Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China

Shaoying Li, Dijiang Lyu, Guanping Huang, Xiaohu Zhang*, Feng Gao, Yuting Chen, Xiaoping Liu

School of Geographical Sciences, Guangzhou University, Guangzhou, Guangdong, PR China
Guangdong Guodi Planning Technology Co., Ltd, Guangzhou, Guangdong, PR China
Senseable City Lab, Massachusetts Institute of Technology, USA
School of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, Guangdong, PR China

Abstract

Understanding the relationship between the rail transit ridership and the built environment is crucial to promoting transit-oriented development and sustainable urban growth. Geographically weighted regression (GWR) models have previously been employed to reveal the spatial differences in such relationships at the station level. However, few studies characterized the built environment at a fine scale and associated them with rail transit usage. Moreover, none of the existing studies attempted to categorize the stations for policy-making considering varying impacts of the built environment. In this study, taking Guangzhou as an example, we integrated multi-source spatial big data, such as high spatial resolution remote sensing images, points of interest (POIs), social media and building footprint data to precisely quantify the characteristics of the built environment. This was combined with a GWR model to understand how the impacts of the fine-scale built environment factors on the rail transit ridership vary across the study region. The k-means clustering method was employed to identify distinct station groups based on the coefficients of the GWR model at the local stations. Policy zoning was proposed based on the results and differentiated planning guidance was suggested for different zones. These recommendations are expected to help increase rail transit usage, inform rail transit planning (to relieve the traffic burden on currently crowded lines), and re-allocate industrial and living facilities to reduce the commute for the residents. The policy and planning implications are crucial for the coordinated development of the rail transit system and land use.

1. Introduction

Accelerated global urbanization has facilitated continuous urban population growth and led to a rapid increase in the number of private vehicles, which in turn has resulted in tremendous traffic congestion in global metropolises (Chiou et al., 2015; Sung et al., 2014). Worldwide, urban rail transit is regarded as an efficient form of public transportation that helps alleviate traffic congestion (Li et al., 2018; Zhao et al., 2013). Currently, China is in a rapid growth phase in terms of rail transit planning and construction. According to the “13th Five-Year Plan for the Development and Planning of a Modern and Comprehensive Transportation System” issued in March 2017 by the State Council of the People’s Republic of China, the length of China’s urban railway operating network in 2020 will be double what it was in 2015. As China is undertaking one of the most ambitious rail transit planning projects in the world, it is important to ensure that the planned rail transit stations would meet the commuting requirements of residents of the metropolises (Zhao et al., 2014). Therefore, identifying the determinants of the rail transit ridership plays an important role in rail transit planning and operation (Chiang et al., 2011). In addition, exploring the mechanism of travel demand can provide guidance for the spatial planning of complex land use areas in the vicinity of the rail stations. This is also essential for promoting the coordinated development of transportation and land use in metropolises in China.

The influences of the built environment on rail transit ridership are complex. A substantial body of literature on transportation has looked into the relationship between the travel demand and the built environment which is usually measured by factors such as density, diversity, and design. (Cervero and Kockelman, 1997; Ewing and Cervero, 2001; Ewing and Cervero, 2010; Wang and Zhou, 2017). Empirical

https://doi.org/10.1016/j.jtrangeo.2019.102631
Received 23 July 2019; Received in revised form 3 December 2019; Accepted 28 December 2019
0966-6923/© 2019 Published by Elsevier Ltd.
studies have explored the impacts of the built environment on station-level rail transit by using the direct ridership models (DRMs) owing to their advantages of lower cost and faster response than the four-step models (Walters and Cervero, 2003; Kuby et al., 2004; Loo et al., 2010; Sohn and Shim, 2010; Gutiérrez et al., 2011; Sung and Oh, 2011; Zhao et al., 2013; Jun et al., 2015; Sun et al., 2016a,b). Most of the above-mentioned studies relied on global models which have limitations in examining the spatial heterogeneity of the effects. Hence, the geographically weighted regression (GWR) local model was employed in some cases to overcome this limitation (Blainey, 2010; Cardozo et al., 2012; Jun et al., 2015).

Although some previous studies were conducted to understand the spatial variation in the relationship between the station-level rail transit ridership and the factors influencing these, few studies have examined the impact of fine-scale built environment factors on multi-type transit ridership that can reflect different travel purpose trips. Moreover, none of the existing studies classified stations for policy-making according to the varying influencing mechanisms of the transit ridership. To address this gap, this study attempts to integrate multi-source spatial big data to examine the spatial variations in the influences of fine-scale built environment factors on rail transit ridership in the third biggest rail transit system in China, the Guangzhou metro, and identify the distinct station groups for zoning policy. The rest of this paper is organized as follows: section 2 reviews the existing literature; section 3 presents the study data and introduces the methodology; section 4 presents the results and discusses policy implications; and section 5 summarizes the study and provides suggestions for further research.

2. Related works: measures and methods

Numerous studies have been conducted in recent years to analyze the factors affecting the rail transit ridership. The measurement of transit ridership is diverse in these studies owing to the differences in research purposes or data limitations. The commonly adopted measures included average monthly boardings (Gutiérrez et al., 2011), average weekday boardings (Guerra et al., 2012; Kuby et al., 2004; Loo et al., 2010; Sohn and Shim, 2010; Zhao et al., 2013), daily station boardings (Estupiñán and Rodríguez, 2008; Jun et al., 2015), and average daily ridership including boardings and alightings (Blainey, 2010; Sung et al., 2014; Sun et al., 2016a,b). These studies mainly focused on understanding the impact of certain factors on the total station ridership but neglected the differences in the determinants of the daily, temporal, and directional transit ridership. However, rail transit stations usually have varying ridership patterns depending on day of the week, time of day, or in terms of boardings and alightings (Chen et al., 2009). The importance of exploring the different factors influencing daily, temporal, and directional ridership has been highlighted in some previous studies (Sung and Oh, 2011; Zhao et al., 2014; Iseki, Liu, and Knaap, 2018).

