Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 7345807, 10 pages
https://doi.org/10.1155/2021/7345807

Research Article

Effects of Coverage Area Treatment, Spatial Analysis Unit, and Regression Model on the Results of Station-Level Demand Modeling of Urban Rail Transit

Hongtai Yang,[1] **Chaojing Li**,[1] **Xuan Li**,[1] **Jinghai Huo**,[1] **Yi Wen**,[2] **Emma G. P. Sexton**,[2] **and Yugang Liu**[1]

[1]*School of Transportation and Logistics, National Engineering Laboratory of Integrated Transportation Big Data Application Technology, National United Engineering Laboratory of Integrated and Intelligent Transportation, Institute of System Science and Engineering, Southwest Jiaotong University, Chengdu 610000, China*
[2]*Department of Civil and Environmental Engineering, University of TN, Knoxville, Knoxville, TN 37996, USA*

Correspondence should be addressed to Yugang Liu; liuyugang@swjtu.edu.cn

Received 15 April 2021; Revised 8 August 2021; Accepted 14 August 2021; Published 29 August 2021

Academic Editor: Sheng Jin

Direct ridership models can predict station-level urban rail transit ridership. Previous research indicates that the direct modeling of urban rail transit ridership uses different coverage overlapping area processing methods (such as naive method or Thiessen polygons), area analysis units (such as census block group and census tract), and various regression models (such as linear regression and negative binomial regression). However, the selection of these methods and models seems arbitrary. The objective of this research is to suggest methods of station-level urban rail transit ridership model selection and evaluate the impact of this selection on ridership model results and prediction accuracy. Urban rail transit ridership data in 2010 were collected from five cities: New York, San Francisco, Chicago, Philadelphia, and Boston. Using the built environment characteristics as the independent variables and station-level ridership as the dependent variable, an analysis was conducted to examine the differences in the model performance in ridership prediction. Our results show that a large overlap of circular coverage areas will greatly affect the accuracy of models. The equal division method increases model accuracy significantly. Most models show that the generalized additive models have lower mean absolute percentage errors (MAPE) and higher adjusted R^2 values. By comparison, the Akaike information criterion (AIC) values of the negative binomial models are lower. The influence of different basic spatial analysis unit on the model results is marginal. Therefore, the selection of basic area unit can use existing data. In terms of model selection, advanced models seem to perform better than the linear regression models.

1. Introduction

Urban rail transit is a popular form of urban public transportation because of its large capacity, environmental friendliness, and fast speed. The emergence of the urban rail transit has alleviated the problems of congestion and exhaust pollution caused by private vehicles in the city. Strong evidence shows that urban rail transit stations reduce the private car ownership rates of nearby households [1]. A growing number of mega cities have adopted various measures to develop rail transit systems. In cities such as Seoul and Shanghai, the government promotes the use of urban rail transits through transit-oriented development (TOD) [2, 3]. The station-level ridership is a main factor for determining the operation and planning of the urban rail transit system. Because of this, ridership modeling at station level has consistently been a topic of interest for scholars and practitioners. A reliable ridership model can reflect the underlying factors that influence station-level ridership and facilitate operation and management, formulating measures to promote urban rail transit ridership. An accurate ridership model can assist operators in determining service frequency and best practice in operation.

In ridership modeling, an important factor to consider is the aspect of land use within the station service coverage area. Selecting a suitable station service coverage area is the first step towards building an effective ridership model. Double counting caused by overlapping area is a recognized issue in ridership modeling [4, 5]. To address the potential issues with double-counted data, the use of Thiessen polygons is a commonly proposed measure [6–9]. In brief, Thiessen polygons assign the closest point of the station to that station to tackle this issue. However, issues may still occur when using Thiessen polygons. For example, in places where the station layout is relatively compact, the station service coverage area formed by the Thiessen polygon becomes proportionally smaller, which may underestimate the effects of built environment variables on ridership in a model.

After determining the service coverage area of a station, the next step is to extract the built environment variables (such as population density and employment density) within the service coverage area, which is the basic unit for spatial analysis. Because station service coverage area usually does not completely match with the basic spatial analysis unit, is it often assumed that population and employment data are evenly distributed in the basic spatial analysis unit. Data weight is determined by the ratio of area covered to total area within the analysis unit. For example, the data weight is 1 when the analysis unit is completely covered by the station service coverage area. In existing literature, two basic spatial analysis units are often used: census block groups and census tracts [6, 10–12]. Both census block groups (CBGs) and census tracts (CTs) are geographic units used by the U.S. Census Bureau. CT is a larger geographic unit, consisting of multiple CBG units.

Selection of a fit regression model is the following step to reveal the relationship between station-level ridership and its influencing factors. Linear regression is the most commonly used method in literature [2, 3, 7, 13]. Because of the differences in station locations, ridership can vary significantly by station, resulting in a possible disperse distribution of station-level ridership. To deal with this possibility, negative binomial regression models are suggested in literature and have been adopted [14, 15]. In addition, generalized additive regression models are proposed in tackling the likely nonlinear relationship between the dependent variable and the independent variables in the transportation field [16, 17].

