Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 7538508, 19 pages
https://doi.org/10.1155/2020/7538508

Research Article

Mass Rapid Transit Ridership Forecast Based on Direct Ridership Models: A Case Study in Wuhan, China

Ruili Guo [1] **and Zhengdong Huang** [1,2]

[1]*School of Urban Design, Wuhan University, Wuhan 430072, China*
[2]*Research Institute for Smart Cities, Shenzhen University, Shenzhen 518060, China*

Correspondence should be addressed to Zhengdong Huang; huangitc@126.com

Received 15 August 2019; Revised 14 February 2020; Accepted 19 February 2020; Published 21 March 2020

Academic Editor: Luca D'Acierno

Many large cities rely on Mass Rapid Transit (MRT) to increase passenger mobility. For efficiency, MRT stations should be arranged to attract maximal number of travelers. It is therefore important to develop methods for estimating MRT ridership forecasting models, which are important for policies on land use development or new MRT lines. Direct ridership models (DRMs) at the station level are superior in estimating the benefits of transit-oriented development policies. In this paper, a principal component regression (PCR) is proposed to overcome the issue of multicollinearity that widely occurs in multivariate regression analyses for DRM modeling, especially the ordinary least squares regression. Based on the analysis of 72 MRT stations in Wuhan, China, four principal components are obtained to explain the potential linkage to MRT ridership, which include built-environment related factors, jobs-housing spatial structure related factors, station attributes, and the large compound. Nineteen significant determinants have been identified, among which the four factors of office building area, land use mix, the number of restaurants, and financial institutions are the most influential factors. Built-environment-related factors exert more significant impact on MRT ridership than others. The distance to city center and the number of bus lines around stations have negative association with MRT demand. The proposed PCR-based DRM provides insights for forecasting transit demand brought about by new metro lines and forecasting the consequences of land use development.

1. Introduction

Chinese planners and government agents are actually aware that Mass Rapid Transit (MRT) occupies an important place in urban transport systems. MRT has valuable benefits for densely populated and economically developed cities, such as high capacity, low carbon emissions, and on-time performance. MRT, especially metro transit, has become increasingly popular in China. One subway line, especially in Wuhan, was open every year from 2012 to 2015 (Figure 1). And 14 subway lines were being built at the same time in 2016. By the end of 2021, the total length of rail transit lines in Wuhan will reach 400 km. The introduction of MRT attracts millions of commuters to ride on. The metro transit is responsible for up to 25% trips of public transit in Wuhan, and then the ratio kept increasing to 35% in July 2017, while it was 53% in Shanghai at the same time. MRT plays an important role in shaping travel modal structure and encouraging land development around stations. It is regarded as a major solution to transport problems in megacities.

The issue of urban public transport is concerned with a balance between transit supply and demand, both with complex influencing factors. From the supply side, there are such factors as transit infrastructure, fleet size, timetable, reliability or robustness, and quality of service [1]. These characteristics determine the service capacity to the public. In order to cope with the demand, a capacity optimization and allocation model has been built, with constraints of passenger flow, headway, load factor, and available trains [2]. Many metro systems in megacities stimulate transit travel and may reach its design capacity in a short period upon deployment. Increasing service capacity, mainly by

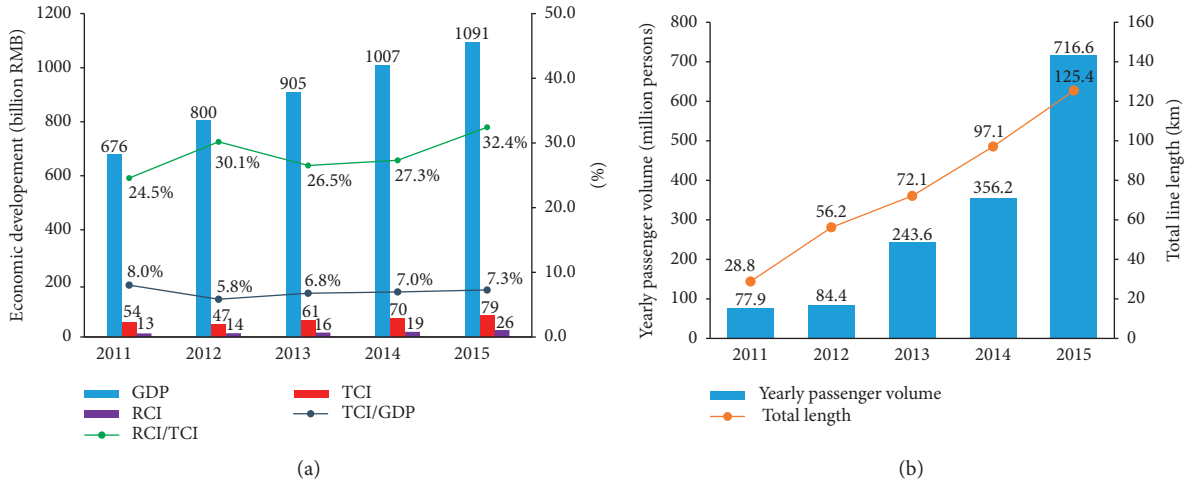

FIGURE 1: Rail transit construction investment (a) and total mileage (b) in Wuhan. GDP, TCI, and RCI mean the gross domestic product, traffic construction investment, and rail transit construction investment in the city.

increasing service frequency, is a possible solution to growing demand. However, higher frequency operations are constrained by many factors, including signaling and train control, station and train crowding, fleet, terminal turnarounds, and service complexity [3]. In many cases, timetabling is to be systematically designed for time-varying passenger demand at congested Stations [4]. The quality of transit service has great influence on transit demand and needs to be considered in demand modeling.

MRT demand models must be established to estimate transit patronage for the development of new MRT lines and adjustments to the overall land use arrangement. The methods of land use planning and development for areas around MRT lines and stations represent a hot spot in global research [5–7]. Many researchers have promoted the smart growth of cities via the construction of MRT systems oriented with land use development (TOD policies). The "3D" principal (i.e., density, diversity, and design) proposed by Cervero and Kockelman [8] represents the basic consideration underlying analyses of the relationships between land use and transit demand. The traditional zone-based travel demand model, which is known as the four-step model, was first developed in the late 1950s as a tool for highway planning. In the four-step model system, the study area is divided into traffic analysis zones (TAZs) that range in size from street blocks to census tracks. In many cases models based on TAZs are too gross to estimate the travel impacts of neighborhood-level development and trips generated by self-selection activity around transit stations [9]. Secondly, the resolution of these models often guides regional highway and transit investment. Lastly, the four-step models lead to differences in the transit ridership with different distances to MRT stations [10].

The Direct Ridership Model (DRM), which is different from the traditional zone-based model, estimates ridership of specific station or stop rather than a line or corridor. DRM is built upon station environments and transit services [11, 12]. These alternative models predict the demands of specific stations and usually have smaller samples because the observations are often from MRT stations. Therefore, the DRM is a precise, quick-response and economical method for estimating potential ridership in the sketch-planning phase. Nevertheless, DRMs are not substitutes for traditional zone-based travel models [13].

In this context, this paper aims to set up a framework for forecasting MRT ridership to explore the casual relationship between the determinants of station environments and transit ridership. First, a review of DRMs is carried out, and factors influencing MRT ridership are examined. Second, an analyzing framework for DRM is presented with a list of candidate variables. Third, a case study in Wuhan, China, is described. Thirty-one factors with potential influence on MRT ridership have been collected and managed in GIS database. Principal component regression (PCR) is applied to reduce dimension of explanatory variables and estimate station-level transit ridership. The model is evaluated in this section, followed by sections of discussion and conclusions.

2. Direct Ridership Forecasting

2.1. Direct Ridership Models. DRMs can be divided into two categories: traditional econometric models and spatial econometric models (Table 1). The ordinary least squares (OLS) regression method, as a basic approach of traditional econometric models, has been widely applied in recent studies. However, the OLS regression model cannot account for collinearity and may result in certain issues, such as independent variables with the wrong sign and certain vital variables with statistically insignificant results. Therefore, other modified models have been developed, such as variable subset selection (VSS), ridge regression (RR), partial least squares regression (PLSR), and PCR. VSS was realized by OLS, although the variable selection procedure should be conducted with empirical evidence and professional knowledge in advance. RR may miss certain vital predictors required for better results, e.g., dummy variables and other continuous variables [28]. Chu [19] obtained more accurate results using the Poisson regression model than those using the linear regression model. Chakour and Eluru [20] proposed a composite marginal likelihood (CML) method with

Table 1: Summary of literature on DRMs.

Authors	Models	Software	Pros and cons
Traditional econometric models			
Kuby et al. (2004) [14]	OLS/heteroskedasticity consistent covariances method (HCCM)	LIMDEP	HCCM assumes that the residual error is irrelevant; however, this assumption is not universal
Cervero (2006) [13]	OLS	—	The log-log form of OLS is applied to describe suitable relations based on the data; however, the multicollinearity is not delivered
Usvyat et al. (2009) [15]	OLS	GIS/ Microsoft excel	The intercept is negative, which does not conform to the actual conditions; additionally, the issue of collinearity has not been addressed
Sohn and Shim (2010) [16]	OLS/structural equation model (SEM))	AMOS	The SEM method needs a large data sample for support; in addition, the selection of latent variables depends on technical and empirical knowledge
Zhao et al (2013) [17]	OLS	SAS	The signs of the four variable coefficients in the final DRM are negative and incorrect, which indicates that multicollinearity is not considered in the analysis
Ramos-Santiago and Brown (2016) [18]	Negative binomial regression (NBR)	Stata	The application condition of NBR is that the mean value is equal to the variance; the data sample should be large
Chu (2004) [19]	Poisson regression (PR)	Stata	The application condition of PR is that the predicted variance should be larger than the mean value
Chakour and Eluru (2016) [20]	Ordered response probit (ORP) model	Matlab	The parameters should be calibrated by the method of composite marginal likelihood (CML) in this paper
Spatial econometric models			
Cardozo et al. (2012) [10]	GWR	GIS	The model could explain the diversity of results for spatial factors; however, it needs a large data sample
Pulugurtha and Agurla (2012) [21]	Spatial proximity method (SPM)/ spatial weighted method (SWM)	SPSS/GIS	The buffer of a stop has been divided into four bandwidths, and the best catchment can be identified based on SPM, but the weight function of SWM model (1/D2) is not continuous because of the defined set of bandwidths
Sung et al. (2014) [22]	Spatial error model (SEM)/spatial lag model (SLM)	GeoDa	This model could describe the relationships between the spatial factor and ridership; however, the definition of the spatial connection matrix is sensitive to the result
Jun et al. (2015) [23]	Multinomial logit model (MNL)/ OLS/MGWR (mixed GWR)	GWR4	The problem of autocorrelation has been addressed; however, the application of GWR needs a larger data sample for support
Ma et al. (2018) [24]	Geographically and temporally weighted regression (GTWR)	—	Explanatory variables are eliminated with the index of Pearson correlation lager than 0.6; however, the multicollinearity may exist among the variables and should be tested with variance inflation factor (VIF) index.
He et al. (2019) [25]	GWR	GWR4	The explanatory variables are selected from 14 factors; the coefficients of population and official land use that resulted from GWR have negative sign in some areas around stations
Zhu et al. (2019) [26]	Bayesian negative binomial regression model/GWR	SPSS/GIS	Station ridership determinants are obtained from the method of Bayesian negative binomial regression; the multicollinearity could be addressed as well
Tang et al. (2019) [27]	GWR/generalized linear model (GLM)	—	The multicollinearity is tested with a Pearson correlation coefficient (PCC) greater than 0.7 and a VIF greater than 7.5