The factors that influence transit ridership were derived differently in earlier studies, as per data availability. The characteristics of the built environment surrounding the stations were usually hypothesized to be associated with station-level transit ridership (Gutiérrez et al., 2011; Jun et al., 2015; Sohn and Shim, 2010; Sun et al., 2016a,b; Zhao et al., 2013; Sung et al., 2014; Sung and Oh, 2011). The uncertainty in built environment measures may sometimes lead to bias conclusion (Sun 2018). High accuracy of measures becomes crucial in built environment research. More accurate measurements improve our confidence when interpreting the mechanism and thus offer more appropriate policy recommendations. Recently, geospatial big data which was regarded as a potential resource for smart city applications (Liu et al., 2015; Yuan et al., 2019) has brought new opportunities for perceiving built environment with fine spatiotemporal resolutions (Doran et al., 2016). The increase in scale opens a new window to investigate the impact of built environment on human behavior. It results in a richer understanding of mobility pattern and its relationship with built environment.

Land use area and diversity, population, and employment density were the most frequently used built environment characteristics for exploring influences of transit ridership (Blainey, 2010; Blainey and Mulley, 2013; Choi et al., 2012; Currie et al., 2011; Jun et al., 2015; Zhang and Wang, 2014). Land use types usually included residential, commercial, and industrial (Sohn and Shim, 2010; Sung et al., 2014; Sung and Oh, 2011; Zhao et al., 2013). These studies, which were limited by coarse land use classifications, could not investigate the impact of more detailed land use types that are closely related to the purpose of travel (e.g., education and sports). As the emergence of geospatial big data brought in new opportunities to perceive the built environment (Liu et al., 2015; Martí Ciriquián et al., 2019), a few studies began to employ POI data from online sources to estimate different functional facilities to infer different travel purposes and link them with the transit ridership (Zhao et al., 2014; Sun et al., 2016a,b; Shi et al., 2018). Shi et al. (2018) paid attention to the number of facilities at fine scales but ignored their land area. Sun et al. (2016a,b) used POI data to classify detailed land use and employed a field investigation to calculate their average built area. Their results confirmed the contribution of joint analyses of POIs and travel purposes in land use and public transport studies. Thus, research on the determinants of the transit ridership would benefit from a comprehensive understanding of the land use as this could impact the travel purpose.

The population and employment density in the earlier studies were mainly from government statistics at the administrative district or from a travel survey at the traffic analysis zone (TAZ) level (Gutiérrez et al., 2011; Cardozo et al., 2012; Zhao et al., 2013; Jun et al., 2015). The lack of spatial details forced the researchers to assume a uniform distribution of population and employment within each administrative district or TAZ (Sohn and Shim, 2010; Zhao et al., 2013). However, the spatial densities of population and employment are uneven and are usually higher around the rail stations. The assumption of uniform distribution may lead to a potential bias of these variables which affects the credibility of the findings. The emergence of big data has made it possible to obtain the real-time dynamic population distribution for extracts the detailed population and employment density. For instance, mobile phone signals (Ratti et al., 2006; Toole et al., 2012; Zhou et al., 2018) and social media data (Cai et al., 2017; Kaplan and Haenlein, 2010; Yao et al., 2017; Song et al., 2019) were used and their effectiveness was demonstrated for mapping the dynamic population distribution at fine spatiotemporal resolutions. Hence, it is particularly essential to employ such big data to acquire accurate spatial population and employment densities to explore their influences on the rail transit ridership.

The models adopted in the existing literature fall into two main categories, namely global and local regression models. The global models such as the ordinary least squares (OLS) (Kuby et al., 2004; Walters and Cervero, 2003; Loo et al., 2010; Sung and Oh, 2011; Zhao et al., 2013; Zhao et al., 2014), two-stage least squares (2SLS) (Taylor et al., 2003), distance-decay regression model (Gutiérrez et al., 2011), and structural equation model (SEM) (Sohn and Shim, 2010) were adopted to investigate the differences in the factors influencing the rail transit ridership at a station level. However, as the parameters were assumed to be stable in a global model, the calculated coefficients do not have significant differences in space (Cardozo et al., 2012). In fact, the estimated coefficients of the influencing factors of the transit ridership may be inconsistent in different spatial units. Recently, some studies have been undertaken to explore the spatial variations using local models. A frequently adopted model is the geographically weighted regression (GWR) model (Blainey, 2010; Cardozo et al., 2012; Jun et al., 2015). It is a local regression method based on the geographical location and has been proven to have better fitting results than the global OLS model in these studies. Also, GWR which takes account the spatial instability and can reveal the spatial heterogeneity of parameters across the space has demonstrated its usefulness in the rail transit ridership research (Clark, 2007; Jun et al., 2015). By
compared to GWR and OLS, Cardozo et al. (2012) found that GWR provides a better fit than OLS for estimating the boarding ridership at Madrid metro stations. The mixed GWR model that allows some variables to be local and other variables to be global has been proposed in some recent studies to effectively separate the independent variables that exert global effects and local effects (Jun et al., 2015; Yang et al., 2017). Furthermore, Shi et al. (2018) employed the geographically and temporally weighted regression (GTWR) model to examine the spatial and temporal variation in the relationship between the hourly ridership and policy implications.

