Research Article

Classification of Housing Submarkets considering Human Preference: A Case Study in Shenyang, China

Xinrui Liu,1 Rongfei Guo,1 Guoping Lei,1 and Na Liu2

1Institute of Land Resource Management, Northeastern University, Shenyang, China
2Liaoning Ecology & Environment Monitoring Center, Shenyang, China

Correspondence should be addressed to Xinrui Liu; xliu1984@qq.com

Received 12 April 2022; Revised 5 June 2022; Accepted 30 June 2022; Published 30 July 2022

Academic Editor: Hongtai Yang

Copyright © 2022 Xinrui Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Housing submarket is affected by the quality of public facilities in neighborhood. In this article, a three-step framework is proposed to classify the housing market in Shenyang city, China. The saliences of point of interest, POI, are used to quantify the degree of activities using social media data and merged into the model of density-field hotspot detector to extract high-quality of urban facilities, aiming at exploring the influences on the housing price. The multiscale geographically weighted regression model, MGWR, is applied to investigate relations between the housing price and attribute variables geographically and to identify multiscale impacts on the housing price with flexible bandwidths. The spatial “k”luster analysis by the tree edge removal method, SKATER, is applied to delimit urban housing submarkets based on the MGWR betas, where ten groups of the housing submarkets are recognized. The proposed classification method captures multiscale spatial heterogeneity of the housing price and predominant influencing factors for each housing submarket in Shenyang, China. It also offers suggestions for urban planners to restrain the unbalanced development of the housing market.

1. Introduction

Housing market is influenced by socioeconomic diversity and heterogeneous distribution of public facilities. Urban housing price is spatially varied based on unbalanced urban distribution and neighborhood differences related to human preference [1–3]. In this research, human preference is defined as public sensibilities and perceptions used for evaluations of urban situations. The urban area with high human preference reflects a high level of attention and visit, implying popularity and convenience to public. For instance, school district houses are preferential to parents pursuing superior education for children. The housing price is higher within high-rank school districts. Schools are high-rank hotspots exhibiting high visit frequencies and public recognitions based on human preferences. Limited research studies have addressed the relations between high preference-triggered urban environment and the housing price. In general, the classification of the housing submarkets is more focused on the housing characteristics, including location, structure, and neighborhood, without considering human preferences [4, 5].

In this article, we propose a three-step framework to classify the housing submarkets considering human preference in Shenyang, China. The framework involves social media information, hedonic model, and cluster analysis. A model of density-field hotspot detector data is proposed based on the POI and social media, in order to identify the hotspots of different types of human activities and evaluate the corresponding influences on the housing price. Traditional methods based on POI data are insufficient to confirm whether the site is related to distance or popularity [6]. Location-based social media data have been widely used to show social preferences of each spot [7], where people share locations and activities using smart phones or tablets [8]. For instance, people use check-in data to show daily activities posted on Twitter or WeChat. The check-in data from the microblog of Sina Weibo record the number of visitors at specific sites or POIs in a certain period. The search volumes of “Baidu index” (http://index.Baidu.com/) reveal the search
frequencies of specific keywords (spots) in different periods (mostly daily basis) and regions. These social media provide data related to individual activities and updated viewpoint to examine human spatial preferences in urban space [9, 10]. According to the survey report of Chinese home buyers in 2018, more than 85 percent of house buyers are in the 20–49 age group. People in this age group are the largest users using smart electronic devices. Therefore, the housing submarkets are analyzed by exploring the human preferences revealed by check-in and Baidu search index data.

Hedonic models have been extensively applied to evaluate the housing value since the 1960s [11, 12]. This model investigates the housing price, structure, and neighborhood condition. However, the hedonic model becomes less reasonable to delineate the housing submarkets when considering bundles of price determinants. It is necessary to explain the spatial differentials of the housing value and estimate accurate coefficients in the hedonic housing price model of the submarkets [13, 14]. Spatial hedonic models are introduced for biased estimations due to spatial effects, highlighting the influence of variables on the housing price in a prescribed place [15]. As a popular spatial hedonic model, the geographically weighted regression, GWR, is effective to characterize spatial nonstationarity and heterogeneity of the housing price, through varied coefficients with spatial locations [16]. The application of GWR in empirical housing hedonic studies becomes widespread [17–21]. Nevertheless, only limited relations between the housing price and some covariates are spatially varied in these research studies. The influence of the variables may be local or global, introducing bias to the parameter estimation [22]. The GWR model is further improved by considering both the spatial stationarity and nonstationarity, namely, a semi-parametric GWR model, SGWR [23, 24]. The SGWR distinguishes local and global variables; however, all of the local variables are still assumed to operate at the same bandwidth or the same spatial scale.

Therefore, a multiscale geographically weighted regression model, MGWR, is used to estimate fixed and varied coefficients at different spatial scales [25, 26]. The MGWR offers a more flexible option to analyze the housing price and has been successfully applied in real estate and urban studies [27, 28]. In this paper, the MGWR method is selected to identify globally fixed elements and locally varied factors, characterize the spatial clustering effects of the influential factors, and analyze the corresponding determinants for the submarkets.

The factors affecting the housing price are complex and diverse. The influential factors to each house may be different, but similarities exist between houses in a prescribed area. The spatially homogeneous submarkets and the major contribution factors are detected using the coefficients of the MGWR model. Cluster analysis based on statistical geodata analyses is superior and objective for the classification of the housing submarkets. This method simultaneously considers spatial clusters of the housing structure and neighborhood, allowing examinations of spatial features at the household level [29]. The cluster algorithms for the housing submarket include factor analysis [30], principal component analysis, PCA [31], partitioning algorithms [32], K-means clustering [33], and hybrid methods of clustering algorithms, GWPCA, and DBSCAN [34]. Most algorithms consider no geographical spatial context. Representations of the clustering pattern may be diverse when the multivariate cluster analysis method is used to analyze spatial data. Spatial locations should be considered in the classification process by setting spatial constraints or spatial weights. Since the housing market is related to urban spatial space, the housing submarkets should be classified using spatial clustering methods based on big geodata. In addition, the classification may involve many attribute data with varied ranges, types, and measurement units. The cluster method should allow single and multiple factors to explore the underlying pattern. Spatially constrained multivariate clustering method based on the SKATER algorithm is available to cluster and visualize multiple variables considering geographic locations and cluster size constraint.

This article selects spatially constrained multivariate clustering method to detect homogeneity and divide the housing market of Shenyang into spatially contiguous submarkets based on the coefficients of the MGWR. In comparison with K-means and DBSCAN with uneven density clustering distribution, the method uses the spatial “k”luster analysis by the tree edge removal method, SKATER, to recognize natural clusters [35]. The SKATER uses graph partitioning for efficient division of spatial entities with different densities while preserving neighborhood connectivity, balancing the computational costs and sensitivity [36]. This study aims to detect how the high preference-triggered housing facilities affect the housing price and propose a mixed method to classify the housing submarkets based on the coefficients of factors. The planning strategies are suggested according to the driving factors of the housing submarkets in Shenyang.

2. Study Area and Data

2.1. Study Area. Shenyang is the central city of Liaoning province, serving as an important industrial center in the northeast of China (Figure 1). Shenyang is facing the challenges of industrial transition and sustainable development, similar to other heavy industry cities in the world. In 2019, the average housing price is less than 10,000 CNY/m² (Chinese yuan/m²), significantly lower than the housing prices in other large cities in China. Nevertheless, local overheating and polarization exist in the housing market. The core area of Shenyang is selected to analyze key influential factors to the housing prices and classify the housing market, in order to optimize distributions of urban public resources.

A total of 4,671 dwelling units of the housing transaction records, with real estate attributes of HomeLink Real Estate website from January to December in 2017, are collected by the crawler technology. An apartment may be simultaneously transacted in different agencies, resulting in differences in the data structure and repeated records. These resources are not used in this study. In this manner, we randomly select 706 community samples with the housing...
prices close to the average of each community as training samples, and 698 samples with the housing prices as the median of each community as testing samples. The check-in data from the social network are limited; therefore, a short period of the trading data is used and the influences of the time are neglected. The spatial pattern of the housing price in the core area of Shenyang is shown in Figure 2. The data are analyzed by the exploratory spatial data analysis of geostatistical analyst in ArcGIS 10.3, showing that the housing price decreases from south to north, and the high price is localized in the central zone along the east-west direction.

