RETAIL COMMERCIAL SPACE CLUSTERING BASED ON POST-CARBON ERA CONTEXT

A Case Study of Shanghai

QINYU CUI1, SHUYU ZHANG2 and YITING HUANG3
1,3School of Architecture and Urban Planning, Shenzhen University, Guangdong, China.
2College of Architecture and Urban Planning, Tongji University, Shanghai, China.
1cuiqinyu2020@email.szu.edu.cn, 0000-0001-6687-9992
2hadyyu@gmail.com, 0000-0002-0459-4200
3906125150@qq.com, 0000-0001-5441-1612

Abstract. In the post-carbon era, it has become a development and research trend on adjusting commercial locations to help achieve resource conservation by using big data. This paper uses multi-source urban data and machine learning to make reasonable evaluations and adjustments to commercial district planning. Many relevant factors are affecting urban commercial agglomeration, but how to select the appropriate ones among the many factors is a problem to be considered and studied, while there may be spatial differences in the strength of each influencing factor on commercial agglomeration. Therefore, this paper takes Shanghai, a city with a high economic and commercial development level in China, as an example and identifies the influencing factors through a literature review. Next, this paper uses the machine learning BORUTA algorithm of features selection to screen the influencing factors. It then uses multi-scale geographically weighted regression model (MGWR) to analyse the spatial heterogeneity of factors affecting retail spatial agglomeration. Finally, based on the background of the changing transportation modes and the unchanged social activities in the post-carbon era, the future spatial planning pattern of retail commercial space is discussed to provide particular suggestions for the future location adjustment of urban commerce.

Keywords. Business District Hierarchy; Agglomeration Effect; Spatial Variability; Multi-scale Geographically Weighted Regression Model; Machine Learning; Big Data Analysis; SDG 8; SDG 12.

1. Introduction

The post-carbon era is an ecologically harmonious, green, low-carbon, and sustainable society, and its main technical characteristics are digitization and intelligence (Rifkin,
2013). In this context, the transformation of the agglomeration characteristics of urban commercial space is an inevitable requirement to meet the needs of future human life. It is also an important guarantee for developing a city with high efficiency and low consumption.

Retail commercial space is the main place for human leisure activities. In the early days, empirical case studies were mainly carried out on the central place theory. With the development of space measurement technology, the combination of quantitative and qualitative research methods has been widely used. In quantitative research, effective selection based on influencing factors is very critical. The appropriateness of variable selection in previous studies was mostly tested by covariance and bivariate (Jia and Zhang, 2021). In addition, the spatial heterogeneity of influencing factors has always been a problem in the quantitative research of geospatial types. In urban quantitative research, in the areas of spatial regression models such as geography and economics, the fitting effect of linear regression models is usually inferior to that of geographically weighted regression (Zhang et al., 2019), mainly because of the uneven distribution of urban economic, social environment and the functional space of various entities in the city. Multi-scale geographic weighted regression models can give each relationship exclusive spatial bandwidth (Oshan et al., 2019), which can reflect the local, regional, or global spatial scale effects of different influencing factors and achieve a more accurate result model fitting effect.

In commercial space agglomeration research, although a lot of research has been conducted on the influencing factors of accumulation, there are few studies on the feature selection of influencing factors based on machine learning. In addition, quantitative research on urban commercial agglomeration based on geographically weighted regression is still extremely lacking. In the post-carbon era, people try to use clean energy as much as possible to reduce carbon emissions. This will change the way people travel, and there is a strong coupling between transportation stations and commercial clusters. In summary, in the context of the post-carbon era, with the changes in people’s lifestyles, business agglomeration in new and old spaces has changed. This is also the purpose of this research.

Therefore, this paper selects Shanghai, which has relatively complete commercial development in China. Machine learning feature selection, multi-scale geographic weighted regression model, and spatial analysis technology are carried out on multi-source data to conduct in-depth analysis and research on the spatial heterogeneity of factors affecting commercial agglomeration in Shanghai.

