Article

Estimating Rail Transit Passenger Flow Considering Built Environment Factors: A Case Study in Shenzhen

Wenjing Wang [1,2] **, Haiyan Wang** [1,]*** **, Jun Liu** [2] **, Chengfa Liu** [3] **, Shipeng Wang** [1] **and Yong Zhang** [4]

1   School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan 430063, China;
    fannywang@whut.edu.cn (W.W.); wangsp1228@163.com (S.W.)
2   Institute of Transport Management, Guangdong City Technician College, Guangzhou 510520, China
3   Production Management Center, Shenzhen Metro Operation Group Co., Ltd., Shenzhen 518000, China
4   Shenzhen Research Institute, Northwestern Polytechnical University, Shenzhen 518063, China
*   Correspondence: hywang777@whut.edu.cn

Abstract: This paper aims to analyze the influence mechanism of built environment factors on passenger flow by predicting the passenger flow of Shenzhen rail transit in the morning peak hour. Based on the classification of built environment factors into socio-economic variables, built environment variables, and station characteristics variables, eight lines and one hundred sixty-six stations in Shenzhen Railway Transportation are taken as research objects. Based on the automatic fare collection (AFC) system data and the POI data of AMAP, the multiple regression model (OLS) and the geographically weighted regression (GWR) model based on the least squares method are established, respectively. The results show that the average house price is significantly negatively correlated with passenger flow. The GWR model considering the house price factor has a high prediction accuracy, revealing the spatial characteristics of the built-up environment in the administrative districts of Shenzhen, which has shifted from the industrial structure in the east to the commercial and residential structure in the west. This paper provides a theoretical basis for the synergistic planning of house price regulation and rail transportation in Shenzhen, which helps to develop effective management and planning strategies.

Keywords: rail transit; spatial heterogeneity; passenger flow forecast; built environment factor; ordinary least squares (OLS) regression model; geographically weighted regression (GWR) model

1. Introduction

As the global urbanization process continues to advance, rail transit systems have become a key means to alleviate urban traffic congestion, improve travel efficiency, and promote sustainable urban development. In China, in large cities like Shenzhen, rail transit plays a crucial role in carrying the demand for short-distance commuting and cross-city travel within the city. However, with the increasing improvement in the rail transit network, the spatial distribution characteristics of passenger flow and its complex interaction with the built environment around the stations have become increasingly important research topics in the field of transportation planning and management. Therefore, this study focuses on the Shenzhen rail transit system, exploring the influence mechanism of built environment factors around rail transit stations on the generation and distribution of passenger flow, which will provide theoretical support and a decision-making basis for the scientific planning and efficient management of urban transportation systems. At present, the research on the influence mechanism of built environment factors on the passenger flow of rail transit stations mainly focuses on analyzing influencing factors [1,2] and methods [3]. Influencing factors mainly include socio-economic and land use diversity [4]. Housing prices, as an important socio-economic influencing factor, affect the distribution and density of the urban built environment, which indirectly affects the generation of rail transit passenger flow. Meanwhile, land use diversity is a more commonly used built

environment influencing factor for exploring transit patronage. In previous studies, POI data have often been used to describe land diversity for passenger flow prediction [5] or to analyze the spatial heterogeneity of the influencing factors [6,7]. Some studies have also used POI data to validate the accuracy of the model, but few studies have used the spatial characteristics of the influencing factors of the rail traffic flow to further analyze the urban zoning policy.

Regarding research methodology, many current models examine how built environment factors affect passenger flow at rail transit stations. The four-stage model is a widely used method [8,9] in urban traffic flow prediction, which studies the traffic volume in the global scope, and the workload is large, it is difficult to respond to the changes in the built environment around the station in time, and the model prediction results are inaccurate. Although neural network models [10,11] are suitable for dealing with large-scale, high-dimensional datasets due to their strong fitting abilities, they face challenges in achieving good results because they require a large amount of high-quality data for training, and their high costs make medium- and long-term predictions challenging. At present, the OLS (Ordinary Least Squares) model considering global regression is also commonly used in such research [12,13]. Such models, while providing an analysis of overall trends, fail to adequately account for spatial heterogeneity in spatial parameters. In fact, the degree of influence of the influencing factors is not necessarily the same for different rail transit stations [14]. In this paper, we address the shortcomings of traditional models in analyzing spatial heterogeneity by introducing a geographically weighted regression (GWR) model [4,15]. The GWR model takes into account the spatial instability of passenger flow, and the model has been widely used in rail transit passenger flow prediction [16]. Therefore, it can be used to study the impact of built environment factors on passenger flow at urban rail transit stations.

Therefore, this paper focuses on the Shenzhen rail transit system as the research object and aims to analyze the influence of built environment factors on the passenger flow of rail transit stations. First, spatial autocorrelation tests and multicollinearity tests were conducted on the independent variables to ensure the validity of the model inputs. Second, housing prices were introduced into the OLS model and GWR model, and the mechanism of house price influence on rail transit patronage was explored through accuracy comparison. To further enhance the accuracy of the analysis, the K-means clustering algorithm was used to classify the rail transit stations into six types, including servicing businesses, scenic spots, public services, government and corporate offices, commercial housing, and transportation hub. Based on this classification, the performances of the OLS model and the GWR model were compared, and the advantages of the GWR model in capturing spatial heterogeneity were verified. Finally, through the visualization and analysis of the GWR model, this paper reveals the differences in the characteristics of each rail station under the influence of different built environment factors, and based on the analysis results, targeted recommendations are provided for the development of Shenzhen's administrative districts.

The innovations of this paper are as follows: (1) Unlike previous studies, this paper systematically explores for the first time the influence of economic factors such as housing prices on passenger flow in the Shenzhen rail transit system. By incorporating housing prices factor into the OLS model and the GWR model and comparing the accuracy of the performance of the two models, this study reveals the important role of housing prices in the generation of urban rail transit passenger flow. It also provides a new perspective for understanding the influence mechanism of economic factors on transportation travel decisions. (2) This paper reveals the spatial structural transformation characteristics of the built environment in Shenzhen. The GWR model is used to visualize and analyze rail transit stations, showing the shift in the built environment from the eastern industrial areas to the commercial and residential areas in the west. The analysis provides insights into the spatial characteristics of the built environment in Shenzhen from the industrial structure in the east to the commercial and residential structure in the west from the perspective of each administrative district.

Following the introduction, Section 2 provides a literature review of the built environment influences and modeling of rail transit patronage. Section 3 describes the research methodology used in this paper. Section 4 presents a case study of rail transit in Shenzhen. Section 5 presents the results and discussion of this study. Section 6 summarizes the full paper.

2. Literature Review

To more deeply analyze the spatial heterogeneity of various built environment factors on the passenger flow of Shenzhen's rail transit stations, two key aspects must be considered: the selection of factors influencing the passenger flow of rail transit stations and the use of modeling methods. This section reviews and summarizes both the built environment factors affecting passenger flow at rail transit stations in previous studies and the modeling approaches used in this study.

2.1. Selection of Built Environment Factors

Built environment factors affecting rail transit flow vary from city to city and from dataset to dataset. Kuby et al. (2004) [17] examined the number of people employed and residing within the buffer zones of rail transit stations in nine U.S. cities and found that there was a significant correlation with light rail patronage. Sohn and Shim (2010) [18] conducted a model analysis and comparison using the passenger flow data of Seoul rail transit and Kuby's research. It was found that the results of the two studies were similar, and it was found that employment, commercial building area, office building area, net population density, transfer times, and the number of feeder bus lines were significantly correlated with passenger flow. In China, city size is usually measured by population size [19]. In megacities, He et al. (2018) [20] found that commuting activities generate weekday metro traffic and leisure activities, such as weekend metro traffic generated by shopping using influencing factors such as the population distribution and number of office locations in Taipei City. Zhang et al. (2023) [21] used POI and AFC data to find that tourist attractions, other types of POI, bus stations, and station accessibility are significantly related to the passenger flow of Nanjing Metro rail transit. Among megacities, Li et al. [2] found that population density and common residential land use were the main factors affecting morning and evening commuter flows in Guangzhou by comparing the different effects of day, time, and directional passenger flows. An et al. (2019) [22] used OLS regression analysis and found that the effects of commercial land use, bus stops, and tourist attractions on rail transit patronage in Shanghai are independent of weekday time and all have significant positive effects. Wang et al. (2022) [23] analyzed the "7D" built environment variables and found that the density of office facilities, the density of sports and leisure facilities, the density of medical service facilities, the density of buildings, and the plot ratio have a significant effect on the passenger flow out of each metro station in Beijing.

Although these studies have analyzed in detail the mechanisms by which built environment factors affect rail transit patronage from the perspective of the characteristics of the respective cities, they have generally neglected the role of house prices as a key economic variable. House prices not only determine the spatial distribution of the built environment but also affect the geographic trend of patronage. Although Wang et al. (2023) [24] and Yang et al. (2023) [25] considered the factor of house price in their studies in Beijing and Chengdu, respectively, they did not bring house price into different models for comparative analysis to explore the significance of its impact on passenger flow. In addition, there is an even greater lack of studies on short-distance commuting-cum-cross-city access in megacities.

2.2. The Model Method Used

Currently, the passenger flow prediction models used by scholars for rail transit stations are mainly divided into global regression models and local regression models. The

traditional four-stage model has shortcomings in data accuracy and model application range. Its research area is global, and it is difficult to refine the impact of the building environment around the station on passenger flow. Given the limitations of the traditional four-stage model for predicting passenger flow, more and more scholars have begun to focus on the OLS (Ordinary Least Squares) model. He et al. [26] studied direct passenger flow forecasting models and concluded that the most widely used is the OLS linear regression model. However, the OLS regression model is also a global regression model, which does not take into account the spatially localized influences affecting station passenger flow or the spatial heterogeneity of the independent variables. Fotheringham (1996) [27] proposed the GWR (geographically weighted regression) model. The GWR model is based on the regression analysis of the OLS model taking into account the local characteristics of the independent variables, including spatial heterogeneity as well as spatial correlation, which can better deal with spatially variable data. Cardozo et al. (2012) [28] and Tu et al. (2018) [29] compared the effectiveness of fitting OLS and GWR models for rail transit passenger flow prediction, and they found that GWR was more suitable. They combined geographic and economic factors to explore the relationship between subway passenger flow and land use characteristics, providing a new research direction that justifies further research using the GWR model.