2.2. Housing Data Processing and Analysis. Both the global and local Moran's I coefficients are applied to detect and measure spatial effects on the housing price. The global Moran's I value is 0.467, and the z-score is 21.6, with p value <0.000 (Figure 3(a)). The housing price is positively auto-correlated (clusters), showing a moderately high spatial clustering pattern. In extreme conditions, this kind of pattern may be attributed to a random choice. When the housing price in the neighborhood is high, the price in the surrounding areas is also high.

The local indicator of spatial association, LISA, is selected to examine the variation of the spatial auto-correlation in the study area (Figure 3(b)). The high-high (HH) clustering points are concentrated along the Hunhe River in Shenyang, the Golden Corridor of Youth Street in Shenhe District, the central commercial region in Tiexi District, and scattered in the new city Hall of Hunnan District. The low-high (LH) clustering points appear near the HH clustering points, and no high-low (HL) clustering points are observed. The high housing price shows positive effects on the spatial spillover. The low-low (LL) clustering points are predominantly observed in the northwestern regions. The housing prices in the related communities are essentially homogeneous.

2.3. POIs and Social Network Data Processing and Analysis. External factors affecting the housing price, including local and surrounding attributes, are calculated by ArcGIS 10.3 according to 167,892 POI data and 144,915 data derived from the social media of the Sina Visitor System between April 2016 and June 2017, and 102,987 Baidu search index data from January to December 2017. The social media data
have been removed when the following criteria are satisfied: (i) the check-in data are located outside the core area of Shenyang; (ii) the key words of the social network data are not corresponding to POI data; (iii) the POI salience is not included in the social network; (iv) the Baidu search index needs to be transformed from average daily data to annual data based on the name and searching time; and (v) the human preferences are rare in between April 2016 and December 2017. Following these criteria, 116,692 POIs, 18,379 check-ins, and 15,459 Baidu search indexes are finally filtered. These POIs, check-ins, and Baidu search indexes are then merged when the data records include all three data kinds, where 11,971 POIs related to eight categories are achieved, as listed in Table 1. Three categories of POIs with 5047 records—the commerce centers, elementary and secondary schools, and Three-A hospitals, are eventually derived. In addition, 119,135 points of bus stops and subway stations, and 879 regional data of river and park are collected.

3. Methodologies

3.1. Framework. Figure 4 shows the framework of this research. Similar houses are selected as training and validation data. Social media data are used to quantify the POI’s popularities and measure the performance of different urban facilities. The MGWR model is applied to detect local variables with spatial heterogeneity. The coefficients of these variables are meanwhile used in a spatially constrained multivariate clustering analysis to divide the housing market into submarkets. The OLS results of validation data within the submarkets are used to verify the rationality and reliability of the proposed classification method.

3.2. Salience-Weighted Hotspot Detection

3.2.1. POI Salience. The urban areas attracting high levels of attention and visit are hotspots for economic, educational, or
social activity. Urban planners should consider how to effectively identify these hotspots based on human preference, and characterize the attractiveness and assess the quality of these hotspots. The POI salience index is established based on social media data to represent the service ability of public facilities. The POI salience is calculated by public cognition from social network data to adjust the weight of hotspot detection of POIs, as follows:

\[
S_{POI} = w_A A + w_B B + w_C C, \tag{1}
\]

where \(S_{POI}\) is the POI salience; \(A\) is the public attention calculated by Weibo and Baidu search index data; \(B\) is different levels of POI; \(C\) is the levels of dependence on the distance to downtown; and \(w_A\), \(w_B\), and \(w_C\) are the weights of \(A\), \(B\), and \(C\), generally recommended as 0.4, 0.4, and 0.2, respectively [37]. All indicators of \(A\), \(B\), and \(C\) are normalized at a scale of 0 to 1 using min-max normalization:

\[
z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{2}
\]

where \(\max(x)\) signifies the maximum value in all observed values of \(x\), \(\min(x)\) is the minimum value, and \(z_i\) is the \(i\)th normalized data.

### Table 1: Top level of POI saliences of Shenyang.

| Name                                               | Type       | Baidu1   | Baidu2   | Baidu3   | Ckin | Indi | Center | \(S_{POI}\) |
|----------------------------------------------------|------------|----------|----------|----------|------|------|--------|-------------|
| Northeastern University                            | Edu        | 416765   | 156457   | 1568168  | 25125| 5    | 0.9    | 0.9770      |
| Shenyang Railway Station                           | Trans      | 97704    | 40964    | 73385    | 23288| 5    | 1      | 0.8551      |
| Shenyang North Railway Station                     | Trans      | 118192   | 58401    | 129440   | 25443| 5    | 1      | 0.8850      |
| Middle Street                                      | Comm       | 63355    | 7757     | 23476    | 11228| 5    | 1      | 0.7248      |
| The First Affiliated Hospital of China Medical University | Med       | 129447   | 62368    | 85042    | 5302 | 5    | 1      | 0.6984      |
| Taiyuan Street                                     | Comm       | 93400    | 25689    | 64057    | 3403 | 5    | 1      | 0.6631      |
| Shenyang Palace Museum                             | Sicht      | 49906    | 11361    | 25131    | 2643 | 5    | 1      | 0.6409      |
| Beiling Park                                       | Park       | 138401   | 29276    | 49126    | 2656 | 5    | 0.9    | 0.6487      |
| Fantawild Adventure                                | Ent        | 132657   | 73164    | 49126    | 2576 | 5    | 0.7    | 0.6157      |
| Liaoning Province Shiyan High School               | Edu        | 49805    | 4491     | 4716     | 704  | 5    | 0.9    | 0.6010      |

Baidu1 is based on searches within Shenyang, Baidu2 is based on searches within other cities of Liaoning Province, and Baidu3 is searches based on other cities outside Liaoning Province.
The parameter $A$ used in (1) is calculated by

$$A = w_u \sum \text{user}_{w} \cdot \text{Search}_I + w_i \cdot \text{Check}_I,$$

(3)

where $\text{user}_{w}$ is the weight of living regions. In general, local residents are more familiar with the neighborhood, since public cognitive abilities and attentions are constantly improved with experience accumulation. The search frequencies from the local residents are more specific, where the Baidu search index and check-in record user behaviors. Three sources of $\text{user}_{w}$ are derived by the Baidu search index from Shenyang city, other cities in province, and other cities outside the province. The weights are recommended as 0.7, 0.2, and 0.1, respectively; $w_u$ and $w_i$ are recommended as 0.4 and 0.6, respectively; parameter $B$ is related to the POI type.

The original POIs acquired from the Baidu map and check-in data have 16 categories. Some of the POIs such as convenience stores and residential areas are less prominent, and some of the POIs show similar functions. We choose the POIs with high influences and popularity to the public, such as hospitals and commercial centers. The 16 categories of the POIs are, therefore, reclassified as 8 categories, including transportation, education, sight, medical facility, park, commercial center, entertainment, and enterprise.

Parameter $C$ is ranked by the distance to the city center. In Shenyang, most roads and streets are connected by several major ring roads. These ring roads almost share the same center located near the Palace Museum. The ring roads are numbered from inner to outer of the city. The center degree of POI is the highest, located at the area of the 1st ring road, and decreases with the distance to the city center.

To avoid overly small or zero weight, the weighting formula of POI follows:

$$W_{POI} = \text{round} (S_{POI} \cdot 100 + 1).$$

(4)

According to equations (1)–(4), the POI saliences are calculated and 10 items greater than 0.6 are listed in Table 1.

### 3.2.2. Weighted Hotspot Detection

The quantitative characteristics of POIs are the prerequisite to identifications of the relations between public facilities and housing values. The distances from these POIs are calculated as explanatory variables to analyze the related housing price using MGWR. The Kernel Density tools in ArcGIS 10.3 are used to analyze the densities of POIs. The classification of the kernel density output raster is important for the results. The common method of natural breaks is insufficient to identify the differences between hotspots, and the traditional kernel density method is hard to quantitatively extract and express the hotspot peaks [38, 39].