Though GWR was employed widely in rail transit ridership analyses, there were still some shortcomings in earlier studies. Firstly, few studies integrated spatial big data to extract fine-scale spatial information on land use and population, and limited studies associated these variables with multi-type transit ridership that represent daily, temporal and directional transit ridership. The emergence of the geospatial big data brings new opportunities to identify fine-scale urban factors and perceive the built environment (Liu et al., 2015). In this study, multi-source geospatial big data such as Tencent social media, building footprints, POIs, and high-resolution images were combined to estimate the land use, population, and employment in fine detail to quantify the characteristics of the built environment around stations in a better way. This was combined with the GWR model to explore which fine-scale factors contribute more to different types of rail transit ridership and how their effects vary across the study region.

Secondly, the previous studies mainly focused on proving the advantages of local models in getting better fitness than global models (Blainey, 2010; Cardozo et al., 2012; Shi et al., 2018). Their modeling results were of great significance for transit ridership forecasting in new stations. As some scholars have emphasized, studies on factors of transit ridership can help formulate such zoning policy. Still, none of the existing studies quantitatively grouped the stations based on the coefficients of influencing factors that reflect the affecting mechanisms of station-ridership. To solve this problem, we employed k-means clustering to classify the stations according to the coefficients of explanatory variables, and further explore the policy implications by characterizing the influencing mechanisms of multi-type rail transit ridership for different groups of stations.

3. Data and methodology

3.1. Study area

Guangzhou, the third largest city located in Guangdong Province, southern China was selected as our case study area. Guangzhou is one of the metropolises with rapid urbanization and motorization, where rail transit was developed relatively early. Guangzhou metro’s first line, line 1, was opened in 1997, making Guangzhou the fourth city in China to operate a subway system. The Guangzhou metro network, with a total length of 390.5 km in 2017, was ranked third in the country, and among the top ten in the world. In 2017, its total passenger volume reached 2.8 billion, with a daily average of 7.678 million passengers and was ranked first in China. With the continuous expansion of the urban scale and the increasing traffic demand, Guangzhou plans to construct new subway lines. By the year 2025, the length of Guangzhou metro is expected to increase to 1000 km according to Guangzhou metro network planning. These make Guangzhou a good candidate for studying the influencing mechanism of rail transit ridership for station planning and passenger flow management. As shown in Fig. 1, 124 subway stations in Guangzhou, opened before January 2016, were included in our case study.
The pedestrian catchment area (PCA) is derived using a buffer area. 800 m is usually considered as an acceptable walkable distance, which is thus, employed to delineate the PCAs of the rail stations (Cardozo et al., 2012; Sohn and Shim, 2010; Zhao et al., 2013). However, these buffers between some stations may overlap, especially in the central area. The Thiessen polygon was used to deal with the overlaps. First, an 800 m buffer and a Thiessen polygon for each station were created. Next, the buffer and the Thiessen polygon were intersected. Finally, some areas close to the Pearl River were removed to eliminate the effects of the river.

3.2. Data

3.2.1. Rail transit ridership data

Rail transit ridership is classified into weekday ridership, weekend ridership, average of morning-peak boardings and evening-peak alightings on weekdays (henceforth called 'morning boardings & evening alightings'), the average of morning-peak alightings and evening-peak boardings on weekdays (henceforth called 'morning alightings & evening boardings') in this study. These four, capturing different types of rail transit ridership, were chosen as the dependent variables in our models. The weekday and the weekend ridership were considered to reflect largely the ‘working-purpose’ trips and the ‘leisure-purpose’ trips, respectively (Sung and Oh, 2011). Morning boardings & evening alightings ridership is a great indicator for the travel demand of the locally resident commuters, while the morning alightings & evening boardings ridership is more associated with travel for commercial purposes within the PCA of each station.

The boardings and alightings per hour from January 4 to January 10, 2016 were summed up from the smart card data acquired from Guangzhou Metro Company. The rail transit ridership was the highest during 17:00–19:00 (about 20%), while the ridership during 8:00–10:00 was about 19% on weekdays. Hence, this paper defines 8:00–10:00 and 17:00–19:00 as the morning peak time and the evening peak time, respectively. The four types of rail transit ridership mentioned above were aggregated from the hourly boardings and alightings.

3.2.2. Built environment data

The candidate independent variables are population and employment, land use density and diversity, and station characteristics (see Table 1), which were chosen on the basis of the findings of the previous researchers (Cardozo et al., 2012; Gutiérrez et al., 2011; Jun et al., 2015; Kuby et al., 2004; Sohn and Shim, 2010; Sung et al., 2014; Sung and Oh, 2011; Zhao et al., 2013; Blainey, 2010; Blainey and Mulley, 2013; Choi et al., 2012; Currie et al., 2011; Zhang and Wang, 2014; Zhao et al., 2014). In this study, multi-source geospatial big data including high resolution images, POI data, Tencent user data, and building footprint data were used to measure the built environment.

