received the B.S. degree in transportation from Southeast University in 2009 and the M.S. degree in management science and engineering from the Dalian University of Technology in 2012. He is currently pursuing the Ph.D. degree in transportation engineering with Southeast University.

His research interests include green transportation, big data analytics, and intelligent transportation systems.

XIAOMENG SHI was born in Suqian, China, in 1991. He received the B.S. and Ph.D. degrees in transportation engineering from the School of Transportation, Southeast University, Nanjing, China, in 2013 and 2018, respectively.

His research interests include the analysis of pedestrian flow characteristics, urban mobility analytics, and intelligent transportation systems.

HAO ZHANG was born in 1982. He is currently pursuing the Ph.D. degree in transportation engineering with the School of Transportation, Southeast University, Nanjing, China. He is also an Associate Professor.

His research interests include the analysis of traffic safety, intelligent transportation systems, and traffic big data processing.