The existing density-field detection method cannot use human preferences sufficiently; therefore, this article uses a modified method of salience-weighted density-field hotspot detection, SWDF-HD. This model is achieved by using the spatial analysis of kernel density and the local max-value extraction considering human preference. It is more effective to adjust the weight of the POIs and derive spatial distributions of high visits and qualities. The salience of POIs is initially defined as the population field value of the traditional kernel density tool, in order to construct a weighted kernel density surface as the focus statistics of the raster data. The neighborhood statistics method is then used to extract maximum value from the raster data and generate the extreme-value surfaces. The neighborhood statistics is the focal statistics used to calculate local density maxima of all cells in the neighborhood. The raster calculator tool is selected to execute map algebra calculation between the density and the extreme-value surfaces to extract the local maximum value. Finally, spatial structures of local density maxima are derived at different levels. Figure 5 shows the entire modeling process.

Not all POIs are required to adjust the saliences and analyze the housing price. The subway or bus station, for instance, requires no adjustment on saliences. The subway or bus station with high visits is mostly due to its location to the resident destination, not its attractiveness. In contrast, parents prefer to purchase school district houses to provide better education, healthcare, or life convenience for children. Therefore, the SWHD-HD method is used to extract local high-quality spatial locations of primary and secondary schools, large polyclinic hospitals, and commercial centers. A total of 5047 POIs are acquired with social network media, and the impacts of distance to these locations on the housing price are analyzed.
3.3. Multiscale Geographically Weighted Regression. The ordinary least squares linear regression, OLS, is applied to establish the relation between an explained variable \( y \) and a set of explanatory variables \( x \), as follows:

\[
y_i = \sum_{j=0}^{m} \beta_j x_{ij} + \epsilon_i, \tag{5}
\]

where \( y_i \) is the response variable of dependent variable \( i \); \( \beta_j \) is the \( j \)th parameter estimate; \( x_{ij} \) is the \( j \)th independent variable of the \( i \)th observation variable \( j = 1 \) to \( m \); and \( \epsilon_i \) is the normally distributed error. \( \hat{\beta} \) is calculated by the minimum sum of the squared differences between the observed and predicted values, as follows:

\[
\hat{\beta} = \left( X^T X \right)^{-1} X^T Y, \tag{6}
\]

where \( X \) is the explanatory variables matrix and \( Y \) denotes the vector composed of the \( n \) observed values of the explained variable.

GWR provides a local model to capture the spatial heterogeneity. In case there are \( n \) observations, \( i \in \{1, 2, \ldots, n\} \) is the observation number at location \((u_i, v_i)\), and the estimation of the predicted dependent variable \( y \) for the \( i \)th observation is given by

\[
y_i = \sum_{j=0}^{m} \beta_j (u_i, v_i) x_{ij} + \epsilon_i, \tag{7}
\]

where \( j \in \{0, 1, 2, \ldots, m\} \) denotes the feature number, \( y_i \) is the explained variable of the \( i \)th observation, and \( \beta_j (u_i, v_i) \) denotes the \( j \)th coefficient related to given spatial coordinates \((u_i, v_i)\); \( x_{ij} \) denotes the \( i \)th observed value of the \( j \)th feature and \( \epsilon_i \) signifies the error item.

The SGWR allows a subset of constant parameters across the space and a subset of variables, which can be regarded as a specific case of MGWR, as follows:

\[
y_i = \sum_{j=1}^{k_a} a_j x_{ij}(a) + \sum_{l=k_a+1}^{k_b} b_l (u_i, v_i) x_{il}(b) + \epsilon_i, \tag{8}
\]

where \( y_i \) is the \( i \)th observed explanatory variable; \((u_i, v_i)\) is spatial coordinates for each point \( i; j \in \{1, 2, \ldots, k_a\} \) is global feature number; \( l \in \{k_a+1, k_a+2, \ldots, k_b\} \) is local feature number; \( a_j \) is the estimation of the \( j \)th global parameter of fixed independent variable; \( b_l (u_i, v_i) \) denotes the \( l \)th local coefficient; and \( \epsilon_i \) represents the error term.

The MGWR model allows the conditional relations between the responses, and each explanatory variable occurs locally or regionally. Both GWR and SGWR assume that all relations vary at the same or the particular spatial scale. The MGWR relaxes implicit assumptions and identifies the relations between dependent variables and independent variables at various scales [40]. The MGWR model is given as follows:
\[ y_i = \sum_{j=0}^{k} \beta_{bwj}(u_i, v_i)x_{ij} + \varepsilon_i, \]  
(9)

where \( bwj \) denotes the bandwidth used for calibrating the \( j \)th conditional relation, \( (u_i, v_i) \) is the coordinates for each location \( i \), and \( \varepsilon_i \) is the error term.

The optimal bandwidth selection of MGWR is determined in a trial-and-error manner. A trial bandwidth is fitted using either GWR or SGWR, associated with corrected Akaike information criteria (AICc). A suitable bandwidth is the one minimizing the AICc, as follows:

\[ AICc = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + \frac{n + tr(S)}{n - 2 - tr(S)}, \]  
(10)

where \( n \) denotes the observation number, \( \hat{\sigma} \) represents the estimated standard deviation of residuals, and \( tr(S) \) denotes the trace of MGWR hat matrix \( S \).

### 3.4. Spatially Constrained Multivariate Clustering Analysis

The spatially constrained multivariate clustering tool in ArcGIS Pro 2.8 uses a connectivity graph and an algorithm called spatial cluster analysis by tree edge removal, SKATER, to construct natural clusters with space. SKATER is a method using the graph theory, with a pruning of the minimum spanning tree technique. A minimal spanning tree, MST, is initially constructed based on the connectivity graph representing the neighborhood relations among the geographic features. The classification method allows one or more variables that are used to calculate an integer optimal number of groups and control cluster size constraints. We take the set of coefficients of MGWR as analysis attributes and the housing price as the cluster size constraining parameter to cluster submarkets.

The MST reflects both spatial structure and associated field values of the features. A series of areal entities are related to a set of quantitative attributes \( \{A_1, \ldots, A_n\} \). The areal entities have an attribute vector \( x = (a_1, \ldots, a_n) \), where the attribute of \( A_1 \) has the possible value of \( a_0 \). A connected graph \( G = (V, L) \) consists of an empty set \( V \) of vertices and a set \( L \) of pairs of vertices termed edges. If \( s_i \) and \( s_j \) are adjacent, an edge is then constructed to connect vertices \( v_i \) and \( v_j \). The cost \( d(i, j) \) is connected with the edge of \( v_i \) and \( v_j \) vertices and calculated based on the attribute vectors \( x_i \) and \( x_j \) of real entity \( i \) and \( j \). A general choice is the Euclidean square distance, as follows:

\[ d_{ij} = d(x_i, x_j) = \sum_{l=1}^{n} (x_{il} - x_{jl})^2. \]  
(11)

The MST is built through Prim’s algorithm [41]. The algorithm constructs a graph \( G \ast \), including a group of trees \( T = (T_1, \ldots, T_n) \). All these trees are linked, but no edge or vertices are shared. \( G \ast \) includes one tree as the MST in the first iteration. In each iteration, one edge is pruned until the number of specified clusters is obtained. The selected edge minimizes the dissimilarity in the resultant clusters, avoiding (if possible) singletons (clusters with only one feature) measured by a minimum sum of the intracluster square deviations, as follows:

\[ Q(\Pi) = \sum_{j=0}^{k} \text{SSD}_j, \]  
(12)

where \( \Pi \) indicates a division of entities into \( k \) trees; \( Q(\Pi) \) denotes a parameter related to the \( \Pi \) partition quality, which needs to be minimized; and \( \text{SSD}_j \) summarizes the square deviations in region \( i \), as follows:

\[ \text{SSD}_k = \sum_{j=1}^{m} \frac{n_k}{nc} (x_{ij} - \bar{x}_j)^2, \]  
(13)

where \( n_k \) denotes the spatial entity number in tree \( k \); \( x_{ij} \) denotes the \( j \)th attribute of the entity \( i \); \( m \) is the attribute number; and \( \bar{x}_j \) is the mean magnitude of the \( j \)th attribute.