High-resolution images with a 0.55 m resolution, collected from Google Earth and POIs acquired from the Baidu map APlIn October 2015 were used to identify the fine land use types. Unlike most of the previous studies, which only focused on the residential, commercial, and industrial land uses (Jun et al., 2015; Sohn and Shim, 2010; Sung et al., 2014), in this study the land use was classified into 19 categories following the national urban land use classification standard (GB 50137–2011). Fine-scale urban land use information can be derived by combining the rich surface information provided by the high-resolution remote sensing images with the implicit socio-economic information from the POI data. First, we identified water (E1), green space (G), and urban land, based on the spectral information, shape, and texture of the high-resolution spatial image. Then, we classified the urban land by combining the categorical attributes of the POIs, which gives us administration and public services land use (A1–A6), commercial land use (B), industrial land use (M), residential land use (R1–R3), urban transport land use (S1, S2, S3), logistics and warehouse land use (W), and utility land use (U). Further, three types of residential land were distinguished by the height, area, shape, distribution pattern, and the shadow of the buildings. These three types of residential land include the first-level (R1; e.g., villas), second-level (R2; e.g., common residences), and third-level (R3; e.g., urban villages). These three types reflect the economic profile of the residents. Villas tend to be the residences of high-income people, while urban villages tend to be inhabited by low-income people or young workers who have just worked for a short time. Finally, a common residence is usually a high-density residential unit inhabited by middle-level income people.

Land use diversity was calculated on the basis of the Shannon entropy index (Shannon, 1948), as shown below:

\[ H = -\sum p_i \ln(p_i) \]  

where \( H \) is the entropy and \( p_i \) is the percentage of land use by the \( i \)th category. The spatial distributions of land use diversity and examples of fine land use in the PCAs are shown in Fig. 2.

In this study, we used Tencent user density (TUD) dataset provided by Tencent (http://www.qq.com), to calculate the population and employment densities within the PCAs of the rail stations. TUD records the locations of smart phone users who are using Tencent products, such as Tencent QQ (a messenger software), WeChat (a mobile chatting software), Tencent Map (a desktop and web mapping services and navigation software) and other LBS services (Chen et al., 2017; Yao et al., 2017). Tencent is currently the largest messaging and social media platform in China and has the broadest coverage, especially in metropolitan cities. The ratio of Tencent users to total population has exceeded 93% in China’s first-tier cities, including our study area, Guangzhou (http://bigdata.qq.com, 2016). Although TUD can be regarded as the bias sampling of the total population dynamic distribution, TUD have been successfully used in pioneering studies on fine-scale population distributions mapping (Yao et al., 2017), building functions inferring (Niu et al., 2017; Chen et al., 2017), urban park use analysis (Chen et al., 2019), due to its large user base. These studies illustrated that TUD data could be a useful proxy of dynamic population distribution in the metropolitan cities of China. Therefore, TUD were used to describe the population and employment around the rail transit stations in this study. As 10:00 and 22:00 on weekdays are usually regarded as the times when people are at work and at home, respectively (Liu et al., 2016), we defined the average TUD from 10:00 to 11:00 on weekdays as the employment density and the average density from 22:00 to 23:00 on weekdays as the population density. Two different areas are used as examples to show the spatial distributions of population and employment densities within the PCAs (Fig. 3).

Building footprint and story data as in October 2015 were obtained from Gaode Amap online service to compute the average floor area ratio (FAR). The number of feeder bus lines was acquired from Baidu map online service. In addition, the number of station entrances and exits, information on whether the station allows transfers or not, and the number of years of operation of the station still January 2016 were acquired from the Guangzhou rail transit company website.

