Spatiotemporal Data Mining of Real Estate Registration

Yidong Hu[1], Fenglei Mei[2], Chenling Pan[3], Tao Nie[4*]

[1]huyidong@vip.sina.com,[2]librachn@foxmail.com, [3]39304911@qq.com
[4*] Corresponding author: tntshuang@aliyun.com

Wuhan Natural Resource and Planning Information Center, 13 Sanyang Road, Wuhan, China

Abstract:Real estate registration is an important part of the field of natural resource management. Mining the value of registration data can help to improve the government's real estate asset management level, optimize urban planning and land use strategies. The data of real estate registration includes multi-dimensional information such as the right holder, area of property right, registration time and space location. Using registration data mining real estate registration space-time value can reveal the economic law of real estate transfer. Taking the transfer registration of existing houses in Jiang'an District of Wuhan as an example, analyze the spatial distribution characteristics from two spatial dimensions: natural buildings and cadastral sub-areas.We use econometrics and GIS spatial analysis methods to explore the overall spatial distribution of real estate through global autocorrelation, and analyze regional distribution characteristics through local autocorrelation. And researched the spatial clustering of cadastral sub-areas to study the spatial differentiation. Research has shown that dual factor clustering is more accurate in characterizing spatial clustering patterns and features than single factor clustering.

Keywords: spatiotemporal data mining; real estate registration; clustering; spatial distribution

1 INTRODUCTION

Since 2015, China has established a real estate property registration system with Chinese characteristics, integrating the country, provinces and cities into one unified system. The real estate registration department has established the real estate registration information basic management platform by using the information technology, and realized the real estate registration in the whole process of land, housing and so on. The real estate registration information records the attribute information of the subject of the right holder and the object of the real estate in detail. Mining registration information can well reveal the spatial distribution pattern and value transfer law of real estate registration.

In the use of real estate registration information, the relevant research of domestic scholars mainly focused on real estate registration inventory data cleaning, integration, standardization and so on. ZHANG Tiehong analyzed some aspects from database structure optimization, integration method selection, comprehensive data collection, full analysis of data, land and real estate registration information collation and drawing, to optimize and integrate land registration and real estate registration data and ensure their accuracy [1]. LIU Wei researched on optimization of key technologies for real estate registration data integration, optimized the real

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341903

estate registration data integration process, and established the real estate registration data extraction and integration system [2]. Li Linhui introduced feature similarity technology to propose a multi-level fast fusion model to realize batch fusion of real estate multi-source heterogeneous data. similarity threshold parameters and limited range parameters are used to further improve the accuracy and efficiency of multi-source heterogeneous data fusion. A lot of research and practice are of great significance to improve the quality and efficiency of registration. In the era of digital economy, how to make full use of registration data elements to play the potential of data has great research value [3].

Most international scholars focus on price estimation, and prediction of real estate mainly. Efthymiou studied the impact of accessibility on real estate prices based on the location of transportation infrastructure, confirming the correlation between accessibility and price [4]. James examined the expected rates of return of the commercial real estate market in major US metropolitan areas to confirm that there is a positive correlation between the commercial real estate markets in major US cities [5]. Li studied the spatial-temporal patterns of housing prices in China's large urban agglomerations by using big data analysis, confirmed the spatial correlation between cities, and draws the conclusion that there is clustering and differentiation in housing value space. These studies mainly from the perspective of the real estate market, study the relevance of real estate space, but lack of support for government management decision-making [6].

However, it is difficult to find the inherent characteristics of a lack of real estate registration with out data mining. This article researches spatial features from a big data perspective which is based on the real estate registration of Wuha’s central urban area. The spatial clustering analysis is used to find the spatial clustering characteristic area of real estate registration and explore the structural characteristics of different types of registration data in urban space. The results of this article can help to analyze the fairness of public service facilities and the rationality of resource distribution based on real estate registration data.

2 RESEARCH METHODS

2.1 Global Moran’s I index

Firstly, we use exploratory spatial analysis to reveal spatial autocorrelation of real estate data for confirming spatial distribution patterns with a global view. According to registration data of cadastre subareas, Global Moran’s I index can evaluate the spatial distribution pattern and test significance. Global Molain I measures the spatial autocorrelation of real estate. Given market attributes such as price and transaction volume, or the number of new and second-hand homes registered within the cadastral subarea, it is possible to assess the characteristics of spatial distribution patterns, whether clustered, dispersed or random. We use the ArcGIS spatial statistical tool to calculate Moran's I index value, and we needed to use both Z value and p value to evaluate the significance of the index I index value is between [-1, 1]. When the I index is greater than 0, the registration data have a spatial positive correlation. When the I index is less than 0, the registration data are uncorrelated, then, through z score to finish the significance test in statistics.