2. Materials and methods

2.1. STUDY AREA

For this research, we selected the central district within the outer ring line of Shanghai, of around 660 km², as the study area (Figure 1). From the perspective of industrial structure, Shanghai’s tertiary industry accounts for the largest proportion of its GDP. As one of the main forces driving and supporting the development of Shanghai’s tertiary industry, retail business reflects the development and expansion of Shanghai’s business. Therefore, Shanghai is a suitable case city to study China’s urban commercial
2.2. VARIABLES AND DATA

Agglomeration of retail spaces was considered the dependent variable, and a series of urban indicators were considered independent variables. The data sources and quantification methods of the variables are introduced below (Figure 2).

The dependent variable for this study was the accumulation of retail spaces, from now on referred to as “RETAIL AGGLOMERATION”. We collected data from the Baidu Map Open Platform (http://lbsyun.baidu.com/), crawling the names, commercial types, and spatial locations using Python programming language. Then we cleaned up the data, deleted the repeated and problematic data, and reclassified the data combined with the map to obtain the final data.

Based on the relevant research results at home and abroad and combined with research questions, this paper determines 27 variables affecting retail agglomeration in 4 types, 8 categories, among which the variables of 4 types are: transport properties,
supporting facilities, population flow, and others. 27 variables and their processing methods are detailed in figure 2.

2.3. METHODOLOGY

Step 1: Analysis of retail agglomeration by Kernel Density Estimation (KDE)

Commercial kernel density is a value used to evaluate commercial spatial density, influenced by neighborhood division. The Kernel density estimation (KDE) calculates the density of discrete nodes in different threshold ranges (h) of the output grid cells of the research region (Fei et al., 2019). A larger kernel density value indicates a stronger concentration—i.e., a higher degree of commercial space agglomeration. The kernel density at the center of the grid is the sum of the densities within the projection area of a kernel function $K(\cdot)$, as shown in Formula (1):

$$f^x = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

where $f^x$ is the kernel density at the center of the research region; $h$ is the threshold value (i.e. the searching radius); $n$ is the number of nodes in the threshold range; $d$ is the dimension of the calculating data; $K(\cdot)$ is a non-negative kernel function with the property of the probability density (Chen et al., 2013).

Step 2: Analysis of Commercial Spatial Agglomeration Characteristics Using Moran’s I and Getis-Ord Gi* algorithm

In our study, global Moran’s I is used to evaluate the spatial clustering pattern of retail commercial space.

Next, we used the Getis-Ord Gi* algorithm to measure retail commercial space distribution’s hot or cold regions. When the Gi* statistic is higher than the mathematical expectation and passes the hypothesis test, it is a hot spot, otherwise a cold spot.

Mathematically, Getis-Ord Gi* is expressed as

$$G_i^* = \frac{\sum_{j=1}^{n} W_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{j=1}^{n} W_{ij}}$$

Step 3: Using machine learning BORUTA to select suitable variables and make correlation analysis

This study considers several variables that affect business accumulation, some of which have significant effects. In contrast, others have no significant effects, so references’ point is needed to help distinguish the really important variables from the unimportant ones. We have applied the BORUTA algorithm that provides criteria for selecting relevant features to cope with this problem.

The BORUTA algorithm applies random forests for feature relevance estimation. This method mainly compares real predictive variables’ importance with random shadow variables through statistical tests and multiple RF runs. A copy of each variable is added in each run, doubling the set of predictors. Variables with significantly larger or smaller importance values are declared important or unimportant, respectively. Then delete all unimportant variables and shadow variables, and repeat the previous steps until all variables are classified, or the pre-specified number of runs is performed.
After screening the significant variables by Boruta, the collinearity of independent and dependent variables was detected, and Spearman’s rank correlation coefficient was carried out by SPSS tool.