This study selects five cities in the United States as case studies: New York City, Philadelphia, Boston, Chicago, and San Francisco. The urban rail transit systems in these cities are among the largest in USA. The inclusion of these five cities is to yield a more generalizable conclusion, as they vary across many aspects [18]. This study aims to determine which regression method is the most reliable in dealing with station coverage area overlapping issues, to explore if there exist significant differences between modeling CBGs versus CTs as spatial analysis units, and to provide insights into which model performs the best when modeling direct ridership at the station level.

The rest of this article is structured as follows. The next section reviews the existing literature in contributing factors of transit ridership, measures to address the overlapping issues of station coverage area, and applications of various spatial analysis units and models. The third section describes the data of the study and the three proposed methods to overcome the overlapping issue of station coverage area. It also explores the differences in the ordinary least square regression, negative binomial regression, and generalized additive models using two spatial analysis units (CBG and CT) and different station coverage treatment methods. The fourth section assesses the accuracy and reliability of the models and proposes directions for model improvement. Finally, a conclusion of our study findings, merits, and limitations is presented.

2. Literature Review

A large number of studies on urban rail ridership are based on station level. It is important to study the impact of the built environment on station-level ridership. Many studies have found that higher population density and employment density increase the ridership [3, 19–21]. Built environment factors around the station such as density, diversity, and design have a significant effect on the ridership of urban rail [14, 22–24]. The attributes of the station itself also have a significant impact on ridership, with transfer and terminal stations associated with higher ridership [25–27].

To determine the built environment attributes around a station, buffers are usually drawn around the stations to represent coverage areas of the stations. Generally, a circular buffer zone with the radius of 800 meters is commonly used as the service coverage area of a station [25, 28, 29]. However, this naive approach tends to ignore the problem of overlapping buffers, especially when stations are close to each other. Some scholars have proposed solutions to such problem. The Thiessen polygon method is the most widely used method [6, 7, 9]. However, in places where the stations are close to each other, the use of Thiessen polygons can make the coverage area of stations be very small, which introduces errors when estimating the values of the built environment variables. As a result, neither of the two methods seems to address the issue of overlapping buffer well. Therefore, this paper proposes a new method to deal with this issue: calculate the values of the variables such as population, employment, and number of bus stops in the overlapping area of buffers, divide the value of such variables by the number of overlapping buffers, and assign the result to each overlapping buffer.

Different spatial analysis units are used by different studies. These units usually include CBG and CT. As the coverage area of stations does not completely match the spatial analysis unit, when estimating the values of built environment variables around stations, the common practice is to treat those variables as evenly distributed in the spatial analysis unit and calculate the values of the variables within the coverage area. Therefore, the choice of spatial analysis unit affects the results of the direct ridership model. Studies using CT as the spatial analysis unit include [6, 10], and studies using CBG as the spatial analysis unit include [11, 12].

Different studies also use different regression models to construct the relationship between independent variables and the station-level ridership. The linear regression model is the most common method [24, 26, 27]. Another widely used model is the negative binomial regression [14, 15]. Recently, nonlinear models have been extensively used. The nonlinear models include machine learning models and polynomial statistical models. Compared to statistical models, machine learning models have no significance inference for the independent variable and are prone to overfitting [30, 31]. GAM is an advanced statistical model that captures the nonlinear relationship between the independent and dependent variables through a smoothing function [16, 17, 31, 32]. Ding et al. [16] found the nonlinear association between almost all built environment variables and electric bike ownership. Hu et al. [17] used the generalized additive mixed effects model (GAMM) to investigate the nonlinear relationship between determinants and the attractiveness of car sharing. Since spatial autocorrelation is always observed when dealing with spatial data, spatial econometric models are usually used to deal with this issue [15, 33, 34]. Gan et al. [33] applied the spatial error model and found the factors that significantly influence station-level ridership while controlling for spatial autocorrelation.

In conclusion, previous studies have used different spatial analysis units, treatments of overlapping buffer areas, and regression models when estimating the station-level ridership of urban rail transit system. However, these methods and models have not been compared yet. So far, there is no guideline on which spatial analysis unit, treatment of overlapping buffer areas, and regression model should be used. As a result, this study will analyze the effect of different spatial analysis units, treatments of overlapping buffer areas, and regression models on the results so as to provide guidelines on which method or model should be used.

3. Research Design

3.1. Study Area. We collected the 2010 demographic variables from latest smart location database (SLD) and urban rail transit data of 2010 for the five selected U.S. cities: New York, San Francisco, Chicago, Philadelphia, and Boston. The five cities are selected because they provide ridership data that is available to the public. In addition, the numbers of urban rail transit stations in these five cities are relatively diverse. New York city has the largest urban rail transit system among the five cites, with 421 stations spreading across Manhattan, Brooklyn, Queens, and Bronx. San Francisco has the least number of stations, with 44 stations connecting cities in the Bay Area. The numbers of stations in Chicago, Philadelphia, and Boston are 136, 156, and 153, respectively.

3.2. Variables and Data Sources. Our dependable variables are the urban rail transit station-level ridership in 2010, which are in form of average weekday station ridership. The data sources and information are shown in (Table 1).

Our independent variables are derived from smart location database (SLD) 2010 and open data in five cities. SLD is available through the US Environmental Protection Agency. We obtain demographic data from SLD with CBG as our basic unit of spatial analysis and cluster them into CTs through GEOIDs. GEOIDs are numeric codes that uniquely identify geographic units. Shapefiles of roads, bus stops, bus lines, and urban rail transit stations are retrieved from the open data. Using ArcGIS spatial analysis tools, the variable data used for modeling are divided into three categories: demographic, land use, and station characteristics. Variable information is described in Table 2.