The Calinski–Harabasz pseudo-\( F \)-statistic is a common method to measure the clustering effectiveness, requiring an appropriate number of clusters, behaving as a ratio of the between-cluster variance to the within-cluster variance, as follows:

\[ F = \frac{\left( R^2/n_k - 1 \right)}{\left( 1 - R^2/m - n_k \right)}. \]  
(14)

where \( R^2 = \text{SST} - \text{SSE}/\text{SST} \) and SST reflect between-cluster differences and SSE reflects intracluster differences; \( n \) denotes the entity number; and \( n_k \) is the cluster number.

### 4. Results of the Housing Submarket Classification

#### 4.1. Predictor Variables and Global Modeling

In the application of spatial hedonic price model, a linear function is applied to directly reflect the influences of different factors. The dependent variable is the housing price in prescribed communities. To avoid multicollinearity, twelve factors are selected to explain the response variables of the housing price, including basic condition, traffic, environment, and living condition of the community. The meaning and classification of the twelve variables are shown in Table 2.

For living conditions, the distance to the nearest facilities is used by the method of SWDF-HD. The numbers of POIs related to healthcare and education are less than the total numbers of the commercial facilities, such as shopping mall, restaurants, and supermarkets. Therefore, the kernel bandwidth for each three POI types are different in the estimation of the kernel density. The average edge of the streets of Shenyang is 0.224 km, and the multihotspots of the three types are calculated with the bandwidth from 0.1 km to 1 km. For commercial centers, the hotspots in the first level are over-detected below 0.2 km and under-detected above 0.6 km. Consequently, an average of 0.4 km is used as the optimal bandwidth from 0.3 km to 0.6 km. For medical and educational centers, the optimal bandwidth is 0.7 km. Similarly, the education hotspots and commercial facilities are 182 and 498, respectively. Figure 6 shows the preliminary
Table 2: Selection and quantification of housing price and influential factors.

| Category                     | Variables       | Definition                                                                 | Min   | Max   | Mean  | Estimate | VIF  |
|------------------------------|-----------------|-----------------------------------------------------------------------------|-------|-------|-------|----------|------|
| Dependent variables          | Price           | Transaction price of the house per community (10,000CNY/m²)                | 0.319 | 4.254 | 0.953 | —        | —    |
| Basic condition              | Area            | Living area (m²)                                                           | 30.300| 481.940| 93.400| 0.108    | 1.127|
|                              | Age             | The age of buildings till December 2017                                    | 0.333 | 10.094| 29.000| −0.057   | 2.595|
|                              | Property management fee (Fee) | Community management fee (1.0CNY/m²/month)                              | 0.030 | 5.500 | 1.413 | 0.604*   | 2.256|
| Community condition          | Floor area ratio (FAR) | The ratio of building’s useable area to the total area in China | 0.200 | 12.000| 2.581 | −0.045   | 1.220|
|                              | Greenspace ratio (GR) | The ratio of the green land to the area of the residential area         | 0.050 | 0.800 | 0.335 | −0.011   | 1.281|
| Public transport             | Bus             | Distance to the nearest bus stops (km)                                    | 0.010 | 0.949 | 0.179 | 0.004    | 1.191|
|                              | Subway          | Distance to the nearest subway station (km)                               | 0.089 | 10.429| 2.088 | −0.021   | 1.541|
| Environment condition        | Park            | Distance to the nearest park (km)                                         | 0.022 | 8.123 | 1.829 | −0.180*  | 2.319|
|                              | River           | Distance to the nearest river or reservoir (km)                          | 0.010 | 3.096 | 0.910 | −0.087*  | 1.201|
|                              | Commercial center (CC) | Distance to the nearest commercial centers of shopping malls, supermarkets, and restaurants (km) | 0.022 | 2.685 | 0.452 | 0.028    | 1.471|
| Living condition             | School          | Distance to the nearest schools (km)                                      | 0.071 | 4.812 | 0.856 | −0.231*  | 1.691|
|                              | Hospital        | Distance to the nearest hospitals (km)                                    | 0.041 | 6.928 | 1.175 | −0.077*  | 2.370|

N = 706, AICc = 1567.071, intercept = 0, *five percent level of significance.
data of the POIs and the hotspot maps of the hierarchical scale distribution for commercial, medical, and educational service based on the SWDF-HD model, respectively.

All data are processed using data cleaning and collinear processing methods before estimation, so that abnormal values are eliminated. A total of 706 effective samples are collected. Table 2 also lists the parameter estimates and inference results using the ordinary least squares method, OLS. In order to compare and distinguish the dominant factors to the housing price, the response and predictor variables are standardized so that they are concentrated at zero when the same variation range is used. The property management fee is the most significant factor with a coefficient of 0.604, and the second one with 0.231 is the shortest distance to educational hotspots. In contrast, the shortest distance to bus stations is less important. Residents prefer high-quality of living and education, and the bus system is not the primary transportation in Shenyang. The significance of the model variables is examined by a robust test and marked by star. All of the variance inflation factors (VIF) are below 3, implying no global significant multicollinearity or redundancy among the variables.

The $R^2$ of OLS regression is 0.483, and Adj. $R^2$ (adjusted $R^2$) is 0.474; therefore, the OLS model explains 48% of the variation in the housing price in the research region. The statistical results such as the Koenker (BP) statistic (105.977...
and $p < 0.001$) and Jarque–Bera statistic (1824.814 and $p < 0.001$) are statistically significant, indicating nonstationary in the OLS model.

### 4.2. Local Modeling with MGWR

Twelve variables are involved in the GWR and MGWR models with the Python package MGWR [42]. Table 3 shows performance evaluation of five models of OLS, GWR, and MGWR with fixed and adaptive spatial kernel using 12 variables, in terms of Adj. $R^2$, AIC-corrected (AICc) values, residual sum of squares, and Moran’s I test for residual spatial auto-correlation. In contrast to the global model of OLS, both the GWR and MGWR models show significant improvements for Adj. $R^2$, residual sum of squares, and AICc. The MGWR shows better performance than GWR, irrelevant to the selected kernel function. The best performance is the MGWR model using an adaptive kernel function. It has an AICc, $R^2$, and Adj. $R^2$ of 1035.531, 0.862, and 0.820, respectively. The residual square sum is 102.809, Moran’s index for the model residuals is −0.003, and the $p$ value is 0.726. The MGWR model is, therefore, selected with the adaptive kernel function for estimation after comparing the results of these models.

Table 4 lists the MGWR covariate-specific bandwidths, GWR bandwidths, and bandwidth confidence intervals, adjusted critical $t$ values, and Monte Carlo test for spatial variability with 95% confidence intervals. The bandwidth in the MGWR is from 43 to 705. In contrast, the optimal GWR bandwidth is 105, implying that the conditional relations corresponding to the housing price are spatially different. These scales are categorized into three groups at global, regional, and local scales. The global variables (larger than 600) include housing age, floor area ratio, green space ratio, and the shortest distance to bus stops and subway stations, demonstrating consistent effects and no significant spatial heterogeneity across the study region. Local variables (bandwidth less than 50) are 6% of the total number of the housing sampling points. Three local variables include area and the shortest distances to the park and river. The study region covers a total area of 1,254 km$^2$, and 6% of the area is 74.7 km$^2$. The remaining parameters are also at regional scales, and spatial distributions of these parameters are relatively similar, exhibiting spatial nonstationarity at the regional scales.