Previous studies have focused on the comparison of model accuracy, but few studies have combined the spatial heterogeneity of passenger flow influencing factors to give policy advice. Li et al. (2020) [30] classified Guangzhou rail transit stations through K-means clustering and gave policy advice for each type of station but did not give policy advice on a macrolevel. Zhang et al. (2023) [21] verified that the GWR model was more accurate than the OLS model by comparing the model accuracies but did not further utilize this model to give opinions on transportation policies in Nanjing. Analyzing the influence mechanism of the factors affecting passenger flow, taking administrative districts as an analytical perspective, and giving policy advice is of guiding significance to public transportation planning and urban land use, etc., and there is a relative lack of research on such.

3. Methods

3.1. Spatial Autocorrelation Analysis

To study the spatial heterogeneity of dependent variables and independent variables, we must firstly perform spatial autocorrelation analysis on each variable to ensure the spatial aggregation of variables. It is necessary to verify the existence of spatial positive correlation in the spatial distribution, using Moran's index to measure whether the data distribution of the variable has spatial autocorrelation or not, as shown in Equation (1):

$$I = \frac{\sum\limits_{i=1}^{s} \sum\limits_{j=1}^{s} \omega_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum\limits_{i=1}^{s} (x_i - \overline{x})^2} \tag{1}$$

where I is the Moran's index, and the value range is $[-1,1]$. When $I > 0$, it indicates that the variables show positive spatial correlation and the variable data are spatially aggregated, and the larger the value is, the more obvious the degree of spatial aggregation of variables is. When $I < 0$, it means that the variables show negative spatial correlation and the variable data are spatially discrete, and the smaller the value is, the more obvious the degree of spatial discretization of the variables is. When $I = 0$, it indicates that the variables are spatially uncorrelated, and the variable data are spatially randomly distributed. s is the number of rail transit stations; x_i and x_j are the independent variables of the ith subway station and the jth subway station, respectively; ω_{ij} is the spatial weight between the ith and jth subway station.

3.2. Multicollinearity Test

To avoid the existence of multicollinearity between independent variables, which affects the prediction effect of the regression model, the Pearson correlation coefficient is generally used to determine whether there is the existence of multicollinearity, as shown in Equation (2):

$$r = \frac{\sum_{i=1}^{s} (x_{ia} - \overline{x_a})(x_{ib} - \overline{x_b})}{\sqrt{\sum_{i=1}^{s} (x_{ia} - \overline{x_a})}\sqrt{\sum_{i=1}^{s} (x_{ib} - \overline{x_b})^2}} \tag{2}$$

In the formula, r is the Pearson correlation coefficient, whose positive and negative values represent positive and negative correlation, and the absolute value size represents the degree of correlation; when $|r| > 0.8$, it is considered that there is a high degree of multicollinearity between the two explanatory variables, and at this time, the variables with relatively high $|r|$ and high frequency of covariance with other independent variables are excluded until $|r| < 0.8$ among all independent variables; x_{ia} and x_{ib} represent the sample mean values of a and b, two types of explanatory variables in the ith subway station; $\overline{x_a}$ and $\overline{x_b}$ represent the sample means of the a and b explanatory variables, respectively.

3.3. Regression Model

The GWR model takes into account the spatial heterogeneity of the distribution of each rail transit station (as shown in Figure 1), and the regression coefficients for each subway station take different values due to the differences in geographic location, and the model expression is shown in the following Equation (3):

$$\hat{y}_i = \beta_0(u_i, v_i) + \sum_{t=1}^{m} \beta_{it}(u_i, v_i)x_{it} + \varepsilon_i \tag{3}$$

where $\hat{y}_i$ is the fitted value of the dependent variable passenger flow y at the station i, (u_i, v_i) are the spatial coordinates of the station i, and $\beta_{it}(u_i, v_i)$ is the coefficient of the regression analysis of the tth independent variable of the continuous function $\beta(u,v)$ at station i. The GWR model provides a measurable measure of spatial heterogeneity.
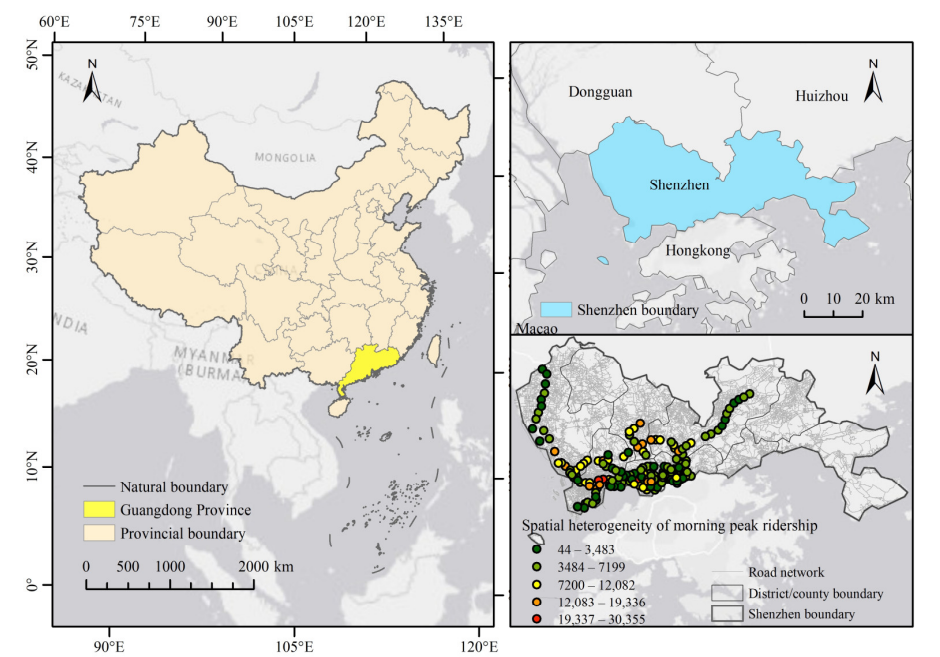

Figure 1. Map of the administrative area of Shenzhen, China.

In the GWR regression analysis, the basic principle of calculating the weights is as follows: the closer the distance, the higher the weight value or, conversely, the lower the weight value. The weights that can realize the monotonic reduction function of spatial distance are called kernel functions. The two most commonly used kernel functions are the Gaussian function and the Bi-square function. Considering the overall spatial distribution of urban rail transit stations in Shenzhen, the kernel function of the GWR model adopts the Bi-square function, which is shown in the following Equation (4):

$$\omega_{ij} = \begin{cases} \left[1 - (d_{ij}/b)^2\right]^2, & d_{ij} \leq b \\ 0, & d_{ij} > b \end{cases} \tag{4}$$

where ω_{ij} is the spatial weight of the estimated point j when fitting the model for subway station i; d_{ij} is the distance between the two; and b is the kernel bandwidth, which is the key control parameter and can be specified either by a fixed distance (i.e., fixed bandwidth) or a fixed number of nearest neighbors (i.e., adaptive bandwidth). The setting of the bandwidth plays a decisive role in the regression model in terms of both prediction accuracy and stability.

The optimal bandwidth can be found by minimizing a number of model fit goodness-of-fit diagnostics, and the choice of bandwidth values can be made by either CV (cross-validation) [31] or the AIC (Akaike information criterion) [32,33]. CV generally considers only the accuracy of the prediction, whereas the AIC will consider parsimony, which balances accuracy with complexity. It is mainly used to test and calibrate the level of fit of a regression model, and in experiments, a modified AIC, AIC_C, is usually used, which, unlike the basic AIC, is a function of sample size [34], i.e., the AIC_C will be infinitely close to the AIC when the sample size is large. The AIC_C formula [35] is shown in Equation (5) below:

$$AICC = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n + tr(S)}{n - 2 - tr(S)}\right\} \tag{5}$$

where n is the sample size; $\hat{\sigma}$ is the estimated standard deviation of the error term; and $tr(S)$ denotes the trajectory of the hat matrix S. The hat matrix is the projection matrix from the observation y to the fitted value $\hat{y}$, where for the GWR, each row r_i of this hat matrix is shown in Equation (6):

$$r_i = X_i(X^T W_i X)^{-1} X^T W_i \tag{6}$$

where X_i is the ith row of the matrix of independent variables X, and W_i is a diagonal matrix indicating the geographic weight of each observation at station i.

4. Research Case

4.1. Research Object

As a sub-provincial city in China's Guangdong Province, Shenzhen has been identified by the State Council as a special economic zone in China and is one of the most important cities in China, which is also an economic, transportation, and logistics center. It is connected to Hong Kong in the south and borders the cities of Dongguan and Huizhou in the north. Compared with other megacities such as Guangzhou, Shanghai, and Beijing, Shenzhen's average one-way commuting distance by rail is relatively short. The statistical report for 2023 shows only 8.5 km [36], which not only reflects Shenzhen's high degree of intensification in land use but also highlights the important role of the rail transit system in supporting short-distance travel. In addition, as an immigrant city with a large number of migrant workers, Shenzhen has a large population. At the same time, high housing price is another socio-economic attribute of Shenzhen, which has brought about the phenomenon of cross-city commuting in the city, increasing the need for convenient and efficient transportation connections. Therefore, it is of special significance for this paper to take Shenzhen as the research object and deeply explore how the built environment factors (including house prices) affect the generation and distribution of rail transit passenger flow.

Meanwhile, the relevant research data in this paper mainly come from Shenzhen Railway Transportation AFC data, totaling 25,267,919 credit card records, including information such as in/out time, in/out line, in/out station name, etc., as shown in Table 1.

Table 1. Shenzhen Metro AFC data.