4 periods in a day. Additionally, the geographically weighted regression (GWR) was constructed, and it references a family of "spatially adjusted" regressions to solve the problem of spatial autocorrelation, which is common in spatial data [10]. Chow et al. [29] developed GWR models in the level of region with 2000 TAZs, and two subregions with 555 and 215 TAZs for Broward Country, Florida. Chow found that GWR models had better performance than the OLS method, and subregional GWR models had better fit than regional GWR model. Yu et al. [30] analyzed the spatial patterns of industrial agglomeration using the local indicator of Spatial Association Statistics and proved industrial agglomeration has close relationship with rail transit access but varies across industrial sectors and different distance ties from Dart stations. Nevertheless, the performance of GWR depends heavily on the selected scale, which is subjective. Meanwhile, the data sample size of GWR method should be sufficiently large (usually more than 100 independent samples) because a small-sized data sample is not appropriate for this approach.

In addition to the abovementioned models, transit demand is also estimated with other methods. Huang et al. [31]

presented a GIS-based weighted accessibility approach for estimating light rail transit peak-hour boarding, considering both the potential travel demand around a station and the attractiveness of target stations. A natural line of thought for ridership estimation is to make use of such data as smart card, automatic fare collection (AFC), and point of interest (POI). Rail transit smart card data are collected with precise location (origin and destination) and swiping time. There are two approaches: one is to forecast weekday demand based on the same time slot recorded in previous weeks; the other is to forecast demand using continuous time series data prior to the forecasted time slot [32]. AFC data may serve as a proxy for land use type and help to build station-level ridership growth model [33]. A univariate state-space model with AFC data clustering has also been developed to estimate short-term ridership at the station level [34]. POI data may help identifying characteristics of built environment so as to implement ridership forecast [35]. With availability of temporal transit data, intelligence-related technologies have also been explored. Ivanov and Osetrov [36] built a passenger flow forecasting model based on artificial neural networks (ANNs). Combing with seven different statistic methods, the model has achieved reasonable accuracy. Combing ANNs with other intelligent methods (such as genetic algorithm) has also proved promising [37, 38]. Many of the above approaches may effectively respond to short-term forecasting requirements. Some indicators may also be incorporated into regression models for exploring the causal effect between built environment or land use development and MRT ridership.

2.2. Factors Influencing MRT Ridership

2.2.1. Definition of Pedestrian Catchment Area. According to the description by Cervero [13], station-level transit ridership was a function of the station environment and transit service; thus, defining the catchment area of the station was critical. According to the surveys of traffic modes connecting to metro stations conducted by transport department in China, more than 60% of passengers accessed the metro stations by walking, and the percent was 80% in Madrid [39]. Walking was the main mode for the population around stations to access metro network. Therefore, the catchment area was always known as the pedestrian catchment area (PCA) and usually determined by the maximum walking distance or the area within which major users arrived by foot [17]. However, the value of the walking distance varied according to different studies, e.g., 300 m [23], 400 m [13, 14], 500 m [16], and 800 m [10, 17].

In general, most variables assumed to affect ridership were calculated within the PCA. Land use had strong effects on ridership at the station level, and it required accurate estimations of the mutually exclusive PCA of each station. Therefore, the relevant distance should be the shortest network path rather than the Euclidean distance [10, 14].

2.2.2. Influencing Factors. The factors influencing ridership on public transit are critical for scholars, policy makers, engineers, and planners and represent a research topic in the field of transportation and urban planning. Chu [19] addressed six categories of factors to estimate ridership: social-demographics in the catchment area around the station, service supply, street environment, accessibility, interaction with other modes, and competition for other stations. Cervero [13] illustrated the factors as the built environment and transit service within the immediate areas of prospective stations. Gutiérrez et al. [40] built a ridership forecasting model based on two types of variables: service area characteristics and station characteristics. Considering these factors described in the previous literature, the influencing factors can be divided into 4 categories: (1) built environment, (2) external connectivity, (3) intermodal connections, and (4) transit characteristics of stations.

(1) Built environment. Many studies have examined the effect of the built environment, such as land use variables, on transit ridership [14–17, 19, 23, 40]. Jun et al. [23] found that the density of the population and employment around stations showed a gradient descent, which means that a higher density of population and jobs were closer to stations. However, extremely precise geographic information could not be acquired, and data were always aggregated at a certain scale. The land use type and diversity also have a significant effect on metro use. Generally, residential, office, and commercial land use has a credible influence on ridership [16, 17, 22, 23].

(2) External connectivity. Studies have used the road length, the distance to centers, and centrality as predictors for describing the interconnection between stations and the external network [8, 13, 16, 17]. Cervero [13] showed that the station distance to CBD was negative with transit ridership in Charlotte and significant at the 0.001 level. Sohn and Shim [16] designed three types of centrality indices, closeness, betweenness, and straightness centralities, although the three variables were found to be insignificantly associated with transit ridership. In addition, Sohn and Shim [16] assumed that the automobile-dominated road length had a negative influence on station-level ridership.

(3) Intermodal connections. With the development of multimodal transit, travel demand forecasts are needed to consider the impacts of other trip choices. Many researchers have introduced variables, e.g., bus services, Park and Ride spaces (P&R), and the walking cost around stations [8, 13–19, 40, 41]. These previous studies showed that the number of bus lines had a significant impact on transit patronage [8, 16]. Few studies have focused on the walking cost to a station except for Cervero and Chu [13, 19] because the PCA expresses the maximum walking distance. Zhao et al. [17] chose bicycle P&R space rather than auto P&R space in the Chinese context.

(4) Transit characteristics of stations. The characteristics of stations include dummy variables for the terminal stations, interchange stations, intermodal stations, and station spacing [8, 13, 17, 18, 40]. The CBD dummy variable has

been frequently discussed in association with DRMs. Usvyat et al. [15] stated that employment and land use were quite distinct based on the location. Therefore, stations were divided into CBD and non-CBD stations, and they were used to test the DRMs accordingly. Kuby et al. [14] discussed the influencing factors from the perspective of LRT and underground stations. The result was that the CBD dummy variable was insignificant in LRT stations, whereas the opposite was true in underground stations. Zhao et al. [17] also chose the CBD dummy variable to illustrate the powerful relationship with transit ridership, whereas Sohn and Shim [16] did not consider this variable. Chu [19] noted that the distance between adjacent stations should also be tested because a competitive ridership relationship between neighboring stations might occur.

2.3. Challenges of DRM Forecasting. In sum, a number of earlier studies have explored the causal model between the station environment and transit ridership. However, certain limitations are observed with DRMs. First, a variety of DRM expressions have been developed by traditional econometric models and spatial econometric models. OLS has been widely used but is also criticized as not being able to account for collinearity [13–15, 17]. Spatial econometric models have been used in recent years, although they are limited by the definition of scale and requirement of large sample data [10, 22, 23]. Although PCRs have the advantage of nonparameter estimation and can solve the problem of multicollinearity, they have rarely been applied in MRT demand forecast. Second, the cited studies are limited to available data. Although population and employment are contained in these data, aggregation of the data occurs at a certain scale. Scholars often managed the issue of aggregating data by including the assumption that population and employment are uniformly distributed [16, 17] or using the areal interpolation method [10]. Additionally, this treatment will reduce the effects of DRM. The land use statistics are calculated by the floor area, whereas the development intensity is neglected. Although the built environment has been actively discussed, traffic attraction spots are of less concern. Accessibility is introduced as a job-over-time or simply as standardized travel time in calibrating DRMs [14, 40]. However, the attractiveness of working people is omitted from the MRT station ridership. Third, the PCA statistical cell is usually delimited by GIS via a buffer analysis based on a straight-line distance, which generates a nonexclusive or exclusive circular area [13, 16, 17, 19]. With the buffer approach, barriers (buildings, water, etc.) that are not being crossable by travelers are ignored in the process of generating PCAs. As an amendment some studies have considered the road net and delineated the PCA based on network distance [14, 16, 18].

This study details thirty-one of the influencing factors categorized according to the built environment, external connectivity, intermodal connection, and transit characteristics. These data are expected to achieve four main goals. The first is to establish a GIS database based on a road network, transit network, buildings, and land use types. Considering the influence of barriers, the PCA is delimited by the cost distance tool in the GIS. The influencing factors are calculated within the PCA from the 3D perspective. The second goal is to verify the hypothesis that influencing factors have a potential influence on station-level ridership. Multicollinearity is diagnosed with the Pearson correlation coefficient index for all pairwise combinations among the independent variable set. The third goal is to explore the feasibility of modeling station-level transit demand via DRM based on a PCR in terms of method improvement. The principal components are expected to explain the generation mechanism of MRT demand. The last goal is to provide meaningful recommendation for MRT planners and policymakers based on the causal relation of MRT demand and land use development.

3. Methodology

3.1. Analysis Framework. A conceptual framework is conceived based on ArcGIS software built by Environmental Systems Research Institute (ESRI), based in Redlands, California, as the market leader in mapping software and findings from previous studies (Figure 2). The framework can be divided into three parts. First, the GIS database is established, and factor statistics are calculated for the PCA. The primary data types include a road network, transit network (metro and bus), land use distribution, buildings, population and employment, and ridership at MRT stations. Second, the correlation test between candidate factors and ridership is used to identify the independent variable set for DRM. Third, the PCR method is applied to generate several principal components. These principal components are cited from a professional viewpoint, which may reveal potential influencing links between explanatory variables and MRT ridership. With this framework, transit demand at the station level is estimated by the PCR-based DRM.

