Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 2763304, 19 pages
https://doi.org/10.1155/2021/2763304

Research Article

How Do Different Treatments of Catchment Area Affect the Station Level Demand Modeling of Urban Rail Transit?

Hongtai Yang, Xuan Li, Chaojing Li, Jinghai Huo, and Yugang Liu

School of Transportation and Logistics, Institute of System Science and Engineering,
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

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

Received 14 April 2021; Accepted 11 June 2021; Published 30 June 2021

Academic Editor: Maria Vittoria Corazza

Direct demand modeling is a useful tool to estimate the demand of urban rail transit stations and to determine factors that significantly influence such demand. The construction of a direct demand model involves determination of the catchment area. Although there have been many methods to determine the catchment area, the choice of those methods is very arbitrary. Different methods will lead to different results and their effects on the results are still not clear. This paper intends to investigate this issue by focusing on three aspects related to the catchment area: size of the catchment area, processing methods of the overlapping areas, and whether to apply the distance decay function on the catchment area. Five catchment areas are defined by drawing buffers around each station with radius distance ranging from 300 to 1500 meters with the interval of 300 meters. Three methods to process the overlapping areas are tested, which are the naïve method, Thiessen polygon, and equal division. The effect of distance decay is considered by applying lower weight to the outer catchment area. Data from five cities in the United States are analyzed. Built environment characteristics within the catchment area are extracted as explanatory variables. Annual average weekday ridership of each station is used as the response variable. To further analyze the effect of regression models on the results, three commonly used models, including the linear regression, log-linear regression, and negative binomial regression models, are applied to examine which type of catchment area yields the highest goodness-of-fit. We find that the ideal buffer sizes vary among cities, and different buffer sizes do not have a great impact on the model's goodness-of-fit and prediction accuracy. When the catchment areas are heavily overlapping, dividing the overlapping area by the number of times of overlapping can improve model results. The application of distance decay function could barely improve the model results. The goodness-of-fit of the three models is comparable, though the log-linear regression model has the highest prediction accuracy. This study could provide useful references for researchers and planners on how to select catchment areas when constructing direct demand models for urban rail transit stations.

1. Introduction

Transit-oriented development (TOD) plays a pivotal role in urban planning. TOD refers to a planning and design method that maximizes the use of public transportation for both residential and commercial areas. For example, a customized urban center with a radius of 400–800 meters can be built to integrate commerce, education, culture, employment, and residence facilities adjacent to the public transportation stations. In TOD planning, an important component is to establish an accurate and reliable direct ridership model to facilitate transportation operators to formulate urban rail transit operation strategies and assist urban planners to design more efficient and convenient urban and transportation plans.

In direct ridership models, land use characteristics within the catchment area are indispensable. A crucial question in building a ridership model is the choice of an appropriate size for catchment area. Some previous studies used 800 meters as the radius of the circular buffer to predict the ridership at the transit station level [1, 2]. Others set up direct ridership models with 500 meters as the buffer radius. In addition, many introduce the concept of a half-mile walking distance as the radius of the catchment area [3, 4]. Cervero [5] studied the commuting options of people living within 0.5 and 3 miles of the San Francisco Bay Area with the results showing that people who work near transit stations are more likely to live within 0.5 miles of urban rail transit stations and use public transit to commute. He also studied the factors affecting travel demand based on a total of 261 light rail stations in the United States and Canada [6]. He suggested that both population and employment densities within 0.5 miles of the station positively correlate with daily transit ridership. Pan et al. [7] created buffer areas of 500 meters, 1000 meters, and 2000 meters to model Shanghai subway ridership and found similar results. There is a wealth of studies on direct ridership modeling which use different radius distances for the catchment area, and it appears that the choice criteria of buffers have not yet reached a consensus.

The task of determining an appropriate size for catchment area is not limited to choosing the radius of the catchment area but includes how to apply the catchment area. Many studies directly used the circular catchment area to obtain values of influencing variables within the catchment area after determining the radius [7–9]. The circular catchment area does nothing with the overlap area, we regard it as the first method. The most common and straightforward approach to generate the circular area is by using the ArcGIS software. However, in cities with dense stations, there will be excessive overlaps between the catchment areas, with some areas being repeatedly counted, which may impose a negative impact on the results. Therefore, some studies employed the Thiessen polygons to tackle the overlapping areas by assigning the closets points of a station to a polygon around that station [1, 10–12]. In our study, Thiessen polygon is the second method to deal with the overlapping catchment area. Yet, the Thiessen polygon method has its limitations. For example, it does not perform well for densely urbanized areas where the short distances among stations may generate clusters of small polygons, which lead to potential calculation inaccuracy of variable values. To overcome the overlapping issue, we propose to improve the circular buffer area method by dividing the overlapping area by the number of times of overlapping. So, we take the approach of dividing the overlap area equally as the third method to process the coverage overlap area. Besides the treatment of the overlapping area, some scholars used the distance decay method to represent the fact that the attraction of an urban rail transit station decreases as the distance to the station increases [13, 14]. In distance decay approach, the weights of variables change with the distance. Some different weighting methods of distance decay have appeared one after another, and we will compare two distance decay methods with different weighing methods.

Once the catchment area has been determined and properly represented, the next task is to establish the direct ridership models. Regarding the regression models in transportation demand research, most studies in recent years used ordinary multiple linear regression techniques [2, 4, 15]. In this line of studies, there are also some attempts to apply logarithmic transformation on the dependent variable to form a log-linear regression model [16]. Last, some papers adopt the use of negative binomial regression models [9, 17, 18].

In light of this gap in the literature, our paper establishes direct ridership models for five cities in the United States, aiming to answer the following four questions. (1) What is an appropriate size of the catchment area? (2) Among the three methods to process the overlap of catchment area, i.e., the naïve implementation (ordinary circle), combining circle, and Thiessen polygon into new catchment area, and dividing the overlap of the circular buffer area equally into adjacent circles, which method performs the best? (3) Which weighting method is better when using the distance decay approach? (4) Which model performs better among linear regression, log-linear regression, and negative binomial regression techniques? To avoid drawing biased conclusion from one city, we included five different cities in the U.S. to obtain more generalizable findings.

This paper consists of five parts. Section 1 reviews the existing literature in the following aspects: influencing factors of transit ridership, buffer radius selection, coverage overlapping area treatment methods, and model selection. Section 2 describes the research objectives and data sources. Section 3 presents the main research design and methodology, including the treatment of the overlapping area of a catchment area, model treatment, and two simplified weighting methods. Section 4 discusses the model results, interpretation, and analysis. Section 5 concludes our study findings.

2. Literature Review

Understanding the influencing factors of transit ridership has been a recent research focus. In general, the influencing factors can be divided into the following three categories: socioeconomic, land use, and traffic attributes. Former studies show that socioeconomic characteristics such as population and employment are positively correlated with transit ridership [4, 17, 19–21]. Land use characteristics include but are not limited to 3Ds, land use density, design, diversity, and mixed land use levels [22–26]. Scholars found land use density and diversity have a positive impact on ridership [21]. The traffic attributes and station characteristics are such as bus routes, road density, and accessibility; station types may also have significant influence on transit demand [2, 27–30]. Some authors found road density is positively related to ridership; they also noted that transfer station and terminal station increased ridership [2, 8].

In the studies of land use and transportation demand, land use variables are usually measured on the basis of arbitrarily defined areas, and various methods for defining catchment areas lead to distinct results. Ruiz-Pérez and Seguí-Pons [31] compared the effect of four different geographical spatial units (neighborhood, census section, cadastral block, and 400 × 400 m mesh) on bus service level analysis. Results showed that different zones led to significant differences of the service level and combining zonings of different sizes simultaneously was recommended. Guerra et al. [15] studied 6 buffer bands with increments of 0.25 mile from 0.25 mile to 1.5 mile and concluded that different catchment areas have little influence on a model's predictive power. After analyzing the data of 1,449 rail transit stations in 21 cities in the United States, Jun et al. [21] studied the land use characteristics of the Seoul metropolitan area and the impact of land use characteristics on station-level ridership with different buffer bands for the buffer area of 300 meters, 300–600 meters, and 600–900 meters. Their results show that the impact of population density and mixed land use on ridership is only significant at the 300 meters and 300–600 meters buffer band levels, recommending that a compact urban pedestrian catchment area like Seoul should be defined using a radius of 600 meters. Mitra and Buliung [32] established 4 buffer areas at different scales (250 meters, 400 meters, 800 meters, and 1000 meters) around children's homes and schools, measured the effects of the built environment characteristics in 4 different scales, and found that the goodness-of-fits of the four models do not exhibit much different though as the distance increases, the model fits slightly worse. In addition, the magnitudes of effects of the individual built environment characteristics are inconsistent with different scale buffers. Relying on survey data, Li et al. [9] identified seven buffer zones of 50 meters, 100 meters, 200 meters, 400 meters, 800 meters, 1200 meters, and 1600 meters and established a regression model on resident's travel behavior. As can be seen from these studies, no conclusion has been drawn with respect to the size of the catchment area and the effect of the size of the catchment area on the modeling results is still not clear.

The handling of the buffers can significantly affect the results and should be treated with caution. When processing the buffer zone, most articles adopt a naïve circular buffer method, which means that the issue of overlapping areas of station coverage is simply ignored [8, 9, 19, 21, 28, 33–35]. Still, there are some other methods. For example, Li et al. [22] used Thiessen polygons to deal with the overlapping issue with the 800 meters circular buffer. When Sun et al. [12] divided the multilevel water catchment area, the radius of the pedestrian and traffic area was determined by the residents' travel survey, and the division of the potential catchment area was determined by the Thiessen polygon. Gutiérrez et al. [13] generated Thiessen polygons to specifically divide service areas based on multiple circular bands in order to consider competition between stations. The cropped area used by Guerra et al. [15] is similar to the buffer area generated by Thiessen polygon. Kuby et al. [4] used a grid-based connection on-off network method to improve the standard buffer delimitation method of ArcGIS to redefine more accurate service area. The equal division method adopted in this article is not covered by the previous studies, and its merits and limitations are not compared to other buffer delimitation methods. Considering that the Euclidean buffers may overlap, Corazza and Favaretto [36] used the network buffers which were determined based on polygons covering all the edges that are within 400-meter area of the stop.