Inbound Station	Inbound Line	Inbound Time	Outbound Station	Outbound Line	Outbound Time
Minzhi	Shenzhen Metro Line 5	18 June 2019 8:33	Honghu	Shenzhen Metro Line 7	18 June 2019 9:08
Qiaocheng North	Shenzhen Metro Line 2	18 June 2019 8:57	Lianhua West	Shenzhen Metro Line 2	18 June 2019 9:14
---	---	---	---	---	---
Lianhuacun	Shenzhen Metro Line 3	18 June 2019 15:38	Fumin	Shenzhen Metro Line 4	18 June 2019 15:55

In this paper, a total of 10 districts in Shenzhen are selected as the study area, and 8 lines and 166 stations in Shenzhen are selected as the study objects. The study sample includes the operable railroad subway stations within the administrative area of Shenzhen in June 2019 (see Figure 1 below).

4.2. Dependent Variable Data Processing

In this paper, Shenzhen Metro's AFC data for one consecutive week from 15 June to 21 June 2019 are selected. Figure 2 depicts the proportional distribution of the total passenger flow of all stations for each hour of the day for a week (considering the operating hours of the rail stations from 06:00 to 24:00), and the greater the total number of passengers in a given hour during operating hours, the larger the proportional value and the darker the color. The peak hour of passenger flow is mainly from 08:00 a.m. to 09:00 a.m. on weekdays. Therefore, the average passenger flow in and out of the station at 08:00–09:00 during the peak hours of the working day of Shenzhen Metro is taken as the dependent variable.
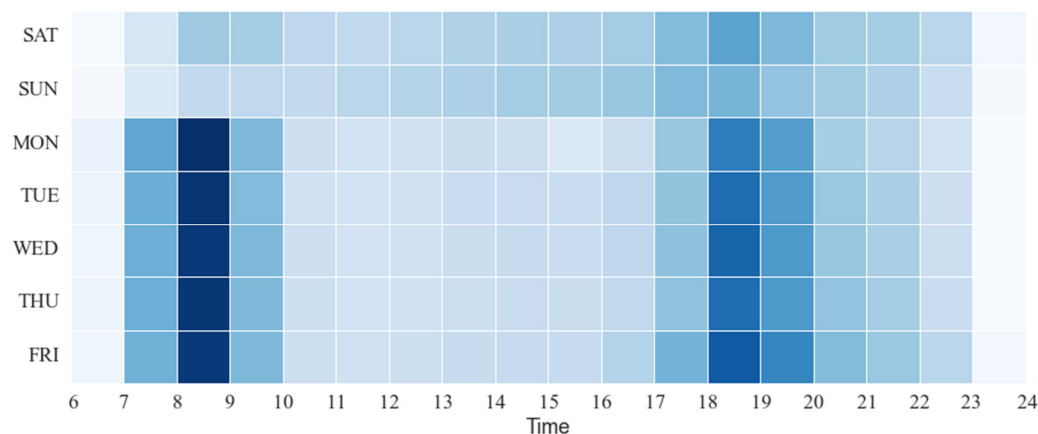

Figure 2. Percentage distribution of passengers per hour during operating hours.

Considering the passenger travel characteristics and regularity, the data were cleaned by Python version 3.10 to obtain the final data of the dependent variable, and the main processing methods were as follows:

(1) Extract the required fields for this study, delete the useless fields, and delete the records that do not match the station name properly.
(2) Delete records whose inbound time is later than the outbound time.
(3) Delete records whose inbound and outbound times are outside the range of subway operating hours.
(4) Delete records where the inbound and outbound dates are not on the same day (some lines operate across days).

(5) Delete records where the difference between the time of entry and exit is greater than 3.5 h; because the Shenzhen Metro has a supplemental charge of CNY 14 for use for more than 3.5 h, very few passengers use the subway for more than 3.5 h.

(6) Delete records with the same station in and out.

(7) Delete some of the fields missing records, such as fields with 0 or null records.

4.3. Independent Variable Data

4.3.1. Independent Variable Data Processing

Transportation-Oriented Development (TOD) plays a crucial role in urban construction and development. TOD is a method of public transportation planning and design that maximizes the use of residential and commercial areas [37]. The rail station buffer zone is the core area of TOD and is also the urban space where close urban development and mixed land use are concentrated. Rail station buffers are important nodes in urban development and key components of urban rail transit growth. Defining the buffer zone is the premise and basis for studies, such as assessing the built environment around stations and forecasting passenger flow at rail stations.

The area accessible to people within a 10 min walk from a rail transit station is generally regarded as the attraction area for rail transit passengers [38]. Combined with previous studies, the buffer zone radius of the rail transit station involved in this paper is set to 800 m [39–41], as shown in Figure 3a below. Considering the high-density characteristics of the built environment in Shenzhen, to avoid the overlapping of buffer zones, this paper utilizes the Tyson polygon to divide the buffer zone range [24], as shown in Figure 3b below. Through the division of the buffer zone, the factors affecting passenger flow at subway stations and related data are obtained.
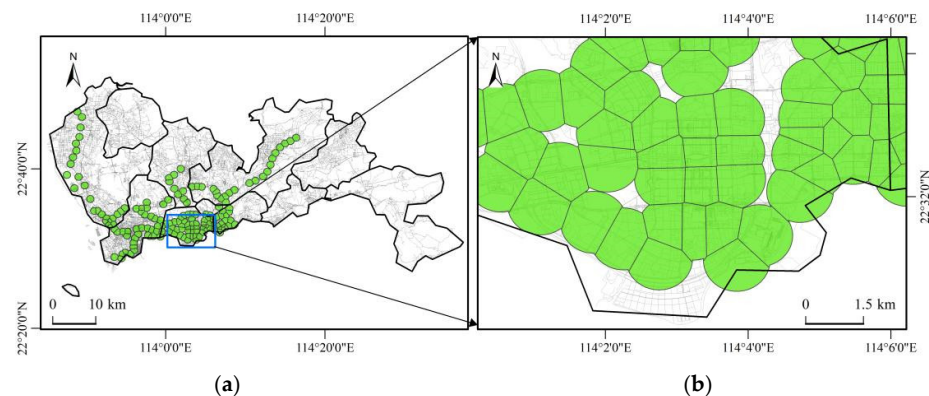

(**a**) (**b**)

Figure 3. Tyson polygon division radius is set to 800 m buffer. (**a**) Shenzhen subway 800 m buffer diagram; (**b**) Thiessen polygon partition buffer zoom chart.

Based on the multi-center characteristics of Shenzhen's urban areas and the commuting patterns of urban rail transit, and drawing from previous studies [21], this paper selects 15 built environment factors as the independent variables $X_1–X_{15}$. These variables are classified into three categories: socio-economic variables, built environment variables, and station characteristics [37]. To examine whether the average housing price affects the morning peak passenger flow of Shenzhen rail transit, the average housing price is included as an economic variable, and Table 2 describes these independent variables.

4.3.2. Socio-Economic Variables

Shenzhen has a large number of migrant workers and a small city area. The population size within the buffer zone is closely related to the passenger flow of urban rail transit, so it is an important factor affecting the passenger flow of rail transit stations. Reasonable house prices help to optimize the allocation of urban resources, improve the efficiency of land use, and optimize the built environment and then affect the station passenger flow [42].

Meanwhile, high house prices in Shenzhen bring about cross-city commuting. Housing prices can be another socio-economic factor in the influence of passenger flow.

Table 2. Description of statistics of independent variables.

Independent Variable	Variable Name
Socio-economic variables	
X_1	Population
X_2	Average house price
Built environment variables	
X_3	Servicing businesses
X_4	Scenic spots
X_5	Public services
X_6	Government and corporate offices
X_7	Commercial housing
X_8	Transportation hub
X_9	Floor area ratio
X_{10}	Land use mix
X_{11}	Road network density
Station characteristics variables	
X_{12}	Number of entrances and exits of rail transit stations
X_{13}	Accessibility
X_{14}	Bus lines
X_{15}	Bus stops

The population size was estimated from a website to determine the number of residents around each rail transit station, denoted as variable X_1. Variable X_2 represents the unit price of housing in residential neighborhoods around rail transit stations, which is calculated based on the number of resident households, as shown in Equation (7). Python was used to obtain data on the number of residential households and the average housing prices from Shell.com(accessed on 1 December 2023), a Chinese real estate service platform.

$$X_2 = \frac{\sum\limits_{i=1}^{t} P_i \times N_i}{\sum\limits_{i=1}^{t} N_i} \qquad (7)$$

where P_i is the average house price of cell i out of a total of t cells within the buffer zone, CNY/m^2, and N_i is the number of households in cell i out of a total of t cells within the buffer zone.

4.3.3. Building Environment Variables

Land use type POI is a major component of the built environment. In this study, we used Python to obtain POI data in Shenzhen, China, through the AMAP open platform. AMAP divides POI data into 23 categories. According to the type of POI and the standard of urban land use [43], land use is divided into six categories: servicing businesses, scenic spots, public services, government and corporate offices, commercial housing, and transportation hub, corresponding to variables X_3 to X_8.

The floor area ratio reflects the intensity of land use development. To some extent, it reflects the economic level and residential activities in the area, denoted as X_9. The formula is shown in Equation (8):

$$X_9 = \frac{F}{S} \qquad (8)$$

where F is the total area of all above-ground buildings within the rail site buffer zone area, m^2, and S is the area of the rail site buffer zone, m^2.

Land use mix can explain land use, and land use will affect residents' travel [44]. According to the existing research [45,46], the Shannon–Wiener index was used to characterize the land mixing degree as the independent variable X_{10}. The formula is shown in Equation (9):

$$X_{10} = \frac{-(\sum\limits_{j=1}^{n} p_j \times InP_j)}{InA} \tag{9}$$

where X_{10} is the land use mix within the buffer zone of the rail transit site; P_j is the proportion of POI type j within the buffer zone of the rail transit site; A is the total number of POI types, and according to the previously mentioned land use classification, A is denoted as 6 in this paper.

The road network density reflects the external accessibility of the rail transit station, reflecting the convenience of the residents to travel; it also affects the passenger flow of the rail transit station. The road network density is recorded as X_{11}, as shown in Equation (10).

$$X_{11} = \frac{\sum L}{S} \tag{10}$$

where S is the area of the buffer zone range of rail transit stations, km^2, and L is the length of each road section within the buffer zone range, km.