3.3. Overlapping Issue of Station Coverage Area. This study applies three different methods to deal with the overlapping issue of the station service coverage area: naive method (i.e., no treatment), equal division method, and the Thiessen polygons method. The circular buffer zone is a circular area with a station as the center and a radius of 800 meters as shown in Figure 1. However, when stations are located densely, a large amount of overlapping can occur with this method. As a result, this may significantly affect the model accuracy. For this reason, two other methods are used. The equal division method is similar to the naïve method. But the overlapped population, employment, and bus stops will be equally divided within a spatial unit. However, it should be noted that other variables (Table 1) do not involve equal splits under this method. As highlighted in the red square in Figure 1, when the two circular buffer areas overlap, the data in the overlapping part will be evenly assigned to the service coverage area of the two nearby stations. Another possible method is the use of Thiessen polygons, where any data point within the polygon is assigned to its closest station measured in distance, as illustrated in Figure 2.

3.4. The Modeling Approaches. In addition to the multiple linear regression, negative binomial regression, and spatial models that are common in the literature, this research also adds a generalized additive model. When building a direct ridership model for urban rail transit stations, linear regression is the most commonly used model by scholars. However, one assumption of linear regression is that the relationship between the dependent variable and the independent variable(s) can only be linear, which is rarely the case for some types of data. We thus introduce a generalized additive model to perform nonlinear fitting to capture the possibility of any possible nonlinear correlation between the dependent variable and the independent variable(s) through smoothing function. The station-level ridership difference among stations can be large. Through calculations, it is found that the coefficient of variation averages 1.49 from 0.83 (Chicago) to 3.07 (Philadelphia), an indication of overdispersion. Therefore, the negative binomial regression is proposed and used to overcome this overdispersion pattern of the dependent variable in the study [18]. Spatial variability in station-level ridership, for which a spatial error model is applied to address the spatial autocorrelation. All modeling and analyses are performed with the *R* software

Table 1: Urban rail transit system overview in selected cities.

City	Number of stations	Data sources	Opening year	Service lines
New York	421	New York City Transit Authority (MTA)	1904	27
San Francisco	44	Bay Area Rapid Transit District (BARTD)	1972	5
Chicago	136	Chicago Transit Authority (CTA)	1892	8
Philadelphia	153	Southeastern Pennsylvania Transportation Authority (SPTA)	1907	13
Boston	156	Massachusetts Bay Transportation Authority (MBTA)	1901	5

Table 2: Variable description.

Symbol	Description
Demographic variables	
Housing unit	Number of housing units in CBG/CT
Population	Population in CBG/CT
Employment	Total employment in CBG/CT
Household	Number of households (occupied housing units) (CBG/CT)
Zero-car	Percent of zero-car households
One-car	Percent of one-car households
Worker	Number of workers in CBG (home location) (CBG/CT)
LowWageH	Percent of workers earning $1250/month or less (home location)
MedWageH	Percent of workers earning between $1250/month and $3333/month (home location)
Land-use variables	
Bus stop	Number of bus stops in the station service coverage area
Bus route	Number of bus routes within a 200-meter radius of the urban rail transit station
Distance to CBD	Distance to Commercial Business District
Station characteristic variables	
Transfer	Coded as 1 if it is a transfer station and 0 otherwise
Terminal	Coded as 1 if it is a terminal station, and0 otherwise

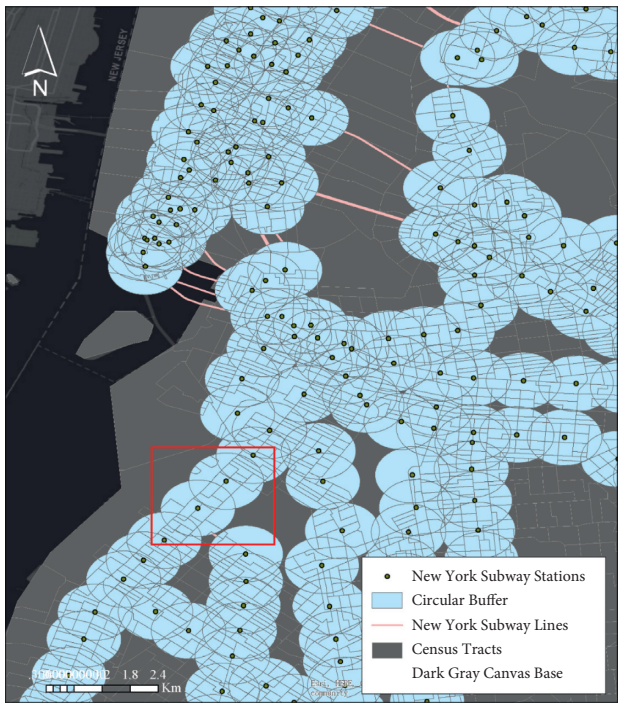

Figure 1: Illustration of the naive method.