3.3. Method

3.3.1. Geographically weighted regression model

GWR model is a commonly adopted regression model that takes into account the spatial autocorrelation. In contrast to the OLS model, which estimates a set of global parameters, GWR incorporates spatial variation into the coefficient estimation of the explanatory variables in regression models (Fotheringham et al., 2000; Blainey and Mulley, 2013; Chiou et al., 2015). The coefficients of variables \( b_k \) vary for each location and are defined by the location's spatial coordinates \( (u, v) \). The value of a typical explanatory variable \( y_i \) is estimated according to the equation shown below (Fotheringham et al., 1998):
<table>
<thead>
<tr>
<th>Category and Employment</th>
<th>Variables</th>
<th>Details</th>
<th>Mean</th>
<th>Std.Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population and Employment</td>
<td>Population density (PD)</td>
<td>The average Tencent user density from 22:00 to 23:00 on weekdays within the PCA of each rail station</td>
<td>58,075.89</td>
<td>32,630.88</td>
</tr>
<tr>
<td></td>
<td>Employment density (ED)</td>
<td>The average Tencent user density from 10:00 to 11:00 on weekdays within the PCA of each rail station</td>
<td>65,442.55</td>
<td>35,578.02</td>
</tr>
<tr>
<td>Land use density and diversity</td>
<td>Floor area ratio (FAR)</td>
<td>Average floor area ratio within the PCA of each rail station</td>
<td>1.69</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Administrative LU (A1)</td>
<td>The area of administrative office land use within the PCA of each rail station (m²)</td>
<td>9343.68</td>
<td>23,377.62</td>
</tr>
<tr>
<td></td>
<td>Cultural LU (A2)</td>
<td>The area of cultural facility land use within the PCA of each rail station (m²)</td>
<td>6293.77</td>
<td>25,080.02</td>
</tr>
<tr>
<td></td>
<td>Educational LU (A3)</td>
<td>The area of educational research land use within the PCA of each rail station (m²)</td>
<td>97,389.55</td>
<td>225,038.84</td>
</tr>
<tr>
<td></td>
<td>Sports LU (A4)</td>
<td>The area of sports land use within the PCA of each rail station (m²)</td>
<td>19,149.33</td>
<td>64,192.72</td>
</tr>
<tr>
<td></td>
<td>Medical LU (A5)</td>
<td>The area of medical land use within the PCA of each rail station (m²)</td>
<td>10,261.22</td>
<td>22,566.24</td>
</tr>
<tr>
<td></td>
<td>Other administration and public services LU (A6)</td>
<td>The area of other administration and public services land use within the PCA of each rail station (m²)</td>
<td>660.04</td>
<td>2110.79</td>
</tr>
<tr>
<td></td>
<td>Commercial LU (B)</td>
<td>The area of commercial land use within the PCA of each rail station (m²)</td>
<td>192,335.75</td>
<td>152,129.70</td>
</tr>
<tr>
<td></td>
<td>Green space LU (G)</td>
<td>The area of green space within the PCA of each rail station (m²)</td>
<td>130,338.24</td>
<td>146,319.15</td>
</tr>
<tr>
<td></td>
<td>Industrial LU (M)</td>
<td>The area of industrial land use within the PCA of each rail station (m²)</td>
<td>139,583.66</td>
<td>195,718.59</td>
</tr>
<tr>
<td></td>
<td>Water LU (E1)</td>
<td>The area of water within the PCA of each rail station (m²)</td>
<td>28,549.69</td>
<td>52,925.26</td>
</tr>
<tr>
<td></td>
<td>The first-level residential LU (R1)</td>
<td>The area of the first-level residential land use (e.g. villas) within the PCA of each rail station (m²)</td>
<td>1203.12</td>
<td>6015.10</td>
</tr>
<tr>
<td></td>
<td>The second-level residential LU (R2)</td>
<td>The area of the second-level residential land use (e.g. common residence) within the PCA of each rail station (m²)</td>
<td>303,027.25</td>
<td>238,507.54</td>
</tr>
<tr>
<td></td>
<td>The third-level residential LU (R3)</td>
<td>The area of the third-level residential land use (e.g.urban villages) within the PCA of each rail station (m²)</td>
<td>205,066.21</td>
<td>184,140.52</td>
</tr>
<tr>
<td></td>
<td>Urban road LU (S1)</td>
<td>The area of urban road land use within the PCA of each rail station (m²)</td>
<td>216,692.62</td>
<td>95,384.56</td>
</tr>
<tr>
<td></td>
<td>Urban rail transit LU (S2)</td>
<td>The area of urban rail transit land use within the PCA of each rail station (m²)</td>
<td>881.74</td>
<td>29,644.47</td>
</tr>
<tr>
<td></td>
<td>Special traffic facilities LU (S3)</td>
<td>The area of traffic facilities including airports, high-speed railway station site, bus station sites and the parking lots around rail stations within the PCA of each rail station (m²)</td>
<td>45,134.04</td>
<td>161,242.23</td>
</tr>
<tr>
<td></td>
<td>Utility LU (U)</td>
<td>The area of utility land use within the PCA of each rail station (m²)</td>
<td>3129.91</td>
<td>21,782.28</td>
</tr>
<tr>
<td></td>
<td>Logistics and warehouse LU (W)</td>
<td>The area of logistics and warehouse land use within the PCA of each rail station (m²)</td>
<td>5561.86</td>
<td>17,130.57</td>
</tr>
<tr>
<td></td>
<td>Other land use (Other)</td>
<td>The area of other land use within the PCA of each rail station (m²)</td>
<td>142,360.77</td>
<td>255,344.77</td>
</tr>
<tr>
<td>Land use diversity (LUD)</td>
<td>Land use diversity (LUD)</td>
<td>The Shannon entropy index within the PCA of each rail station (m²)</td>
<td>0.51</td>
<td>0.12</td>
</tr>
<tr>
<td>Station characteristics</td>
<td>Bus lines (BL)</td>
<td>Number of feeder bus lines within the PCA of each rail station (m²)</td>
<td>98.57</td>
<td>55.14</td>
</tr>
<tr>
<td></td>
<td>Transfer (T)</td>
<td>Transfer or not (1 = yes, 2 = no)</td>
<td>1.16</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Entrees/exits (EE)</td>
<td>Number of entrances or exits of each rail station</td>
<td>3.65</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>Operation time (OT)</td>
<td>The number of years of each rail station operation until January 2016</td>
<td>8.23</td>
<td>4.49</td>
</tr>
</tbody>
</table>
\[
y_i = \beta_{i0}(u_i, v_i) + \sum_{k=1}^{p} \beta_{ik}(u_i, v_i)x_{ik} + \epsilon_i
\]  

where

\( \beta_{i0}(u_i, v_i) \) – The estimated intercept for the \( i \)th observation at \((u_i, v_i)\);
\( p \) – The number of independent variables;
\( x_{ik} \) – The \( k \)th independent variable for the \( i \)th observation;
\( \beta_{ik}(u_i, v_i) \)–The estimated coefficient of independent variable \( x_{ik} \).
\( \epsilon_i \)– The random error for the \( i \)th observation.