3.2. PCR-Based DRM. The essence of PCR is a regression analysis that uses a principal component analysis to identify possible components relating the influencing factors to MRT ridership. PCR is a procedure to overcome the problems of multicollinearity when explanatory variables are close to being collinear. The most striking characteristic of PCR is the nonparametric analysis. The method of dimensional reduction decreases a complex data set to a low dimension to reveal the hidden and simplified structure. PCR is a mainstay of modern data analyses, and it is widely used in near-infrared diffuse reflectance spectroscopy, artificial neural networks, and so on [42]. However, PCR is rarely used as an analysis method in DRMs. Therefore, the feasibility of using this procedure must be verified and a possible alternative for predicting DRMs with a small data sample must be provided.

The outline of the mathematics underlying the PCR method is illustrated, and it shows that PCR can overcome the disturbance of multicollinearity via orthonormality. A typical multiple regression equation is as follows:

$$Y = \mathrm{XP} + \varepsilon, \tag{1}$$

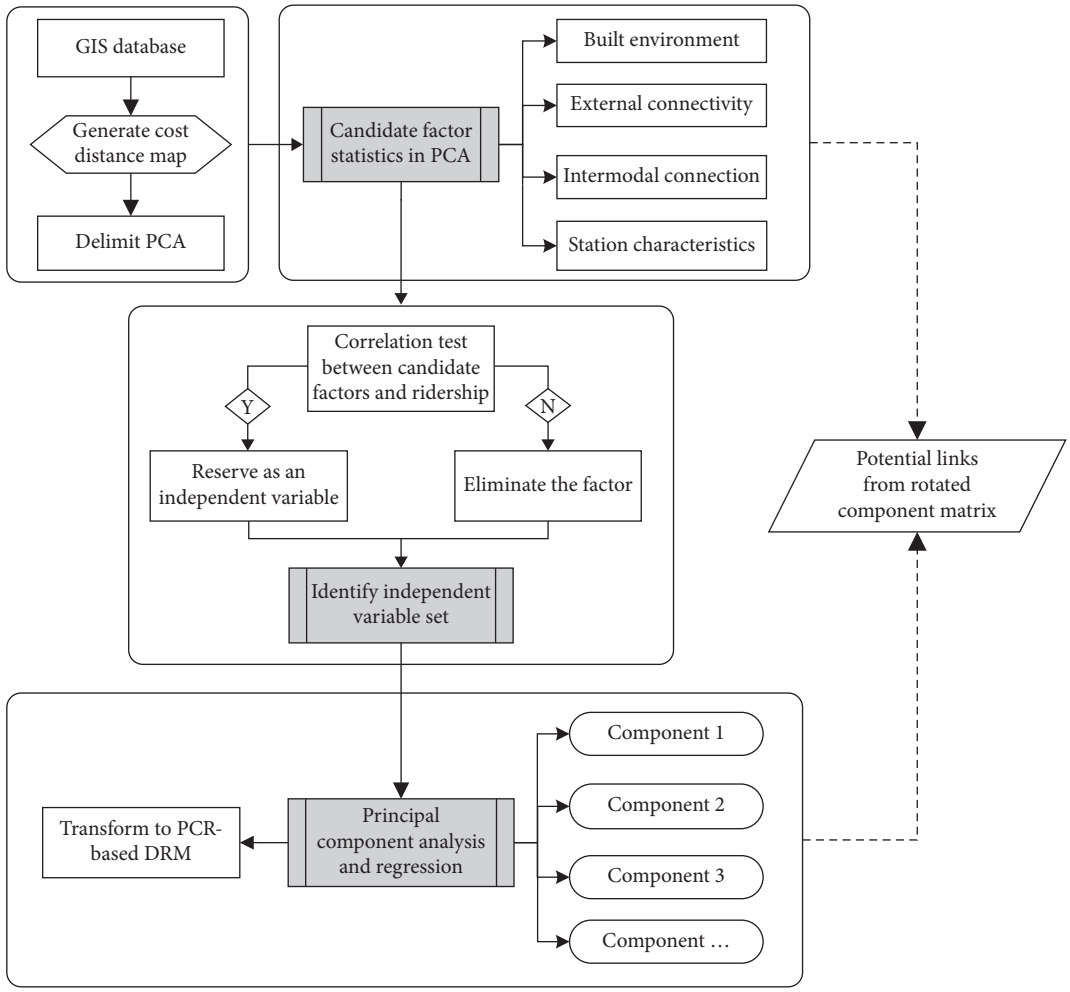

Figure 2: Framework of the PCR-based DRM identification.

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}_{m\times 1},$$
$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix},$$
$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}_{n\times 1}, \quad (2)$$
$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1},$$

where X represents the original influencing factors that have an obvious relationship with MRT ridership at the station level, P is the column vector of the coefficients, ε is the disturbance vector, which is normally distributed, and Y is the ridership at MRT stations.

The eigenvalues and eigenvectors of the covariance matrix are calculated via the Jacobi iteration method. All the eigenvectors create matrix U, which is orthogonal and normal with $UU^T = I$ (identity matrix):

$$U = \begin{bmatrix} \mu_1^{(1)} & \cdots & \mu_n^{(1)} \\ \vdots & \ddots & \vdots \\ \mu_n^{(n)} & \cdots & \mu_n^{(1)} \end{bmatrix}_{n\times n}. \quad (3)$$

The component c_i can be obtained with the algorithm $C = U\tilde{X}$. The matrix C retains the same statistical information as the standardized X matrix:

$$\begin{aligned} c_1 &= \mu_1^{(1)}\tilde{x}_1 + \mu_2^{(1)}\tilde{x}_2 + \cdots + \mu_n^{(1)}\tilde{x}_n, \\ c_2 &= \mu_1^{(2)}\tilde{x}_1 + \mu_2^{(2)}\tilde{x}_2 + \cdots + \mu_n^{(2)}\tilde{x}_n, \\ c_n &= \mu_1^{(n)}\tilde{x}_1 + \mu_2^{(n)}\tilde{x}_2 + \cdots + \mu_n^{(n)}\tilde{x}_n, \end{aligned} \quad (4)$$

where $\tilde{x}_i$ is the standardized variable and c_i represents the principal components, which are arranged in descending order according to the contribution to variance and decorrelation. The previous k components will remain when the accumulated variance contribution ratio is large enough, e.g., more than 80%.

In the next step, the multiple linear regression is run between Y and the principal component matrix C. The independent variable is standardized, and the standardized value must be transformed to the original value. The transformation can be executed based on

$$\widetilde{x}_i = \frac{(x_i - \overline{x})}{\sigma}, \tag{5}$$

where $\widetilde{x}$ is the standardized variable; x_i is the original variable; $\overline{x}$ is the mean value of original variables X; and σ is the standard deviation of the original variables X.

The final direct model estimating the daily ridership at the MRT station level is obtained and shown as follows:

$$\widehat{y} = p_0 + \widehat{p}_1 x_1 + \cdots + \widehat{p}_n x_n, \tag{6}$$

where $\widehat{y}$ is the estimate of ridership at MRT stations, p_0 is the constant, and $\widehat{p}_i$ is the estimate of the regression coefficient of the explanatory variable in the final PCR-based DRM.

3.3. Candidate Variables. In this study, the many candidate factors are calculated in GIS database. All potential factors are described in Table 2. We hypothesize that candidate factors have a close relationship to station-level MRT ridership.

3.3.1. Accessibility. Accessibility is considered as the ease of reaching spatially distributed opportunities from a given origin with a particular transportation network [43]. Previous researchers have explored the definition and measures of accessibility by the activity-to-cost ratio [8, 40, 44], e.g., the opportunities-to-population ratio and the relative value of cost [14]. One of the most popular measures is generally expressed by the function of $A_i = \sum_j S_i / f(C_{ij})$, in which S_i is the size of the activity in zone i, C_{ij} is the cost from zone i to j and $f(C_{ij})$ is the impedance function measuring the spatial separation between zones i to j. Existing approaches focus on measuring potential jobs to estimate the transit ridership. Less attention has been given to the attractiveness of the working-age population from other stations to the target station. In this research, a supplementary accessibility index is applied to measure the attraction power as indicated in equations (7) and (8):

$$\text{J\_Accessbility}_i = \frac{\sum_{j=1}^{n} E_j}{f\left(C_{ij}\right)}, \tag{7}$$

$$\text{P\_Accessbility}_i = \frac{\sum_{j=1}^{n} P_j}{f\left(C_{ij}\right)}, \tag{8}$$

$$f\left(C_{ij}\right) = \frac{\exp\left(\beta t_{ij} - \alpha\right)}{1 + \exp\left(\beta t_{ij} - \alpha\right)}, \tag{9}$$

$$t_{ij} = t_i^{\text{wait}} + t_{ij}^{\text{metro}} + t_{ij}^{\text{transfer}}, \tag{10}$$

where J_Accessbility_i is the potential of station i to reach all jobs at station j, P_Accessibility_i is the potential of station i to transfer the entire working-age population at station j, E_j denotes the jobs in station j, P_j is the working-age population at station j, t_{ij} is the total travel time through the MRT network from stations i to j (including t_i^{wait} is waiting time, t_{ij}^{metro} is metro ride duration between station i and j and t_{ij}^{transfer}—transfer time (if there is a transfer)), and α and β are impedance parameters.

Generally, the transit demand exhibits a distance decay nature. And the impedance function $f(C_{ij})$ has three normal forms: gravity function, exponential function, and logistic function [45]. From the view of trip production, when the cost from the origin to the destination is beyond sufferance, the trip may be canceled. Therefore, a ceiling of distance should be given in the impedance function. The logistic function $f(C_{ij}) = \exp(\beta t_{ij} - \alpha)/1 + \exp(\beta t_{ij} - \alpha)$, which is a "S" curve and suitable to explain the mechanism of traffic cost and transit demand. The logistic form is applied to estimate the impedance from origins to destinations (equation (9)). Different combination of parameter α, and β may cause various curves. In general, the data of mode choice and origin to destination matric are required to calibrate the parameters of α and β in the model of accessibility.

The working-age population is also needed for the variable of employment attraction. According to Chinese conventions, the ages range from 16 to 59 for males and 16 to 54 for females. For Chinese demographic census data of age distribution that were counted by the unit of census tract, the areal interpolation method is used to calculate the working-age population based on the study of Gutiérrez et al. [40].