**Step 4: MGWR was used to analyse the spatial heterogeneity of variables of retail agglomeration**

Multi-scale Geographic Weighted Regression (MGWR) was proposed by Fotheringham (2017) and Yu et al. (2019); they improved the statistical inference, which is superior in replicating parameter surfaces with different levels of spatial heterogeneity and provides valuable information on the scale at which different processes operate. In this study, we use MGWR to analyze the spatial heterogeneity and scale difference of the influencing factors of commercial spatial agglomeration. Mathematically, MGWR is expressed as

\[
Y_i = \sum_{l=1}^{k} \beta_{bwj}(u_i, v_i) x_{ij} + \varepsilon_i
\]

Where \(\beta_{bwj}\) represents the regression coefficient of local variables, \(bwj\) represents the bandwidth used by the regression coefficient of variable \(j\), \((u_i, v_i)\) represents the geographical spatial coordinates of sample point \(i\), \(x_{ij}\) represents the observed value of variable \(j\) at sample point \(i\), and \(\varepsilon_i\) represents the random disturbance term.

3. Previous Analysis

3.1. RETAIL COMMERCIAL SPACE CLUSTERING EXTRACTION

**Kernel Density and hot spot analysis**

After comparing and analyzing the four kernel density search radii of 400m, 800m, 1200m, and 1600m, respectively, the search radius was set to 400m for retail commercial kernel density analysis by combining the common Shanghai neighborhood size (Figure 3). Subsequently, by dividing neighborhoods and 400×400m grids by major roads as the study units, we made an exploratory attempt to use the Getis-Ord Gi* index to explore the spatial distribution hierarchy of retail businesses in Shanghai. Since the area size of the neighborhoods in the study area varies greatly, the hot spot analysis effect is different from common sense, so the grid was finally adopted as the study unit for hot spot analysis (Figure 3).

![Figure 3. Retail commercial space clustering form map and hot spot distribution](image-url)
Study unit extraction

Spatial sets with Z-values greater than 1.65 are hotspot areas selected as the study units of this paper, and the kernel density values are extracted vectorized into the study units. Spatial visualization was performed in ArcGIS Pro 2.5 by the natural breakpoint method (Figure 4-a), and the spatial distribution of the extracted study units was roughly mostly clustered within the inner ring and scattered on the outer side. The high-value regions are also mostly concentrated in the inner ring, with some local high-value regions formed sporadically on the outer side.

In addition, since some of the variables in the previous paper (e.g., traffic and activity space) need to be considered under the 5/15-minute living circle scale, the study units were made into 300m and 900m buffers, and the values of the corresponding variables were counted separately (Figure 4-b/c/d).

Figure 4. Study unit related features: (a) average kernel density value for each unit; and(b) study unit; and(c) study unit (300m buffer zone); and(d) study unit (900m buffer zone).

3.2. MODEL VARIABLE SELECTION

Machine learning BORUTA and Spearman correlation analysis

According to the results of BORUTA variable screening, there are 11 definite acceptance variables, 2 tentative variables, and the rest are rejection variables (Figure 5).

This paper uses SPSS software to perform Spearman correlation analysis on the independent variables determined to be accepted and the dependent variable and visualize them. We can see that the correlation between the independent and dependent variables is strong, showing significance at the 1% statistical level. In addition, the correlation coefficients of the dependent variables are all less than 0.8 with each other, there is no multicollinearity, and all the variables determined to be accepted can be put into the model operations (Figure 6).
4. Result Analysis

4.1. MODEL BUILDING

Table 1. Model diagnostic information description

<table>
<thead>
<tr>
<th>Model</th>
<th>Adj. $R^2$</th>
<th>RSS</th>
<th>AICc</th>
<th>RSS Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLR</td>
<td>0.442</td>
<td>359.305</td>
<td>1420.926</td>
<td>0.077***</td>
</tr>
<tr>
<td>GWR</td>
<td>0.563</td>
<td>283.667</td>
<td>1367.037</td>
<td>-0.017</td>
</tr>
<tr>
<td>MGWR</td>
<td>0.598</td>
<td>270.749</td>
<td>1339.839</td>
<td>-0.024</td>
</tr>
</tbody>
</table>

MGWR model has the highest fit, the smallest residual sum of squares and AICc, and the best overall fit among the three models (Table 1). In addition, the residuals of the OLR model had significant positive spatial autocorrelation. In contrast, the residuals of the GWR and MGWR models did not have significant spatial autocorrelation, indicating that the geographically weighted class of models considering spatial heterogeneity outperformed the global model in dealing with the spatial autocorrelation of residuals (Harris et al., 2018). In summary, the Shanghai business agglomeration can be modeled by the influencing factors selected in this paper to build the MGWR model for regression analysis.