The weighted least squares method is usually used to estimate the local parameters \( \beta_{i0}(u_i, v_i) \), and the spatial weighted function and the bandwidth are commonly required. The Gaussian function was adopted to compute the spatial weighted matrix in this study, which can be expressed as:

\[
w_{ij}(i) = \exp[-(d_{ij}/b)^2], j = 1, 2, \ldots, n
\]  

where \( d_{ij} \) is the distance between the \( i \)th and the \( j \)th observations and \( b \) is bandwidth. GWR model gives larger weight to the nearest observations when fitting the equation. Referring to Cardozo et al. (2012)’s study and considering the irregular distribution of the rail stations, we used the adaptive kernel function and the Akaike information criterion (AIC) to determine the optimal bandwidth. The AIC is a standard for measuring the goodness of a statistical model fit. It is based on the concept of entropy and can measure the complexity of the estimated model and the goodness of fit of its data. The lower the AIC value, the better is the model.

3.3.2. K-means clustering

K-means algorithm is used to classify a sample set of objects into \( k \) different clusters according to the similarities of their attributes. The purpose of clustering is to minimize the distance of the samples within the same cluster, and at the same time, maximize the distance between the clusters. The sum of the squared errors within the group, denoted by \( E \), is usually adopted as the standard measure function, which can be calculated below:

\[
E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2
\]  

where \( x \) represents the value of the sample object, and \( \mu_i \) is the mean value of the cluster \( C_i \) calculated as follows:

\[
\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x
\]  

4. Results and discussions

4.1. Stepwise linear regression and variable selection

As multi-collinearity exists in the variables, we first used backward stepwise regression (Chen et al., 2017; Mitsuda, Yoshida, & Imada, 2001) to eliminate the multi-collinearity and select significant variables with strong explanatory power for four models. Before applying stepwise linear regression, candidate variables were standardized with zero-mean standardization to ensure that the mean was 0 and the standard deviation was 1. Table 2 showed the obvious differences in the

Fig. 2. The spatial distributions of land use diversity (LUD) and the examples of fine land use in the PCAs.
Fig. 3. The spatial distributions of population and employment density.
The signiﬁcant Inﬂuencing variables included in the four models. Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weekday</th>
<th>Weekend</th>
<th>Morning boardings &amp; evening alightings</th>
<th>Morning alightings &amp; evening boardings</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD</td>
<td>* * *</td>
<td>* * *</td>
<td>* * *</td>
<td>* * *</td>
</tr>
<tr>
<td>ED</td>
<td>* * *</td>
<td>* * *</td>
<td>* *</td>
<td>* *</td>
</tr>
<tr>
<td>FAR</td>
<td>* * *</td>
<td>* * *</td>
<td>* *</td>
<td>* *</td>
</tr>
<tr>
<td>A3</td>
<td>* * *</td>
<td>*</td>
<td>* *</td>
<td>* *</td>
</tr>
<tr>
<td>A4</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>* *</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>G</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>E1</td>
<td>* * *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>R2</td>
<td>* * *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>R3</td>
<td>* *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>S3</td>
<td>* * *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Other</td>
<td>* *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>LUD</td>
<td>* * *</td>
<td>* *</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>OT</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>EE</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

*a* * means the variable is retained in the model; * * * * p < 0.01; * * * p ≤ 0.05; * p ≤ 0.1.

Independent variables retained in the four models. The notable difference between Weekday and Weekend models is that ED (employment density), B (commercial LU) and the R2 (second-level residential LU) were incorporated in Weekend ridership model, while the PD (population density), M (Industrial LU) and E1 (water LU) were included in Weekday ridership model. The Morning boardings & evening alightings ridership was found to be related to residential functional factors (R2-the second-level residential LU and R3-the third-level residential LU), A3 (Educational LU) and LUD (Land use diversity) which were eliminated in Morning alightings & evening boardings model. The transit ridership in Morning alightings & evening boardings was signiﬁcantly associated with the explanatory variables of ED (employment density), FAR (Floor area ratio), A4 (Sports LU) and B (commercial LU), which were excluded in Morning boardings & evening alightings model.

The results showed that PD (population density), ED (employment density), FAR (Floor area ratio), LUD (Land use diversity), Transfer (T), Entrances/Exits (EE) are highly important in cultivating transit ridership which is consistent with the results of the existing studies. However, this study identiﬁed ﬁne-scale land use by using high-resolution remote sensing and POIs, and revealed some results that have not been found in other studies without big data based BE measures. For example, this study subdivided the residential land use into three different types which can reﬂect the general social-economic characteristics of residents. The results showed that R1 (eg. Villas) is extremely non-associated with rail transit ridership which indicated that such kind of residence around rail stations do not play an active role in promoting transit ridership. R2 (e.g. common residence) and R3 (e.g. urban villages) are considerably important in Weekday ridership and Morning boardings & evening alightings ridership. Among them, R3 also signiﬁcantly affects weekend ridership. Moreover, the results also revealed the effects of other ﬁne-scale land use, such as educational research (A3), sports (A4) on different types of ridership.

4.2. Spatial variation of coefﬁcients from GWR models

OLS can evaluate the inﬂuence of explanatory variables, but the estimated parameters in the global OLS model were identical for each rail transit station (Chiou et al., 2015). An important characteristic of GWR model differing from global model is that the spatial differences of estimated local parameters can explain the spatial heterogeneity of relationships between dependent variable and the independent variables (Yang et al., 2017; Li et al., 2019). In the current study, we attempted to obtain the local influence coefﬁcients of transit ridership at each rail station based on the GWR model. Initially, the spatial autocorrelation tests for the four dependent variables and their possible predictors were made by using Moran’s I index in OpenGeoDa software. As showed in Table 3, the spatial autocorrelation was found for all of the variables and the p-value were lower than 0.05, except for A4 variable. This denotes that most of the variables are spatially clustered.