3.3.2. Building Areas. Recent studies have found that the floor area or density of certain land use types was positively correlated with station ridership [16, 17, 46, 47]. Additionally, Loo et al. [46] considered the mixed use of commercial and residential buildings. However, most of these studies selected the floor area of land use as a limitation of the data source [16, 17, 20]. In this study, the detailed land uses and buildings with a plot ratio throughout the entire city are obtained. The land use types selected are determined based on the existing distribution of the building area along MRT lines. The building area is divided into four types: residential areas, commercial areas, office areas, and other areas. The four types are independent and have no intersection with each other.

3.3.3. Land Use Diversity. The land use type and diversity within station catchment area play key roles in transit use. Cervero and Kockelman [8] built a method of calculating land use diversity using the concept of entropy. Bhat and Guo [48] proposed a function that was popular with many later studies, although it can yield a negative result if certain land use areas were small. Studies have calculated land use diversity by using the reciprocal of the variation coefficient of the area covered by different land uses [40]. Sung et al. [22] proposed a formula for measuring the land use diversity for two different land uses; however, the function was difficult to use. This paper proposed a diversity model to

Table 2: Description of candidate variables.

Explaining variables	Description	Expected sign
Population	Total population within the PCA	+
Employment	Total employment within the PCA	+
Resi_Area	Building area of residence within the PCA	+
Com_Area	Building area of commerce within the PCA	+
Office_Area	Building area of office land use within the PCA	+
Other_Area	Building area of other uses within the PCA	+
Land_use_Mix	Land use mix index (equation (9))	+
Restaurant_Num	Number of restaurants within the PCA	+
College_Num	Number of colleges within the PCA	+
Hospital_Num	Number of hospitals within the PCA	+
Shopping_Num	Number of shopping stores within the PCA	+
Financial_Num	Number of financial institutions within the PCA	+
Scenic_spot_Num	Number of scenic spots within the PCA	+
Ext_hub_Num	Number of external traffic hubs within the PCA	−
Parking_Num	Number of parking spaces within the PCA	−
Recreational_Num	Number of recreational spots within the PCA	+
Gov_Agency_Num	Number of government agencies within the PCA	+
Hotel_Num	Number of hotels within the PCA	+
Road_Len	Length of road within the PCA	−
Auto_domi_Len	Length of automobile-dominated road (freeway and arterial road) within the PCA	−
Dis_to_centers	Minimum linear distance from a station to all city centers	−
Dis_to_station	Average linear distance between stops and stations within the 500 m station buffer	−
Bus_line_Num	Number of bus lines within the 500 m station buffer	+
Bus_stop_Num	Number of bus stops within the 500 m station buffer	+
Dummy_terminal	Whether the station is a terminal station (yes = 1; no = 0)	+
Dummy_line_transfer	Whether the station is a transfer station between metro lines (yes = 1; no = 0)	+
Dummy_modal_transfer	Whether the station is a transfer station between different traffic modes (yes = 1; no = 0)	+
Dummy_CBD	Whether the station is located in CBD (yes = 1; no = 0)	−
Dis_between_adj_station	The adjacent linear distance between stations	+
J_Accessibility	The accessibility of jobs from other stations (equation (7))	+
P_Accessibility	The accessibility of working-age people from other stations (equation (8))	+

estimate the degree of land use mix. The function is modified as follows:

$$Land\_use\_Mix = 1 - \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{\left(A_i - A_j\right)}{\left(A_i + A_j\right)}\right|, \quad (11)$$

where A_i and A_j are the building areas of two different uses of land and n is the total number of land use types.

3.3.4. Other Factors. The frequency of the MRT service is not considered since there are only three MRT lines in this study, and the frequency of the various fleets has little difference with 6 min, 4 min, and 4 min. For indicating transfer between bus and rail transit, the number of bus lines within 500 m buffer area of MRT stations is collected, considering the maximum transfer distance. The variable of the distance to the city center is calculated. In case there is more than one city center, only the distance to the closest center is recorded. The length of high-grade streets is calculated with a freeway and arterial roads within the PCA, and it is assumed to be an impedance for transit ridership. The P&R space variable is not included as it is not relevant in this case study. The shopping store refers to single and small shops rather than shopping malls or markets. For large shopping malls, the number of the stores inside is accumulated. The hospitals in this study exclude the drugstores, private clinics, and community hospitals that mainly attract nonmotorized trips.

4. Data and Model Results

4.1. Study Context. Wuhan, the capital city of Hubei Province, is a fast-growing metropolitan city in central China. It is in the middle reaches of the Yangtze River with a jurisdiction area of 8494 km^2 and a population of more than 10 million. The Yangtze River and the Han River (the largest branch of the Yangtze River) divided the city into three towns and shaped a "triple-town" spatial morphology. Meanwhile the presence of large lakes along with the two great rivers has resulted in a spatially dispersed urban form [49].

In this study, Mass Rapid Transit (MRT) and rail transit are used interchangeable, which include both light rail transit (LRT) and metro transit in the case of Wuhan. MRT Line 1 is an elevated light rail (LRT), which was open to traffic in July 2004. There were 26 stations along the line with a total length of 28.5 km. MRT Line 2 (metro), with a total length of 27.33 km and 21 stations, was opened in December 2012. It passes through the Yangtze River and is linked to the Hankou Railway Station. This line has two stations that can transfer to MRT Line 4, which transports passengers to the Wuchang Railway Station and the Wuhan High-speed

Railway Station. The MRT Line 4 (metro) was open in December 2014. It has 27 stations and a total length of 32.7 km.

The 1000 m walking distance is adopted for the following reasons. First, a survey of traffic modes connecting to the metro station including revealed preference (RP) and stated preference (SP) is conducted in Wuhan. The statistical results show that the walking time in 85% cumulative probability of RP is 9 min (approximately equal to 750 m), and that of SP is 12 min (about 1000 m). Second, with respect to demand, appropriate overestimation of standard threshold of MRT station coverage distance is important [39]. Third, people are willing to walk much longer distances to use MRT, which are not commonly found in Wuhan. Combining the different given cost values of various land use types, the cost distance of the spatial analysis tool in ArcGIS is used to generate the PCA (Figure 3).

4.2. Data Source and Preprocessing. The station-level ridership, including the boarding and alighting weekday data of MRT lines in April 2015, is acquired from the metro operator. Four city centers, including the Wangjiadun-CBD, the Hankou Business Center, the Shuiguohu Administrative Center, and the Optical Valley Science and Business Center, are selected based on the master plan of Wuhan. The disaggregated employment and population data in 2015 are obtained from the 6th national demographic census data in 2010, the 3rd Wuhan economic census data in 2015, and the social insurance information in 2016, which are drawn from the statistics department of Wuhan. A data disaggregation procedure conducted using C# in Microsoft Visual Studio, which is combined in the GIS Engine, transforms population and employment data into raster cells based on a procedure by Huang et al. [31]. The data of different land uses with the floor area and plot ratio of the architecture in 2014 are also provided by the urban planning and information department of Wuhan. The building area and diversity degree of the four basic categories of land uses are calculated via GIS and Microsoft Excel. The bus, metro, and road networks have been accumulated by our research group and improved according to the Baidu map and Smart City, an app developed by the Wuhan government. The road network with road class information in GIS was completed after a year of data accumulation. Traffic attraction spots (colleges, shops, hospitals, etc.) around the MRT stations are determined by POI (position of interest) data collected from Baidu maps. The various factors are calculated by the indices of the maximum, minimum, average, and standard deviation in Table 3.

The mode choice of the public transit system and origin to destination matric of metro network in Wuhan were obtained from the traffic department of Wuhan. These data are useful to define the integrated accessibility model. The logistic function $f(C_{ij}) = \exp(\beta t_{ij} - \alpha)/1 + \exp(\beta t_{ij} - \alpha)$ is used in this study. The metro mode shared 25% trips in the public transit system. And travelers have equal probabilities (0.5 each) of choose metro and other public transit modes at the distance of 9 km, i.e., the travel time of 18 minutes. The β is 0.098 and the α is 1.772, calibrated through a regression procedure based on the data of public transit trips. In this process, the waiting time (${t_i}^{\text{wait}}$) is set to half of departing interval time of different metro lines. The in-vehicle time of metro (${t_{ij}}^{\text{metro}}$) is calculated by the origin-destination matric of metro by the software of GIS. The transfer time in metro network (t_{ij}^{transfer}) of 5 min is included at every transfer station. The ridership and accessibility of rail transit stations are shown in Figure 4.

The monthly average weekday boarding is selected as the dependent variable. Boarding and alighting are considered to have a strong relationship with each other [19, 20]. Alighting is not considered because it is highly correlated with boarding and presents a correlation coefficient of 0.938. The potential variables with summary statistics are listed in Table 3.

4.3. Independent Variable Selection. To test for the correlation between each candidate variable and metro ridership, a bivariate correlation analysis is conducted (the second column in Table 4). The result shows that 19 predictors are clearly significant in explaining ridership at a significance level of 5%, and the predictor of the distance to the city center is the only variable with a negative Pearson correlation coefficient. The correlation coefficients range from 0.316 to 0.788. The commercial area, the office area, the land use mix, the number of restaurants in PCA, and the transfer hub have the highest correlation among 19 factors. The value of the restaurants (0.788) is the largest, after which the office area ranks second (the value is 0.777). The residential land use area has much weaker association with MRT ridership than the commercial and office land uses within PCA. The land use mix variable is the third highest factor affecting the ridership ranking behind the variable of the number of financial institutions. The correlation coefficients of population and employment have a moderate effect, although the population has a slightly stronger relationship to ridership than employment. However, from the perspective of transit network, the job accessibility that expressed the degree of attractiveness of employment in the corridor of rail transit network has higher correlation coefficient than working-age population accessibility.

Another bivariate correlation analysis is conducted among the 19 variables (Table 4). The results indicate that numerous variable pairs have medium and high correlations, and the possible reasons for these results are analyzed. In large cities in China, many large communities are observed in locations that show an aggregation of commercial and office activities, which will help absorb a large number of people, workplaces, and fundamental facilities. Therefore, the higher positive or negative correlations emerge when we use the bivariate correlation analysis. This situation decreases the suitability of the linear regression method. The PCR-based model is adopted to solve the multicollinearity problem.

Pearson's correlation coefficients with values of ±0.1, ±0.3, and ±0.5 represent the boundary points of a low, moderate, and high correlation, respectively [50].