2.2 High-Low Clustering

Secondly, High-Low Clustering (Getis-Ord General G) analysis is used to identify the hotspot, coldspots, and outliers in the local region. After we found the possibility of clustering property-related attributes, we also want to know what kind of data is clustering, which kind of data can determine which kind of value is clustering, this is called “High/low value clustering” analysis. The General G Index, like the Moran's I Index, is a statistical corollary, which is the next step after you get the data. The so-called inferential statistics, is the use of limited data to the overall situation of the characteristics of the process of estimation. Getis-Ord General G will calculate two indices, one called the observed value, which is the actual index, and one called the expected value, which is the hypothetical index. The observed values are compared with the expected values. If the observations are large, then the Z score is a positive number, indicating a high aggregation. If the expected value is large, then the Z score is negative, indicating low aggregation.

Global spatial autocorrelation can reveal the distribution pattern of a single attribute in global space, but difficult to find the distribution pattern and association pattern of a single feature in the local space. The advantage of local spatial autocorrelation analysis is that it can explore local spatial distribution patterns, detect the location of clusters and the degree of regional correlation. Some common methods can be used, like Moran scatterplot, LISA Anselin Local Moran’s I, etc. Based on the division of cadastral subareas, the weight relations among cadastral subareas are measured by spatial adjacency. This article takes Anselin Local Moran’s I method to determine cadastral subareas based on similar characteristics and uses the Local Moran I index to classify and identify the analysis results of cadastral subareas. The results were divided into three categories. In the first category, results with positive statistical significance are identified as high-value or low-value clustering. Second, the results with negative statistical significance are identified as outliers. The third category results are not statistically significant.

2.3 K-Means clustering

Finally, the K-Means clustering method is used to find the spatial structure features. K-means clustering is one of the unsupervised learning algorithms, and it is a kind of iterative clustering method. In K-means clustering, the number of categories to be divided (K) is determined first, then each cluster center (means) is determined, then the distance from each sample to the cluster center is calculated, and finally the classification is made according to the distance.

The basic steps for K-means clustering are as follows: 1. Determine the number of categories to be divided into K. In practical application, we need to try again and again according to the actual problem, get different classification and compare, finally get the optimal number of categories. 2. Determine the initial cluster center for K categories. After determining the number of categories K, we need to select K samples as the initial clustering center of each category, and according to practical problems and experience to consider. 3. The Euclidean distances of each sample to the K cluster centers were calculated sequentially according to the identified K initial cluster centers, and all samples were divided into the predefined k categories according to the principle of nearest distance. The objective function is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{1}$$

where $\left\| x_i^{(j)} - c_j \right\|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center, c_j is an indicator of the distance of the n data points from their respective cluster centers.

The K-Means clustering method is adopted to cluster the cadastral subareas of the central urban area of Wuhan according to the K value to form the K class. According to the number of real estate registration business types in each cadastral subarea, the study area was divided into K areas with similar attributes.

3 SPATIAL ANALYSIS

3.1 Test Data

The research data is sourced from data produced by the Wuhan Real Estate Registration System. Wuhan City includes 7 central urban areas, 5 remote urban areas, and 2 functional zones. According to the statistical analysis of registration data, from 2016 to 2022, the transaction volume of second-hand houses in Jiang'an District, Wuchang District, and Jianghan District was greater than that of new houses, belonging to a typical stock housing market. This article takes Jiang'an District as the research area, and the registration business is the transfer registration of existing houses. The spatial scope covers 120 cadastral sub areas and 6433 natural buildings. The residential space is characterized by natural buildings, which contain the total number of units per building.

3.2 Global Spatial Autocorrelation Analysis

The spatial autocorrelation analysis results of residential units are shown in Figure 1, and the spatial autocorrelation analysis results of second-hand housing transaction volume are shown in Figure 2.

The Global Moran index value of total residential units is 0.2125 which is greater than 0, and the z score is 65.3925 which is greater than 2.58 for inspection, showing that only 1% of the real estate registration data distribution may be random distribution. The results have statistical significance, showing positively related to the spatial pattern. So, we can further explore the hot spot of real estate registration data in space distribution characteristics. This indicates that the global autocorrelation of real estate registration data in the spatial distribution pattern conforms to the clustering pattern characteristics of statistical significance, and the clustering characteristics of real estate registration data in the spatial clustering pattern can be further explored.The spatial autocorrelation analysis of second-hand housing transaction volume also shows that the space has global clustering characteristics.