After getting the local coefﬁcients by using GWR model, local coefﬁcients of ED (employment density), PD (population density) and S3 (special trafﬁc facilities land use) were mapped as shown in Fig. 4 because of space limitation. Variable ED has signiﬁcant eﬀect on two types of rail transit ridership, weekday and morning alightings & evening boardings. The spatial distributions of the eﬀects were shown in Fig. 4. The coeﬃcients of ED on these two types of ridership that contain most work trips are lower for the southern and eastern stations than for the central, northern and western ones. This may imply that employment population in the southern and eastern areas, which represent the suburban areas in Guangzhou, are less reliant on metro for commuting. This mainly results from the lower density of metro stations at suburban areas, as well as the serious congestion at the central areas.

![Fig. 4](image)
Although other studies have explored the spatial varying effect of ED and PD (Cardozo et al., 2012; Jun et al., 2015), they focused on the total station ridership but neglected the varying ridership patterns depending on day of the week, time of day, or in terms of boardings and alightings. In this study, ED and PD data were based on the dynamic population distribution data obtained from TUD. With this big-data-based BE measures, the results can help us understand how the impacts of ED and PD vary on different types of ridership (day of the week, time of day, or in terms of boardings and alightings), as well as how the impacts vary across the region. These results can enrich the understanding of relationships between BE and mobility pattern (e.g., morning boardings & evening alightings ridership represents go-to-work trips of residents).

The local coefficients of S3 in the four types of ridership were shown in Fig. 6. The special traffic facilities (S3) in Guangzhou mainly include the airport located in Baiyun district; the high-speed railway station in Panyu district; the railway station in Yantu district; and the connecting bus station sites, and the parking lots around the rail stations. Small coefficients of S3 were observed in the central regions, as opposed to large values in Panyu and Nansha districts in the south far away from the central regions in the four models. Because of the low density of the rail stations in these areas, bus station sites and parking lots have been provided to reach the rail stations. The finding implies that rail transit travelers in the outer city are more reliant on connecting facilities than in the inner city. The airports and railway stations have not played a significant role in explaining the impact of S3 on these four types of rail transit ridership. In addition, the stations in Huangpu and Baiyun districts showed different effects in different models. The effects of S3 on eastern regions were almost as large as those on the southern regions in the morning boardings & evening alightings model (see Fig. 6(g, h)). Huangpu and Baiyun districts, being the fringe regions, also have some connecting facilities around the rail stations. This reason, together with the agglomeration of residential population, led to the high impact of S3 on the morning boardings & evening alightings ridership in the eastern regions. Similarly, the agglomeration of employment population further caused the S3 to have a greater impact in the northern areas on the morning alightings & evening boardings ridership.

4.3. Station classification and policy zoning

To further analyze the spatial variations of the influencing mechanisms of the station ridership and explore the policy implications for different regions, the k-means clustering method was used to classify the stations based on their coefficients in each GWR model. The sum of the squares within the groups was computed to determine the optimal number of clusters. The “turning point” at which the sum of the squares decreases slowly and tends to be stable, is usually considered the point that gives the optimal number of clusters. This optimal number of clusters is five in the morning boardings & evening alightings model, and four in the other three models. The rail stations were classified into a corresponding number of groups.

In each model, the average coefficients of the influencing factors of the stations in the same cluster were calculated, and are shown in Fig. 7. Except for a few variables (e.g., Transfer (T)) whose coefficients were similar in different clusters, the coefficients of all the variables varied significantly in different clusters. For example, the coefficients of S3 (Special traffic facilities LU), Entrances/exits (EE), land use diversity (LUD), and Floor area ratio (FAR) in cluster 3 were much larger than those in the other three clusters in weekday model (see Fig. 7(a)). It is noticed from the figure that the effect of a few variables on the same transit ridership in different clusters was in the opposite direction. As Fig. 7(c) showed, S3 presented significant negative influence on the morning boardings & evening alightings ridership at the stations in clusters 2 and 4, while it showed a positive effect on the same ridership at...
stations of other three clusters.

We visualized the clustering results of the stations and showed the averages of three variables with the highest values in Fig. 8 because of space limited. There were some differences in the clustering results of the four models. According to the clustering results in four models (Fig. 8 (a-d)), the whole area is demarcated into several zones for policy implication (Fig. 9) based on the following principles: 1) If two stations were classified into the same cluster in two or more clustering results, they will be classified into the same cluster; 2) Further considering the street administrative boundary, the area were zoned into regions, each one with the same classified stations.

Zone 1 includes the northern Nansha and the southern Panyu, considered as the outer suburban areas of Guangzhou. The coefficients of S3 was much higher in Zone 1 than other zones, for both weekday ridership and weekend ridership (see Fig. 8 (a,b)). That is to say, the same S3 land area would generate more metro trip in zone 1 than other regions. It implies that for the outer suburban areas where the station density is relatively low, planners should pay more attention to improving the metro usage by enhancing last-mile connectivity via land allocation for bus station sites and parking lots. In addition, the effect of FAR in this zone was much higher than that in the other regions for three types of ridership (see Fig. 8 (a,b,d)). This suggests that densification is encouraged at the station areas in outer suburbs. These policy implications which help the planners increase the transit ridership are consistent with Jun et al. (2015)’s study in Seoul. On the other hand, the coefficient of LUD in higher in this zone than other zone for morning boardings & evening alightings ridership (see Fig. 8(c)). In addition, R2 (for middle-income residents) was significantly associated with morning boardings & evening alightings ridership for the stations in this zone. These imply that the land use diversity plays a positive role in influencing the go-to-work trips from residential land. The diversified land use around the stations suggesting the existence of comprehensive facilities may attract many people to live here, though many of them do not work in these areas and are reliant on metro for working purpose trips. This may imply that the spatial mismatch of the workplace and the residence in such outer suburbs would put burdens on the traffic systems. The planning authorities may consider adopting mix land-use development that provides more job opportunities for local residents in the new town planning.