Table 5 summarizes the statistics of the parameter estimations using the MGWR method. The housing area, property management fee, green space ratio, and the shortest distance to commercial centers impose positive influences on the housing price, and the other parameters show negative influences. The housing price is higher along with larger size, higher property management fee, more green space and closer distance to the bus and subway stations, and high-rank hospitals and schools. In contrast, the housing price is lower in old or dense-populated communities. Furthermore, the housing price decreases when the house is near commercial centers, largely due to external noises in these areas. One of the most significant variables is the property management fee, showing that high management quality is the most concerned factor for residents, followed by environment and education. The shortest distance to bus stops has the lowest impact on the housing price. Although the bus

| Parameter | MGWR bandwidth | GWR bandwidth | MGWR CI (95%) | Adj. $t$-val (95%) | $p$ value |
|-----------|----------------|---------------|---------------|-------------------|----------|
| Intercept | 95             | 105           | (79.0, 116.0) | 2.726             | 0.000    |
| Area      | 43             | 105           | (43.0, 45.0)  | 3.223             | 0.000    |
| Age       | 683            | 105           | (549.0, 692.0) | 2.038             | 0.451    |
| Fee       | 79             | 105           | (65.0, 102.0) | 2.990             | 0.032    |
| FAR       | 705            | 105           | (549.0, 705.0) | 2.031             | 0.824    |
| GR        | 705            | 105           | (549.0, 705.0) | 2.056             | 0.966    |
| Bus       | 705            | 105           | (549.0, 705.0) | 2.123             | 0.875    |
| Subway    | 705            | 105           | (609.0, 705.0) | 2.011             | 0.996    |
| Park      | 45             | 105           | (44.0, 51.0)  | 3.121             | 0.000    |
| River     | 45             | 105           | (44.0, 47.0)  | 3.131             | 0.000    |
| CC        | 68             | 105           | (560.0, 79.0) | 3.066             | 0.001    |
| School    | 109            | 105           | (102.0, 139.0) | 2.850             | 0.000    |
| Hospital  | 146            | 105           | (102.0, 162.0) | 2.721             | 0.000    |

Table 3: GWR and MGWR results.

| Measure   | OLS   | GWR Model1 | GWR Model2 | MGWR Model1 | MGWR Model2 |
|-----------|-------|------------|------------|-------------|-------------|
| $R^2$     | 0.483 | 0.835      | 0.755      | 0.862       | 0.837       |
| Adj. $R^2$| 0.474 | 0.777      | 0.687      | 0.820       | 0.786       |
| AICc      | 1567.071 | 1232.397  | 1401.538   | 1035.531    | 1170.130    |
| RSS       | 365.353 | 116.499    | 173.308    | 97.391      | 115.194     |
| MIR       | 0.133 ($p = 0.000$) | −0.002 ($p = 0.897$) | 0.014 ($p = 0.006$) | −0.003($p = 0.726$) | −0.008 ($p = 0.383$) |

Abbreviation: RSS, residual sum of squares; MIR, Moran’s I test for residual; Model1 is the adaptive type kernel function and Model2 is used for the fixed-type kernel function.

Table 4: Optimal bandwidths for parameters derived from MGWR and GWR.
transportation is normally cheaper and widespread, disadvantages such as overload and traffic congestion make this kind of transportation less attractive. With the improvement of the transportation system, residents prefer to live far away from the noises of the commercial centers and the street traffic in the study region.

The MGWR results are visualized to analyze spatial differences of the individual parameter estimates. The negative parameter estimates are expressed in blue and positive values in red. The insignificant estimates less than the absolute value of critical $t$-value at a significance level of 95% are set to zero and displayed in grey.

Three significant global factors are the housing age, floor area ratio, and shortest distance to subway stations, as shown in Figures 7(a)–7(c). The same color denotes similar spatial impacts on the housing price. The cold blue colors corresponding to the house age, floor area ratio, and the shortest distance to subway stations are negative impacts on the housing price. According to Figure 7(a), the coefficients of the area are slightly changed, and the maximum value is in the center and the minimum value is in the west. Figure 7(b) shows that the floor area ratio in south is dark blue, implying larger population density and building density. Figure 7(c) shows that the traffic effects in the north are less important than the effects in the central and southern areas.

Figures 7(d)–7(f) show three local estimated coefficients, which are spatially diversified, corresponding to the house area and the shortest distances to park and river. According to Figure 7(d), the area shows both positive (red colors) and negative effects (blue colors) on the housing price. The local area estimates show significant positive effects in two zones. The first zone is located in the central business district, namely, Tiexi square. The second zone is the Heping District near Hunhe River. The houses with larger areas are more popular in these zones due to better public facilities and good view. In contrast, small houses are more popular for temporary residences near the Taiyuan Street marked blue. As the most popular commercial pedestrian street, the Taiyuan Street includes a large number of stores and shopping malls. A large number of visitors make this region less suitable for long-term living, due to noises and lighting pollution.

Table 5: Statistical description of parameter estimates for the MGWR model.

| Parameter  | Mean  | Std   | Min   | Median | Max  |
|------------|-------|-------|-------|--------|------|
| Intercept  | −0.083| 0.259 | −0.505| −0.183 | 0.487|
| Area       | 0.084 | 0.271 | −0.433| 0.057  | 1.663|
| Age        | −0.138| 0.006 | −0.152| −0.140 | −0.124|
| Fee        | 0.404 | 0.147 | 0.070 | 0.431  | 0.656|
| FAR        | −0.067| 0.004 | −0.073| −0.068 | −0.056|
| GR         | 0.025 | 0.002 | 0.029 | 0.035  | 0.038|
| Bus        | −0.008| 0.004 | −0.016| −0.009 | 0.001|
| Subway     | −0.096| 0.001 | −0.099| −0.096 | −0.094|
| Park       | −0.307| 0.511 | −2.378| −0.241 | 0.380|
| River      | −0.184| 0.295 | −1.191| −0.158 | 0.375|
| CC         | 0.049 | 0.163 | −0.397| 0.039  | 0.593|
| School     | −0.134| 0.203 | −0.523| −0.104 | 0.272|
| Hospital   | −0.036| 0.306 | −0.557| −0.125 | 0.523|

The blue colors of Figures 7(e) and 7(f) show negative relations between the housing price and the shortest distance to park and river. Residents in Shenyang tend to live in good ecological environments. Figure 7(e) shows significant local parameter estimates corresponding to the shortest distance to parks. The housing price is higher near the parks of Changbaidao and Shenshuiwan located along the Hunhe River. The zone showing a similar tendency is located in the Taiyuan commercial zone near Zhongshan Park. These areas show positive demand for urban green space. Figure 7(f) shows the significant effects of the shortest distance to river. Blue colors imply that a shorter distance to the river corresponds to a higher house price in Nanhu Street near the South canal and Wulihet Street near the Hunhe River.

Figure 8 shows the regional parameter estimates for the variables. Figure 8(a) show the coefficient distribution of the property management fee. Red colors indicate that higher property management fee is in general related to higher housing price, where better community management is served. The central part in Shenyang is more sensitive to the property management fee. Most families prefer high-quality property management.

Figure 8(b) shows the estimated coefficients for the nearest distance to large commercial centers. This is opposite to the property management fee. The central sections along the Hunhe River show dark red, where a longer distance to commercial centers corresponds to a higher housing price. Residents in this zone prefer high-quality and quiet living environments.

Figure 8(c) shows the coefficients of the nearest distance to school. Most communities near Tiexi square are in dark blue colors. The higher housing price is related to the shorter distance. Similar trends are observed in the Taiyuan commercial zone, Changbai and Shenshuiwan Streets of the Heping District. Based on Figure 6, three well-know, eight very popular, and nine popular educational hotspots are observed in these regions. Therefore, the shorter distance to these hotspots equals higher prices.

Figure 8(d) shows different distributions of the shortest distance to hospital. In the red zone, the hospital is a positive factor, where the housing price is higher when it is far away from the hospital. According to Figure 6, the living environment possibly becomes crowded and less attractive, since high-rank hospitals are frequently visited by many patients each day. On the contrary, the blue zone around Changbai Street has the opposite trend, where the hospital becomes a negative factor to the housing price. A higher housing price is expected when the house is closer to hospitals. High-quality medical resources in this zone are positive demands, due to the lack of high-rank hospitals. Figure 8(d) also highlights the coefficients of the medical resources in the northern area are larger than those in the southern area. High-rank medical resources are concentrated in central and northern areas, and the medical resources are insufficient in the southern area.

Figure 9 shows the coefficient distribution for the intercept of the housing price with the MGWR model. The average housing price in each region is observed while all covariates are constants, regarded as the localized effects for
the housing price. The positive effects are restricted in the central business district of Tiexi square, attributed to convenient public transportation and excellent educational institutions and medical facilities. Lower housing prices are localized in old residence areas of Nanta street and northern suburb of Shenyang.