4.3.4. Station Characterization Variables

The number of entrances and exits, accessibility, bus lines, and bus stops of the rail stations within the buffer zone affect the transfer (connection) capacity of the stations with other modes of public transportation. They have certain impacts on the generation of passenger flow at the stations, which are noted as X_{12} to X_{15}. The raw data for X_{12}, X_{13}, X_{14}, and X_{15} were obtained from the OSM (2021) [21,47] websites, and ArcGIS was applied to trim the processing of the data in the buffer. The data for X_{13} are represented by using ArcGIS to calculate the average travel time for residents traveling by subway from other stations in the network to this station. It reflects the level of transportation economy as well as transportation facilities within the buffer zone of the rail transit station, which illustrates the opportunity, ability, and willingness of residents to travel.

By applying the data of each independent variable within the buffer zone of a rail transit station, the K-means clustering algorithm is used to calculate their cluster importance for station classification, as shown in Figure 4.
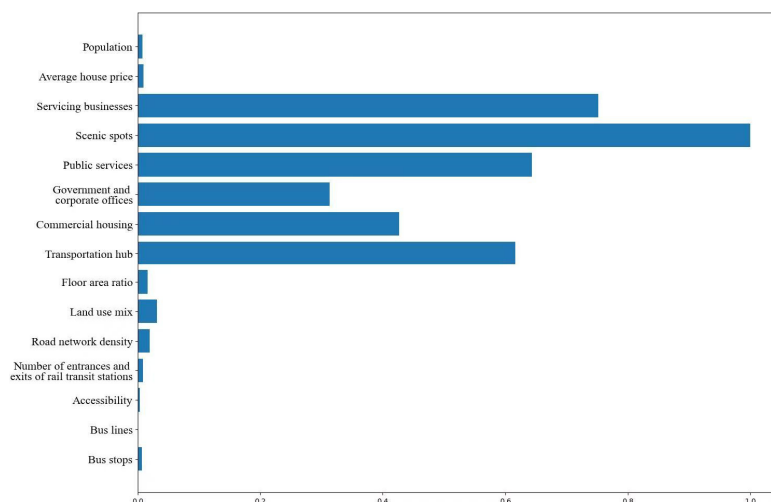

Figure 4. Clustering importance of independent variables affecting station passenger flow.

As can be seen from the analysis in Figure 4, there are six independent variables with high clustering importance, which are, in order from largest to smallest, as follows: scenic spots, servicing businesses, public services, transportation hub, commercial housing, and government and corporate offices. Therefore, the stations are classified into six types, which are servicing businesses, scenic spots, public services, government and corporate offices, commercial housing, and transportation hub.

5. Results and Discussion

To explore the degree of influence of average house price on the morning peak passenger flow of Shenzhen rail transit, the research ideas in this section are as follows: by adding average house price to the explanatory variables of the OLS model and GWR model, we compare the model fit before and after the addition of average house price to the two models and obtain the comparative analysis results. Due to the similarity of the calculation methods, this paper only lists the calculation process of calculated house prices, and the calculation results of uncalculated house prices will be listed in the model comparison.

5.1. Spatial Autocorrelation Test

The dependent variable and independent variable data are imported into ArcMap10.8, and the Moran's index of each variable is obtained through the "spatial autocorrelation" function, and the spatial autocorrelation test is performed. The results are shown in Table 3.

Table 3. Moran's I of the candidate variables.

Types of Variables	Moran's I	Expectation's Index	Variance	Z Score	*p* Value
Dependent variable					
Y	0.149558	−0.006061	0.001300	4.315885	0.000016
Independent variable					
X_1	0.365038	−0.006061	0.001325	10.195376	0.000000
X_2	0.150144	−0.006061	0.001326	4.289490	0.000018
X_3	0.258392	−0.006061	0.001059	8.128176	0.000000
X_4	0.220944	−0.006061	0.001097	6.854306	0.000000
X_5	0.257723	−0.006061	0.001024	8.245217	0.000000
X_6	0.272543	−0.006061	0.001204	8.028907	0.000000
X_7	0.241150	−0.006061	0.000914	8.178440	0.000000
X_8	0.327639	−0.006061	0.001025	10.421534	0.000000
X_9	0.469518	−0.006061	0.001333	13.025460	0.000000
X_{10}	0.087219	−0.006061	0.001329	2.558424	0.010515
X_{11}	0.495300	−0.006061	0.001332	13.739109	0.000000
X_{12}	0.090308	−0.006061	0.001282	2.691288	0.007118
X_{13}	0.454277	−0.006061	0.001308	12.729156	0.000000
X_{14}	0.063647	−0.006061	0.001254	1.968457	0.049015
X_{15}	0.127684	−0.006061	0.001324	3.675660	0.000237

According to the analysis of Table 3, the Z value of the dependent variable and the independent variable is greater than 1.96. The Moran's index is greater than zero, and the *p* value is less than 0.05, indicating that there is a significant spatial positive correlation between the dependent variable and the independent variable. It shows that variable data can be used to analyze the GWR model for explaining spatial heterogeneity.

5.2. Multicollinearity Analysis

Before the regression model analysis, the independent variables must pass the multicollinearity test. The independent variables with collinearity are screened by the Pearson correlation coefficient and variance expansion factor to ensure the effect of passenger flow prediction. Through correlation analysis, when the Pearson correlation coefficient is greater than 0.8, it shows that there is multicollinearity between the independent variables. After

the multicollinearity test, the correlation coefficients between variables such as servicing businesses (X_3), public services (X_5), government and corporate offices (X_6), commercial housing (X_7), and the transportation hub (X_8) are high (show Figure 5 below, where "*" denotes significant correlation, and "**" denotes highly significant correlation), indicating that there is multicollinearity between them.
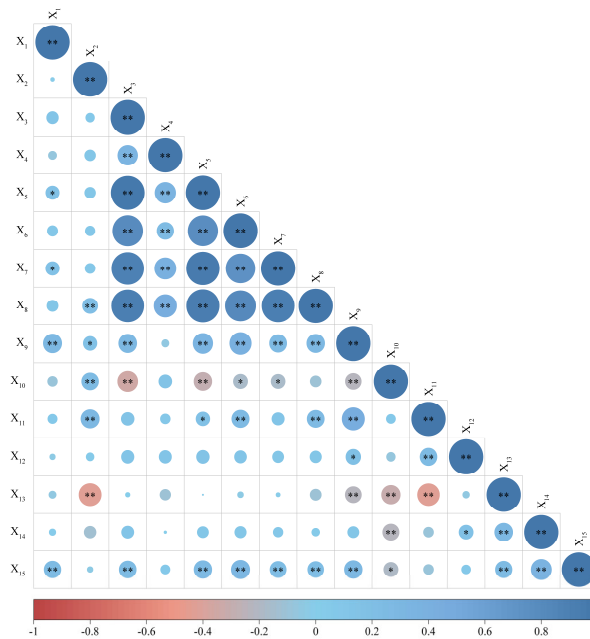

Figure 5. Correlation coefficients of independent variables.

In addition to considering multicollinearity, the independent variables that have a significant impact on the dependent variables are also selected. In this paper, the stepwise regression method is used to optimize the model by gradually increasing or decreasing the independent variables to find the most effective set of independent variables for the prediction of dependent variables. Finally, five independent variables are obtained: average house price, government and corporate offices, commercial housing, floor area ratio, and accessibility.

5.3. Comparison of Model Results

The OLS and GWR models were used to predict station ridership using weekday peak-hour inbound ridership as the dependent variable. Due to the different scales of the independent variables and the large quantitative differences between the scales, the independent variables are standardized in this paper to facilitate the comparison of the magnitude of the explanatory impact of the model's independent variables on the dependent variable. The processed results are shown in Table 4.

Table 4. Descriptive statistics of independent variable dimension.

Variable Type	Variable Name	Min	Max	Mean	STD
Dependent variable	Y	44	30,355	6441	5078
Independent variable	X_2	0.00	154,239.00	47,452.65	25,850.26
	X_6	4.00	4412.00	473.26	593.65
	X_7	3.00	3008	196.19	289.95
	X_9	0.03	4.68	1.75	0.82
	X_{13}	23.44	83.78	34.15	12.37

5.3.1. OLS Regression Analysis

After the spatial autocorrelation test, multicollinearity analysis, and standardization of the data of independent variables and dependent variables, IBM SPSS Statistics 27 software[48,49] was used for multiple linear regression modeling analysis. The results of OLS model regression analysis were obtained, as shown in Table 5 below.

Table 5. Regression results of the OLS model.

Variable	Unstandardized Coefficients				Collinearity Statistics	
	Coefficient	Standard Error	t	P	Tolerance	VIF
Constant	0.000	0.057	4.364	0.000	-	-
X_2	−0.163	0.064	−2.545	0.012	0.8	1.25
X_6	0.601	0.090	6.66	0.000	0.402	2.488
X_7	−0.183	0.085	−2.148	0.033	0.449	2.226
X_9	0.278	0.065	4.297	0.000	0.78	1.281
X_{13}	−0.229	0.065	−3.522	0.001	0.777	1.286
Residual sum of squares	86.949					
Log-likelihood	−181.871					
R^2	0.476					
Adjusted R^2	0.460					
AIC_C	378.451					

Note: Sig. = significance level of a variable. If sig is <0.05, it is significant.

As analyzed in Table 5, the *p* value of the t-test of regression coefficients of the five independent variables is less than 0.05, indicating that there is a significant correlation between these independent variables on the dependent variable. Meanwhile, the VIF values of all independent variables are lower than 7.5, indicating that there is no multicollinearity among the five independent variables. The R^2 value of the model is 0.476, which indicates that the model has some correlation and the model fit is average. The t-value of X_5 is 6.66, and the regression coefficient B is 0.601, both of which are the highest, indicating that the government and the corporate offices' high significance influences the passenger flow of the railroad transportation. The regression coefficient of X_2 has a negative value (B = −0.163), which indicates that the average house price is globally negatively correlated to the passenger flow.