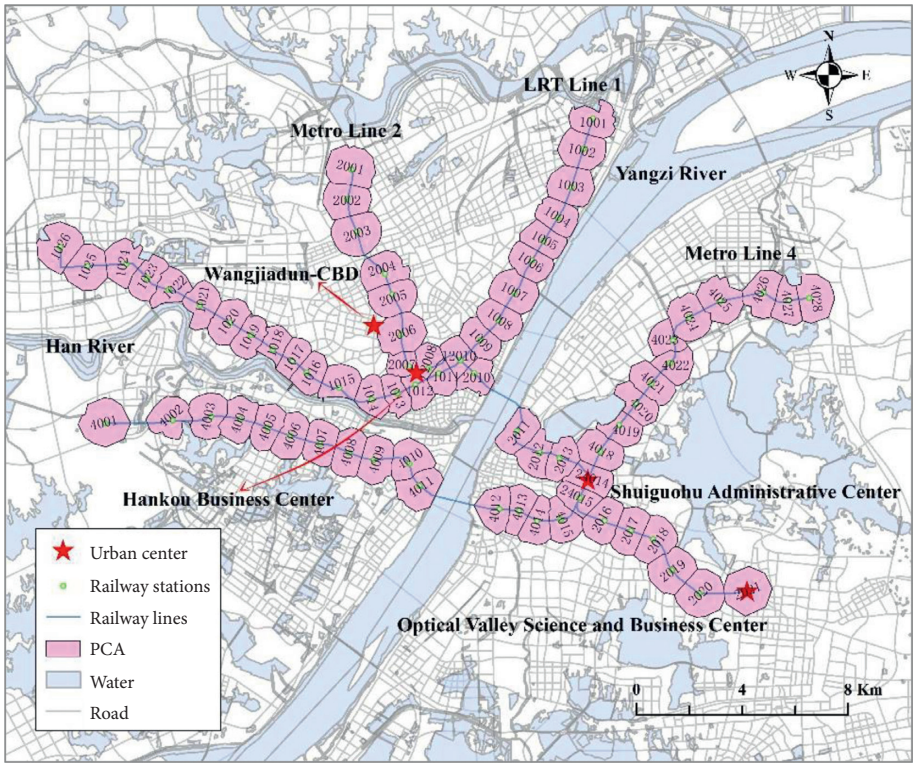

Figure 3: Study area showing the PCA of MRT stations in Wuhan. The values filled in PCA refer to the station order of rail transit lines in Wuhan.

Table 3: Summary statistics for candidate variables.

Variables	Mean	Max. value	Min. value	Standard deviation
Ridership	17051	68093	2011	12917
Population (person)	38965	105822	3268	21109
Employment (employee)	22054	113886	472	21441
Resi_Area (m^2)	1296607	3162713	47770	637214
Com_Area (m^2)	229387	1284208	2716	245010
Offi_Area (m^2)	83528	419806	1035	97021
Oth_Area (m^2)	472042	1194859	28566	257450
Land_use_Mix	0.358	0.681	0.154	0.123
Restaurant_Num	77	441	0	80
College_Num	2	14	0	2
Hospital_Num	2	6	0	1
Shopping_Num	213	1190	6	208
Financial_Num	25	73	0	19
Scenic_spot_Num	3	50	0	8
Ext_hub_Num	0	2	0	0
Parking_Num	10	32	0	8
Recreational_Num	55	200	0	41
Gov_Agency_Num	17	94	0	20
Hotel_Num	20	84	0	20
Road_Len (m)	9384	17422	3745	2708
Auto_domi_Len (m)	2800	5794	140	1332
Dis_to_centers (m)	4503	11670	283	3027
Bus_line_Num	19	52	1	13
Bus_stop_Num	5	15	1	3
Dis_to_station (m)	270	437	15	74
Dummy_terminal	0	1	0	0
Dummy_line_transfer	0	1	0	0
Dummy_modal_transfer	0	1	0	0
Dummy_CBD	0	1	0	0
Dis_between_adj_station (m)	1201	2550	755	329
J_Accessibility (employee/minute)	5989	10235	1199	2453
P_Accessibility (working person/minute)	7256	12035	2155	2708

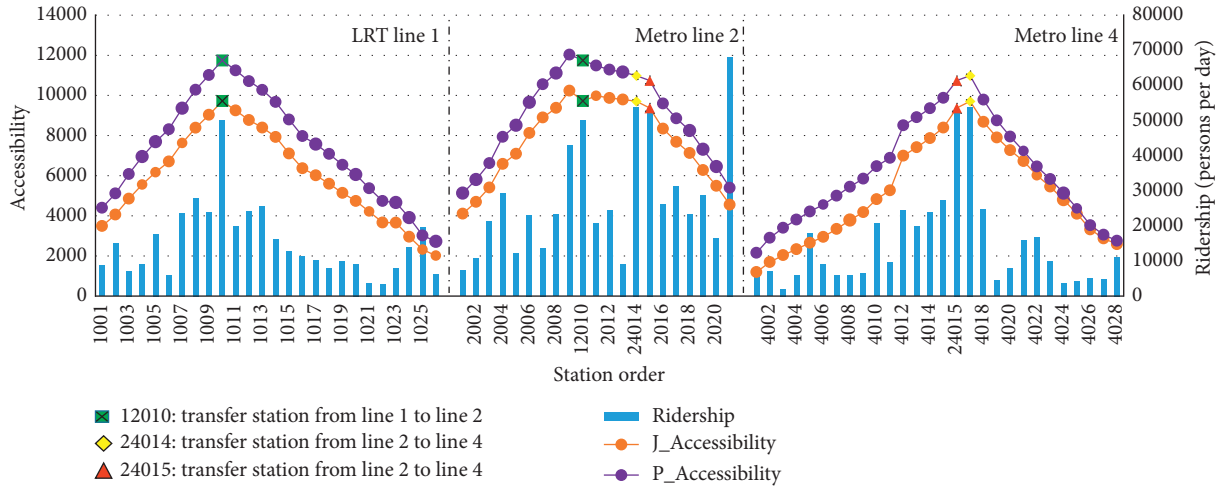

Figure 4: The integrated accessibility and ridership of MRT stations in Wuhan.

4.4. Results and Interpretation. The PCR procedure is processed via SPSS 17.0. According to the results of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (0.873) and Bartlett's test of Sphericity (sig = 0.000), the factor analysis is suitable for the PCR procedure. Four components are generated by extracting eigenvalues with values greater than 1. More than 75% of the total variance can be explained using the four principal components.

A rotation using the max-variance method is used to label every component properly and professionally. From the perspective of the mechanism of travel behavior, eleven variables are combined into four components (Table 5). The first component is named factors related to the built environment, and it explains approximately 31% of the variance of the original variables. The second component is named factors related to jobs-housing spatial structure, and it explains 24% of the variance. The third component is called factors related to station attributes, and it explains 10% of the variance. The last component is called factors related to the large compound, and it occupies 10% of the variance. The results indicate that the MRT ridership at the station level should be seriously considered based on built environment and jobs-housing distribution, which account for more than 50% of the variance. The transfer hub attribute is also a vital factor indicated by a number of previous studies [14, 16, 18, 23]. The large compound has vital effect on MRT ridership. The large compounds (such as the college campus and hospital) have various forms, but they typically have similar elements in common, located in urban central areas, covering a comparatively larger area, and being foreign seal and internal opening in physical space. Therefore, the large compound would produce or attract plenty of trips and exert an obvious effect on the transit demand. Because the four components are orthogonal, they are mutually independent. Therefore, the estimation of transit ridership demand developed based on these components can prevent multicollinearity.

The four components are served as variables in the data view and variable view of SPSS. A linear regression between ridership and components 1–4 is performed. The initial set of regression results are summarized in Table 6. The model has an R^2 value of 0.885 (adjusted R^2 of 0.878) and an F-statistic value of 126.988, which is obviously significant at the 0.000 level. The four components are all significant at the 0.000 level, and the t-statistic values are all greater than 6. To test the severity of the multicollinearity, the variance inflation factor (VIF) index is calculated in the model. VIFs are close to the value of 1, which indicates that the regression coefficients estimated by PCR have little bias compared with the factual value.

The final relationship between the monthly average MRT station ridership and the 19 initial influential factors is described in Table 7.

4.5. Model Evaluation. The fit analysis between actual and predicted values is carried out in the left part of Figure 5. The value of 2021 is abandoned for being greatly underestimated. The fitting equation with an R^2 value of 0.881 shows that the relationship between the actual and predicted values is strong and achieves a higher level of compliance. The standardized residuals exhibit random characteristics and remain stable in the range from −2 to 2, which indicates that the PCR-based DRM has homogeneous variances.

The indicator of relative errors (RE) is a ratio of the predicted ridership minus actual ridership to the actual ridership and is calculated for evaluating the results at individual MRT stations (Figure 6). The range of RE is between -0.78 and 1.13. It shows that 80% of stations with the absolute value of RE lower than 0.5. Transfer stations, i.e., 12010, 24014, and 24015, are slightly underestimated, respectively, by 3%, 10%, and 19%. This finding indicates that the attribute of transfer hub has more crucial and undiscovered influence. Meanwhile, the boarding of station 2021, which is located in local CBD area of Optical Volley Business and Science Center, is underestimated by 39%. A main reason for this might lie in the fact that this terminal station has a much larger service area. For various MRT lines, the station ridership of Line 1 and Line 2 has been estimated better than that of Line 4. Actually, the first phase of line 4 in Wuchang (stations 4012–4028) was opened in May 2013, and

Table 4: Matrix of a bivariate correlation test between independent variables and ridership.