4.2. ANALYSIS OF SPATIAL HETEROGENEITY OF INFLUENCING FACTORS OF COMMERCIAL SPACE AGGLOMERATION

On the whole (Figure 7):

1) the population size formed by the residential and leisure behaviours of the population is the primary factor constituting retail commercial accumulation;

2) the impact of two different quantitative scales of social activities, movie watching and KTVs, on urban retail spatial agglomeration, showed complementary spatial distribution;

3) the intensity of the impact of land price, commercial mixing degree, public transportation and commercial buildings on commercial accumulation did not differ geographically and spatially, and the remaining relevant factors of The uneven spatial
distribution of the other relevant factors are the main reason for the differences in the spatial concentration of urban retail businesses;

4) In addition, the higher the density of the road network, the greater the accessibility of crowd activities, which is more conducive to the concentration of businesses.

5. Conclusion and Discussion

In this paper, we took retail commerce within the central city of Shanghai as an example. Based on machine learning BORUTA screening variables and regression analysis using the MGWR model, we determined the positive effects of relevant influencing factors on retail commercial accumulation and the reasons for spatial differences.

In the context of future urban development in the post-carbon era, some of the lifestyles of city dwellers will change. The transformation of new energy, the further development of high technology, the Internet of Things, the way urban residents travel, and their consumption habits bring changes, such as new energy vehicles, driverless technology, and online order consumption. The offline, face-to-face activities of urban residents are also a typical example. Despite the rapid development of technology in the past 20 years, it has not been greatly affected or replaced, and it is still an important
part of urban residents' lives.

Figure 8 shows that some retail businesses used to serve people who traveled in both directions from home or work to public transportation and clustered in this space. In the future, the clustering characteristics of retail commerce will favor the new public transportation space near the work area, such as the charging piles or parking lots of shared trams, and the realization of a 5-minute service circle centered on various types of transportation nodes.

Some factors affecting retail business agglomeration are inferred in chapter 4.2 of the paper. Based on this conclusion, the agglomeration characteristics of retail business in the post-carbon era are discussed.

1) The model analysis shows that the accumulation of retail business is closely related to the crowd's activities, and the current business agglomeration is still clustered around residential and entertainment areas.

2) Living, commuting, and work are the main components of a working day for urban residents. The areas from residence to transportation stations and from work to transportation stations are the key areas for small businesses to gather.

3) The transportation mode of commuting in the post-carbon era will change. For example, electric vehicles need to be replenished by charging piles, which may become new traffic sites. At present, the retail business layout in this part of the area is still blank.

4) The Covid-19 epidemic has changed people's consumption habits and promoted the new retail model of online shopping-offline distribution. In the future, some offline retail stores will be transformed into non-core links responsible for sorting, packaging, distribution, etc., and the rest will be transformed into experience centers, exhibition centers, and service centers. This business model transformation can ensure the upgrade of consumer experience and minimize operating costs.

5) The service space generated by large-scale retail business clusters is the main place for social and leisure activities at work and weekends. Its functions include high-end brand image display, mid-to-high-end catering consumption experience, and physical leisure space such as KTV and cinema. This space has a high degree of commercial mixing, rich commercial types, grades, and functions. The function and service radius will not change greatly and will remain in the future.

To sum up, commercial gatherings of different scales will undergo different changes, but commercial gatherings are always process-oriented by crowd activity. Figure 8 presents our summary and assumptions about the characteristics of retail business accumulation in the current and post-carbon era.
6. Application and Production

The results of this paper will assist commercial real estate investment companies in evaluating the investment value and appreciation potential of projects in various business districts in Shanghai. This paper obtained data and platform support from Metrodata Technology Company to transform research into a platform product, as shown in figure 9 below.

Figure 9. Platform support from Metrodata Technology Company

Acknowledgements

Metrodata Technology Company supports the data and analysis platform of this paper. (https://www.metrodata.cn/)

References