5.3.2. GWR Regression Analysis

Using MGWR4.0 software, the regression results of GWR are obtained (see Table 6 below). The specific operation is as follows: Due to the faster convergence speed of GWR, this paper chooses GWR estimation as the initial estimation. The bandwidth length is adjusted according to the sample point adaptive kernel bandwidth determination method. The bandwidth selects the golden section mode and the spatial kernel function, and the bandwidth selection criterion selects the quadratic kernel function and the modified Akaike information criterion (AIC_C). Since the convergence criterion selects SOC-f to converge more strictly than SOC-RSS, SOC-f convergence is selected. To improve the prediction accuracy of the model, the order of magnitude of the convergence error is set to 10^{-5}. When the fluctuation range of the fitting coefficient is less than 10^{-5}, the iteration is completed.

From the analysis of Table 6, it is evident that the bandwidth of the independent variables is small, indicating that the explanatory variables have a significant spatial heterogeneity effect on passenger flow in Shenzhen's rail transit system as a whole, and the regression coefficients of each variable are significantly different. Log-likelihood measures the probability of data occurrence given model parameters. The larger the log-likelihood, the more likely the observed data are to occur under these parameters, meaning the model can better explain the data, resulting in a better fit. The log-likelihood (−131.691) of the GWR model is higher than that of the OLS model (−181.871). At the same time, the R^2 (0.714) adjusted R^2 (0.655) value and logarithmic probability of the GWR model are higher than those of the OLS model. The residual sum of squares and AIC_C are smaller

than those of the OLS model, indicating that the GWR model has a better fitting ability for describing the peak passenger flow of Shenzhen's rail transit system.

Table 6. Regression results of the GWR model.

			Regression Coefficient			
Variable	**Bandwidth**	**Minimum**	**Maximum**	**Mean**	**Median**	**STD**
Constant	71.000	−2.221	0.689	−0.317	−0.057	0.734
X_2	71.000	−0.314	−0.028	−0.150	−0.154	0.073
X_6	71.000	−0.209	1.257	0.482	0.603	0.315
X_7	71.000	−0.363	0.987	0.173	0.215	0.436
X_9	71.000	−0.121	0.597	0.334	0.337	0.170
X_{13}	71.000	−3.548	0.477	−0.799	−0.371	1.002
Residual sum of squares			47.501			
Log-likelihood			−131.691			
R^2			0.714			
Adjusted R^2			0.655			
AIC_C			334.509			

5.3.3. Comparative Analysis of Model Regression Results Considering Average House Prices

To explore the impact of average house price on the peak hourly passenger flow within a weekday of Shenzhen rail transit, in the study in this paper, the impact factors without average house price (X_2) are re-substituted into the OLS model as well as the GWR model to be analyzed against the methods of Section 5.3.1 and 5.3.2, and the results of the comparative analyses before and after the substitution of the average house price (X_2) are finally summarized and obtained (as shown in Table 7 below).

Table 7. Comparison of regression results of models considering average house price.

	Regression Coefficient			
	OLS Without X_2	**OLS With X_2**	**GWR Without X_2**	**GWR With X_2**
Residual sum of squares	88.188	86.949	49.916	47.501
Log-likelihood	−185.192	−181.871	−135.807	−131.691
R^2	0.463	0.476	0.699	0.714
Adjusted R^2	0.455	0.460	0.640	0.655
AIC_C	380.912	378.451	339.701	334.509

From the analysis in Table 7, it can be seen that the average house price has a significant negative correlation effect on the spatial trend of Shenzhen rail transit passenger flow, and the consideration of the average house price as an influencing factor can significantly improve the accuracy of the model in predicting the peak hour passenger flow of Shenzhen rail transit. In particular, the GWR model, which describes spatial heterogeneity, shows higher prediction accuracy than the traditional OLS model and can better capture the influence mechanism of built environment factors on the passenger flow of urban rail transit stations.

5.3.4. Comparative Analysis of Prediction Accuracy of OLS and GWR Model

In order to further verify the accuracy of the GWR model in predicting the passenger flow of rail transit, this paper uses the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) as the evaluation indexes and compares the prediction and simulation accuracy of the GWR model with the OLS model, as shown in Table 8 below.

Table 8. Comparison of model prediction results.

Type	Station	OLS		GWR	
		AE	RE	AE	RE
Commercial housing	Tangkeng	1310	17.33%	686	9.07%
	Meicun	820	11.65%	661	9.39%
Commercial services	Gangxia	2807	15.89%	1044	5.91%
	Shixia	1648	12.40%	224	1.69%
Government and corporate offices	Science Museum	2273	20.21%	1066	9.48%
	Shawei	2116	14.07%	314	2.09%
Public services	Caopu	1507	16.87%	875	9.80%
	Henggang	1270	13.18%	216	2.24%
Scenic spots	Children's Palace	1017	10.10%	62	0.61%
	University Town	1999	12.58%	904	5.69%
Transportation hub	Qianhaiwan	840	18.26%	667	14.50%
	Airport	509	12.94%	73	1.85%
MAE		4266		3021	
MAPE		77.57%		46.63%	
RMSE		5610		4147	

The lower accuracy of passenger flow prediction for commercial housing stations indicates that Shenzhen, as a special economic zone and an immigrant city [50], has high population mobility. The travel behavior of residents is affected by multiple factors, including the frequency of commercial activities, diverse transportation options, and the fast pace of urban life. This complex travel behavior increases the difficulty of prediction, resulting in the relatively low accuracy of passenger flow forecasts for residential stations. Passenger flow predictions for transportation hub stations show instability. Transportation hubs in the city center have higher prediction accuracy, whereas those further from the center have lower accuracy. This indicates that transportation hubs in central Shenzhen are surrounded by well-developed business districts, where passenger flow tends to be relatively concentrated and stable. Areas farther away from the city center are not as developed as the city center, and travel choices are more homogeneous with uncertainties. The higher accuracy and stability of passenger flow forecasts for servicing business and scenic spot stations indicate that the built environment around commercial stations in Shenzhen is strongly developed, attracting a large number of consumers and tourists. Traffic in these areas is more stable and predictable, as commercial activities tend to have a fixed regularity (e.g., commuting peaks, holiday shopping, etc.), which is consistent with the urban nature of Shenzhen as a national economic center city. Although there are natural landscapes in the scenic spot stations, the surrounding built environment has formed a more stable tourism economic circle due to the influence of urban special economic zones, and the passenger flow prediction model has a better fitting effect. The citywide passenger flow forecasts for both government and corporate offices and public service stations are relatively stable, illustrating Shenzhen's multi-cluster urban structure, with more balanced office and public service support facilities. Shenzhen has achieved a good balance between urban planning and administrative area development.

The above analysis illustrates that, compared with the OLS model, adding the average house price as the solution independent variable and applying the GWR model that takes into account spatial heterogeneity improves the accuracy in station ridership prediction and captures the spatial heterogeneity of the factors affecting the passenger flow at the station more effectively.

5.4. Spatial Heterogeneity of Local Influence Factor Coefficients in GWR Model

The GWR model analysis allows us to obtain the coefficients for each site, and we can observe the spatial distribution of and variation in these coefficients through the spatial structure map. To study the spatial heterogeneity of each influencing factor on the morning peak passenger flow in Shenzhen, this paper visualizes the distribution

of the local coefficients of each influencing factor by ArcMap10.8, which reacts to the spatial heterogeneity of the influencing factors' impacts on the rail transit passenger flow. Since Shenzhen is a multi-cluster center city, and its structure is mainly divided into regions [51], this paper discusses each administrative district of Shenzhen as a unit, as shown in Figures 6–10.
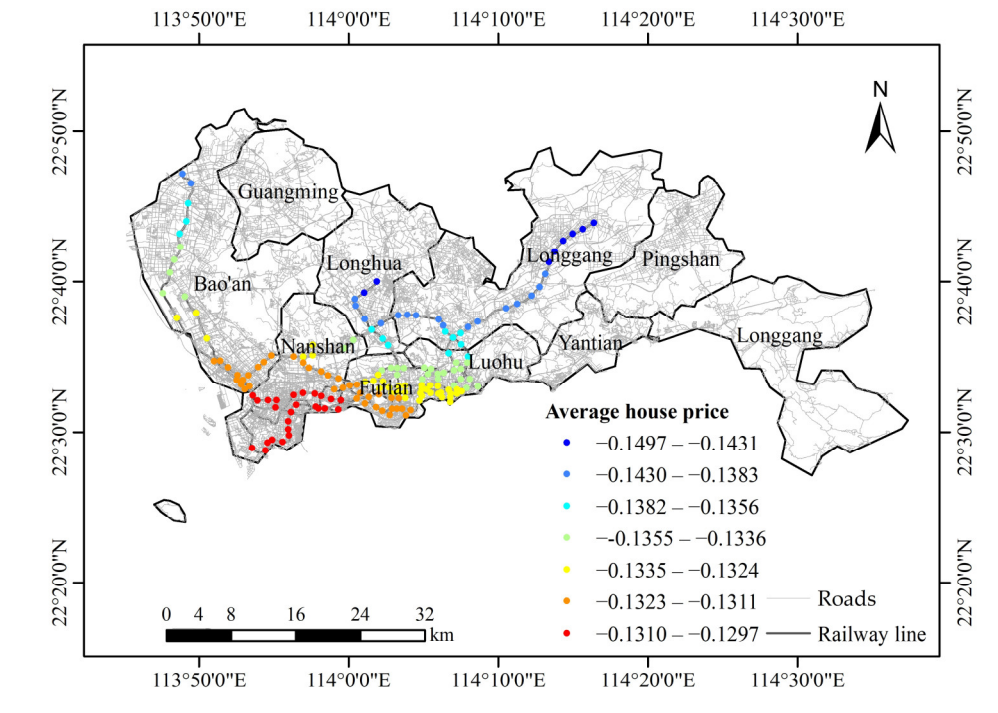