	Ridership	Population	Employment	Com_Area	Offi_Area	Land_use_Mix	Restaurant_Num	College_Num	Hospital_Num	Shopping_Num	Financial_Num	Parking_Num	Recreational_Num	Hotel_Num	Dis_to_centers	Bus_line_Num	Dummy_line_transfer	Dummy_CBD	J_Accessibility	P_Accessibility
Ridership	1	0.618**	0.591**	0.695**	0.777**	0.722**	0.788**	0.316**	0.507**	0.689**	0.755**	0.714**	0.632**	0.609**	−0.585**	0.425**	0.569**	0.537**	0.561**	0.555**
Population		1	0.539**	0.588**	0.554**	0.486**	0.694**	0.374**	0.387**	0.733**	0.816**	0.784**	0.840**	0.671**	−0.669**	0.599**	0.176	0.637**	0.653**	0.705**
Employment			1	0.567**	0.577**	0.501**	0.438**	−0.013	0.184	0.436**	0.682**	0.649**	0.511**	0.347**	−0.528**	0.427**	0.445**	.465**	0.593**	0.588**
Com_Area				1	0.531**	0.659**	0.736**	0.091	0.282*	0.783**	0.752**	0.683**	0.586**	0.564**	−0.533**	0.486**	0.223	0.448**	0.492**	0.506**
Offi_Area					1	0.797**	0.587**	0.287*	0.260*	0.505**	0.725**	0.641**	0.576**	0.521**	−0.511**	0.223	0.399**	0.339**	0.455**	0.443**
Land_use_Mix						1	0.515**	0.222	0.307**	0.510**	0.628**	0.584**	0.454**	0.465**	−0.514**	0.313**	0.349**	0.294*	0.497**	0.479**
Restaurant_Num							1	0.429**	0.354**	0.874**	0.771**	0.773**	0.765**	0.745**	−0.497**	0.412**	0.133	0.481**	0.410**	0.426**
College_Num								1	0.183	0.317**	0.252*	0.290*	0.405**	0.429**	−0.287*	0.144	0.043	0.235*	0.130	0.143
Hospital_Num									1	0.315**	0.359**	0.330**	0.334**	0.338**	−0.271*	0.211	0.314**	0.366**	0.322**	0.344**
Shopping_Num										1	0.746**	0.698**	0.788**	0.708**	−0.457**	0.478**	0.165	0.584**	0.455**	0.491**
Financial_Num											1	0.871**	0.807**	0.596**	−0.638**	0.522**	0.276*	0.521**	0.576**	0.600**
Parking_Num												1	0.721**	0.645**	−0.686**	0.587**	0.173	0.544**	0.624**	0.640**
Recreational_Num													1	0.675**	−0.515**	0.388**	0.146	0.594**	0.488**	0.527**
Hotel_Num														1	−0.436**	0.368**	0.064	0.394**	0.370**	0.401**
Dis_to_centers															1	−0.565**	-0.246*	-0.442**	-0.830**	-0.847**
Bus_line_Num																1	0.232*	0.513**	0.470**	0.531**
Dummy_line_transfer																	1	0.348**	0.309**	0.303**
Dummy_CBD																		1	0.498**	0.539**
J_Accessibility																			1	0.988**
P_Accessibility																				1

*Significant at the 0.05 level. **Significant at the 0.01 level.

Table 5: Eigenvectors of the principal components* and the correlation matrix.

Principal components	Variables	Component 1	Component 2	Component 3	Component 4
Component 1: factors related to the built environment	Restaurant_Num	0.837			
	Shopping_Num	0.791			
	Financial_Num	0.774			
	Com_Area	0.768			
	Hotel_Num	0.716			
	Parking_Num	0.716			
	Recreational_Num	0.715			
	Offi_Area	0.679			
	Land_use_mix	0.629			
Component 2: factors related to jobs-housing spatial structure	P_Accessibility		0.889		
	J_Accessibility		0.859		
	Dis_to_centers		−0.781		
	Bus_line_Num		0.681		
	Population		0.613		
	Dummy_CBD		0.564		
	Employment		0.510		
Component 3: factors related to station attributes	Dummy_line_transfer			0.833	
Component 4: factors related to large compound	College_Num				0.734
	Hospital_Num				0.595

*Correlation coefficients between variables and components greater than 0.5 are shown in the table.

Table 6: Results of the PCR. Dependent variable: Ridership[a].

Independent variables	B	Std. error	Beta	t-statistic	Sig.	VIF
(Constant)	16650.298	476.101		34.972	0.000	
Component 1	6744.863	520.778	0.543	12.952	0.000	1.008
Component 2	4397.415	479.999	0.383	9.161	0.000	1.004
Component 3	6969.519	475.314	0.612	14.663	0.000	1.000
Component 4	2937.604	482.842	0.255	6.084	0.000	1.006
Degree of freedom (df): 68						
$F = 126.988$ (sig = 0.000)						
R: 930						
R^2: 0.885						
Adjusted R^2: 0.878						

[a]One station with a standardized residual greater than 4 is eliminated. B denotes the final coefficient in the regression function. Beta denotes the standardized coefficient in the regression function.

the second phase in Hanyang (stations 4001–4011) was put into operation in December, 2014. Many of the 2nd phase metro stations are still in the period of passenger cultivation. In total, 61% of Line 4 stations have been overestimated with the percent difference varying between 3% and 113%.

To validate the direct ridership model based on PCR, a new MRT line (Line 3) is utilized (Figure 7). Line 3 was opened in December 2015, with 24 stations and a total length of 30 km. The metro line is the first MRT that passes through Han River and links Hankou (located at the north bank of Han River) and Hanyang (located at the south bank of Han River). The average weekday ridership at stations was based on data in April 2016. There are three transfer stations connecting Lines 1, 2, and 4, which are, respectively, coded by 13015, 23011, and 34016. The other stations are coded between 3001 and 3024 from north to south. It needs to be noted that station 3008 is removed from the statistics due to its exceptional result. The relative errors (RE) is shown on the right side of Figure 7. The range of RE for all stations of Line 3 is between -0.57 and 0.94. Results show that 70% of stations come up with absolute values of RE less than 0.5. About 65% of stations have been overestimated and the mean value of RE is 0.15. The analysis carried out indicates that PCR-based DRM has better explanatory power to estimate the rail transit demand.

5. Discussion

A bivariate correlation analysis identified nineteen influencing factors that have a close relationship to MRT demands. The factors of restaurants, office building area, and land use mix within PCA have the highest correlation coefficient. However, the number of bus stops is weakly related to ridership at the metro station level. It is possibly related to the locations of the stations, which were all in the central urban area where bus stops around MRT stations are more common. Jun et al. [23] also noted that the local ridership effect of bus stops was statistically insignificant in the city center but significant in the suburban area. Meanwhile, most of the activity attractors, such as restaurants, shopping

Table 7: Direct ridership model (DRM) estimating daily ridership at the station level in Wuhan, China.

Independent variables	B	Beta
Intercept	−1215.182	—
Population	0.086	0.014
Employment	0.154	0.098
Com_Area	0.002	0.043
Offi_Area	0.023	0.192
Land_use_mix	16941.145	0.177
Restaurant_Num	6.950	0.039
College_Num	280.784	0.062
Hospital_Num	1408.537	0.181
Shopping_Num	1.477	0.018
Financial_Num	47.943	0.071
Parking_Num	43.409	0.026
Recreational_Num	10.473	0.032
Hotel_Num	21.267	0.029
Dis_to_centers	−0.071	−0.022
Bus_line_Num	−37.498	−0.039
Dummy_line_transfer	14960.126	0.272
Dummy_CBD	1396.449	0.059
J_Accessibility	0.151	0.039
P_Accessibility	0.092	0.028
Model statistics		
F-statistic	126.988 (sig = 0.000)	
Std. Error	4004.668	
R^2	0.885	
Adjusted R^2	0.878	

B denotes the final coefficient in the regression function. Beta denotes the standardized coefficient in the regression function.

stores, and hospitals, are significant for metro ridership in Wuhan. This result is unexpected but likely reflects the importance that concentrating distribution of special generators around stations will increase the MRT ridership at the station level. The length of roads and the length of automobile-dominated roads within the PCA have a minimal impact on MRT ridership.

The PCR method is acceptable for the conceptual process of excavating potential components to draw latent links between influencing factors and the ridership target. Four components are obtained, and they are properly and professionally named based on the mechanisms of trip generation. Each component has several subfactors. There are nine factors related to the built environment. The variables (the number of restaurants, shopping stores, financial institutions, parking spaces, recreational spots, hotels, the building area of commercial land and office land, and the mixed land use) are relatively more highly related to component 1. The correlation coefficient of restaurants is 0.837, and it is higher than that of other special generators. The commercial building area has a larger correlation value (0.768) than that of the office building area and land use mix. The related jobs-housing spatial structure includes seven subfactors (i.e., working-age population accessibility, job accessibility, distance to city centers, number of connected bus lines, population, employment, and dummy CBD). The integrated accessibility has the highest correlation coefficients (0.889 and 0.859), which indicates the balance level of jobs and housing in the measure of metro system. The population and employment have a little less correlation values (0.613 and 0.510, respectively) because they are also subordinate to component 1, built environment to some extent. The initial characteristics of MRT stations are also crucial for the contribution of ridership. The correlation coefficient between the number of transfer lines and component 3 is 0.804. The attribute of the transfer station plays more of a key role than other station characteristics, e.g., terminal stations and the distance between adjacent stations. The component of large compound has two subfactors, colleges and hospitals. The number of colleges has a higher correlation with the Large Compound component, although a station located close to a general hospital can attract a large number of riders.

The regression coefficients from the PCR-based DRM indicate the elasticity degree of the predictors (Table 7). Based on the standardized coefficients (Beta in Table 7), the office building area, the land use mix, the number of hospitals, and dummy transfer hub are the four most influential variables. The employment variable ranks behind them. The employment has higher impacts on transit ridership than the population, which is consistent with the integrated accessibility that the job accessibility has a larger elasticity than the working-age population accessibility. This finding indicates that existing MRT lines in Wuhan mainly serve as commuting purpose, and the jobs-intensive distribution plays a crucial role in MRT ridership. The number of connected bus lines and the distance to city centers are the only two variables that have negative influence on the rail station demand. The number of bus routes around MRT stations is a strongly significant and robust factor for influencing rail transit ridership. This negative regression coefficient indicates that the bus transit has potential competition with MRT in Wuhan. For better integrating the two transit modes, the bus lines layout should be adjusted properly when the metro system is introduced. On the one hand, the short distance feeder bus lines will gather people into or distribute from the MRT stations and will improve the ridership of the metro transit. On the other hand, the long-distance bus lines, which are deployed along the corridor of metro lines, should be adjusted and referential measures to encourage bus and metro transfers should be introduced so as to reduce competition. For example, the transport agency has stopped bus line 519 for its patronage drop due to overlapping with Metro Line 2.