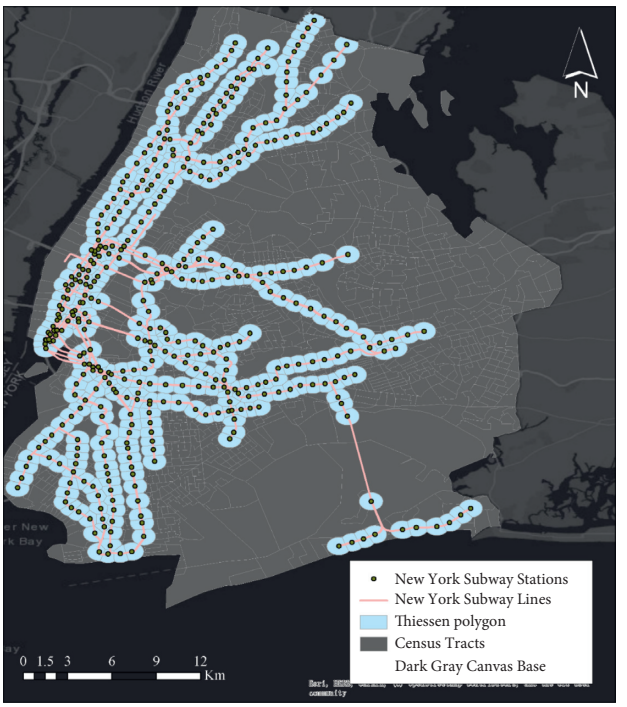

Figure 2: Illustration of the Thiessen polygon method.

(version 3.6. 1), in which the generalized additive model is activated through the "mgcv" package [35], negative binomial regression is provided by the "MASS" package [36], and spatial error model is provided by the "spatialreg" package [37].

The negative binomial, generalized additive, and spatial error models used in this study can be expressed as follows.

3.4.1. Negative Binomial Regression.

$$Y = \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \varepsilon\right), \quad (1)$$

where Y represents the station-level ridership and β_0 is the intercept. $x_1, x_2, \ldots, x_i$ represent the independent variables, which are population, employment, and other variables in our context, $\beta_1 \ldots \beta_i$ are the coefficients of each independent variable, and ε is the residual term.

3.4.2. Generalized Additive Regression. The generalized additive model uses a spline function to capture the nonlinear relationship between the independent and dependent variables, with the following formula:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + s(x_1) + \cdots + s(x_k) + \varepsilon, \quad (2)$$

where Y represents the station-level ridership and β_0 is the intercept. $x_1, x_2, \ldots, x_n$ are the independent variables used for linear fitting, $\beta_1 \ldots \beta_n$ are the coefficients of independent variables, $s(x_1) \ldots s(x_k)$ represent the independent variables of nonlinear fitting, and ε is the residual term.

3.4.3. Spatial Error Model. The spatial error model assumes that only the effect from the error term in the spatial autocorrelation process of the elements is captured by the following formula:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \lambda W_\varepsilon + \mu, \quad (3)$$

where W is the weight matrix, ε is the coefficient vector of the random error term, λ is the spatial error coefficient, and μ is the random error.

3.5. Model Performance Indicators. To evaluate the accuracy of the model, this study used adjusted R^2, MAPE, and AIC. Among them, adjusted R^2 is used to indicate goodness of fit of the models [38]. MAPE is used to evaluate the accuracy of the model [39]. AIC is used to compare the quality of different models [40]. These three indicators can be expressed mathematically as

$$\text{Adjusted } R^2 = 1 - \left(\frac{\text{SS}_{\text{residuals}}/n - p - 1}{\text{SS}_{\text{total}}/n - 1}\right),$$

$$\text{MAPE} = \frac{100\%}{n} \sum\left(\frac{|y - \widehat{y}|}{y}\right), \quad (4)$$

$$\text{AIC} = -2 \ln(L) + 2k,$$

where $\text{SS}_{\text{residuals}}$ is the residual sum of squares and $n - p - 1$ is the degree of freedom of the $\text{SS}_{\text{residuals}}$. SS_{total} is the total sum of squares, and its degree of freedom is $n - 1$. y represents the true values of the dependent variable in the model, while $\widehat{y}$ is the predicted value of the dependent variable. L denotes the likelihood function and k represents number of parameters.

4. Results

4.1. Model Outputs and Evaluations. In this section, we discuss differences in the use of various spatial analysis units, the processing methods for the overlapping parts of the station coverage areas, and the results of the three types of models. A total of 90 models have been built in this study. For model evaluation, we first use the variance inflation factor (VIF) value to examine potential collinearity between the variables and filter out the variables with a VIF value exceeding 10 [41]. The variables "housing unit" and "household" are removed. With generalized additive models, we use the AIC to determine the value of k. The model results for New York based on the spatial analysis unit of CBG and the naive treatment method for the overlapping buffer area are given in Table 3.

We then obtain and summarize the adjusted R^2, MAPE, and AIC values of all final fitted models, as shown in Tables 4–7 . In addition, we calculate the size of the overlapping area and the overlap rates in the five cities using circular buffers (Table 8). With this calculation method, New York and Boston have a relatively high overlap rate, close to half of the area size, and this reflects the stations layout compactly.