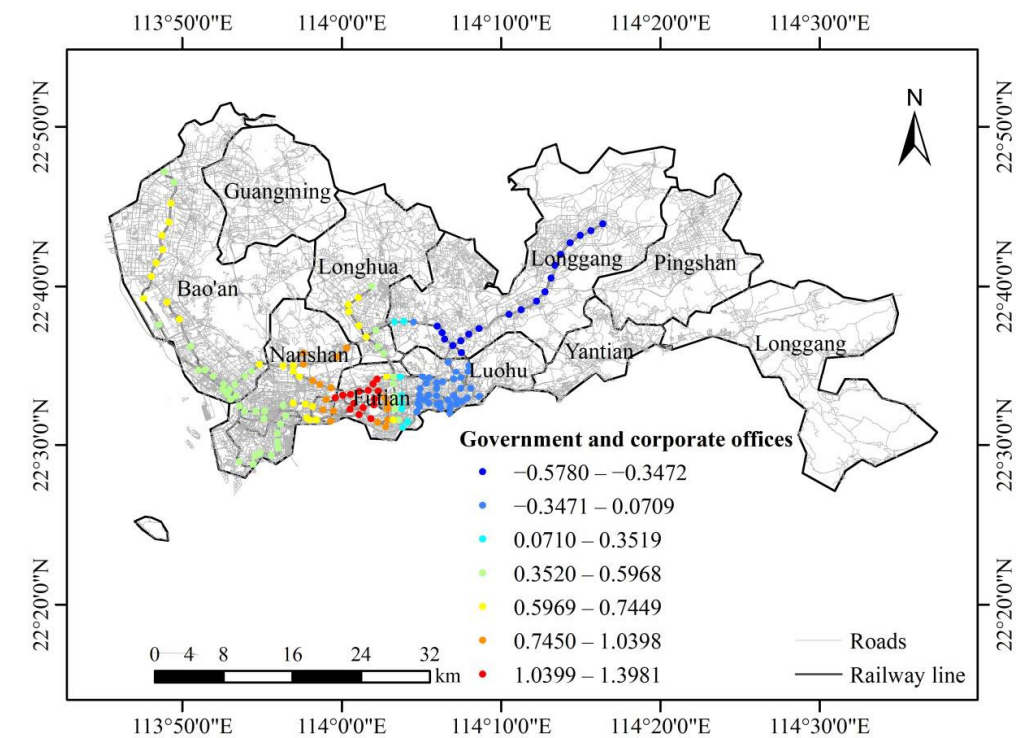**Figure 6.** Spatial distribution of local coefficients of average house price.

Figure 7. Spatial distribution of local coefficients of government and corporate offices.
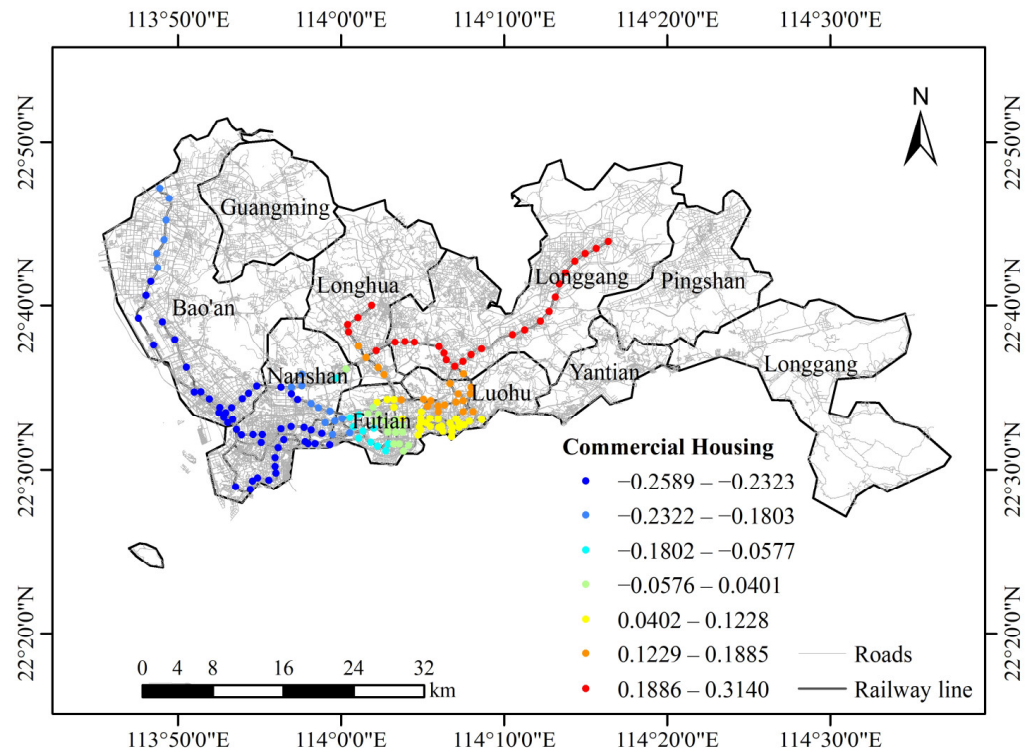
Figure 8. Spatial distribution of local coefficients of commercial housing.
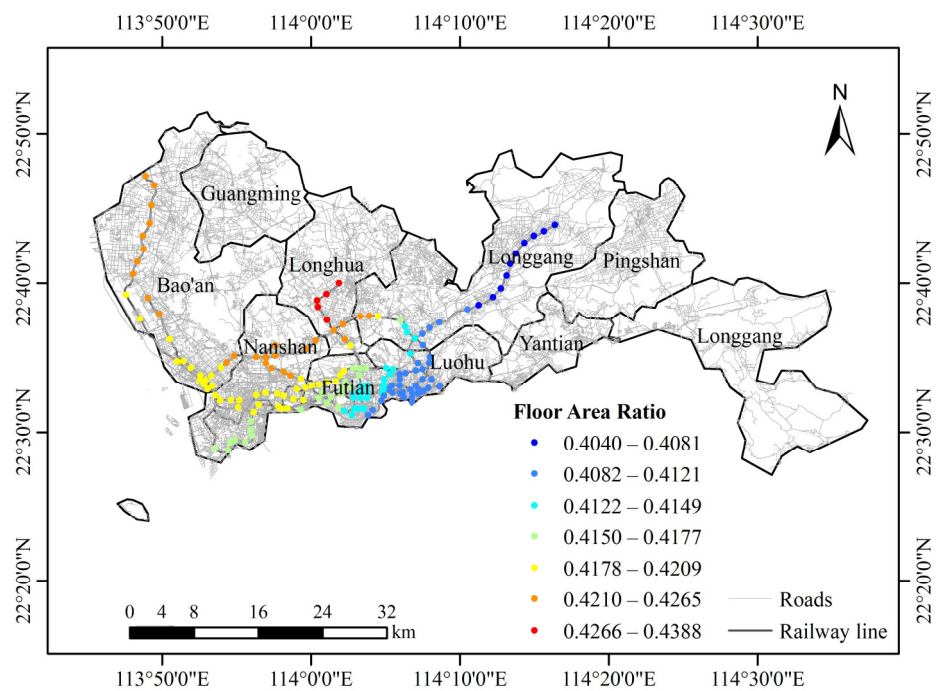
Figure 9. Spatial distribution of local coefficients of floor area ratio.
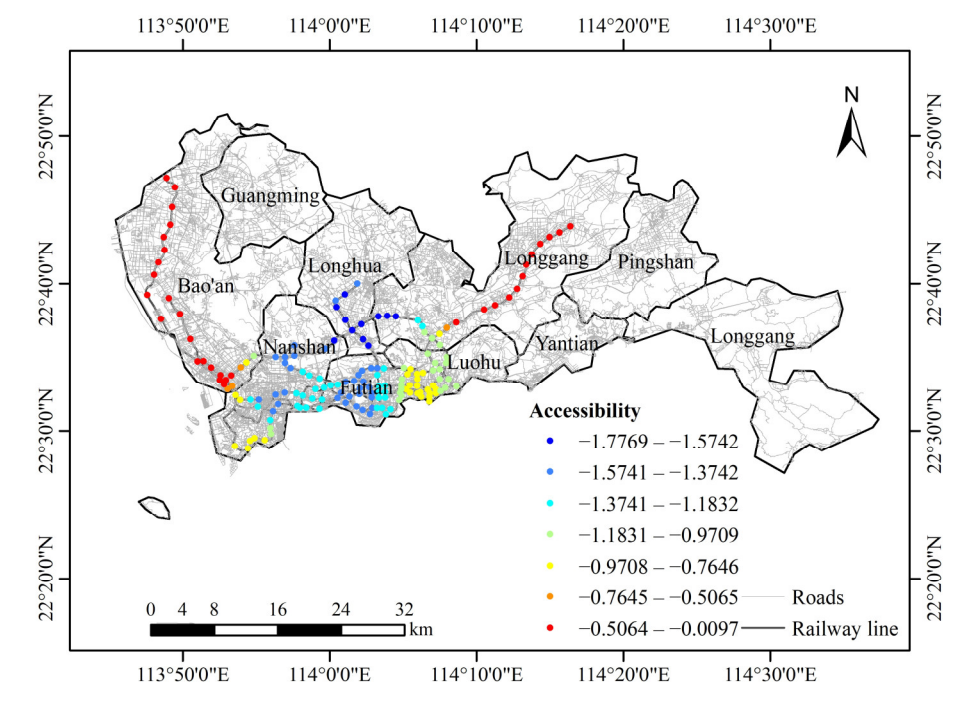
Figure 10. Spatial distribution of local coefficients of accessibility.

From the analysis in Figure 6, it is clear that the average house price has a globally negative impact on the morning peak passenger flow. This may be because there is little structural difference in the built environment of Shenzhen regions in general. Regions with high house prices have diversified transportation facilities, and residents may choose to travel by private car; according to the short one-way average morning peak distance in Shenzhen, those with a close morning peak distance may choose to travel on foot or by bicycle. The majority of migrant workers with low and middle income levels will choose to live in areas with low house prices, which can reflect the current situation of residents' transportation in Shenzhen.

The analysis in Figure 7 shows the spatial heterogeneity of the influence of government and corporate office locations on the passenger flow of Shenzhen rail transit. The western part of the Futian District has a higher influence coefficient due to its proximity to the core commercial CBD, while the eastern industrial area is less affected by the passenger flow despite its proximity to the commercial area. Extending southwestward to the Nanshan, Longhua, and Bao'an Districts, the positive impact coefficient decreases, and the impact of the northern section of Bao'an picks up due to its proximity to Dongguan, reflecting the attractiveness of the office-dense area to commuter flows. Extending northeastward to the Luohu and Longgang Districts, the negative impact coefficient increases, indicating that the high level of automation and fewer office workers in these industrial districts have a limited impact on rail traffic flow.