Distance to the station is another important consideration in the model. Some studies found that the probability of passengers choosing a station is related to the distance to the station. Untermann found that most people are willing to walk up to 500 feet (152.4 meters), 40% are willing to walk 1,000 feet (304.8 meters), but only 10% are willing to walk up to 1 mile [37]. This uncertainty of distance inspires innovations in methodology such as the distance decay method. Gutiérrez et al. [13] combined the distance decay function with the multiple regression analysis to establish a rapid response passenger prediction model using different distance thresholds to improve the model results (3.4% on the 800 m threshold). Yet, the distance decay method is not without its limitations. The studies considering distance decay only infer conclusions based on their own results without comparing their results with other weighting scales or verifying their results using data of other cities.

Model improvement techniques on this topic have been developed and applied over the recent years. In most previous studies, linear regression is frequently used [2, 4, 7, 11, 12, 15, 28, 38]. However, skewed distribution of the dependent variables can be an issue for the linear regression approach. To solve this, logarithmic transformation of dependent variables can be applied [16]. Concerning overdispersed data, some chose negative binomial regression to reduce standard errors [9, 17, 18]. For example, Thompson et al. [17] used a negative binomial regression model to study factors affecting transit ridership in Florida and found that some variables such as population, total household income, and total employment can explain rise in bus ridership.

In summary, although the existing studies have analyzed the impact of the built environment on ridership from many aspects, the effects of the size of the catchment area, processing methods of the overlapping area, whether to consider distance decay and model selection on the results are still not clear. This article tries to study those effects by analyzing the urban rail transit ridership data from five American cities.

3. Methods

3.1. Study Area. We chose five American urban rail transit systems for our analysis (i.e., New York, Chicago, Philadelphia, Boston, and San Francisco Bay Area) as these urban rail transit systems are well developed and serve a large metropolitan population. The spatial analysis unit is census block groups (CBG). Among these cities, New York has the largest urban rail transit system with a total of 36 lines and 472 stations. Its urban rail transit system also has the longest history among the five systems. The rail transit systems of Chicago, Philadelphia, and Boston have similar number of stations, more than 100. The Bay Area Rapid Transit system has 50 stations. The urban rail transit systems of New York and Chicago are shown in Figure 1 as examples.

The buffer areas defined in this paper are circular buffer areas with radii of 300 meters, 600 meters, 900 meters, 1200 meters, and 1500 meters. On this basis, bands of 300–600 meters, 600–900 meters, 900–1200 meters, and 1200–1500

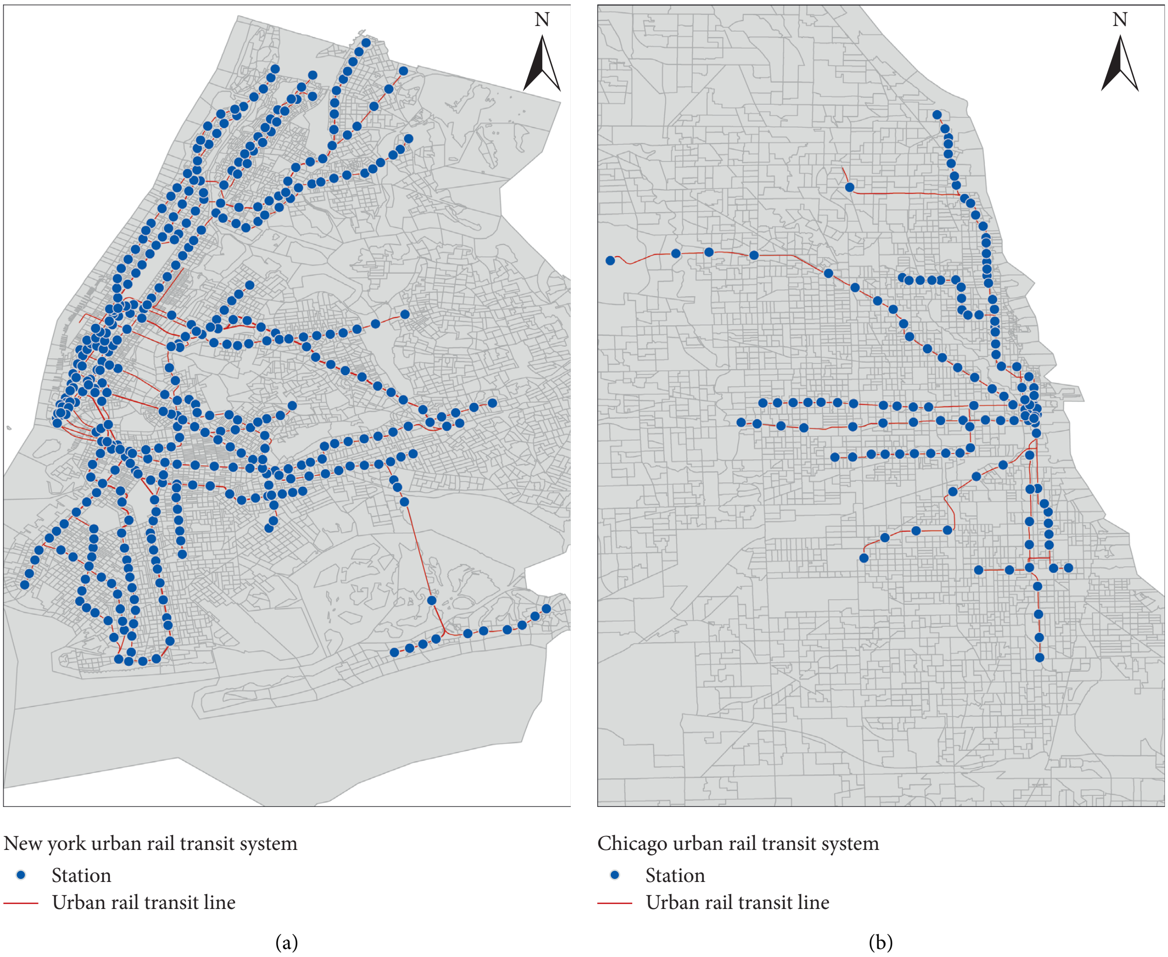


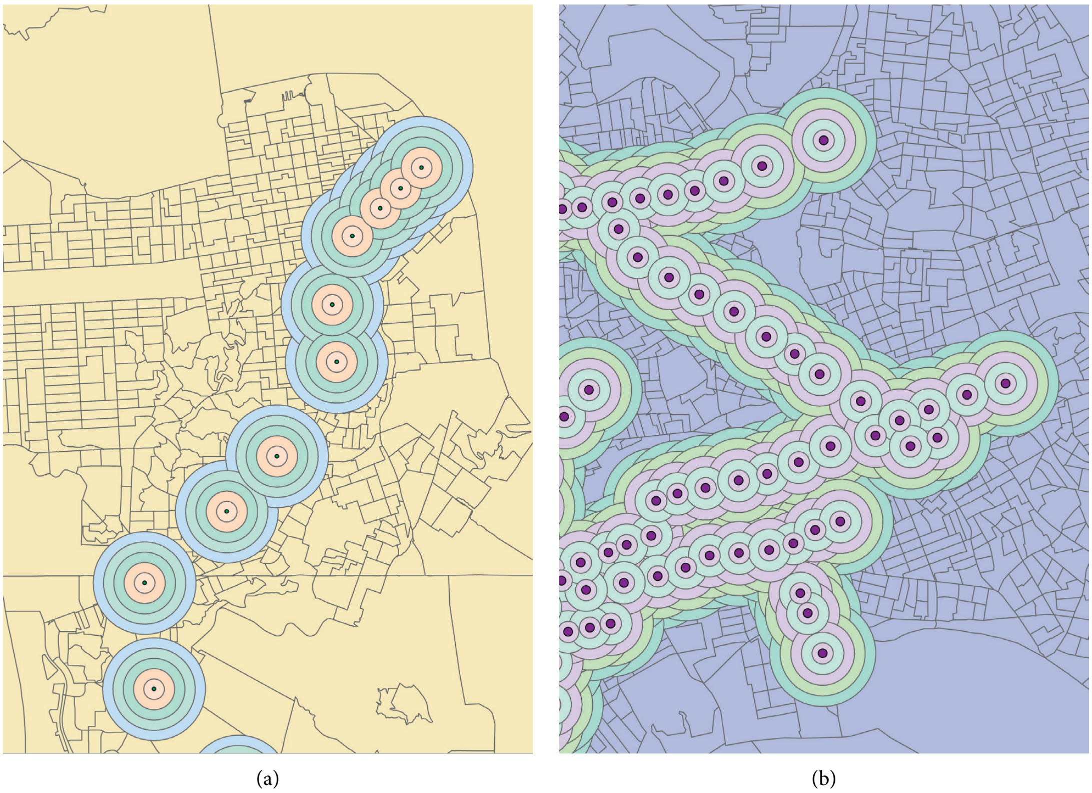
Figure 1: The urban rail transit systems of New York and Chicago.

meters were developed as inputs for the distance decay function. The distribution of the urban rail transit stations in the five cities is different: for example, the urban rail transit stations in San Francisco Bay Area are relatively scattered, and the overlapping buffer area is small. In contrast, urban rail transit stations in New York are denser and therefore there is much higher overlapping buffer area (Figure 2).

3.2. Transit Ridership Data. The dependent variable of our study is the average weekday ridership of the five urban rail transit systems in 2010. The year of 2010 is selected because the values of built environment variables of 2010 are very accurate. We obtained ridership data from the New York City Transit Authority (MTA), the Chicago Transit Authority (CTA), the Massachusetts Bay Transportation Authority (MBTA), Southeastern Pennsylvania Transportation Authority (SEPTA), Port Authority Transit Corporation (PATCO), and the Bay Area Rapid Transit (BART), respectively.