4.2. Other Methods to Address the Overlapping Issue. In cities with compact station layouts, such as New York and Boston, the coverage overlap rates are close to half. With the same model parameters and analysis units, the equal division method in processing the overlapping area appears to be the most effective method overall with average adjusted $R^2 = 0.6$, MAPE = 112.078, and AIC = 8020.195, which is better than the naive method (0.454, 164.758, and 8094.405) and the Thiessen polygon method (0.560, 126.657, and 8046.428), respectively. The method using Thiessen polygons performs better than the naive method as well. This is likely because when urban stations are densely located, the use of naive method recalculates a large amount of data, while the other two methods can better dilute the situation and reflect it around the station more closely. In cities with sparsely located stations, such as San Francisco, Philadelphia, and Chicago, the overlap rates are 0.034, 0.08, and 0.254, respectively. The naive method, equal division method, and Thiessen polygon method do not seem to show significant differences.

4.3. Model Comparison. The goodness of fit of a model can be assessed using the AIC value. Controlling for the same processing on the overlapping part of the station service coverage area, the same spatial analysis unit, and the same city, our results show that the AIC values of most negative binomial regression models are lower than those of the

Table 3: Results of the four models.

	Linear estimate		Negative binomial estimate		Spatial error estimate			Generalized additive estimate	
(Intercept)			12.610	***			(Intercept)		
Population	0.063	*	0.000	*	0.058	***	Zero-car	0.002	**
Employment	0.682	***	0.000	***	0.683	***	One-car	0.062	.
Zero-car	0.006	**	2.329	***			Worker		
One-car			1.918	.			LowWageH		
Workers	0.245	.					MedWageH		
LowWageH	0.075	***	−1.588	.	0.074	*	Bus stop	−0.114	**
MedWageH							Bus route	0.194	***
Bus stop	−0.196	***	−0.019	***	−0.111	**	Road density		
Bus route	0.113	**	0.064	***	0.194	***	Transfer	0.018	**
Road density			0.023	*			Terminal	0.030	.
Distance to CBD	−0.080	***	−0.003	***			Smooth terms:		
Transfer	0.036	**	0.218	*	0.056	*		Edf	
Terminal	0.056	.			0.037	**	S (population)	1.24	**
							S (employment)	8.806	***
							S (distance to CBD)	1.077	***

Table 4: Ordinary least squares regression model outputs.

		Equal division		Naive method		Thiessen polygons	
		CBG	CT	CBG	CT	CBG	CT
New York		0.549	0.517	0.422	0.432	0.462	0.542
San Francisco		0.617	0.591	0.733	0.745	0.719	0.706
Chicago	Adjusted R^2	0.445	0.428	0.431	0.443	0.369	0.382
Philadelphia		0.891	0.919	0.887	0.912	0.886	0.921
Boston		0.634	0.57	0.424	0.409	0.543	0.513
New York		72.739	79.516	89.695	92.069	98.572	81.518
San Francisco		50.642	44.164	42.775	42.226	49.094	47.054
Chicago	MAPE	59.19	57.231	62.261	60.258	64.85	63.097
Philadelphia		177.094	155.92	185.698	164.094	188.526	161.804
Boston		148.806	157.567	233.008	265.39	158.748	159.404
New York		13896.030	13923.020	14000.460	13993.310	13972.550	13903.600
San Francisco		885.742	886.852	868.122	866.078	872.189	874.187
Chicago	AIC	4201.277	4204.414	4203.672	4200.771	4216.351	4212.398
Philadelphia		4062.948	4019.140	4068.485	4032.550	4070.100	4015.053
Boston		2491.448	2511.762	2551.117	2554.565	2521.508	2528.015

Table 5: Negative binomial regression model outputs.

		Equal division		Naive method		Thiessen polygons	
		CBG	CT	CBG	CT	CBG	CT
New York		0.67	0.662	0.652	0.62	0.624	0.640
San Francisco		0.71	0.69	0.713	0.721	0.733	0.723
Chicago	Adjusted R^2	0.441	0.450	0.431	0.462	0.378	0.385
Philadelphia		0.607	0.633	0.594	0.615	0.619	0.628
Boston		0.473	0.438	0.273	0.251	0.428	0.416
New York		68.168	69.302	69.302	78.57	80.5	75.684
San Francisco		36.268	36.01	35.252	34.88	34.649	35.743
Chicago	MAPE	56.809	56.802	59.602	56.513	65.188	63.648
Philadelphia		144.898	126.649	157.06	137.508	137.094	129.996
Boston		156.525	173.617	249.653	270.122	169.63	180.475
New York		13072.860	13082.100	13082.100	13134.810	13130.290	13111.300
San Francisco		828.829	828.838	827.508	826.25	825.107	826.703
Chicago	AIC	4104.980	4101.555	4106.677	4099.243	4119.761	4117.203
Philadelphia		3769.013	3758.913	3774.007	3766.500	3766.702	3761.777
Boston		2390.601	2398.985	2437.039	2440.725	2402.503	2404.642

Table 6: Generalized additive regression model outputs.

		Equal division		Naive method		Thiessen polygons	
		CBG	CT	CBG	CT	CBG	CT
New York	Adjusted R^2	0.683	0.706	0.539	0.529	0.689	0.653
San Francisco		0.98	0.984	0.971	0.974	0.965	0.967
Chicago		0.572	0.587	0.733	0.734	0.655	0.75
Philadelphia		0.982	0.987	0.99	0.989	0.987	0.987
Boston		0.635	0.66	0.456	0.427	0.559	0.647
New York	MAPE	69.19	64.682	79.478	84.718	78.939	85.246
San Francisco		12.224	10.923	13.763	14.431	15.845	16.421
Chicago		43.985	48.746	39.846	39.057	51.182	45.021
Philadelphia		113.19	103.895	79.543	85.854	92.655	104.403
Boston		150.868	133.957	201.751	263.342	169.054	182.116
New York	AIC	13761.000	13732.330	13919.390	13921.050	13756.010	13807.900
San Francisco		764.308	753.966	779.888	776.057	787.817	784.301
Chicago		4174.682	4171.752	4118.913	4118.002	4148.942	4112.932
Philadelphia		3813.419	3761.131	3733.719	3742.744	3769.477	3766.686
Boston		2491.280	2490.923	2546.696	2551.598	2517.802	2501.020

Table 7: Spatial error model outputs.