The analysis in Figure 8 shows the spatial heterogeneity of the influence of commercial housing on the passenger flow of Shenzhen rail transit. Starting from the eastern part of Futian District, radiating to the northeast towards the Luohu, Longgang and Longhua Districts, the positive correlation is enhanced, indicating that the proportion of residences in these districts rises and the residents rely on the rail transit due to the distance from the city center, the increase in industrial facilities, and the inadequacy of the transportation facilities. Radiating southwestward to the Nanshan and Bao'an Districts, the negative correlation is enhanced due to the concentration of high-tech enterprises and schools in the Nanshan District and the high residential density and inadequate transportation facilities in the Bao'an District, leading to the possibility that residents may choose other modes

of travel. The northern section of Bao'an is close to Dongguan, with a high number of cross-city commuters, and the negative correlation is weakened.

From the analysis in Figure 9, it can be seen that the positive effect of floor area ratio on the morning peak passenger flow of Shenzhen rail transit has a coefficient that is relatively consistent across stations. This reflects the fact that a higher floor area ratio in Shenzhen implies denser land use, which attracts more rail transit passenger flows. This effect shows a slow decreasing trend from the eastern to the western administrative regions. The extension of Shenzhen from east to west and the shift from industrial to residential and commercial structures are related to the evolution of the urban spatial structure of Shenzhen, which is "led by industry, followed by the construction of residential and commercial facilities" [52]. This spatial change may have an impact on traffic flow and the structure of passenger trips.

The analysis in Figure 10 shows the impact of rail transit station accessibility on the morning peak passenger flow in Shenzhen. The accessibility coefficients of the stations are generally negative, indicating that the accessibility of the stations is inversely proportional to the passenger flow, which shows that improving the accessibility of the stations can increase the passenger flow. The impact coefficient is higher in the western part of Futian District due to its proximity to the CBD business district, while it gradually decreases to the east and west and north to the Longhua District. This reflects that the transportation facilities in the center of Shenzhen are perfect, while the peripheral areas need to be upgraded. The impact coefficient of accessibility to the Longhua District rises, probably due to the large number of migrant workers, the relative lag of rail transportation facilities, and the outstanding traffic congestion problem [53]. It shows that the attractiveness and efficiency of rail transit can be improved by optimizing the transportation network and station connections.

5.5. Analysis and Recommendations

5.5.1. Impact of Housing Price on Morning Peak Passenger Flow

The *South China Morning Post* predicted on 4 July 2024 that in the second half of 2024 the number of migrant workers in Shenzhen would increase significantly [54]. Meanwhile, the official website of Shenzhen Metro noted on 18 July 2024 that in the first half of 2024 the average daily passenger traffic in Shenzhen reached 8,037,400, a record high for a single day's historical passenger flow [55]. According to news reports, Shenzhen housing prices are still experiencing a downward trend in the second half of the year [56]; combined with the previous analysis, the average housing prices have a significant negative correlation impact with the spatial trend of the average daily passenger flow in the morning peak of Shenzhen rail transit, and the hourly passenger flow in the morning peak of Shenzhen rail transit may continue to increase in the second half of 2024. The relevant management and operation departments should pay attention to this: the housing management department should reasonably regulate the level of house prices, and the rail transportation management department should control the passenger flow in a timely and reasonable manner.

5.5.2. Shenzhen City and Regional Synergy Optimization

As a multi-centered city, Shenzhen generally follows the development pattern of "taking industry as the first step, followed by the gradual construction of residential and commercial facilities". Each region has its own unique development characteristics and functional positioning, and the following conclusions and recommendations on the development of administrative districts can be drawn from the perspectives of urban development and regional planning to ensure the synergistic development of Shenzhen's regions and cities.

As the core commercial CBDs of Shenzhen, the Nanshan District and Futian District remain the backbone of the city's economic development. These two districts are not only the concentration of financial, commercial, and high-tech enterprises but also the core area of urban development. North of the Bao'an District, due to its border with Dongguan City,

government and corporate offices, commercial residences, and other influences in this area show a more pronounced spatial heterogeneity, which may be mainly due to the differences between Shenzhen and Dongguan Cities in terms of their economic development patterns and industrial structures. The two regions should engage in economic cooperation on many fronts to promote industrial synergy and optimize regional resource allocation to achieve more balanced and sustainable development. The Longgang District is mostly industrial land and faces a demand for industrial upgrading and urban renewal. By increasing the proportion of office land and raising the population density per unit of office land, it can effectively promote the diversification of the regional economy and increase passenger traffic. The south side of the Nanshan and Bao'an Districts, due to the large number of residential areas, office locations, and schools, needs to take a series of measures to solve the problem of large passenger flow in the morning peak promptly, such as optimizing the public transport network, increasing investment in transport facilities, and improving traffic management facilities to improve the travel efficiency of the residents. For administrative districts farther away from the downtown commercial CBD area, especially the Longhua District, the traffic congestion situation is a concern; the Longhua District and other districts farther away from the downtown commercial CBD area can be considered within the scope of the opening of a new rail station or other modes of transportation to improve the accessibility of the site to improve the region's subway station passenger flow, which has a very significant role.

6. Conclusions

Based on the POI data and AFC data around the stations, this paper explores the influence mechanism of passenger flow and the surrounding built environment factors in Shenzhen rail transit stations. Spatial autocorrelation, multicollinearity, and stepwise regression are used to filter out the key independent variables affecting station patronage. The GWR model, which considers average housing prices as an independent variable, is found to be more accurate in predicting passenger flow. This conclusion is further verified by comparing the prediction results of the OLS and GWR models for different types of stations. Taking administrative districts as the perspective, the spatial heterogeneity of the influence of these factors on rail traffic flow is analyzed by visualizing the local regression coefficients of each influential factor in the GWR model. Based on the results of this study, policy recommendations for urban transportation planning and housing price regulation in Shenzhen are proposed:

(1) Average house prices have a global negative correlation effect on ridership. Lower housing prices are usually accompanied by higher metro ridership. Nowadays, with the increase in migrant workers and passenger flow in Shenzhen, the government should reasonably regulate the level of housing prices to balance the demand of residents for housing and transportation resources.

(2) Based on OLS and GWR model calculations, and visualizing the local regression coefficients of each influential factor in the GWR model, the following is evident: the degree of influence of the influential factors in descending order is the building volume ratio, accessibility, commercial residences, and the average house price. In addition to average housing prices, which have a global negative correlation with the morning peak passenger flow, the building volume ratio has a global positive correlation, while accessibility has a global negative correlation. Government and corporate offices, as well as commercial housing, have significant spatial heterogeneity effects on Shenzhen rail transit passenger flow.

(3) The building volume ratio reflects the intensity of land use and the level of development of the built environment [57]. By visualizing the local coefficients of the building volume ratio, the spatial characteristics of Shenzhen's built environment are revealed, showing a shift from industrial areas in the east to commercial and residential structures in the west, from the perspective of each administrative district.

The shortcomings of this paper are that, although this paper puts forward policy recommendations on urban transportation planning and housing price regulation based on built environment factors, relevant policy factors were not directly quantified or incorporated into the analytical framework during model construction. At the same time, for a city like Shenzhen, which is characterized by both intra-city short-distance commuting and cross-city long-distance commuting, this study does not adequately discuss the impact of the characteristics of travel behaviors on passenger demand under this special commuting mode. In future research, AFC data can be further identified based on POI, and policy factors and travel characteristics can be quantified and included in the analysis to predict passenger flow concerning the regional policies and travel behavior characteristics of Shenzhen. Moreover, the influence of the built environment and other variable factors on passenger flow heterogeneity at rail transit stations during different periods (e.g., morning and evening peaks, weekdays, weekends, holidays) can also be considered to provide a theoretical basis for better urban transportation planning and development.

References

1. Ma, X.; Zhang, J.; Ding, C.; Wang, Y. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Comput. Environ. Urban Syst.* **2018**, *70*, 113–124. [CrossRef]
2. Li, S.; Lyu, D.; Liu, X.; Tan, Z.; Gao, F.; Huang, G.; Wu, Z. The varying patterns of rail transit ridership and their relationships with fine-scale built environment factors: Big data analytics from Guangzhou. *Cities* **2020**, *99*, 102580. [CrossRef]
3. Liu, Q.; Zhao, P.; Xiao, Y.; Zhou, X.; Yang, J. Walking Accessibility to the Bus Stop: Does It Affect Residential Rents? The Case of Jinan, China. *Land* **2022**, *11*, 860. [CrossRef]
4. Jun, M.-J.; Choi, K.; Jeong, J.-E.; Kwon, K.-H.; Kim, H.-J. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *J. Transp. Geogr.* **2015**, *48*, 30–40. [CrossRef]
5. Nyunt, K.T.K.; Wongchavalidkul, N. Evaluation of Relationships Between Ridership Demand and Transit-Oriented Development (TOD) Indicators Focused on Land Use Density, Diversity, and Accessibility: A Case Study of Existing Metro Stations in Bangkok. *Urban Rail Transit* **2020**, *6*, 56–70. [CrossRef]
6. Chen, E.; Ye, Z.; Wang, C.; Zhang, W. Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data. *Cities* **2019**, *95*, 102359. [CrossRef]
7. Liu, Z.; Liu, J.; Hu, R.; Yang, B.; Huang, X.; Yang, L. Calendar events' influence on the relationship between metro ridership and the built environment: A heterogeneous effect analysis in Shenzhen, China. *Tunn. Undergr. Space Technol.* **2023**, *141*, 105388. [CrossRef]
8. Chalumuri, R.S.; Nath, R.; Errampalli, M. Development and evaluation of an integrated transportation system: A case study of Delhi. *Proc. Inst. Civ. Eng.* **2018**, *171*, 75–84. [CrossRef]
9. Zhang, N.; Wang, Z.; Chen, F.; Song, J.; Wang, J.; Li, Y. Low-Carbon Impact of Urban Rail Transit Based on Passenger Demand Forecast in Baoji. *Energies* **2020**, *13*, 782. [CrossRef]
10. Liu, L.; Chen, R.C. A novel passenger flow prediction model using deep learning methods. *Transp. Res. Part C Emerg. Technol.* **2017**, *84*, 74–91. [CrossRef]

11. Li, Z.; Wang, X.; Xu, C.H. Novel Hybrid Spatiotemporal Convolution Neural Network Model for Short-Term Passenger Flow Prediction in a Large-Scale Metro System. *J. Transp. Eng. Part A. Syst.* **2024**, *150*, 04024016. [CrossRef]

12. Walters, G.; Cervero, R. *Forecasting Transit Demand in a Fast Growing Corridor: The Direct-Ridership Model Approach*; Fehrs and Peers Associates: Walnut Creek, CA, USA, 2003.