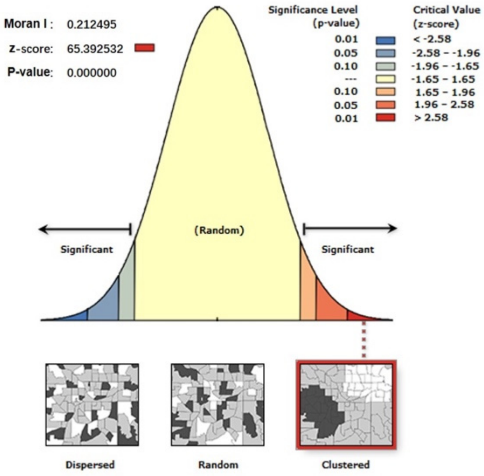

Figure 1 Spatial autocorrelation analysis of residential units

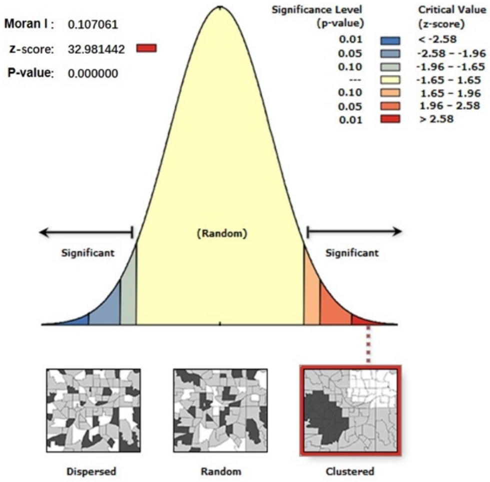

Figure 2 Spatial autocorrelation analysis of second-hand housing transaction

3.3 Local Spatial Autocorrelation Analysis

By using local spatial autocorrelation analysis, it is found that the research area has significant characteristics of high-value clustering or low-value clustering areas, the distribution pattern of outliers, and other areas. Figure 3 shows the results of local autocorrelation clustering. It can be found that the red region is a region with significant high-value clustering, presenting a significant spatial clustering pattern, which is consistent with the overall performance of spatial autocorrelation results. It can be seen from the figure that high-value areas concentrate between the Second Ring Road and the Third Ring Road. These houses were built after 2000, and new residential buildings built after 2012 generally have higher floors.The residential agglomeration characteristics in local areas within the Second Ring Road are mainly reflected in the urban renewal areas.

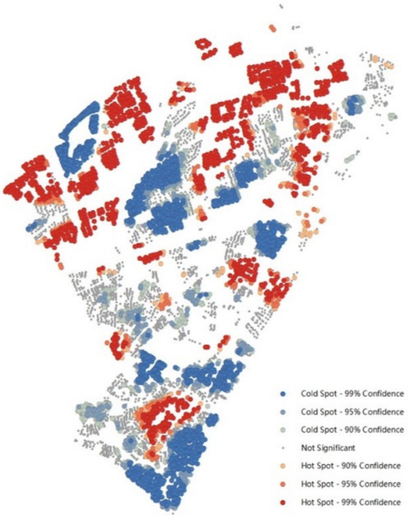
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 3 Hot spots of residential units

The hotspots of second-hand housing transfer are shown in Figure 4, with local distribution characteristics similar to the hotspots of total housing. The Second Ring Road to Third Ring Road are the main trading areas in Jiang'an District, accounting for 63.8%, mainly concentrated in Houhu and Tazihu. The Second Ring Road is mainly concentrated in five areas, including Erqi Road, North of Yangtze River Second Bridge, Shenyang Road, Huangxiaohe Road, and Sanyang Road.

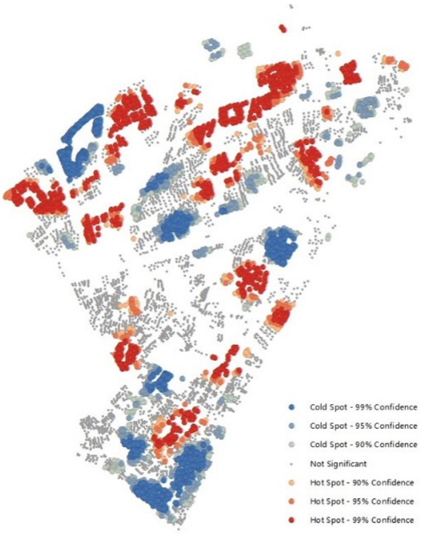
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 4 Hot spot of second-hand housing transaction

The high and low clustering results are shown in Figure 5, which can more clearly display the local clustering characteristics and can be divided into four categories: high high clustering, high low clustering, low high clustering, low low clustering, and obvious spatial differentiation. The main reason is the division of school districts, where small areas of old school district housing and improved school district housing have good market liquidity. Due to the complete living facilities, newer housing, and affordable prices in the Houhu and Tazihu areas, they are the main purchasing areas for rigid demand.