To facilitate model selection, we plot histograms of ridership and logarithmic transformed ridership for each city, as shown in Figure 3. It can be seen that New York has the highest ridership. The ridership of all cities is skewed to the right. After performing the logarithmic transformation on the ridership, their distributions are closer to the normal distribution. Through Figure 3, we noticed that the distribution of ridership in Philadelphia is more uneven: in most days the ridership is below 5000 passengers, with only a few days exceeding this value.

3.3. Independent Variables. Based on the literature review, we selected 18 built environment variables as our independent variables. These variables include socioeconomic variables, built environment variables, and station characteristic variables. Table 1 presents the description of these independent variables. The socioeconomic variables are population, employment, proportion of households with one car or less, proportion of low-income family, and so on. Such variables are drawn from the Smart Location Database (SLD) dataset. The built environment variables include the density of the road network, the distance from the station to the Central Business District (CBD), the number of bus stations within the service area of the station, and the number of bus lines within a 200 meters buffer around the station. The station attribute variables include the number of subway lines available at the station, and two dummy variables that indicate whether the station is a transfer station or a terminal station, respectively.



(a)

(b)

Figure 2: Overlapping buffer area of some stations in the Bay Area (a) and New York (b).

3.4. Processing Method of Overlapping Area of Station Service Coverage. When we draw circular buffers around stations, there could be overlaps between the station buffers if the stations are close to each other (Figure 4). We compare several methods to deal with the overlapping problem here.

3.4.1. Naïve Method. This method ignores this issue and the overlapping areas will be counted multiple times when calculating values of the variables for each buffer. For example, if three buffers intersect in some area, this overlapping area will be counted into all of these three buffers, which means it will be calculated 3 times. As a result, some variables such as population, employment, housing, density of residents, and number of bus stops are repeatedly counted.

3.4.2. Thiessen Polygon. Thiessen polygons are also called Voronoi diagrams or Voronoi polygons. It is one of the basic methods to analyze neighborhood in proximity. Thiessen polygons are used to describe the areas of influence of sample points. For any point in a Thiessen polygon, its distance to the sample point in the polygon is less than its distance to any other sample point. An example of Thiessen polygons based on some stations in Chicago is shown in Figure 5(a) and Figure 5(b).

The specific steps are as follows. First, create circular buffer areas and Thiessen polygons around stations, respectively. Second, use the intersection tool in ArcGIS to intersect the Thiessen polygons with the circular buffer. As we can see in Figure 6(a), it is the boundaries of the Thiessen polygons that cut the overlapping areas of the circular buffers. Finally, we use the fusion function in the data management tool in ArcGIS to form the new catchment area (Figure 6(b)). The advantage of using Thiessen polygons here is that it can handle overlapping areas between buffers to avoid overlapping areas from being double counted.

3.4.3. Equal Division. Equal division method divides the overlapping area by the number of overlapping buffers and applies the division result as a weight to calculate the value of variables. For example, if three buffer areas overlap, the values of some variables such as population, employment, the number of bus stops within the overlapping areas are divided by three and assigned to each buffer area. We use python to implement the aforementioned procedures.

3.5. The Distance Decay Function on Buffer Bands. The theoretical basis of the distance decay function is Tobler's First Law of Geography, which indicated that sample points that are closer have greater impact on the results than the points that are far away. Previous studies found that when the walking distance of potential users increases, the public transport usage decreases [39]. The effect of distance is converted into the weight in the mathematical model. By applying weight to explanatory variables that are within different buffer band, a distance decay weighted regression model is established.

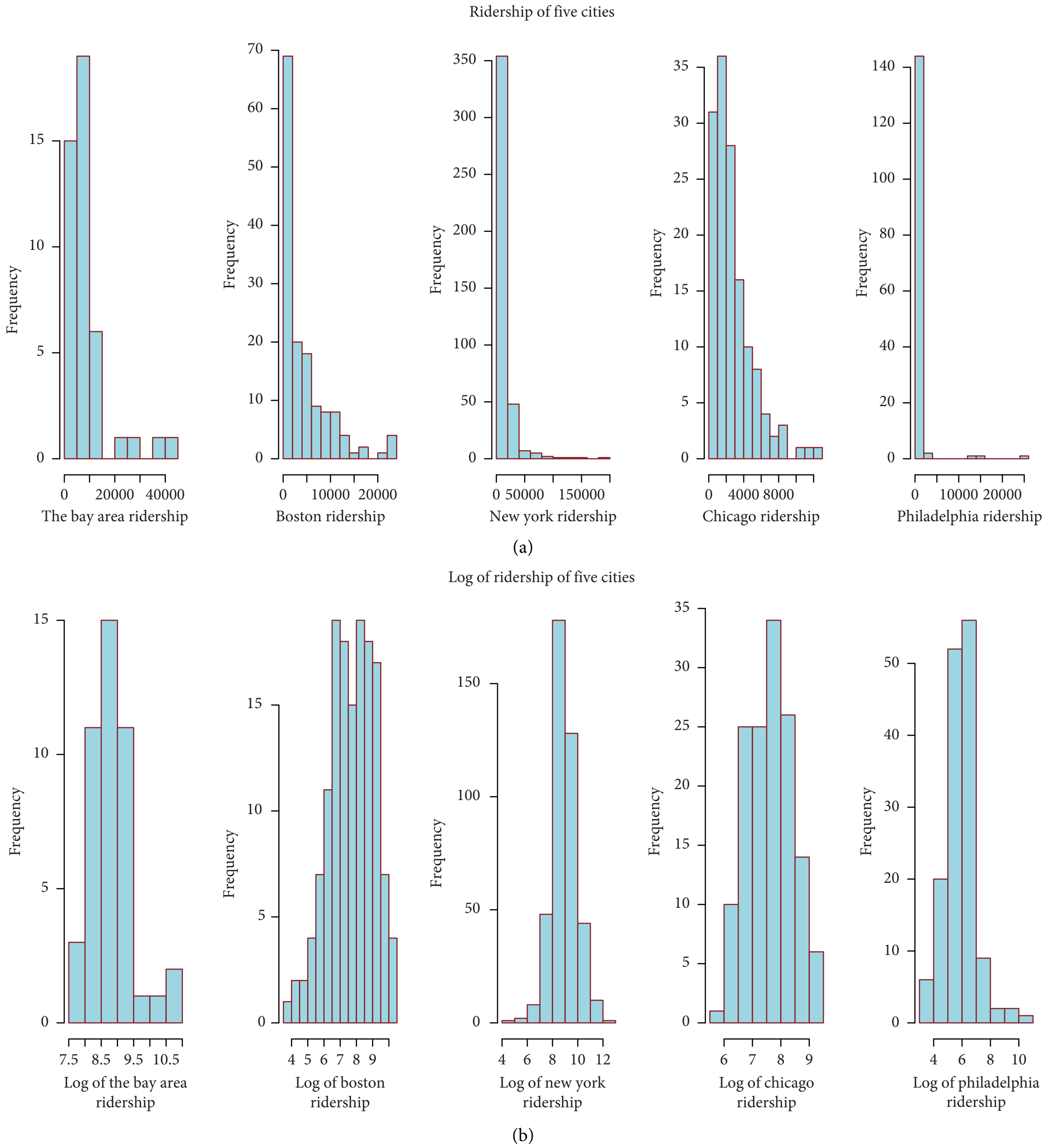


Figure 3: The histograms of ridership distribution of five selected urban rail transit systems.

The buffer bands used in this article are within 300 meters, 300–600 meters, 600–900 meters, 900–1200 meters, and 1200–1500 meters. The weights of these buffer bands are determined based on the distance.

Gutiérrez et al. [13] used a linear distance decay function to perform a weighted regression. We adopt that linear distance decay function and apply the weight of 0.5, 0.4, 0.3, 0.2, 0.1 to the buffer bands of within 300 meters, 300–600 meters, 600–900 meters, 900–1200 meters, and 1200–1500 meters, respectively.

Manout et al. [39] calibrated the distance decay function using the data of the Paris family travel survey. We also adopt this nonlinear distance decay function and apply 0.8, 0.3, 0.1, 0, and 0 to the buffer bands of within 300 meters, 300–600 meters, 600–900 meters, 900–1200 meters, and 1200–1500 meters, respectively. The two weighting methods are shown in Table 2.

4. Model Description

4.1. Multiple Linear Regression. Multiple linear regression is a commonly used regression model [40, 41]. In this article, we employ the multiple linear regression to evaluate the impact of multiple independent variables on station-level

Table 1: The description of independent variables.

Variables	Description
Socioeconomic variables	
HU	Number of housing units
POP	Population
EMP	Number of employments
HH	Number of households (occupied housing units)
AUT	Percentage of households with zero or one automobile
WORKER	Number of workers
RLOW	Percentage of workers earning $1250/month or less who live in the CBG
ELOW	Percentage of workers earning $1250/month or less who work in the CBG
Built environment variables	
GRD	Gross residential density (housing units/acre) on unprotected land
GPD	Gross population density (people/acre) on unprotected land
GED	Gross employment density (jobs/acre) on unprotected land
ROAD	Total road network density
DIST	Euclidean distance from transit stop to CBD (meters)
BS	Number of bus stops within the station service coverage
BR	Number of bus routes within 200 meters buffer of the station
Station characteristic variables	
LN	Number of urban rail transit lines at the station
TRANS	If it is a transfer station, coded as 1, otherwise as 0
TERM	If it is a terminal station, coded as 1, otherwise as 0

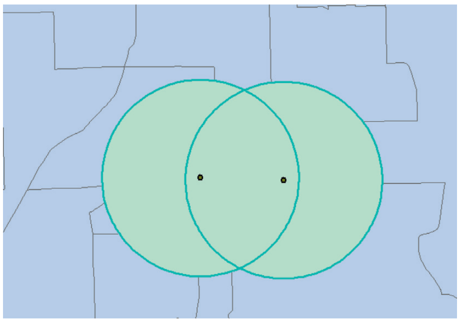


Figure 4: Overlapping area of station service coverage.

ridership. The parameters in linear regression model are determined by minimizing the sum of squared errors. The function of multiple linear regression is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where y_i represents the dependent variable in the model, which is the average annual weekday ridership of an urban rail transit station. $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients. For example β_k represents the average change of the dependent variable for each additional unit of x_{ik} while keeping other variables constant.