13. Loo, B.P.; Chen, C.; Chan, E.T. Rail-based transit-oriented development: Lessons from New York City and Hong Kong. *Landsc. Urban Plan.* **2010**, *97*, 202–212. [CrossRef]

14. Sung, H.; Choi, K.; Lee, S.; Cheon, S. Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership. *J. Transp. Geogr.* **2014**, *36*, 134–140. [CrossRef]

15. Blainey, S. Trip end models of local rail demand in England and Wales. *J. Transp. Geogr.* **2010**, *18*, 153–165. [CrossRef]

16. Yang, H.; Ruan, Z.; Li, W.; Zhu, H.; Zhao, J.; Peng, J. The impact of built environment factors on elderly people's mobility characteristics by metro system considering spatial heterogeneity. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 315. [CrossRef]

17. Kuby, M.; Barranda, A.; Upchurch, C. Factors influencing light-rail station boardings in the United States. *Transp. Res. Part A Policy Pract.* **2004**, *38*, 223–247. [CrossRef]

18. Sohn, K.; Shim, H. Factors generating boardings at Metro stations in the Seoul metropolitan area. *Cities* **2010**, *27*, 358–368. [CrossRef]

19. Qi, W.; Deng, Y. How to define the city size in China? A review over a century from 1918 to 2020. *Cities* **2024**, *144*, 104649. [CrossRef]

20. He, Y.; Zhao, Y.; Tsui, K.L. An Analysis of Factors Influencing Metro Station Ridership: Insights from Taipei Metro. In Proceedings of the 21st IEEE International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 1598–1603.

21. Zhu, Z.; Zhang, Y.; Qiu, S.; Zhao, Y.; Ma, J.; He, Z. Ridership prediction of urban rail transit stations based on AFC and POI data. *J. Transp. Eng. Part A Syst.* **2023**, *149*, 04023077. [CrossRef]

22. An, D.; Tong, X.; Liu, K.; Chan, E.H. Understanding the impact of built environment on metro ridership using open source in Shanghai. *Cities* **2019**, *93*, 177–187. [CrossRef]

23. Wang, Z.; Song, J.; Zhang, Y.; Li, S.; Jia, J.; Song, C. Spatial Heterogeneity Analysis for Influencing Factors of Outbound Ridership of Subway Stations Considering the Optimal Scale Range of "7D" Built Environments. *Sustainability* **2022**, *14*, 16314. [CrossRef]

24. Wang, J.; Wan, F.; Dong, C.; Yin, C.; Chen, X. Spatiotemporal effects of built environment factors on varying rail transit station ridership patterns. *J. Transp. Geogr.* **2023**, *109*, 103597. [CrossRef]

25. Yang, L.; Yu, B.; Liang, Y.; Lu, Y.; Li, W. Time-varying and non-linear associations between metro ridership and the built environment. *Tunn. Undergr. Space Technol.* **2023**, *132*, 104931. [CrossRef]

26. He, Y.; Zhao, Y.; Tsui, K.-L. Geographically modeling and understanding factors influencing transit ridership: An empirical study of Shenzhen metro. *Appl. Sci.* **2019**, *9*, 4217. [CrossRef]

27. Fotheringham, A.S.; Charlton, M.E. Geographically weighted regression: A natural evolution of the expansion method for spatial data. *Environ. Plan. A* **1998**, *30*, 1905–1927. [CrossRef]

28. Cardozo, O.D.; García-Palomares, J.C.; Gutiérrez, J. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Appl. Geogr.* **2012**, *34*, 548–558. [CrossRef]

29. Tu, W.; Cao, R.; Yue, Y.; Zhou, B.; Li, Q.; Li, Q. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *J. Transp. Geogr.* **2018**, *69*, 45–57. [CrossRef]

30. Li, S.; Lyu, D.; Huang, G.; Zhang, X.; Gao, F.; Chen, Y.; Liu, X. Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China. *J. Transp. Geogr.* **2020**, *82*, 102631. [CrossRef]

31. Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836. [CrossRef]

32. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*; Springer: New York, NY, USA, 1998.

33. Bowman, A.W. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **1984**, *71*, 353–360. [CrossRef]

34. Hurvich, C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. B* **1998**, *60*, 271–293. [CrossRef]

35. Lu, B.; Charlton, M.; Harris, P.; Fotheringham, A.S. Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house price data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 660–681. [CrossRef]

36. Sohu. China's Major Cities Commuting Monitoring Report 2023. Available online: https://www.sohu.com/a/714074984_121757514 (accessed on 24 August 2023).

37. Yang, H.; Li, X.; Li, C.; Huo, J.; Liu, Y. How do different treatments of catchment area affect the station level demand modeling of urban rail transit? *J. Adv. Transp.* **2021**, *2021*, 2763304. [CrossRef]

38. Li, L.; Gao, T.; Wang, Y.; Jin, Y. Evaluation of public transportation station area accessibility based on walking perception. *Int. J. Transp. Sci. Technol.* **2023**, *12*, 640–651. [CrossRef]

39. Zhou, J.; Yang, Y. Does yesterday's accessibility shape today's TOD-nesses in metro station areas? A tale of Shenzhen, China. *J. Transp. Land Use* **2024**, *17*, 483–510. [CrossRef]

40. Jiang, Y.; Gu, P.; Cao, Z.; Chen, Y. Impact of Transit-Oriented Development on Residential Property Values around Urban Rail Stations. *Transp. Res. Rec.* **2020**, *2674*, 362–372. [CrossRef]

41. Yang, Y.; Zeng, J.; Yin, J.; Wu, P.; Xu, G.; Jing, C.; Zhou, J.; Wen, X.; Reinders, J.; Amatyakul, W.; et al. Metro Stations as Catalysts for Land Use Patterns: Evidence from Wuhan Line 11. *Sustainability* **2024**, *16*, 6320. [CrossRef]
42. Wang, C.-H.; Chen, N. A geographically weighted regression approach to investigating local built-environment effects on home prices in the housing downturn, recovery, and subsequent increases. *J. Hous. Built Environ.* **2020**, *35*, 1283–1302. [CrossRef]
43. *GB 50137-2011*; Design, C. Code for classification of urban land use and planning standards of development land. State administration of quality supervision. National Standard of the people's Republic of China: Beijing, China, 2010.
44. Zhang, M. The role of land use in travel mode choice: Evidence from Boston and Hong Kong. *J. Am. Plan. Assoc.* **2004**, *70*, 344–360. [CrossRef]
45. Gustavsson, E.; Lennartsson, T.; Emanuelsson, M. Land use more than 200 years ago explains current grassland plant diversity in a Swedish agricultural landscape. *Biol. Conserv.* **2007**, *138*, 47–59. [CrossRef]
46. Dobbs, C.; Kendal, D.; Nitschke, C. The effects of land tenure and land use on the urban forest structure and composition of Melbourne. *Urban For. Urban Green.* **2013**, *12*, 417–425. [CrossRef]
47. Open Street Map. Available online: https://www.openstreetmap.org/ (accessed on 15 September 2024).
48. Dunn, P. SPSS survival manual: A step by step guide to data analysis using IBM SPSS. *Aust. New Zealand J. Public Health* **2013**, *37*, 597. [CrossRef]
49. Hayes, A.F.; Matthes, J. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behav. Res. Methods* **2009**, *41*, 924–936. [CrossRef] [PubMed]
50. An, M.; Zheng, C.; Li, H.; Chen, L.; Yang, Z.; Gan, Y.; Han, X.; Zhao, J.; Shang, H. Independent epidemic patterns of HIV-1 CRF01_AE lineages driven by mobile population in Shenzhen, an immigrant city of China. *Virus Evol.* **2021**, *7*, veab094. [CrossRef]
51. Shen, S.; Wu, C.; Gai, Z.; Fan, C. Analysis of the spatiotemporal evolution of the net carbon sink efficiency and its influencing factors at the city level in three major urban agglomerations in China. *Int. J. Environ. Res. Public Health* **2023**, *20*, 1166. [CrossRef] [PubMed]
52. Sun, H.; Li, X.; Guan, Y.; Tian, S.; Liu, H. The evolution of the urban residential space structure and driving forces in the megacity—A case study of Shenyang City. *Land* **2021**, *10*, 1081. [CrossRef]
53. Plus, S. The Metro Is Crowded with People and the Car Is Jammed! Shenzhen Citizens Complain: 'Centre of the Universe' Longhua Was Abandoned! Available online: https://www.163.com/dy/article/HCAS9I0D05539IXV.html (accessed on 17 August 2017).
54. South China Morning Post. Unemployment Spike in Shenzhen: 'Blip' or a Chip at China's Economic Might? Available online: https://www.scmp.com/economy/economic-indicators/article/3269049/jobless-spike-chinas-shenzhen-temporary-blip-or-chip-economic-might (accessed on 4 July 2024).
55. Metro, S. Shenzhen Metro Carried 1.463 Billion Passengers in the First Half of the Year, an Increase of More Than 17% Year on Year. Available online: https://www.szmc.net/home/xinwenzhongxin/gongsixinwen/202407/104171.html (accessed on 18 July 2024).
56. JIWU. 2024 Shenzhen House Price Trend Is Stable, the Property Market Is Still in a Downward Trend. Available online: https://sz.jiwu.com/news/5618459.html (accessed on 18 June 2024).
57. Lu, S.; Shi, C.; Yang, X. Impacts of Built Environment on Urban Vitality: Regression Analyses of Beijing and Chengdu, China. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4592. [CrossRef]