		Equal division		Naive method		Thiessen polygons	
		CBG	CT	CBG	CT	CBG	CT
New York	Adjusted R^2	0.567	0.565	0.438	0.449	0.477	0.557
San Francisco		0.771	0.763	0.763	0.764	0.768	0.788
Chicago		0.533	0.528	0.514	0.539	0.475	0.497
Philadelphia		0.897	0.903	0.886	0.920	0.898	0.928
Boston		0.663	0.595	0.509	0.474	0.581	0.547
New York	MAPE	74.733	75.294	89.749	91.908	95.056	85.800
San Francisco		43.068	42.702	42.300	42.100	45.948	43.021
Chicago		53.601	52.825	54.406	51.656	52.608	50.653
Philadelphia		138.200	111.300	141.637	117.328	201.967	152.672
Boston		106.656	126.006	183.444	208.402	151.789	173.288
New York	AIC	13897.300	13899.900	14006.800	13999.200	13976.300	13906.600
San Francisco		885.226	885.702	868.122	866.078	881.941	879.407
Chicago		4199.010	4201.250	4202.690	4196.470	4214.380	4209.230
Philadelphia		4041.238	4023.541	4072.510	4038.360	4072.340	4020.310
Boston		2499.250	2523.130	2550.900	2559.020	2527.610	2538.200

Table 8: Overlapping of station service coverage area in five cities.

City	Overlapping area (Km^2)	Total area (Km^2)	Overlap rate
New York	400.566	845.704	0.474
San Francisco	3.019	88.467	0.034
Chicago	71.574	281.44	0.254
Philadelphia	24.025	299.533	0.08
Boston	118.149	265.358	0.445

linear regression and the generalized additive models, indicating a better fit of the model. Besides, generalized additive models yield the highest accuracy rates. Typically, almost all generalized additive models have a smaller MAPE in our result, compared to the linear regression and negative binomial regression models. Negative binomial regression models also perform slightly better than the linear regression models on MAPE. Regarding the generalized additive models, the adjusted R^2 values of the models are higher than those of other models. Yet, smaller AIC values are observed with negative binomial regression models, indicating an overall better fit of data with such models. Smaller MAPE values are observed with the generalized additive models, indicating higher prediction accuracy with these models. Spatial error model also generates better results than the multiple linear regression model based on adjusted R^2.

4.4. Spatial Analysis Unit Comparison. From the results, after controlling for other factors, CBG and CT methods yield distinct performances in different cities, and there is no clear consensus on which one performs better. A possible

explanation is that the data obtained from CBGs are inferred from surveys which are subject to sampling biases. Therefore, when CBG data are aggregated onto CTs, these biases may wash out each other. Therefore, the accuracy of the data at the CT level may not be necessarily worse than that at the CBG level. On the other hand, due to the unobserved heterogeneity of urban rail transit station-level ridership, the room for improving the CBG level analysis can be marginal, despite having a higher accuracy. Both methods have close performance in accuracy. Therefore, for future station-level ridership modeling efforts, the selection of spatial analysis unit may not hold an upmost importance as changes in the overall outcome due to unit selection are minor. In addition, our results show that, for the same city, the model performance indicators AIC and adjusted R^2 carry the same functionality. The larger adjusted R^2, the smaller the AIC. A majority of MAPE and adjusted R^2 values are also the same in terms of functionality. Last, we find that the accuracy of the station-level direct ridership model varies greatly across different cities (ranging from 0.4 to 0.9 in adjusted R^2 values).

5. Discussion

This study discusses the effect of the treatment of service coverage overlap areas, spatial analysis units, and different model choices on the direct ridership modeling results of urban rail transit systems.

First, we investigate the treatment of overlapping buffers. Several previous studies have used the Thiessen polygons approach to deal with overlapping buffers [6–9]. However, they did not compare the Thiessen polygons approach with the naive approach. Although the method of equal division is more cumbersome to use, it can generate better results in dealing with the problem of overlapping buffers.

Regarding the issue of different spatial analysis units (CBG or CT), our results show that there is no evidence as to which one is better. Previous studies have used different spatial analysis units [6, 10–12], and in the future scholars can still use the most readily available spatial analysis unit.

With regard to different regression models, taking into account the overdispersion of ridership, nonlinear relationship between the independent and the dependent variables, and spatial error model could generate better results than the multiple linear regression based on adjusted R^2. In particular, the nonlinear model, GAM, has both lower MAPE value and higher adjusted R^2, which is consistent with the findings of existing studies [17, 23, 42].