The results from the PCR-based DRM are compared to the findings of previous studies. In the result of PCR, the coefficients of the population (0.086) and employment (0.154) are lower than those of the previous studies. It is possible that there are other variables accounted for the built environment and land use development around MRT stations. Kuby et al. [14] studied nine US cities, which ranged from Buffalo to St. Louis to San Diego, where the population density was relatively sparse. The study context in this analysis involves unevenly distributed population. The coefficient of the population is 0.086 in this paper, and it is slightly less than 0.092 found by Kuby. Therefore, the scale of the study area has a slight influence on the structure of the estimating model. In this paper, the coefficient of the

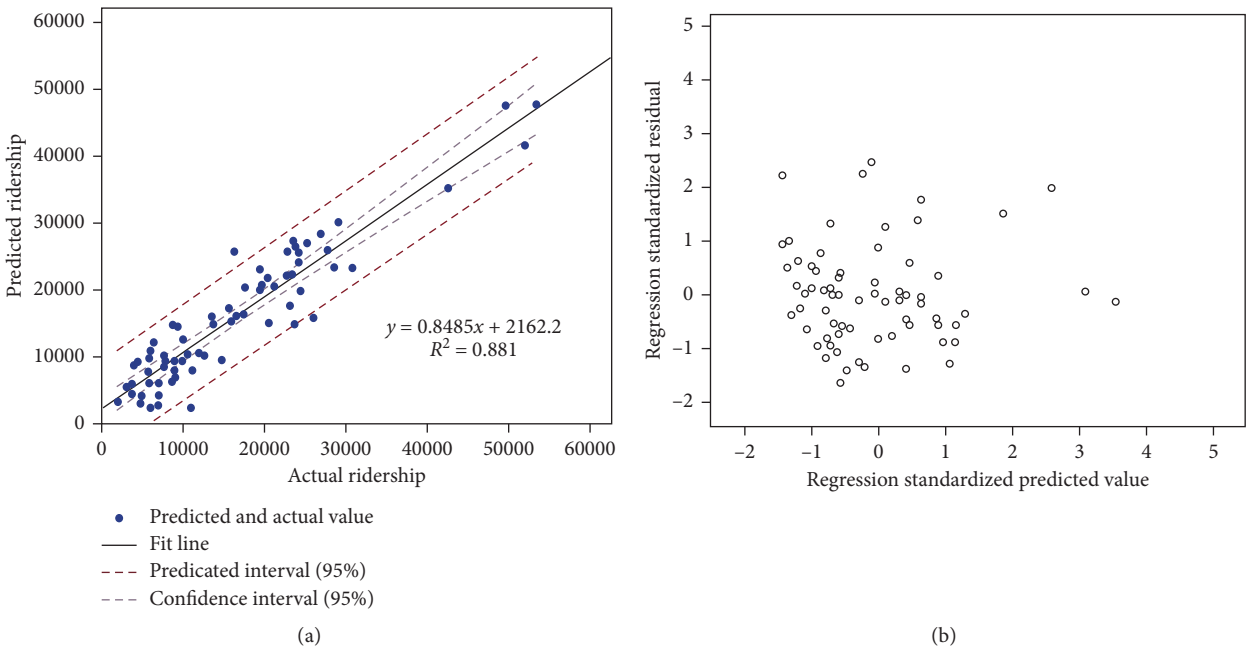

Figure 5: Predicted versus actual ridership (a) and residual analysis (b).

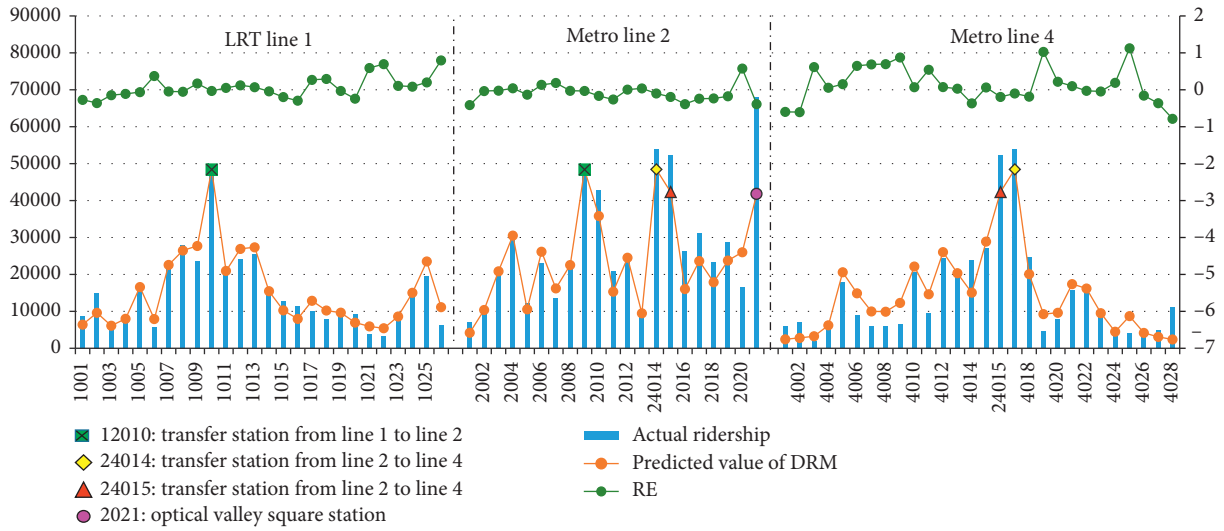

Figure 6: Predicted versus actual ridership and relative errors for MRT stations.

population is much lower than that of employment (0.154). This finding is also supported by the results of the integrated accessibility that the effect of the working-age population accessibility (P-Accessibility) is weaker than that of job accessibility (J-Accessibility). Additionally, this result is opposite to Kuby's research. The main reason for the reverse is possible that the population is more concentrated around stations than employment in our analysis, which is inconsistent with the results of Kuby's study. On the contrary, other studies argued that employment had a stronger effect on ridership than population [14, 17, 22, 23].

In Wuhan, the CBD is closely connected to the entrances or exits of metro stations. Therefore, the CBD dummy variable has a positive coefficient (1396) and is dramatically significant at the 0.000 level. If metro stations are close to the CBD, then approximately 1396 daily riders are added to the station. In addition, adding one more transfer line to the station generates an average of 14960 additional riders every day. The high positive coefficient on the "transfer hub" variable reveals that ridership at MRT stations tends to be much higher than other factors. Meanwhile, the impact of the transfer characteristic is larger than that in Sohn and Shim's study (10092) [16] and Kuby's study (5735) [14]. The probable reason for this difference is that the metro network has not been fully developed in Wuhan, which may generate more transfers. These findings are valuable for policy-makers to improve feeder transit service to MRT.

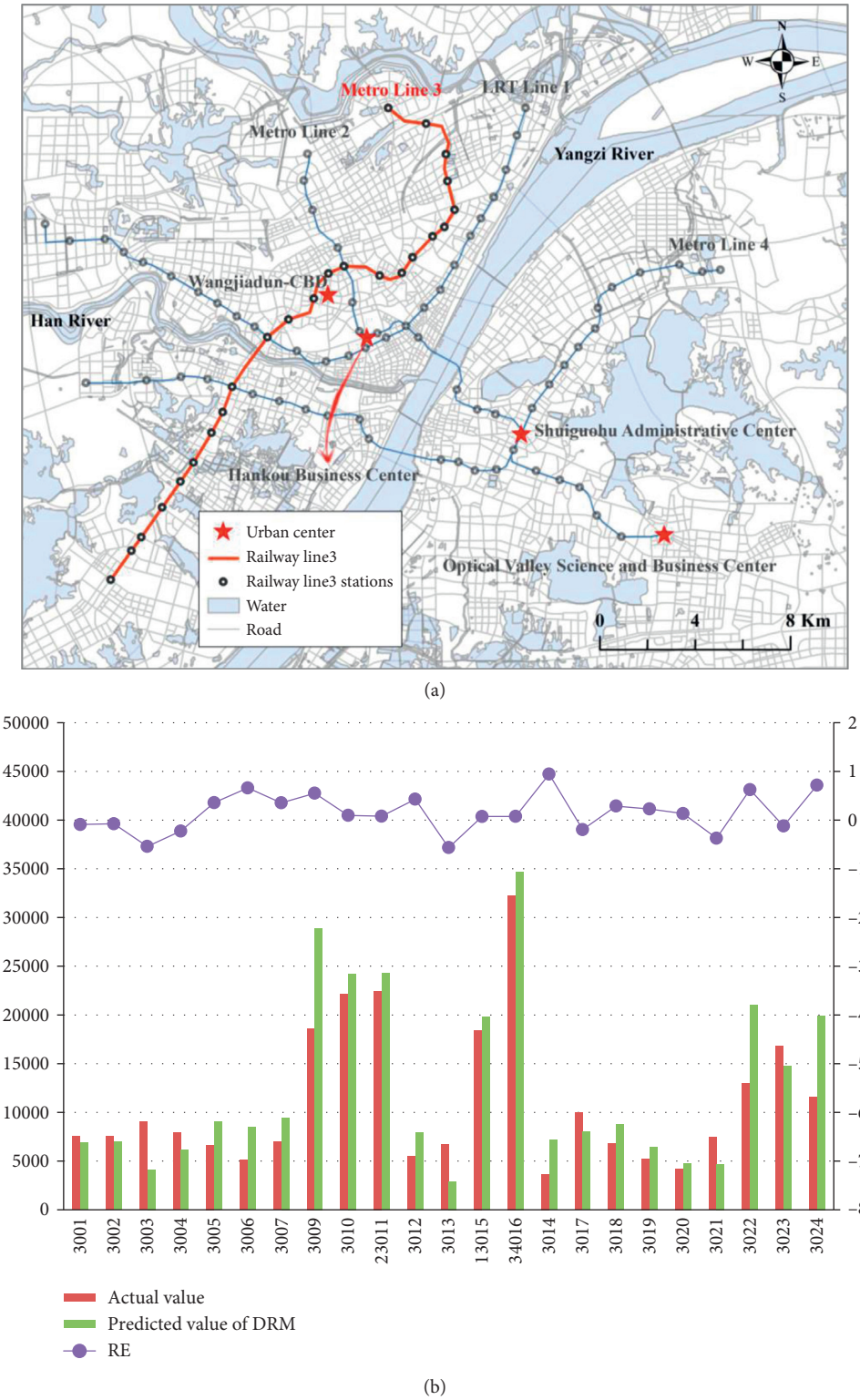

Figure 7: MRT Line 3 in Wuhan (a) and RE analysis (b) (right).

Many researchers did not analyze the possible causal relation between the degree of mixed land use and rail transit use [14, 17, 18, 20]. Cervero et al. [13] pointed out that the mixed land use would generate almost 3750 more weekly ridership than single-use in the DRM of the St. Louis Metrolink. The different expression of land use mix might result in various estimated elasticity. Sung et al. [22] found that land use diversity was obviously associated with rail transit ridership (13545). In this study, the mixed land use has much stronger impact (with the coefficient of 16941) on rail

transit demand from the result of the regression model. Therefore, a more diverse built environment around metro stations corresponds to a greater number of MRT passengers.