4.3. Housing Submarket Structures in Core Region of Shenyang

4.3.1. Spatially Clustering via Spatial Coefficients. The spatially constrained multivariate clustering tool with the unsupervised machine learning method, namely, SKATER, is applied to conduct clustering analysis. The results of MGWR are modeled in 1 km × 1 km grid cells. The factors sharing similar characteristics based on seven significant coefficients of local and regional variables with smaller bandwidths of MGWR (the bandwidth is less than 200 are included) are selected as analysis fields. The housing price is set as the cluster size constraining parameter.

The derivation of the optimal clustering number is the first step for unsupervised clustering algorithms. Figure 10 shows a plot between the pseudo-$F$-statistic vs number of clusters from 2 to 30. The method is used to derive the key point, after which the pseudo-$F$-statistic starts increasing, suggesting that 10 is the optimal choice for running cluster algorithm.

A global spatial pattern of ten submarkets in Shenyang demonstrates that the housing price in the core area is higher than the price in suburban areas, and the housing price in south submarkets is higher than the price in north (Figure 11(a)). Submarket 1 has a relatively high average price, covering both sides of the Hunhe River. Submarkets 2 and 3 cover the central areas of the Tiexi district and partial Heping District, including Tiexi square and Taiyuan commercial center. These areas include popular shopping centers, such as Star Mall, IKEA, and Taiyuan pedestrian street. Submarket 4 involves a cluster of financial services from Wulihe Street to Beiling Street termed Golden Corridor, including the Palace Museum, Beiling park, and Zhongjie pedestrian street. Submarket 5 includes partial Hunnan District and partial Shenhe District. The planning policy affects the classification of the housing submarkets. Submarket 6 is located in south Shenyang, and submarkets 7 and 8 are located in the southwest Shenyang, regarding Tiexi and Yuhong districts. Submarkets 9 and 10 cover a larger area in the north of the city.

Figure 11(b) shows relations between MGWR coefficients of seven local variables. The most significant factors for Submarket 1 include three variables of the shortest distances to the park, hospital, and commercial center. The average housing price becomes higher when the distance to public parks and hospitals is short, and the price is higher when the houses are far away from the commercial center.
Figure 8: Significant local parameter estimates of MGWR for regional variables; (a) property management fee; (b) the shortest distance to commercial centers; (c) the shortest distance to schools; (d) the shortest distance to 3A hospitals.

Figure 9: MGWR estimated coefficients of the intercept.
Residences in these areas, mostly in good financial conditions, prefer quiet and relaxing residential area with good views, educational resources, and infrastructures. The natural surroundings become the most important factor for the housing values. For houses related to submarket 2, the distance to school is clearly more important than others, where the house is more expensive when the distance to schools is shorter. The educational resource has irreplaceable effects in this region, which is opposite to submarket 9. In addition, the housing size is an important factor. Residents in submarket 2 prefer to purchase larger school district apartments with sufficient space for three or four family members. The shortest distance to the hospitals is the most sensitive factor for submarkets 3 and 7. The price of the houses close to the hotspot’s hospitals decreases. The First Affiliated Hospital of China Medical University, as one of the top hospitals in China, is located in this area. With a large flow of patients from different cities, short-term rental is very popular in the surrounding area of the hospital; therefore, safety issues may arise. In submarket 4, the river plays a significant role in influencing the price, where the south canal runs through this region. The housing price is higher when the houses are closer to the commercial center in submarket 5. The influencing levels of each factor are equal, and no factors show obvious significant influences in submarkets 6 and 8. For submarket 7, the property management fee is a major factor in this region, where residents prefer to purchase houses with a high-quality property management level. In contrast, the property management level is less-sensitive factor in submarket 10. This may be attributed to the lack of high-quality property management and inactive real estate development.

4.3.2. Validation. The OLS regression of each submarket is conducted to validate the model and to observe whether the regression results of most submarkets are optimized and $R^2$ is higher. The OLS regression Adj. $R^2$ and $R^2$ of all samples are 0.34 and 0.32, respectively. Table 6 lists the OLS regression of each submarket. Most of the regression results are improved significantly except for submarket 2, possibly due to the unbalanced and over-concentrated samples of this submarket. The classified submarkets mark the spatial heterogeneity of the housing market effectively, meanwhile showing that basic condition, traffic, environment, and living condition of the communities affect the housing submarkets. The absolute values of coefficients greater than 0.20 are significant and highlighted in Table 6. The coefficients of the global variables such as the distance to bus and subway stations have similar influences on all the housing submarkets. In contrast, local variables, such as housing area, property management fee, and the distance to parks, show large differences. For submarkets 1 to 4 in the urban areas, the housing size, the distance to the subway, park, and the popular facilities all influence the housing market characteristics. Meanwhile, high-popularity living conditions of the communities are more significant than the basic condition variables. For submarkets 5 to 10 in the suburban zones, high-quality educational and medical resources are still important factors, although the influences of living conditions obviously reduce. The allocation of public resources strongly affects the housing values and the distribution of the submarkets.

According to the above analysis, unbalanced structure of the housing market exists in Shenyang. This unreasonable situation needs adjustment, according to the characteristics of each submarket. The submarkets in the urban center show
excessive concentration of public facilities, such as commercial centers in submarkets 2 and 3, and the medical resources in submarket 2 and the educational resources in submarkets 1 and 4. Strategies of urban renewal may be adopted, including rehabilitation for existing old buildings and infrastructures, and decentralization for high-quality structures. High-tech and low-carbon industrial clusters may be constructed in the north of suburban areas due to lower costs. Urban green ecological corridor and transport facilities may be planned in the north of the city, in order to increase commuter amount, infrastructure, and land price. Several aspects should be considered before any specific plan is performed, such as urban development policy, financial resource allocation, and land acquisition.
In this article, we use geo-tagged social media data to describe the human preference of urban facilities. We extract geographic patterns and regularities of commercial, medical, and educational facilities based on the visiting numbers and popularity degrees of different types of POIs in the core area of Shenyang. The spatial configuration and attractiveness of the commercial centers, schools, and hospitals are examined, helping urban planners for reallocation and equalization of public resources. Although social media data are widely used, few studies have investigated how popular perceptions of urban critical hotspots affect the housing price. We select these hotspots associated with human preferences as independent variables to analyze spatial effects on the housing price, providing new options to classify the housing submarket.

The selection of the influential factors is an important consideration in this research. Literature review shows the existence of some common factors tightly related to the housing price, such as building or house age, residential area, distance to the commercial center, distance to the river, traffic condition, living facility (submarkets or shopping facilities), education facility, and distance to the park [10, 21, 43]. In addition, the green space ratio, property management, and the distance to sport facility are chosen as independent variables. According to reference [21], selected variables are chosen and divided into four categories, including property structure, basic condition, traffic condition, and living condition. Based on the common factors, the housing area, bedroom number, decoration condition, and orientation are as new property condition variables, the property management fee, floor area ratio, and green ratio as new basic condition variables, and the shortest distance to the movie theater as a new living condition. Factors may be classified into different categories; for instance, the distance to the park is defined as a natural environment condition or living condition. We initially try to involve as many variables as possible in analyzing the housing price, such as the decoration condition or orientation. However, these values are too concentrated, possibly due to data insufficiency. To avoid multicollinearity, we finally select twelve factors to explain the independent variables of the housing price, including basic condition, traffic, environment, and living condition of the community.

We then compare our results with the results of previous studies. The house age, floor area ratio, and distance to subway stations commonly show negative effects [44, 45]. The property management fee and green ratio show an opposite trend, suggesting that the houses near the subway with high-level property management and elegant inner environment are more expensive [46]. The effects of other variables show variance across the space. For instance, the residential area shows spatially changed effects. People prefer to choose small houses with higher housing price when the house is near high-quality schools. People pay special attention to public facilities such as schools, supermarkets, and hospitals in the inner-city area than suburb area [47–49]. Moreover, the characteristics of parks and rivers have different effects on the housing price, where people prefer to live in an environment with a good view to green space and river.