Zone 2 includes the northern Panyu and western Huangpu, which is regarded as the inner suburban area of Guangzhou. Firstly, FAR showed significant effect on all the four types of ridership (see Fig. 8(a-d)), which suggests high-density developments around the rail stations in the inner suburban areas. The influence coefficient of FAR in this zone was slightly less than that in zone 1. Although Jun et al. (2015) encouraged densification in the suburban areas, they did not examine the differences between the outer and the inner suburban areas which are important for policy guidance. Secondly, for the western stations of this zone that include southern line 2 and line 3, LUD, PD, and R2 were significantly associated with morning boardings & evening alightings ridership (see Fig. 8(c)). This indicates that residents, especially those middle-income people living in common residences, were reliant on metro for go-to-work trips. The planning and construction of these two rail lines lead to large-scale development of common residences, thus guiding the agglomeration of middle-income residents. The relatively low house prices and convenient living facilities have attracted many residents, but most of them still work in zone 3, the inner city. In this context, the spatial imbalance of jobs and housings, together with long commuting distance make the residents in zone 1 dependent on metro travel for working purpose. As a result, these two rail lines have serious congestion in the direction from zone 2 to zone 3 during the morning peak hours, and in the opposite direction during the evening peak hours. Hence, more rail lines may be deployed to improve the accessibility of zones 2 and 3. In addition, there is no rail link between the
eastern and the western regions of zone 2. The eastern region of zone 2 includes the middle section of line 4 and the west section of line 5, where some high technology enterprises and Guangzhou university town are located. If people living in the west and working in the east choose to travel by the metro, they need to transfer at zone 3 with a long detour. This would greatly increase the commuting distance.

Zone 3 includes Yuexiu, Liwan, Haizhu, and Tianhe districts in the central city of Guangzhou, while zone 4 includes Baiyun region located in the north of Guangzhou. The influencing factors that affect various types of rail transit ridership are similar in these two zones, except for the morning boardings & evening alightings ridership. A typical characteristic of the rail stations in zone 4 was that PD and R3 had significant influences on the morning boardings & evening alightings transit ridership (see Fig. 8(c)). In Guangzhou, a large number of urban villages are distributed in this region. Because of the lower housing rents, these villages reflected by R3 land use are usually inhabited by low-income people or young workers. Many of them work in the wholesale markets or enterprises in zone 3, with a few working in the manufacturing factories in zone 4. Owing to their relatively low income, this group of residents tends to choose public transportation for commuting. These findings imply that more attention should be paid to planning living facilities around the rail stations to meet the needs of the low-income residents in zone 4. Although altering the land use in the old central city faces great challenges, the interaction between the rail station planning and the urban regeneration in this zone deserves more attentions from planners.

Zone 5 includes Conghua and Zengcheng districts, the outer suburbs farthest away from the city center. There are currently no rail lines going through this zone. The planning and construction of rail lines are needed to improve the transportation accessibility between zone 5 and the central city zone. The influencing mechanisms of transit ridership in zone 1 which is also the outer suburbs, can provide policy reference for the planning of rail stations and the guidance of land use in zone 5.

![Spatial visualization of k-means cluster results based on the coefficient values.](image-url)
5. Conclusions

This study employed the GWR model to examine the spatial variation of the impact of the built environment on various types of rail transit ridership that include daily, temporal and directional. The k-means clustering was used to classify the stations based on the local coefficients of the influencing factors. In addition, policy zoning was suggested, and the planning implications were put forward. The major contributions and findings from this study can be summarized as follows. Firstly, this study explored and revealed the spatial heterogeneity of the influence mechanisms of built-environment variables on different types of ridership at a fine scale. For example, the coefficient of ED (employment density) was lower at the southern and eastern areas than at the northern and western areas. For weekend ridership that contains less work related trips, high values of PD (population density) coefficients were observed in the old city regions. With regards to morning boardings & evening alightings ridership which represents go-to-work trips of residents, high values of PD were found at Panyu suburbs.

Secondly, this study advanced the use of k-means clustering in identifying the distinct station groups according to the coefficients of explanatory variables of multi-type rail transit ridership. The clustering showed that the coefficients varied significantly in different station groups, and some were found to have opposite effects on the same ridership in different station clusters. According to the clustering results, policy zoning was suggested and planning implications were discussed. The results showed quantitative evidence for planners to contemplate strategies of built environment planning around the stations in different zones, with a view to increasing the rail transit ridership. For instance, densifying land development, improving the land use mixture, and enhancing station connectivity via buses are likely to improve the ridership in zone 1, the outer suburban area. Furthermore, the results suggested that the rail transit link between some zones, as well as the
measuring urban park use. Urban For. Urban Green. 31, 130–141.
Sun, G., Webster, C., Ni, M., Zhang, X., 2016b. Measuring high-density built environment for public health research: uncertainty with respect to data, indicator design and
spatial scale. Geospatial health, 2018 13 (1), 35–47.