4.2. Log-Linear Regression. Figure 3 shows that the ridership of the five cities is skewed to the right. As a result, the natural logarithmic transformation of the dependent variable is performed and the transformed variable follows the normal distribution approximately. The function of the log-linear regression is as follows:

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

The interpretation is similar to that of the multiple linear regression. The only difference is that in the log-linear regression model, each additional unit of x_{ik} will make the dependent variable increase by e^{β_k} times.

4.3. Negative Binomial Regression. The dependent variable in this study, ridership, is a count variable. The negative binomial regression model is commonly used to analyze count variable. In addition, it can be observed from the histogram that the ridership is skewed to the right, which satisfies the assumption of the negative binomial regression that the mean is greater than the variance. The function of the negative binomial regression model is shown below:

$$\mu_i = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}}. \tag{3}$$

Among them, $\beta_1, \beta_2, \ldots, \beta_m$ are the regression coefficients and $x_{1i}, x_{2i}, \ldots, x_{mi}$ are the independent variables.

5. Results

By applying the methods and models mentioned in the previous sections, we got the final result. We used three indicators to evaluate the goodness-of-fit of the models: adjusted R^2, mean absolute percentage error (MAPE), and Akaike information criterion (AIC).

Adjusted R^2 means degree-of-freedom adjusted coefficient of determination. In the model results, higher adjusted R^2 indicates better goodness-of-fit. The function of adjusted R^2 is shown below:

$$\text{Adjusted } R^2 = 1 - \frac{\text{SSR}(n - p - 1)}{\text{SST}(n - 1)}. \tag{4}$$

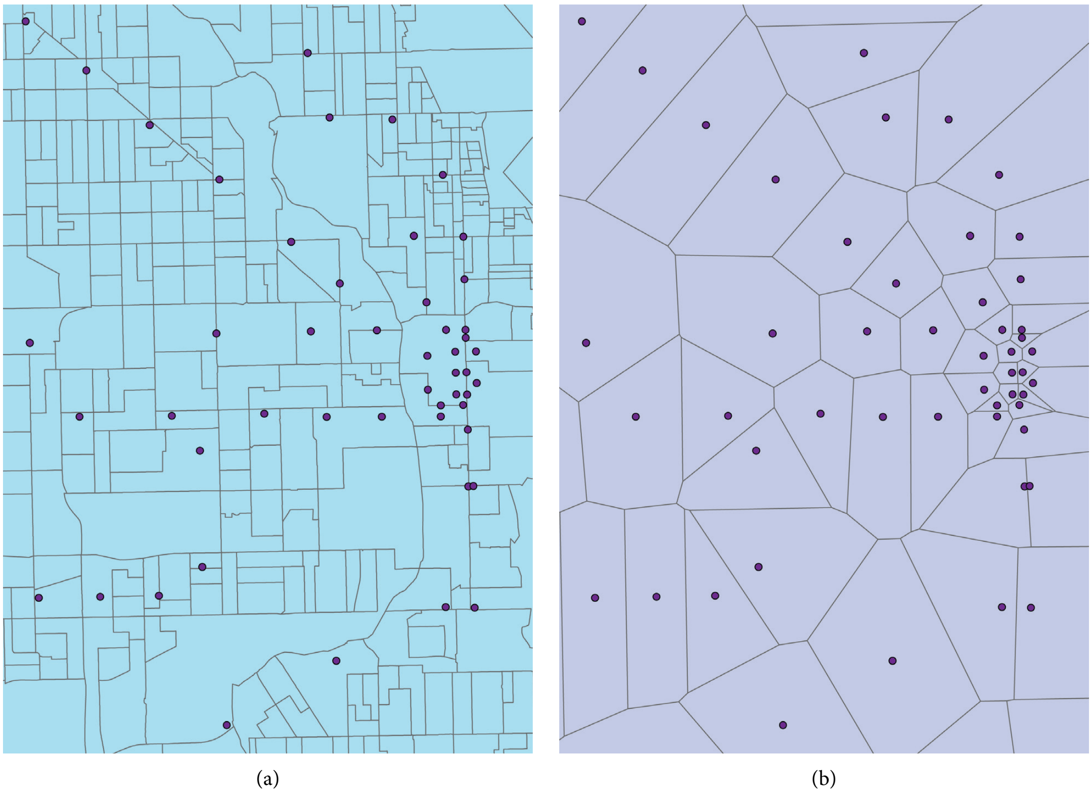


Figure 5: Chicago transit stations (a) and Thiessen polygons (b).

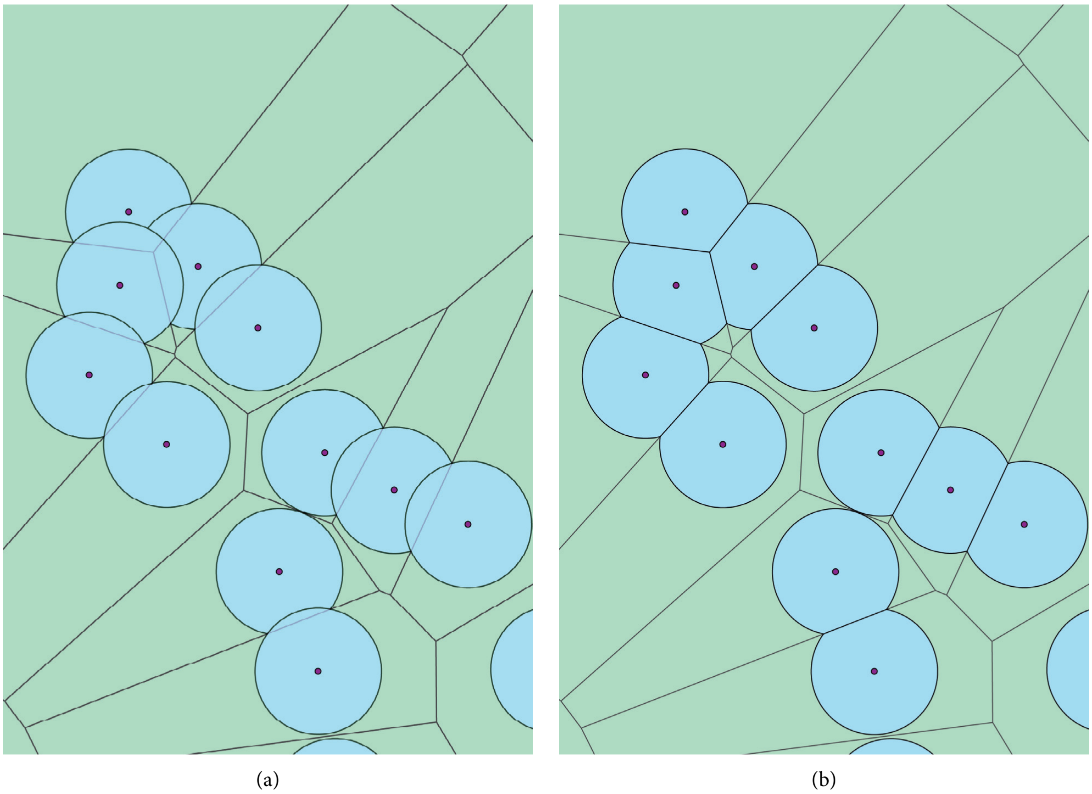


Figure 6: Intersection of circular buffer and Thiessen polygons (a) and new catchment area (b).

Table 2: The weights of each buffer bands.

Buffer bands (meters)	Within 300	300–600	600–900	900–1200	1200–1500
Linear distance decay function	0.5	0.4	0.3	0.2	0.1
Nonlinear distance decay function	0.8	0.3	0.1	0	0

In (4), SSR is the sum of squares of the errors. SST is total sum of squares, which is the sum of squares of the difference between the observed value and the mean.

The formula to calculate the MAPE is as follows:

$$\mathrm{MAPE} = \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \frac{100\%}{n}, \tag{5}$$

where $\hat{y}_i$ is the predicted value of the dependent variable and y_i is the observed value. In the model results, smaller MAPE indicates better model prediction. Its value can be calculated directly with the *forecast* package in R. For log-linear regression, we need to convert the log transformed variable dependent to the original dependent variable to calculate the true MAPE.

AIC is another measure to evaluate the goodness-of-fit of statistical models. Smaller AIC value indicates better goodness-of-fit. The AIC value is calculated by the following function:

$$\mathrm{AIC} = -2\ln(L) + 2k, \tag{6}$$

where k is the number of variables in the model and L is the likelihood function.

Due to the limitation of space, we only illustrate the results of the regression models for the Thiessen polygon with 900 meters radius circle buffer of New York city here, as in Table 3. All the three regression models fit the data well.

It can be observed from Table 3 that population and employment are significant variables that are positively correlated with ridership. In addition, the number of bus lines and the number of urban rail transit lines, whether it is a transfer station or it is a terminal station, are all positively correlated with ridership in the three models. The distance between the station and the CBD is negatively correlated with ridership in the three models, which shows that as the distance increases, people's intention to choose the urban rail transit decreases.

5.1. Comparison of Five Catchment Areas. To explore the most suitable buffer size, we compare the model results of using five different catchment areas.

The adjusted R^2 of stepwise linear regression and the AIC of negative binomial regression for different buffer sizes are shown in Figure 7 and Figure 8, respectively. It can be seen that the goodness-of-fits of models using different buffer sizes are very similar. From Figure 7, we can see that for Chicago and Boston, the buffer of 900 meters performs the best. For New York and Philadelphia, the buffer of 600 meters performs the best. For the Bay Area, the buffer of 1500 meters is the most suitable. Therefore, the most suitable buffer size is different for different cities. We also notice that the difference in goodness-of-fit among various buffer sizes is not much. As a result, researchers could use the handiest buffer size when estimating the station level ridership, which is consistent with the result of Guerra et al.'s study [15].