6. Conclusions

The main contribution of this study is that we explored the effect of different treatment methods for the overlapping buffer area of stations, spatial analysis units, and regression models on station-level demand modeling results. This exploration answers the question of which treatment methods for the overlapping buffer area, spatial analysis unit, and regression model should be used in station-level demand modeling of urban rail transit. To obtain more convincing and universal conclusions, the ridership data of urban rail transit from five major cities in the United States are used to perform the study. We found that the nonlinear model outperforms the linear models in most of the cases. The equal division method usually performs better than the other two methods. Regarding the spatial analysis unit, the choice of CBG or CT does not influence the model results much. Thus, researchers could use either of them to perform the station-level demand modeling study.

Data Availability

The data used in this study are composed of two parts. The ridership data were taken from the transportation agencies of the five cities: New York [43], Philadelphia [44], Chicago [45], Boston [46], and San Francisco [47]. The built environment data were taken from the smart location database [48].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

H. Yang, C. Li, and X. Li conceived and designed the study. C. Li and X. Li performed data collection. H. Yang, C. Li, X. Li, J. Huo, Y. Wen, E. Parks, and Y. Liu performed analysis and interpretation of results. H. Yang, C. Li, X. Li, Y. Wen, E. Parks, and Y. Liu wrote the manuscript. All the authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (Grants nos. 71704145, 51774241, and 71831006), Humanity and Social Science Foundation of Ministry of Education of China (Grant no. 18YJCZH138), China Postdoctoral Science Foundation, and Sichuan Youth Science and Technology Innovation Research Team Project (Grants nos. 2019JDTD0002 and 2020JDTD0027).

References

[1] Y. Zhang, S. Zheng, C. Sun, and R. Wang, "Does subway proximity discourage automobility? Evidence from Beijing," *Transportation Research Part D: Transport and Environment*, vol. 52, pp. 506–517, 2017.

[2] H. Sung and J.-T. Oh, "Transit-oriented development in a high-density city: identifying its association with transit ridership in Seoul, Korea," *Cities*, vol. 28, no. 1, pp. 70–82, 2011.

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

[4] T. J. Kimpel, K. J. Dueker, and A. M. J. U. El-Geneidy, "Using GIS to measure the effect of overlapping service areas on

passenger boardings at bus stops," *Journal R.I.S.A.*vol. 19, no. 1, 2007.

[5] H. Yang, X. Li, C. Li, J. Huo, and Y. J. J. o. A. T. Liu, "How do different treatments of catchment area affect the station level demand modeling of urban rail transit?" *Journal of Advanced Transportation*, vol. 2021, Article ID 2763304, 19 pages, 2021.

[6] D. Zhang and X. Wang, "Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC," *Journal of Transport Geography*, vol. 41, pp. 107–115, 2014.

[7] S. Lee, C. Yi, and S.-P. Hong, "Urban structural hierarchy and the relationship between the ridership of the Seoul Metropolitan Subway and the land-use pattern of the station areas," *Cities*, vol. 35, pp. 69–77, 2013.

[8] 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," *Applied Geography*, vol. 34, pp. 548–558, 2012.

[9] L. S. Sun, S. W. Wang, L. Y. Yao, J. Rong, and J. M. Ma, "Estimation of transit ridership based on spatial analysis and precise land use data," *Transportation Letters*, vol. 8, no. 3, pp. 140–147, 2016.

[10] R. A. Mucci and G. D. Erhardt, "Evaluating the ability of transit direct ridership models to forecast medium-term ridership changes: evidence from san Francisco," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 46, pp. 21–30, 2018.

[11] D. Kim, Y. Ahn, S. Choi, and K. Kim, "Sustainable mobility: longitudinal analysis of built environment on transit ridership," *Sustainability*, vol. 8, no. 10, p. 1016, 2016.

[12] H. Yang, Y. Zhang, L. Zhong, X. Zhang, and Z. Ling, "Exploring spatial variation of bike sharing trip production and attraction: a study based on Chicago's Divvy system," *Applied Geography*, vol. 115, Article ID 102130, 2020.

[13] X. Li, Y. Liu, Z. Gao, and D. Liu, "Linkage between passenger demand and surrounding land-use patterns at urban rail transit stations: a canonical correlation analysis method and case study in Chongqing," *International Journal of Transportation Science and Technology*, vol. 5, no. 1, pp. 10–16, 2016.

[14] G. Thompson, J. Brown, and T. Bhattacharya, "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, 2012.

[15] Y. Zhu, F. Chen, Z. Wang, and J. Deng, "Spatio-temporal analysis of rail station ridership determinants in the built environment," *Transportation*, vol. 46, no. 6, pp. 2269–2289, 2018.

[16] C. Ding, X. Cao, M. Dong, Y. Zhang, and J. Yang, "Non-linear relationships between built environment characteristics and electric-bike ownership in Zhongshan, China," *Transportation Research Part D: Transport and Environment, China*, vol. 75, pp. 286–296, 2019.

[17] S. Hu, P. Chen, H. Lin, C. Xie, and X. Chen, "Promoting carsharing attractiveness and efficiency: an exploratory analysis," *Transportation Research Part D: Transport and Environment*, vol. 65, pp. 229–243, 2018.

[18] J. Huo, H. Yang, C. Li, R. Zheng, L. Yang, and Y. J. J. O. T. G. Wen, "Influence of the built environment on E-scooter sharing ridership: a tale of five cities," *Journal of Transport Geography*, vol. 93, Article ID 103084, 2021.