To further confirm the validity of this research framework and compare results with the original data and coefficients of MGWR, we select the original data of the housing price and the factors, where the optimal number of clusters is derived as 3. In order to compare the results in the same cluster, we still choose 10 as the cluster number, as shown in Figure 12. The differences in the housing prices between the north and south areas of Shenyang are observed in both results. Submarket 1 with the highest price is in the southeast area, where the city government is just relocated, attracting extensive interests. Submarket 2 mostly overlaps submarkets 1, 3, and 4 in Figure 11 and shows the second place of the housing price. The submarket 2 is located in the center of Shenyang with sufficient public facilities, but the house age is old and the green space is low due to poor urban planning. Public facilities in suburban submarkets 8, 9, and 10 are underdeveloped, similar to the submarkets in the suburban areas in Figure 11. The factors demonstrated by Figure 12 show less heterogeneity with smaller intervals between each other than the factors in Figure 11, leading to a

| Table 6: OLS regression results of each submarket. |
|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Submarket     | 1           | 2           | 3           | 4           | 5           | 6           | 7           | 8           | 9           | 10          |
| Area          | 0.378       | 1.429       | -0.178      | -0.275      | -0.007      | -0.088      | -0.024      | -0.068      | 0.072       | 0.162       |
| Age           | -0.595      | -0.007      | -0.207      | -0.337      | -0.184      | -0.199      | -0.025      | -0.122      | -0.327      | 0.182       |
| Fee           | 0.360       | 1.223       | 0.037       | 0.134       | 0.656       | 0.270       | 0.604       | 0.415       | 0.291       | 0.298       |
| FAR           | -0.346      | 0.495       | -0.058      | 0.026       | -0.298      | -0.257      | -0.028      | 0.075       | -0.190      | 0.508       |
| GR            | 0.140       | -0.624      | 0.093       | 0.154       | -0.053      | 0.045       | 0.173       | -0.083      | -0.030      | 0.053       |
| Bus           | 0.092       | -0.167      | -0.028      | 0.109       | 0.003       | -0.112      | 0.201       | -0.044      | 0.015       | -0.054      |
| Subway        | 0.502       | -0.352      | -0.697      | 0.247       | 0.060       | -0.159      | -0.234      | -0.284      | -0.359      | 0.383       |
| Park          | -1.430      | 0.843       | 0.309       | -0.256      | 0.167       | -0.218      | 0.300       | 0.331       | -0.071      | 0.106       |
| River         | 0.698       | -0.275      | -0.218      | -0.565      | 0.028       | -0.161      | -0.064      | 0.021       | -0.074      | 0.083       |
| CC            | -0.083      | 0.271       | 0.102       | -0.130      | -0.135      | 0.103       | 0.194       | -0.012      | 0.094       | 0.051       |
| School        | -0.069      | -0.795      | -0.206      | -0.201      | -0.196      | 0.004       | -0.041      | -0.100      | 0.399       | -0.046      |
| Hospital      | -0.229      | 0.293       | -0.132      | -0.996      | -0.030      | -0.075      | -0.406      | 0.086       | -0.326      | -0.418      |
| R²            | 0.84        | 0.85        | 0.37        | 0.57        | 0.66        | 0.61        | 0.78        | 0.76        | 0.67        | 0.66        |
| Adj.R²        | 0.81        | 0.74        | 0.29        | 0.50        | 0.60        | 0.55        | 0.65        | 0.67        | 0.61        | 0.60        |

Significant coefficients (the absolute values are larger than 0.2) are bolded.

5. Discussion
smaller optimal number of the clusters and contiguous areas in the southwest area. Therefore, the classifications of the submarkets using MGWR coefficients of Shenyang are more reasonable and present higher distinction degree than the results based on the original unprocessed data.

It is worth noting that since the social media data are limited in this research, the scope of the housing price samples is small and the validity of the results needs to be verified with more samples in future. We will use this framework to investigate the spatiotemporal variation of the housing price of each submarket.

Figure 12: Submarkets in Shenyang with original data; (a) spatial distribution of the housing submarkets in Shenyang; (b) the factors and housing price of each submarket.
housing submarkets and to capture spatial variability of the clusters. In addition, only hotspots for commercial, medical, and educational services considered critical factors are analyzed. We will explore and compare the relations between different types of POI hotspots with human preferences and housing prices. More types of geo-tagged data such as food and service comments from online websites or taxi GPS may also be applied to analyze the urban attractive and emotional maps and preferences.

6. Conclusions

This research presents a comprehensive data-driven three-step frame to classify housing submarkets considering human preferences in Shenyang, China. This study uses POI saliences density-field hotspot detector to quantify human preference to public facilities. The MGWR model is applied to investigate spatial relations between the housing price and housing characteristics at multiple scales. The SKATER method is applied to classify the housing submarkets according to the MGWR results. The proposed frame is used to identify the spatial cluster degree of similar factors affecting the housing price. The following conclusions are summarized:

(1) The human preferences to public facilities are spatially heterogeneous and impose different influences on the housing price. Social media data are used to quantify the popularities of public facilities, in order to analyze the influential factors on the housing price in a more accurate and objective manner. High-quality public facilities significantly impact the housing values at local scales; therefore, submarket classification should focus on the effects of the weights of POIs in Shenyang.

(2) The MGWR model provides a more realistic and scalable spatial process model, allowing for different variables with individual suitable bandwidths. The variables of housing size and the shortest distances to park and river impose effects at local scales. In contrast, the variables of housing age, floor area ratio, green space ratio, and the shortest distances to bus and subway stations demonstrate global effects on the housing price. Among all variables, the shortest distance to the bus stop is the least important factor. The spatial heterogeneity of neighborhood environment and high-quality public facilities is more significant in Shenyang.

(3) The SKATER cluster analysis reveals the advantages to identify the spatial characteristics of the housing submarkets based on the MGWR coefficients. Significant unbalance of the housing price exists in Shenyang, where the housing price in the city center is higher than the price in suburban areas. Similar unbalance also exists between south and north areas of Shenyang.

Data Availability

Supplementary data to this article are available upon request to the authors.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the Natural Science Fund of China (Grant no. 71501036).

References

[1] C. Wu and R. Sharma, “Housing submarket classification: the role of spatial contiguity,” *Applied Geography*, vol. 32, no. 2, pp. 746–756, 2012.
[2] M. Helbich, W. Brunauer, J. Hagenauer, and M. Leitner, “Data-driven regionalization of housing markets,” *Annals of the Association of American Geographers*, vol. 103, no. 4, pp. 871–889, 2013.
[3] B. Keskin and C. Watkins, “Defining spatial housing submarkets: exploring the case for expert delineated boundaries,” *Urban Studies*, vol. 54, no. 6, pp. 1446–1462, 2017.
[4] F. Carriazo and J. A. Gomez-Mahecha, “The demand for air quality: evidence from the housing market in Bogota, Colombia,” *Environment and Development Economics*, vol. 23, no. 2, pp. 121–138, 2018.
[5] G. Costello, C. Leishman, S. Rowley, and C. Watkins, “Drivers of spatial change in urban housing submarkets,” *The Geographical Journal*, vol. 185, no. 4, pp. 432–446, 2019.
[6] C. Wu, X. Ye, F. Ren, Y. Wang, P. Ning, and Q. Du, “Spatial and social media data analytics of housing prices in Shenzhen, China,” *PLoS One*, vol. 11, no. 10, Article ID e0164553, 2016.
[7] G. Bello-Orzeg, J. J. Jung, and D. Camacho, “Social big data: recent achievements and new challenges,” *Information Fusion*, vol. 28, no. 3, pp. 45–59, 2016.
[8] F. Zhang, J. Wu, M. Hu et al., “Uncovering inconspicuous places using social media check-ins and street view images,” *Computers, Environment and Urban Systems*, vol. 81, Article ID 101478, 2020.
[9] Y. Shen and K. Karimi, “Urban function connectivity: characterisation of functional urban streets with social media check-in data,” *Cities*, vol. 55, pp. 9–21, 2016.
[10] Y. Wu, Y. D. Wei, and H. Li, “Analyzing spatial heterogeneity of housing prices using large datasets,” *Applied Spatial Analysis and Policy*, vol. 13, no. 1, pp. 223–256, 2020.
[11] K. J. Lancaster, “A new approach to consumer theory,” *Journal of Political Economy*, vol. 74, no. 2, pp. 132–157, 1966.
[12] S. M. Rosen, “Hedonic prices and implicit markets: product differentiation in pure competition,” *Journal of Political Economy*, vol. 82, no. 1, pp. 534–555, 1974.
[13] M. Brennan, D. Olaru, and B. Smith, “Are exclusion factors capitalized in housing prices?” *Case Studies on Transport Policy*, vol. 2, no. 2, pp. 50–60, 2014.
[14] K. Cockx and F. Canters, “Determining heterogeneity of residential location preferences of households in Belgium,” *Applied Geography*, vol. 124, Article ID 102271, 2020.
[15] K. Kopczewska and P. ´Cwiakowski, “Spatio-temporal stability of housing submarkets. Tracking spatial location of clusters of geographically weighted regression estimates of price determinants,” *Land Use Policy*, vol. 103, no. 2, Article ID 105292, 2021.
[16] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, “Some notes on parametric significance tests for geographically weighted regression,” *Journal of Regional Science*, vol. 39, no. 3, pp. 497–524, 1999.
[17] C. Bitter, G. F. Mulligan, and S. Dall’erba, “Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method,” *Journal of Geographical Systems*, vol. 9, no. 1, pp. 7–27, 2007.