5.2. Comparison of Three Processing Methods of Overlapping Area. We take buffer sizes of 300 meters, 900 meters, and 1500 meters as examples to show the ratio of overlapping area to total buffer area (Table 4). It can be seen that, as the buffer size increases, the overlapping ratio increases sufficiently, especially for Chicago, New York, and Boston.

The goodness-of-fits of the linear regression model with the three buffer processing methods are shown in Figure 8. In Figure 9, the naive method that does not deal with the overlapping area is represented as type 1, the Thiessen polygon method is represented as type 2, and the equal division method is represented as type 3.

From Figure 9, we can see from the adjusted R^2 of Chicago, New York, and Boston that the equal division method has better goodness-of-fit than the Thiessen polygon method and the naive method. From Table 4, we can see that the stations in these three cities are densely distributed and the overlap ratio is high. Due to the scattered distribution of stations in the San Francisco Bay Area, the overlap ratio is low. When the buffer radius is 900 meters, the results of the equal division method and the Thiessen polygon method do not outperform that of the naive method. It implies that for cities with densely distributed urban rail transit stations, the equal division method and Thiessen polygon method can generate better results.

5.3. Comparison of Weighting Methods Based on Two Distance Decay Functions. Figure 10 and 11 show the adjusted R^2 of weighting methods based on the linear distance decay function (0.5, 0.4, 0.3, 0.2, and 0.1) and the nonlinear distance decay function (0.8, 0.3, 0.1, 0, and 0) using linear regression and negative binomial regression, respectively. The line in the figure is the average adjusted R^2 of ordinary circular buffer without applying weighting, which we called naïve method.

The trends in Figure 10 and Figure 11 are basically the same. In general, the results of the first weighting method are better than that of the second weighting method in most cases, although the difference is marginal. There is a little difference between the results of the first weighting method and the ordinary circular buffer, which shows that the distance decay function could barely improve the model result.

5.4. Comparison of Three Models. The average values of adjusted R^2 for different buffer sizes are shown in Figure 12. From the figure, we find that the model results of New York

Table 3: Comparison of the results of three models of 900 m Thiessen Polygon buffer area in New York.

Independent variables	Model 1 (linear)	Model 2 (log linear)	Model 3 (negative binomial)
POP	110***	0.00001933***	0.00001864***
EMP	204.3***	0.000003588***	0.000003981***
AUT		1.220·	
RLOW	9242000*		
ELOW	2131000·		
ROAD		21.74·	26.26*
DIST	−97.69**	−0.00003684***	−0.00003926***
BR	2869000***	0.09062***	0.009683***
LN	1096000***	0.2235***	0.2423***
TRANS	4852000***	0.5866***	0.44***
TERM	1339000	0.6746***	0.6436***
Adjusted R^2	0.6592	0.6418	0.73836

Note: $\cdot p < 0.1$; * is $p < 0.05$; ** is $p < 0.01$; *** is $p < 0.001$.

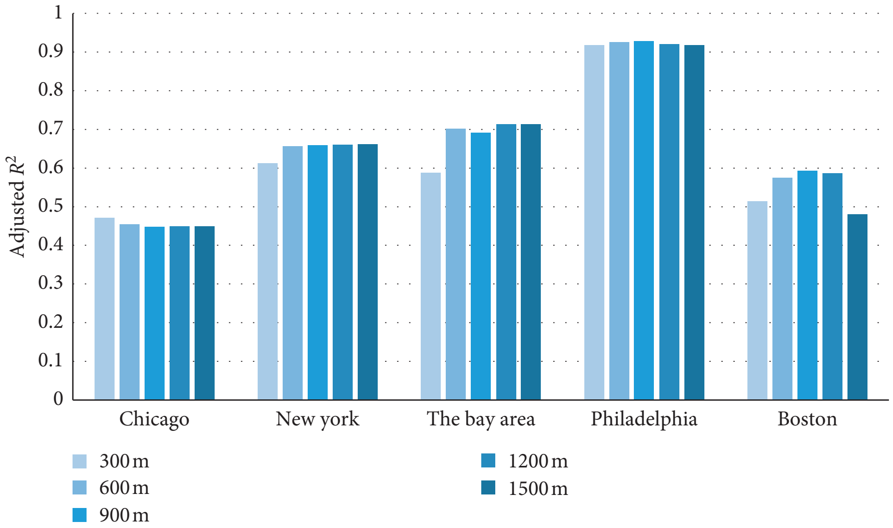


Figure 7: Linear regression result of five different catchment areas.

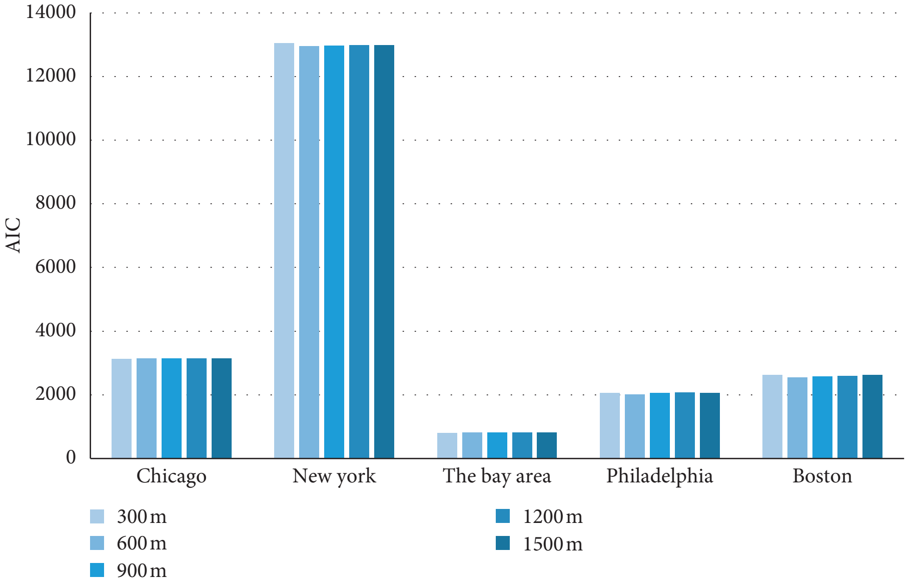


Figure 8: Negative binomial regression result of five different catchment areas.

Table 4: Overlap ratio of buffer area in each city.

Cities	Buffer radius (m)	Overlap ratio
Chicago	300	0.0713
	900	0.309
	1500	0.536
New York	300	0.538
	900	0.539
	1500	0.764
The Bay Area	300	0
	900	0.042
	1500	0.102
Philadelphia	300	0.0329
	900	0.105
	1500	0.273
Boston	300	0.236
	900	0.571
	1500	0.729

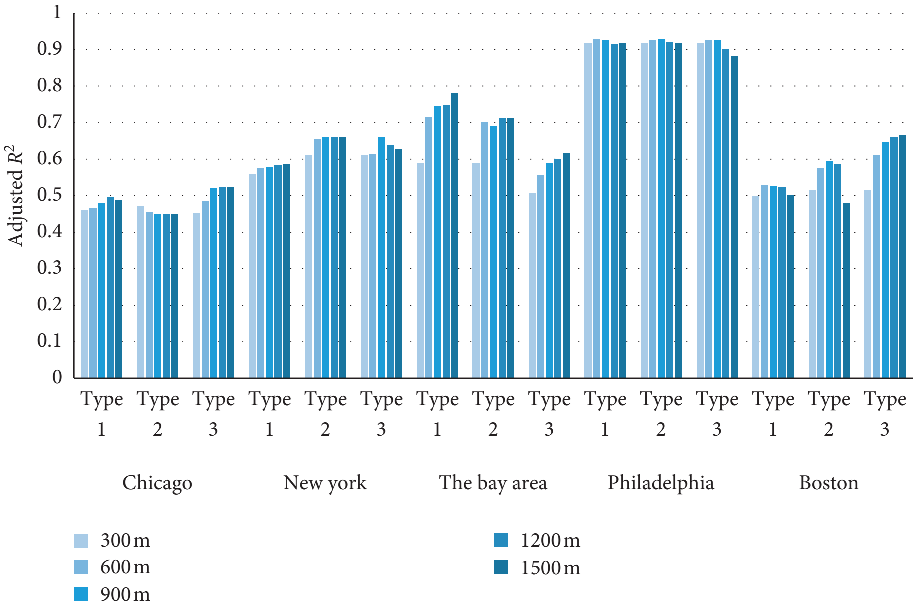


Figure 9: Linear regression result of three processing methods.

and San Francisco Bay Area have similar pattern: the negative binomial regression model has the best performance while the linear regression model is the worst. However, the differences in goodness-of-fit of the three models are not significant. The model results of Chicago and Boston have similar pattern: linear regression has the best performance while negative binomial regression is the worst. Again, the differences in goodness-of-fit of the three models are not significant. But for Philadelphia, the log-linear model has much worse performance than the other two models. This could be due to that in Philadelphia; the ridership of some stations is extremely high (above 20,000) while that of other stations is usually low (below 3,000), which could be observed from Figure 3.

The average values of MAPE for different buffer sizes are shown in Figure 13. We find that the MAPE of log-linear regression is smaller than that of the other two regression models. Therefore, the log-linear regression model could substantially improve the prediction accuracy.

In addition, the detailed results of different models and methods of the five cities are shown in the Tables 5–15. These tables present the same pattern and information.

6. Discussion

This article discusses a series of issues concerning the treatment of the catchment area when studying of the impact of the built environment on urban rail transit.