[19] H. Yang, T. Xu, D. Chen, H. Yang, and L. Pu, "Direct modeling of subway ridership at the station level: a study based on mixed geographically weighted regression," *Canadian Journal of Civil Engineering*, vol. 47, no. 5, pp. 534–545, 2020.

[20] C. Chen and C. E. McKnight, "Does the built environment make a difference? Additional evidence from the daily activity and travel behavior of homemakers living in New York City and suburbs," *Journal of Transport Geography*, vol. 15, no. 5, pp. 380–395, 2007.

[21] M.-J. Jun, K. Choi, J.-E. Jeong, K.-H. Kwon, and H.-J. Kim, "Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul," *Journal of Transport Geography*, vol. 48, pp. 30–40, 2015.

[22] C. Sabrina and M.-M. Luis, "A station-level ridership model for the metro network in Montreal," *Quebec*, vol. 40, no. 3, pp. 254–262, 2013.

[23] Z. Gan, M. Yang, T. Feng, and H. J. P. Timmermans, "Examining the relationship between built environment and metro ridership at station-to-station level," *Transportation Research Part D: Transport and Environment*, vol. 82, Article ID 102332, 2020.

[24] C. Liu, S. Erdogan, T. Ma, and F. W. Ducca, "How to increase rail ridership in Maryland: direct ridership models for policy guidance," *Journal of Urban Planning and Development*, vol. 142, no. 4, Article ID 04016017, 2016.

[25] J. Zhao, W. Deng, Y. Song, and Y. Zhu, "Analysis of Metro ridership at station level and station-to-station level in Nanjing: an approach based on direct demand models," *Transportation*, vol. 41, no. 1, pp. 133–155, 2013.

[26] K. Sohn and H. Shim, "Factors generating boardings at Metro stations in the Seoul metropolitan area," *Cities*, vol. 27, no. 5, pp. 358–368, 2010.

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

[28] A. Galelo, A. Ribeiro, L. M. Martinez, and B. Sciences, "Measuring and evaluating the impacts of TOD measures - searching for evidence of TOD characteristics in azambuja train line," *Procedia - Social and Behavioral Sciences*, vol. 111, pp. 899–908, 2014.

[29] G. Currie and A. Delbosc, "Understanding ridership drivers for bus rapid transit systems in Australia," in *Proceedings of the 33rd Australasian Transport Research Forum*, Canberra, Australia, October 2010.

[30] F. Wang and C. L. Ross, "Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 47, pp. 35–45, 2018.

[31] C. Ding, X. Cao, B. Yu, and Y. Ju, "Non-linear associations between zonal built environment attributes and transit commuting mode choice accounting for spatial heterogeneity," *Transportation Research Part A: Policy and Practice*, vol. 148, pp. 22–35, 2021.

[32] B. Wali, L. D. Frank, J. E. Chapman, and E. H. Fox, "Developing policy thresholds for objectively measured environmental features to support active travel," *Transportation Research Part D: Transport and Environment*, vol. 90, Article ID 102678, 2021.

[33] Z. Gan, T. Feng, M. Yang, H. Timmermans, and J. Luo, "Analysis of metro station ridership considering spatial heterogeneity," *Chinese Geographical Science*, vol. 29, no. 6, pp. 1065–1077, 2019.

[34] R. Pan, H. Yang, K. Xie, and Y. Wen, "Exploring the equity of traditional and ride-hailing taxi services during peak hours,"

Transportation Research Record: Journal of the Transportation Research Board, vol. 2674, no. 9, pp. 266–278, 2020.

[35] S. Wood, "MGCV: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation," Thesis, University of Bath, Bath, UK, 2012.

[36] B. Ripley, B. Venables, D. M. Bates et al., "Package 'mass'," vol. 538, 2013, https://cran.r-project.org/web/packages/MASS/MASS.pdf.

[37] R. Bivand and G. J. R Piras, "Spatial regression analysis," vol. 1, pp. 1–5, 2019, https://cran.r-project.org/web/packages/spatialreg/spatialreg.pdf.

[38] J. J. W. S. S. R. O. Miles, *R Squared, Adjusted R Squared*, Wiley, New Jersey, NJ, USA, 2014.

[39] U. Khair, H. Fahmi, S. A. Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," *Journal of Physics: Conference Series*, IOP Publishing, vol. 930, , Article ID 012002, 2017.

[40] H. Bozdogan, "Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[41] H. Yang, Y. Liang, and L. Yang, "Equitable? Exploring ridesourcing waiting time and its determinants," *Transportation Research Part D: Transport and Environment*, vol. 93, Article ID 102774, 2021.

[42] Y. Xu, X. Yan, X. Liu, and X. Zhao, "Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships," *Transportation Research Part A: Policy and Practice*, vol. 144, pp. 170–188, 2021.

[43] Metropolitan Transportation Authority: 2010, https://www.mta.info/nyct.

[44] Southeastern Pennsylvania Transportation Authority, 2010, http://www.septa.org/.

[45] Chicago Transit Authority, 2010, https://www.transitchicago.com/.

[46] Massachusetts Bay Transportation Authority, 2010, https://www.mbta.com/.

[47] Bay Area Rapid Transit District: 2010, https://www.bart.gov/.

[48] K. Ramsey and A. J. W. Bell, *Smart Location Database*, 2014, https://www.epa.gov/sites/default/files/2014-03/documents/sld_userguide.pdf.