[18] D. M. Hanink, R. G. Cromley, and A. Y. Ebenstein, “Spatial variation in the determinants of house prices and apartment rents in China,” *The Journal of Real Estate Finance and Economics*, vol. 45, no. 2, pp. 347–363, 2010.

[19] A. S. Fotheringham, R. Crespo, and J. Yao, “Geographical and temporal weighted regression (GTR),” *Geographical Analysis*, vol. 47, no. 4, pp. 431–452, 2015.

[20] S. J. Qu, S. G. Hu, W. D. Li, C. R. Zhang, Q. F. Li, and H. Wang, “Temporal variation in the effects of impact factors on residential land prices,” *Applied Geography*, vol. 114, p. 102124, 2020.

[21] D. Wang, Y. J. Li, and H. Yu, “Mass appraisal modeling of real estate in urban centers by geographically and temporally weighted regression: a case study of Beijing’s core area,” *Land*, vol. 9, no. 5, p. 143, 2020.

[22] C. H. Wei and F. Qi, “On the estimation and testing of mixed geographically weighted regression models,” *Economic Modelling*, vol. 29, no. 6, pp. 2615–2620, 2012.

[23] R. Pan, H. Yang, K. Xie, and Y. Wen, “Exploring the equity of traditional and ride-hailing taxi services during peak hours,” *Transportation Research Record*, vol. 2674, no. 9, pp. 266–278, 2020.

[24] H. T. Yang, Y. B. Zhang, L. Z. Zhong, X. J. Zhang, and Z. W. Ling, “Exploring spatial variation of bike sharing trip production and attraction: a study based on Chicago’s Divvy system,” *Applied Geography*, vol. 115, Article ID 102130, 2020.

[25] A. S. Fotheringham, W. Yang, and W. Kang, “Multiscale geographically weighted regression (MGWR),” *Annals of the American Association of Geographers*, vol. 107, no. 6, pp. 1247–1265, 2017.

[26] H. Yu, A. S. Fotheringham, Z. Li, T. Oshan, W. Kang, and L. J. Wolf, “Inference in multiscale geographically weighted regression,” *Geographical Analysis*, vol. 52, no. 1, pp. 87–106, 2020.

[27] Z. Shabrina, B. Buyuklieva, and M. K. M. Ng, “Short-term rental platform in the urban tourism context: a geographically weighted regression (GWR) and a multiscale GWR (MGWR) approaches,” *Geographical Analysis*, vol. 53, no. 4, pp. 666–707, 2021.

[28] Y. Chen, M. Zhu, Q. Zhou, and Y. Qiao, “Research on spatiotemporal differentiation and influence mechanism of urban resilience in China based on MGWR model,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, p. 1056, 2021.

[29] S. J. Smith and M. Munro, “The microstructures of housing markets,” *Housing Studies*, vol. 23, no. 2, pp. 159–162, 2008.

[30] C. Watkins, “Property valuation and the structure of urban housing markets,” *Journal of Property Investment & Finance*, vol. 17, no. 2, pp. 157–175, 1999.

[31] S. C. Bourassa, E. Cantoni, and M. Hoelsi, “Spatial dependence, housing submarkets, and house price prediction,” *The Journal of Real Estate Finance and Economics*, vol. 35, no. 2, pp. 143–160, 2007.

[32] Z. Chen, S. H. Cho, N. Poudyal, and R. K. Roberts, “Forecasting housing prices under different market segmentation assumptions,” *Urban Studies*, vol. 46, no. 1, pp. 167–187, 2009.

[33] S. C. Bourassa, M. Hoelsi, and V. S. Peng, “Do housing submarkets really matter?” *Journal of Housing Economics*, vol. 12, no. 1, pp. 12–28, 2003.

[34] C. Wu, X. Ye, F. Ren, and Q. Du, “Modified data-driven framework for housing market segmentation,” *Journal of Urban Planning and Development*, vol. 144, no. 4, pp. 1–15, 2018.

[35] M. Assuncao, M. C. Neves, G. Camara, and C. Da Costa Freitas, “Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees,” *International Journal of Geographical Information Science*, vol. 20, no. 7, pp. 797–811, 2006.

[36] C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, “Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China,” *Atmospheric Environment*, vol. 203, pp. 172–182, 2019.

[37] M. Weng, Q. Xiong, and M. Kang, “Salience indicators for landmark extraction at large spatial scales based on spatial analysis methods,” *ISPRS International Journal of Geo-Information*, vol. 6, no. 3, p. 72, 2017.

[38] H. Zhang, X. Zhou, G. Tang, G. Ji, X. Zhang, and L. Xiong, “Inference method for cultural diffusion patterns using a field model,” *Transactions in GIS*, vol. 24, no. 6, pp. 1578–1601, 2020.

[39] H. Zhang, X. Zhou, G. Tang, L. Xiong, and K. Dong, “Mining spatial patterns of food culture in China using restaurant POI data,” *Transactions in GIS*, vol. 25, no. 2, pp. 579–601, 2021.

[40] K. Cupido, A. S. Fotheringham, and P. Jevtic, “Local modelling of u.s. mortality rates: a multiscale geographically weighted regression approach,” *Population, Space and Place*, vol. 27, no. 1, pp. 1–11, 2020.

[41] D. K. Smith and D. Jungnickel, “Graphs, networks and algorithms,” *Journal of the Operational Research Society*, vol. 50, no. 12, p. 1286, 1999.

[42] T. M. Oshan, Z. Li, W. Kang, L. J. Wolf, and A. S. Fotheringham, “Mgwr: a python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale,” *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 269, 2019.

[43] H. Wen and Y. Tao, “Polycentric urban structure and housing price in the transitional China: evidence from Hangzhou,” *Habitat International*, vol. 46, pp. 138–146, 2015.

[44] Z. Zhang, I. Li, T. Fung et al., “Multiscale geographically and temporally weighted regression with a unilateral temporal weighting scheme and its application in the analysis of spatiotemporal characteristics of house prices in Beijing,” *International Journal of Geographical Information Science*, vol. 35, no. 11, pp. 2262–2286, 2021.

[45] L. Liu, H. C. Yu, J. Zhao, H. Wu, Z. H. Peng, and R. Wang, “Multiscale effects of multimodal public facilities accessibility on housing prices based on MGWR: a case study of wuhan, China,” *ISPRS International Journal of Geo-Information*, vol. 11, no. 1, p. 57, 2022.

[46] A. Avhand and B. Des, *House prices and relative location Cities*, vol. 95, Article ID 102373, 2019.

[47] H. Li, Y. D. Wei, Y. Wu, and G. Tian, “Analyzing housing prices under different market segmentation frameworks for housing market segmentation,” *ISPRS International Journal of Geo-Information*, vol. 12, no. 1, pp. 12–28, 2003.