First of all, we studied the effect of buffer size on the station level demand modeling results. The results show that, overall, the optimal buffer size varies across the five cities. The impact of buffer size on the goodness-of-fit of the model is trivial, which is consistent with the conclusions of previous studies [15, 32]. This contrasts with the study of Jun et al. [21]

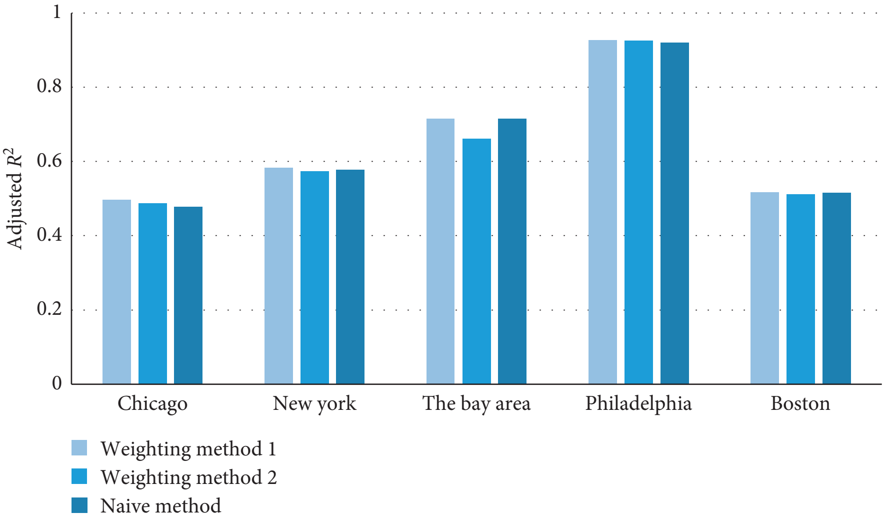


FIGURE 10: Linear regression result of two weighting methods and the naïve method.

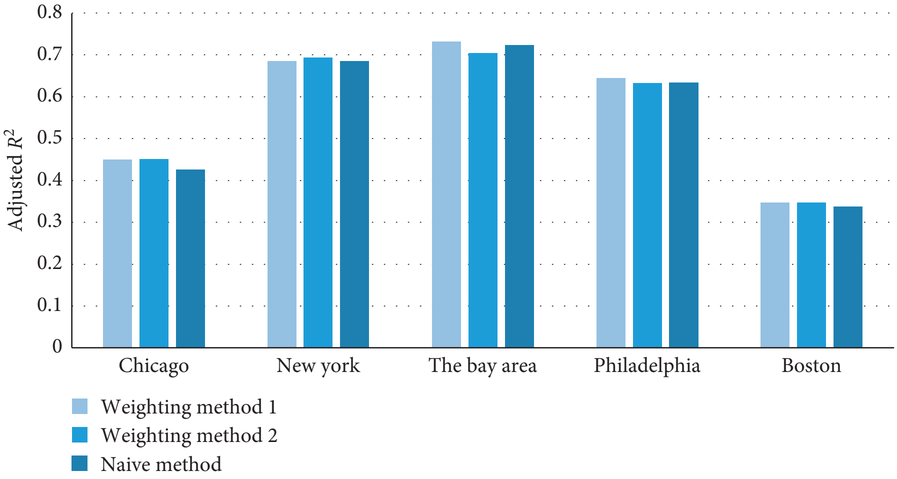


FIGURE 11: Negative binomial regression result of two weighting methods and the naïve method.

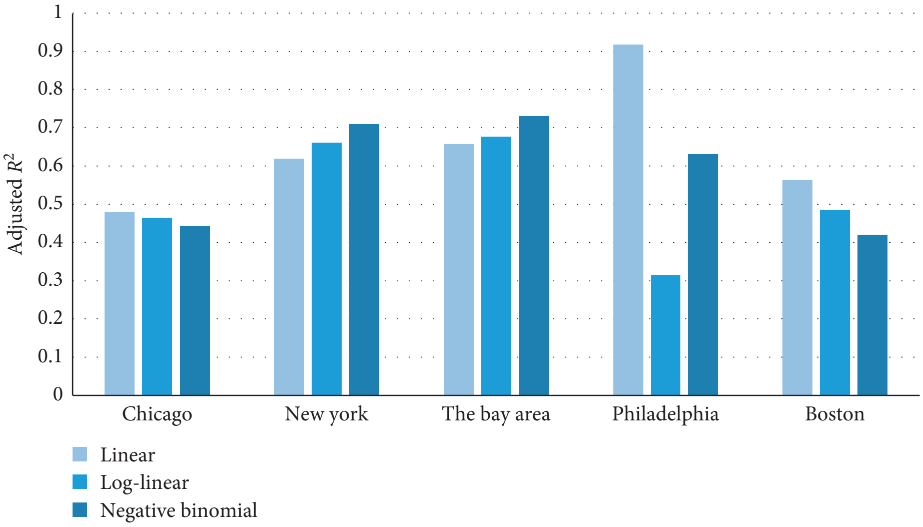


FIGURE 12: The result of three models.

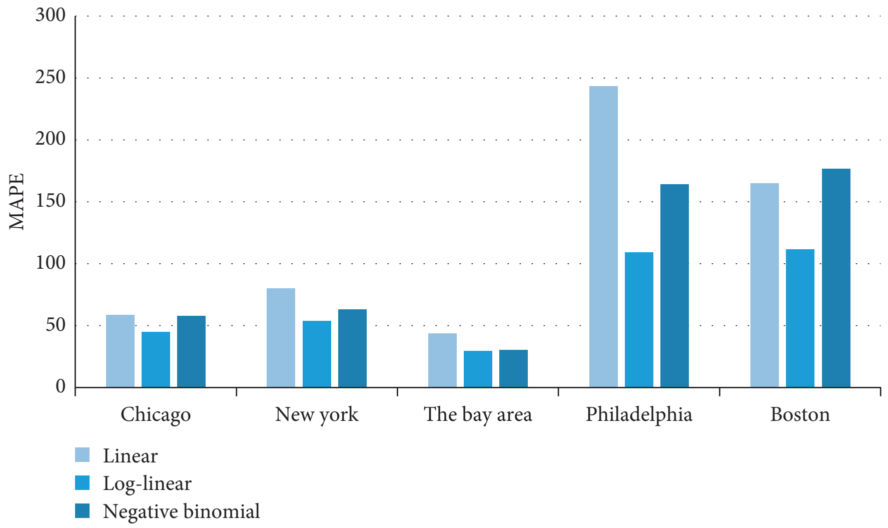


Figure 13: The result of three models.

Table 5: Table of the adjusted R^2 of linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station catchment area (m)				
		300	600	900	1200	1500
Chicago	Circular buffer	0.4593	0.4663	0.4807	0.4951	0.4865
	Thiessen polygon	0.472	0.4547	0.448	0.4493	0.4495
	Equal division	0.4518	0.4835	0.521	0.5234	0.5234
New York	Circular buffer	0.5598	0.5758	0.5776	0.5844	0.5862
	Thiessen polygon	0.6116	0.656	0.6592	0.6595	0.6612
	Equal division	0.6114	0.6128	0.6608	0.6391	0.6268
The Bay Area	Circular buffer	0.5881	0.7151	0.7442	0.7495	0.7812
	Thiessen polygon	0.5881	0.7018	0.6914	0.713	0.7135
	Equal division	0.5881	0.5553	0.5897	0.6	0.6174
Philadelphia	Circular buffer	0.9177	0.9291	0.9244	0.914	0.9172
	Thiessen polygon	0.9172	0.9264	0.9284	0.9206	0.9174
	Equal division	0.9174	0.9251	0.9257	0.8996	0.8816
Boston	Circular buffer	0.4986	0.5298	0.526	0.5238	0.5002
	Thiessen polygon	0.5146	0.5748	0.5936	0.5866	0.4804
	Equal division	0.5138	0.6112	0.6466	0.6613	0.665

Table 6: Table of the adjusted R^2 of log-linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station catchment area				
		300	600	900	1200	1500
Chicago	Circular buffer	0.444	0.4531	0.4463	0.4587	0.4525
	Thiessen polygon	0.4812	0.4495	0.4542	0.4612	0.461
	Equal division	0.444	0.4702	0.4816	0.5042	0.5106
New York	Circular buffer	0.633	0.6467	0.6418	0.6243	0.6222
	Thiessen polygon	0.6356	0.6874	0.6972	0.6913	0.685
	Equal division	0.6441	0.6459	0.6985	0.6868	0.6732
The Bay Area	Circular buffer	0.6181	0.6784	0.6744	0.6885	0.7148
	Thiessen polygon	0.6181	0.6816	0.6699	0.7157	0.7311
	Equal division	0.592	0.6497	0.6706	0.7057	0.7344
Philadelphia	Circular buffer	0.3028	0.3124	0.3189	0.3355	0.3484
	Thiessen polygon	0.3006	0.3173	0.2952	0.2989	0.3101
	Equal division	0.3013	0.3077	0.3027	0.3164	0.3316
Boston	Circular buffer	0.3914	0.4117	0.418	0.412	0.4037
	Thiessen polygon	0.4656	0.546	0.5555	0.5428	0.4637
	Equal division	0.4654	0.5534	0.5533	0.5387	0.5399

Table 7: Table of the adjusted R^2 of negative binomial regression model in five cities.