6. Conclusions

This research aims to build a DRM to estimate newly built MRT demand in megacities. The case study of Wuhan, China, has demonstrated the effectiveness of the framework. Nineteen influencing factors are shown to have a significant relationship with transit ridership, in which the development of land uses and employment play stronger roles. The PCR method has been proposed to descend dimension of explanation variables and manage multicollinearity. Prior to the DRM results, four principal components are obtained and reveal the inherent links between the influencing factors and transit demand. Four independent principal components, namely, factors related to trip generation, factors related station-area built environment, jobs-housing spatial structure, station attributes, and the large compound, explain 31%, 24%, 10%, and 10% of the variance, respectively. The DRM based on the PCR method is able to evaluate the actual ridership with a high degree of fitting (0.878). The results show that variables of the office building area, the land use mix, the number of hospitals, and dummy transfer hub have the strongest relation with metro station-level ridership. Employment has a more significant impact on transit demand than population. These findings are meaningful for policies on development of metro lines and land use. The DRM framework requires observations of features at or around MRT stations and may be rigorously calibrated with real MRT ridership data.

Certain limitations were observed in this study. First, the influence of socioeconomic variables (e.g., the ownership of family cars and household incomes) is scarcely considered due to the lack of data. Second, all MRT stations are addressed independently based on their individual PCAs, which ignores potential influence from adjacent stations. The logical next step of further studies is to take more consideration on spatial effect and additional MRT data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

RL Guo performed literature search and review, meta-analysis, and manuscript writing. ZD Huang performed content planning, meta-analysis, and manuscript writing and editing.

Acknowledgments

The data support from the Wuhan Land Resources and Planning Information Centre (WLRPIC) is gratefully acknowledged. This work was supported by the Research Program of Shenzhen S&T Innovation Committee (Project number. JCYJ20170412105839839) and National Natural Science Foundation of China (no. 41271396).

Supplementary Materials

Building area of residence within the PCA. (*Supplementary Materials*)

References

[1] M. Abril, F. Barber, L. Ingolotti, M. A. Salido, P. Tormos, and A. Lova, "An assessment of railway capacity," *Transportation Research Part E: Logistics and Transportation Review*, vol. 44, no. 5, pp. 774–806, 2008.

[2] B. Wang, J. Huang, and J. Xu, "Capacity optimization and allocation of an urban rail transit network based on multi-source data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 1, pp. 373–383, 2019.

[3] S. Canavan, A. Barron, J. Cohen, D. J. Graham, and R. J. Anderson, "Best practices in operating high frequency metro services," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 9, pp. 491–501, 2019.

[4] K. P. Li, H. Huang, and P. Schonfeld, "Metro timetabling for time-varying passenger demand and congestion at stations," *Journal of Advanced Transportation*, vol. 2018, Article ID 3690603, 26 pages, 2018.

[5] J. B. Ingvardson and O. A. Nielsen, "How urban density, network topology and socio-economy influence public transport ridership: empirical evidence from 48 European metropolitan areas," *Journal of Transport Geography*, vol. 72, no. 1, pp. 50–63, 2018.

[6] M. Zhang, "The role of land use in travel mode choice: evidence from Boston and Hong Kong," *Journal of the American Planning Association*, vol. 70, no. 3, pp. 344–360, 2004.

[7] P. W. G. Newman and J. R. Kenworthy, "Sustainability and cities: extending the metabolism model," *Landscape and Urban Planning*, vol. 44, no. 4, pp. 219–226, 1999.

[8] 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.

[9] R. Cervero and J. Landis, "Twenty years of the bay area Rapid transit system: land use and development impacts," *Transportation Research Part A: Policy and Practice*, vol. 31, no. 4, pp. 309–333, 1997.

[10] 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, no. 1, pp. 548–558, 2012.

[11] R. Cervero, J. Murakami, and M. Miller, "Direct ridership model of bus Rapid transit in Los Angeles county, California," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2145, no. 1, pp. 1–7, 2010.

[12] Z. Shi, N. Zhang, Y. Liu, and W. Xu, "Exploring spatio-temporal variation in hourly metro ridership at station level: the influence of built environment and topological structure," *Sustainability*, vol. 10, no. 12, p. 4564, 2018.

[13] 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.

[14] 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.

[15] L. Usvyat, L. Meckel, M. DiCarlantonio, and C. Lane, "Sketch model to forecast heavy-rail ridership," in *Proceedings of the Presented at 88th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, DC, USA, January 2009.

[16] 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.

[17] J. Zhao, W. Deng, Y. Song, and Y. Zhu, "What influences metro station ridership in China? Insights from Nanjing," *Cities*, vol. 35, no. 1, pp. 114–124, 2013.

[18] L. E. Ramos-Santiago and J. Brown, "A comparative assessment of the factors associated with station-level streetcar versus light rail transit ridership in the United States," *Urban Studies*, vol. 53, no. 5, pp. 915–935, 2016.

[19] X. Chu, "Ridership models at the stop level," in *National Center for Transit Research*, University of South Florida, Washington, DC, USA, 2004.

[20] V. Chakour and N. Eluru, "Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal," *Journal of Transport Geography*, vol. 51, no. 1, pp. 205–217, 2016.

[21] S. Pulugurtha and M. Agurla, "Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods," *Journal of Public Transportation*, vol. 15, no. 1, pp. 33–52, 2012.

[22] 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, no. 1, pp. 134–140, 2014.

[23] 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, no. 1, pp. 30–40, 2015.

[24] X. Ma, J. Zhang, C. Ding, and Y. Wang, "A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership," *Computers, Environment and Urban Systems*, vol. 70, no. 1, pp. 113–124, 2018.

[25] Y. He, Y. Zhao, and K.-L. Tsui, "Geographically modeling and understanding factors influencing transit ridership: an empirical study of Shenzhen metro," *Applied Sciences*, vol. 9, no. 20, p. 4217, 2019.

[26] 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.

[27] J. J. Tang, F. Gao, F. Liu, W. Zhang, and Y. Qi, "Understanding spatio-temporal characteristics of urban travel demand based on the combination of GWR and GLM," *Sustainability*, vol. 11, no. 19, p. 5525, 2019.

[28] L. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[29] L. F. Chow, F. Zhao, H. Chi, and Z. Chen, "Subregional transit ridership models based on geographically weighted regression," in *Proceedings of the Transportation Research Board 89th Annual Meeting*, Transportation Research Board, Washington, DC, USA, January 2010.

[30] H. Yu, J. Jiao, E. Houston, and Z.-R. Peng, "Evaluating the relationship between rail transit and industrial agglomeration: an observation from the Dallas-fort worth region, TX," *Journal of Transport Geography*, vol. 67, no. 1, pp. 33–52, 2018.

[31] Z. Huang, M. Zhang, and X. Liu, "Estimating light-rail transit peak-hour boarding based on accessibility at station and route levels in Wuhan, China," *Transportation Planning & Technology*, vol. 40, no. 2, pp. 624–639, 2017.

[32] Z.-J. Wang, H.-X. Liu, S. Qiu, J.-P. Fang, and T. Wang, "The predictability of short-term urban rail demand: choice of time resolution and methodology," *Sustainability*, vol. 11, no. 21, p. 6173, 2019.

[33] S. Liu, E. Yao, and B. Li, "Exploring urban rail transit station-level ridership growth with network expansion," *Transportation Research Part D: Transport and Environment*, vol. 73, no. 1, pp. 391–402, 2019.

[34] P. Noursalehi, H. N. Koutsopoulos, and J. Zhao, "Real time transit demand prediction capturing station interactions and impact of special events," *Transportation Research Part C: Emerging Technologies*, vol. 97, no. 1, pp. 277–300, 2018.

[35] D. An, X. Tong, K. Liu, and E. H. W. Chan, "Understanding the impact of built environment on metro ridership using open source in Shanghai," *Cities*, vol. 93, no. 1, pp. 177–187, 2019.

[36] V. V. Ivanov and E. S. Osetrov, "Forecasting daily passenger traffic volumes in the Moscow metro," *Physics of Particles and Nuclei Letters*, vol. 15, no. 1, pp. 107–120, 2018.

[37] N. Glisovic, M. Milenković, N. Bojović, L. Švadlenka, and Z. Avramović, "A hybrid model for forecasting the volume of passenger flows on Serbian railways," *Operational Research*, vol. 16, no. 2, pp. 271–285, 2016.

[38] E. Pekel and S. S. Kara, "Passenger flow prediction based on newly adopted algorithms," *Applied Artificial Intelligenc*, vol. 31, no. 1, pp. 64–79, 2017.

[39] J. C. García-Palomares, J. Gutiérrez, and O. D. Cardozo, "Walking accessibility to public transport: an analysis based on microdata and GIS," *Environment and Planning B: Planning and Design*, vol. 40, no. 6, pp. 1087–1102, 2013.

[40] 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.

[41] S. Mavoa, K. Witten, T. McCreanor, and D. O'Sullivan, "GIS based destination accessibility via public transit and walking in Auckland, New Zealand," *Journal of Transport Geography*, vol. 20, no. 1, pp. 15–22, 2012.

[42] S. Alalawi, S. Abdulwahab, and C. Bakheit, "Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone," *Environmental Modelling & Software*, vol. 23, no. 4, pp. 396–403, 2008.

[43] J. M. Morris, P. L. Dumble, and M. R. Wigan, "Accessibility indicators for transport planning," *Transportation Research Part A: General*, vol. 13, no. 2, pp. 91–109, 1979.

[44] W. G. Hansen, "How accessibility shapes land use," *Journal of the American Institute of Planners*, vol. 25, no. 2, pp. 73–76, 1959.

[45] Q. Shen, "Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers," *Environment and Planning B: Planning and Design*, Environment & Planning B Planning & Design, vol. 25, no. 3, pp. 345–365, 1998.

[46] B. P. Y. Loo, C. Chen, and E. T. H. Chan, "Rail-based transit-oriented development: lessons from New York city and Hong

Kong," *Landscape and Urban Planning*, vol. 97, no. 3, pp. 202–212, 2010.

[47] 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.

[48] C. R. Bhat and J. Y. Guo, "A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels," *Transportation Research Part B: Methodological*, vol. 41, no. 5, pp. 506–526, 2007.

[49] Z. Huang and X. Liu, "A hierarchical approach to optimizing bus stop distribution in large and fast developing cities," *ISPRS International Journal of Geo-Information*, vol. 3, no. 2, pp. 554–564, 2014.

[50] R. D. Yockey, *SPSS Demystified: A Step-by -step Guide to Successful Data Analysis*, Prentice Hall PTR, Englewood Cliffs, NJ, USA, 2007.