Cities	Overlapping area processing method	Different scale of station catchment area				
		300	600	900	1200	1500
Chicago	Circular buffer	0.4239	0.4357	0.4206	0.4258	0.4281
	Thiessen polygon	0.4675	0.4265	0.4106	0.4233	0.4421
	Equal division	0.4239	0.4563	0.4617	0.4851	0.5027
New York	Circular buffer	0.6842	0.6945	0.6919	0.6781	0.6787
	Thiessen polygon	0.6891	0.7297	0.7384	0.7358	0.7321
	Equal division	0.6958	0.6983	0.7402	0.7304	0.7205
The Bay Area	Circular buffer	0.6565	0.7318	0.7272	0.7425	0.7597
	Thiessen polygon	0.6565	0.7343	0.7336	0.7616	0.7761
	Equal division	0.6327	0.7224	0.7450	0.7702	0.7849
Philadelphia	Circular buffer	0.6249	0.6298	0.6371	0.6384	0.6424
	Thiessen polygon	0.6261	0.6343	0.6277	0.6270	0.6252
	Equal division	0.6250	0.6278	0.6324	0.6299	0.6282
Boston	Circular buffer	0.3297	0.3533	0.3423	0.3380	0.3232
	Thiessen polygon	0.3606	0.4971	0.5089	0.4868	0.3850
	Equal division	0.3698	0.5180	0.5031	0.4910	0.4902

Table 8: Table of the MAPE of linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station catchment area				
		300	600	900	1200	1500
Chicago	Circular buffer	59.3354	62.7474	60.3440	60.9660	61.0746
	Thiessen polygon	57.9248	57.4009	56.7382	57.2780	58.2991
	Equal division	59.3354	58.6506	55.6506	53.7948	56.3630
New York	Circular buffer	87.3059	84.7205	98.7732	82.9525	87.3304
	Thiessen polygon	80.1300	73.3991	72.7713	73.9423	73.6352
	Equal division	80.7813	78.9289	74.7230	74.8261	75.3502
The Bay Area	Circular buffer	49.1272	43.7829	41.9110	42.1634	40.9523
	Thiessen polygon	49.1272	43.6740	43.0613	41.3585	42.3117
	Equal division	46.5880	44.2976	44.2338	40.3771	43.8683
Philadelphia	Circular buffer	242.6527	241.0024	245.1476	253.3909	243.5712
	Thiessen polygon	240.9156	244.1579	234.3286	243.7437	252.7964
	Equal division	244.9216	236.2143	236.2295	242.2183	253.8006
Boston	Circular buffer	207.5354	210.7967	199.2151	192.1820	218.8120
	Thiessen polygon	176.2811	136.4003	123.7555	121.5683	185.1763
	Equal division	174.0910	138.2898	120.0253	132.2523	139.3389

Table 9: Table of the MAPE of log-linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station coverage area				
		300	600	900	1200	1500
Chicago	Circular buffer	49.3347	48.6815	49.1823	47.7469	48.5808
	Thiessen polygon	46.7290	48.3344	47.8253	47.5446	47.9743
	Equal division	49.3347	46.9591	46.9014	45.6059	45.7798
New York	Circular buffer	56.6489	54.9179	55.8463	58.1294	58.2618
	Thiessen polygon	56.6264	50.2914	49.2053	50.6390	51.9403
	Equal division	55.6161	55.2312	49.4805	51.4516	53.0430
The Bay Area	Circular buffer	33.5826	31.1306	29.6502	29.1024	28.5746
	Thiessen polygon	33.5826	30.9068	29.6526	26.2521	26.1755
	Equal division	34.5532	31.4279	28.8800	26.5521	24.5883
Philadelphia	Circular buffer	110.1204	110.6219	109.0140	105.1705	102.3903
	Thiessen polygon	110.3205	110.0651	116.4305	112.5083	111.0246
	Equal division	110.6485	110.6135	110.8519	104.2393	101.0912
Boston	Circular buffers	124.3840	124.6025	119.9752	121.5094	118.7263
	Thiessen polygons	107.4535	100.4450	97.1663	99.7657	112.3053
	Equal division	110.1274	101.2467	99.0088	137.9086	101.5126

Table 10: Table of the MAPE of negative binomial regression model in five cities.

Cities	Overlapping area processing method	Different scale of station coverage area				
		300	600	900	1200	1500
Chicago	Circular buffer	60.2395	60.0153	59.7718	59.1654	59.4408
	Thiessen polygon	55.8666	58.6037	58.8518	57.7574	56.6678
	Equal division	60.2395	56.8181	55.6457	54.9363	54.5868
New York	Circular buffer	67.2227	65.1532	65.8917	68.7627	68.9214
	Thiessen polygon	67.7463	59.1156	57.6264	58.5905	59.9640
	Equal division	65.3781	64.9620	57.4910	59.3433	61.4158
The Bay Area	Circular buffer	35.7457	30.6730	30.7131	28.9727	27.6419
	Thiessen polygon	35.7439	30.4688	30.2556	27.8836	26.9374
	Equal division	38.2694	32.0858	29.6762	27.4430	26.4143
Philadelphia	Circular buffer	166.1193	161.2752	164.6414	167.9242	157.7943
	Thiessen polygon	165.1357	155.3344	169.5743	170.6572	165.6670
	Equal division	167.0274	160.9760	163.5276	171.4166	157.6668
Boston	Circular buffer	216.3016	204.5145	201.4464	208.7779	210.0852
	Thiessen polygon	190.7709	151.9787	147.7061	150.9345	188.6276
	Equal division	188.4703	148.6984	147.1659	149.9097	149.0657

Table 11: Table of the AIC of linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station coverage area				
		300	600	900	1200	1500
Chicago	Circular buffer	3231.378	3228.641	3225.807	3221.918	3224.244
	Thiessen polygon	3227.165	3230.671	3231.407	3232.981	3232.914
	Equal division	3231.378	3224.135	3212.784	3210.474	3211.014
New York	Circular buffer	13856.92	13841.42	13872.66	13862.94	13862.1
	Thiessen polygon	13835.49	13754.58	13782.49	13781.24	13779.09
	Equal division	13802.9	13802.38	13778.17	13802.37	13816.25
The Bay Area	Circular buffer	850.5072	853.13	866.2529	864.4098	859.3719
	Thiessen polygon	850.5072	855.0916	873.5956	873.8877	873.8073
	Equal division	874.9724	871.3962	887.0369	886.8093	884.8475
Philadelphia	Circular buffer	2307.643	2300.317	2293.576	2328.19	2383.03
	Thiessen polygon	2323.598	2263.626	2302.602	2333.496	2323.215
	Equal division	2308.218	2308.205	2293.92	2349.508	2435.937
Boston	Circular buffer	2763.496	2758.059	2775.692	2757.695	2696.676
	Thiessen polygon	2758.772	2704.434	2733.374	2736.769	2769.475
	Equal division	2759.085	2729.015	2734.338	2729.185	2727.58

Table 12: Table of the AIC of log-linear regression model in five cities.

Cities	Overlapping area processing method	Different scale of station coverage area				
		300	600	900	1200	1500
Chicago	Circular buffer	250.6383	247.4418	248.1906	245.0762	246.6521
	Thiessen polygon	241.0942	246.4564	245.2638	242.5166	243.5422
	Equal division	250.6373	243.0476	238.1658	232.0186	231.1715
New York	Circular buffer	768.1533	752.213	759.7766	779.8623	782.1736
	Thiessen polygon	766.064	703.8898	692.1555	699.2817	707.8272
	Equal division	755.3079	754.2008	688.3685	704.3803	721.2337
The Bay Area	Circular buffer	52.89483	47.65492	48.10873	45.32375	40.58694
	Thiessen polygon	52.89254	47.23284	48.70302	44.42008	41.23757
	Equal division	55.33563	51.33887	49.41588	44.4601	41.42672
Philadelphia	Circular buffer	386.0091	385.7024	382.6974	380.8951	383.7017
	Thiessen polygon	388.2458	377.2787	389.3353	389.3171	385.3118
	Equal division	386.3242	386.6803	387.0131	386.8765	389.3943
Boston	Circular buffer	429.2843	424.4346	424.7866	424.0389	416.1078
	Thiessen polygon	410.9113	382.3009	384.0156	388.025	409.8976
	Equal division	410.7349	385.0343	387.6508	393.2093	393.7696

Table 13: Table of the AIC of negative binomial regression model in five cities.

Cities	Overlapping area processing method	Different scale of station coverage area				
		300	600	900	1200	1500
Chicago	Circular buffer	3146.3	3143.3	3147.1	3145.8	3145.2
	Thiessen polygon	3134.9	3145.7	3149.6	3146.4	3141.7
	Equal division	3146.3	3137.9	3136.5	3130.1	3125.1
New York	Circular buffer	13026	13012	13045	13065	13067
	Thiessen polygon	13049	12958	12974	12978	12984
	Equal division	13010	13006	12971	12987	13003
The Bay Area	Circular buffer	804.72	811.67	830.1	827.51	824.44
	Thiessen polygon	804.72	811.25	829.04	824.09	821.28
	Equal division	825.44	813.17	827.08	822.44	819.48
Philadelphia	Circular buffer	2048.6	2059.8	2043.1	2056.6	2110.2
	Thiessen polygon	2061.6	2018.7	2060.9	2075	2062
	Equal division	2048.6	2060.6	2045.1	2060.2	2116.4
Boston	Circular buffer	2635.4	2629.6	2650.1	2634.4	2572.2
	Thiessen polygon	2625.9	2550.3	2582.8	2589.7	2618.4
	Equal division	2625.5	2583.1	2605.3	2609.1	2609.3

Table 14: Table of the adjusted R^2 of negative binomial regression model of two weighting methods in five cities.

Cities	Two weighting methods	Different regression models		
		Linear	Log linear	Negative binomial
Chicago	Method 1	0.4968	0.4712	0.449121
	Method 2	0.488	0.4675	0.451145
New York	Method 1	0.583	0.6348	0.685609
	Method 2	0.573	0.6469	0.694079
The Bay Area	Method 1	0.715	0.681	0.732852
	Method 2	0.662	0.6637	0.703755
Philadelphia	Method 1	0.927	0.3461	0.644407
	Method 2	0.926	0.321	0.633125
Boston	Method 1	0.517	0.4208	0.347044
	Method 2	0.512	0.4111	0.347405

Table 15: Table of the MAPE of negative binomial regression model of two weighting methods in five cities.

Cities	Two weighting methods	Different regression models		
		Linear	Log linear	Negative binomial
Chicago	Method 1	59.8006	47.0242	57.8561
	Method 2	60.2017	47.7316	58.1439
New York	Method 1	84.5480	56.6961	66.8878
	Method 2	86.4682	54.9704	65.0345
The Bay Area	Method 1	42.0027	29.8307	30.3706
	Method 2	44.35519	32.54356	32.3889
Philadelphia	Method 1	244.0061	105.0997	156.9269
	Method 2	233.2926	108.5531	161.4774
Boston	Method 1	189.1325	119.4015	200.8356
	Method 2	200.0674	122.9177	206.1862

that recommends 600 meters as the radius of the pedestrian catchment area for a compact city like Seoul. We find that for a compact city, such as New York, the size of the buffer still does not have a significant impact on the model results.

Regarding the processing method for the overlapping buffer area, we find that, for cities with densely distributed urban rail transit stations, the equal division method and Thiessen polygon method can generate better results. There have not been any studies that apply the equal division method, but Thiessen polygons have been used by some researchers. For example, Li et al. [22] and Sun et al. [12] used Thiessen polygons to deal with the overlapping area of circular buffers. In the future, the equal division method could also be used to deal with the overlapping buffer area. For the distance decay weighting methods, the weighting method based on linear distance decay function is better than that based on nonlinear distance decay function, although the difference is marginal. The result of applying weighting to buffer bands is similar with that without weighting, which is contradictory to the conclusion of Gutiérrez et al. [13].

Regarding regression models, when the comparison is made based on R^2, the three models have similar performance. Only for the city of Philadelphia, the log linear model has a much lower R^2, which may be because the ridership of Philadelphia has much higher variation. When the comparison is made based on MAPE, the log linear model has the best performance for all the five cities. It indicates that log linear model has higher prediction power than the other two models and that we could obtain different results by using different measures. This also contrasts with the results of Wang et al. [42].

7. Conclusions

This study evaluates the effects of different buffer sizes, treatments of overlapping buffer area, and regression models on the direct ridership modeling results. To compare the performance of all the models and methods, we conducted extensive experiments using the data of five major cities in the U.S. First, the model results of different buffer sizes (300 meters, 600 meters, 900 meters, 1200 meters, and 1500 meters) are compared. We find that different buffer sizes do not have a great impact on models' goodness-of-fit and prediction accuracy. Secondly, we compared the model results of the three methods to deal with the overlapping of the catchment area, which are naïve method, Thiessen polygon method, and equal division method. The results show that, for cities with densely distributed stations and high buffer overlapping ratio, the equal division method is better than Thiessen polygon method, and both outperform the naïve method. However, for cities with more scattered stations and low buffer overlapping ratio, the three methods have comparable performance. Thirdly, we perform weighted regression by applying weights to the variables in the circular buffer band of within 300 meters, 300–600 meters, 600–900 meters, 900–1200 meters, and 1200–1500 meters using two simplified weighting methods. The results show that the weighting method of 0.5, 0.4, 0.3, 0.2, and 0.1 produces better results than the method of 0.8, 0.3, 0.1,0, and 0. However, the weighted regression does not improve the goodness-of-fit much. Finally, three regression models, linear regression, log-linear regression, and negative binomial regression were constructed and their results are compared. Based on the adjusted R-square and MAPE values of the models, we find that the goodness-of-fits of these three models are all satisfactory for all cities except for Philadelphia. The differences of values are within 20%, and the log-linear regression model results in high prediction accuracy.

There are also some limitations of this study. First, the range of the buffer size is between 300 meters and 1500 meters, which comes from the previous studies [13, 15, 21, 32]. The interval of 300 meters is used, which we believe should be small enough to study the optimal buffer size, especially considering the inaccuracy caused by extracting data from the CBG. But a smaller interval, such as 100 meters used by Gutiérrez et al. [13], could generate more detailed results. This is one of the limitations of this study and could be further explored in the future. Secondly, only the linear relationship between ridership and built environment variables is considered. In the future, we will consider the nonlinear relationship between the independent variables and the dependent variable and use machine learning tools to build nonlinear models.

Data Availability

The data used to support the findings of the study are from the New York City Transit Authority (MTA), the Chicago Transit Authority (CTA), the Massachusetts Bay Transportation Authority (MBTA), the Southeastern Pennsylvania Transportation Authority (SEPTA), the Port Authority Transit Corporation (PATCO), and the Bay Area Rapid Transit (BART), respectively.

Conflicts of Interest

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

Acknowledgments

This study was supported by National Natural Science Foundation of China (grant nos. 71704145 and 51774241) and China Postdoctoral Science Foundation, Sichuan Youth Science and Technology Innovation Research Team Project (grant nos. 2019JDTD0002 and 2020JDTD0027).

References

[1] 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.

[2] J. Zhao, W. Deng, Y. Song, and Y. Zhu, "What influences Metro station ridership in China? Insights from Nanjing," *Cities*, vol. 35, pp. 114–124, 2013.

[3] H. Iseki, C. Liu, and G. Knaap, "The determinants of travel demand between rail stations: a direct transit demand model using multilevel analysis for the Washington D.C. Metrorail system," *Transportation Research Part A: Policy and Practice*, vol. 116, pp. 635–649, 2018.

[4] M. Kuby, A. Barranda, and C. Upchurch, "Factors influencing light-rail station boardings in the United States," *Transportation Research Part A: Policy and Practice*, vol. 38, no. 3, pp. 223–247, 2004.

[5] R. Cervero, "Alternative approaches to modeling the travel-demand impacts of smart growth," *Journal of the American Planning Association*, vol. 72, no. 3, pp. 285–295, 2006.

[6] R. Cervero, "Transit-oriented development's ridership bonus: a product of self-selection and public policies," *Environment and Planning A: Economy and Space*, vol. 39, no. 9, pp. 2068–2085, 2007.

[7] 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.

[8] 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.

[9] P. Li, P. Zhao, and T. Schwanen, "Effect of land use on shopping trips in station areas: examining sensitivity to scale," *Transportation Research Part A: Policy and Practice*, vol. 132, pp. 969–985, 2020.

[10] 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.

[11] 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.

[12] 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. 1-8, 2016.

[13] J. Gutiérrez, O. D. Cardozo, and J. C. García-Palomares, "Transit ridership forecasting at station level: an approach based on distance-decay weighted regression," *Journal of Transport Geography*, vol. 19, no. 6, pp. 1081–1092, 2011.

[14] O. Manout, P. Bonnel, L. Bouzouina, and Practice, "Transit accessibility: a new definition of transit connectors," *Transportation Research Part A: Policy and Practice*, vol. 113, pp. 88–100, 2018.

[15] E. Guerra, R. Cervero, and D. Tischler, "Half-mile circle," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2276, no. 1, pp. 101–109, 2012.

[16] 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.

[17] 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.

[18] 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, 2019.

[19] 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.

[20] X. Chu, *Ridership Models at the Stop Level*, 2004.

[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] S. LiD. Lyu et al., "Spatially varying impacts of built environment factors on rail transit ridership at station level: a case study in Guangzhou, China," *Journal of Transport Geography*, vol. 82, p. 2020.

[23] R. Cervero and K. Kockelman, "Travel demand and the 3Ds: density, diversity, and design," *Transportation Research Part D: Transport and Environment*, vol. 2, no. 3, pp. 199–219, 1997.

[24] L. F. Chow, F. Zhao, M. T. Li, and S.-C. Li, "A transit ridership model based on geographically weighted regression and service quality variables," *Transportation Research Record*, no. 1, pp. 105–114, 1972.

[25] C. Lee and A. V. Moudon, "Correlates of walking for transportation or recreation purposes," *Journal of Physical Activity and Health*, vol. 3, no. s1, pp. S77–S98, 2006.

[26] L. Yang, Y. Ao, J. Ke, Y. Lu, and Y. Liang, "To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults," *Journal of Transport Geography*, vol. 94, Article ID 103099, 2021.

[27] B. D. Taylor, D. Miller, H. Iseki, and C. Fink, "Analyzing the determinants of transit ridership using a two-stage least squares regression on a national sample of urbanized areas," *Transportation Research Board 82th Annual Meeting*, 2003.

[28] 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.

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

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

[31] M. Ruiz-Pérez and J. M. Seguí-Pons, "Bus service level and horizontal equity analysis in the context of the modifiable areal unit problem," *ISPRS International Journal of Geo-Information*, vol. 10, no. 3, p. 111, 2021.

[32] R. Mitra and R. N. Buliung, "Built environment correlates of active school transportation: neighborhood and the modifiable areal unit problem," *Journal of Transport Geography*, vol. 20, no. 1, pp. 51–61, 2012.

[33] 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.

[34] 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.

[35] H. Yang, X. Lu, C. Cherry, X. Liu, and Y. Li, "Spatial variations in active mode trip volume at intersections: a local analysis utilizing geographically weighted regression," *Journal of Transport Geography*, vol. 64, pp. 184–194, 2017.

[36] M. Corazza and N. Favaretto, "A methodology to evaluate accessibility to bus stops as a contribution to improve sustainability in urban mobility," *Sustainability*, vol. 11, no. 3, p. 803, 2019.

[37] R. K. Untermann, "Accommodating the pedestrian: adapting towns and neighbourhoods for walking and bicycling," *Transportation Research Board 82th Annual Meeting*, 1984.

[38] R. Guo and Z. Huang, "Mass rapid transit ridership forecast based on direct ridership models: a case study in Wuhan, China," *Journal of Advanced Transportation*, vol. 2020, Article ID 7538508, 19 pages, 2020.

[39] O. Manout, P. Bonnel, and L. Bouzouina, "Transit accessibility: a new definition of transit connectors," *Transportation Research Part A: Policy and Practice*, vol. 113, pp. 88–100, 2018.

[40] L. Yang, K. W. Chau, W. Y. Szeto, X. Cui, and X. Wang, "Accessibility to transit, by transit, and property prices: spatially varying relationships," *Transportation Research Part*

D: Transport and Environment, vol. 85, Article ID 102387, 2020.

[41] L. Yang, X. Chu, Z. Gou, H. Yang, Y. Lu, and W. Huang, "Accessibility and proximity effects of bus rapid transit on housing prices: heterogeneity across price quantiles and space," *Journal of Transport Geography*, vol. 88, Article ID 102850, 2020.

[42] X. Wang, G. Lindsey, S. Hankey, K. Hoff, and Development, "Estimating mixed-mode urban trail traffic using negative binomial regression models," *Journal of Urban Planning and Development*, vol. 140, no. 1, Article ID 04013006, 2014.