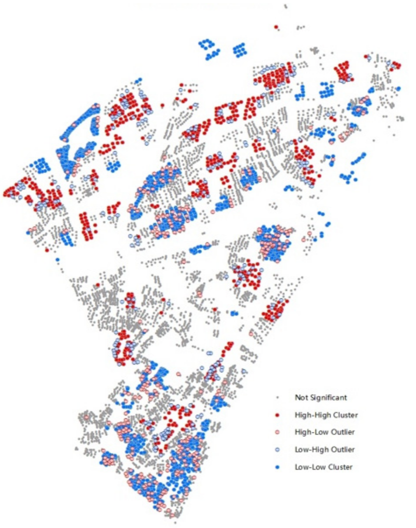

Figure 5 High and low clustering of second-hand housing transaction

3.4 K-Means clustering analysis

- Single factor clustering of transfer quantity

Due to the spatial discontinuity of natural buildings, which are not connected in space, and considering the management requirements of real estate registration, we conducted spatial clustering analysis of elements from the perspective of cadastral sub areas to facilitate comprehensive evaluation of each area, which is consistent with the dimensions of urban planning and land market management.

As shown in Figure 6, the transaction volume of second-hand housing in the cadastral sub district is spatially clustered. The transaction volume is ranked from small to large as Class 1, Class 2, Class 3, Class 4, and Class 5, respectively, including the number of cadastral sub districts 53,34,21,9,3. The annual average transaction volume in each sub district is 21,48,69,91,157, respectively. The clustering calculation results of k=5 are shown in Table 1, between-cluster S.S./total S.S. is 0.951634, indicating good clustering performance.

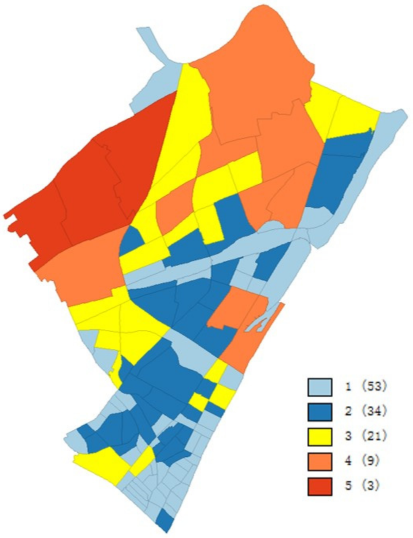

Figure 6 Clustering of second-hand housing transaction quantity

Table 1 Cluster value of second-hand housing transaction

CL	Cluster Center	Within cluster S.S.
C1	-0.727708	1.32491
C2	-0.103806	1.29862
C3	0.676207	1.23809
C4	1.78854	0.85526
C5	3.93358	1.0387
between-cluster S.S./total S.S.=0.951634		

- Single factor clustering of price

The price clustering with k=5 is shown in Figure 7, which presents significantly different characteristics from transaction volume, mainly manifested as large-scale homogenization and local regional polarization. It should be noted that class 4 does not have transaction records and prices. The average price per square meter for Class 1, Class 2, Class 3, and Class 5 is 16324 yuan, 19123 yuan, 36782 yuan, and 51782 yuan. The price gradient is steep and the distribution is extremely uneven. In Table 2, between-cluster S.S./total S.S. is 0.962343, indicating good clustering performance.

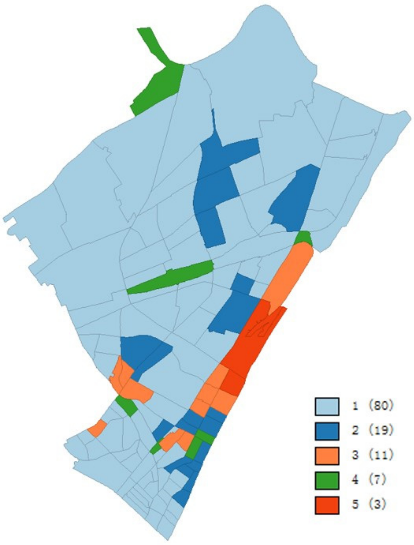

Figure 7 Clustering of price

Table 2 Cluster value of price

CL	Cluster Center	Within cluster S.S.
C1	-0.293372	1.51211
C2	0.323826	0.747795
C3	1.64151	1.98675
C4	-1.87807	1.38051e-030
C5	4.13565	0.234583
between-cluster S.S./total S.S.= 0.962343		

- Volume-Price double factor clustering

Due to the difficulty of accurately characterizing spatial differentiation characteristics with a single factor, the article conducts spatial clustering with a quantity price double factor, and sets the number of clusters to k=5, k=6, k=7 for clustering effect comparison. As shown in Figures 8, 9, and 10, the clustering calculation results are shown in Table 3, Table 4, and Table 5. By comparison, when k=7, it has a good clustering effect. The three clustering results show that the clustering at k=6 divides class 3 (11 cadastral sub-areas) at k=5 into class 6 (3 cadastral sub-areas), which subdivides the improvement area into general improvement areas and top-level improvement areas, which is consistent with the actual situation. Compared to k=6, grouping class 1 and class 2 again with k=7 results in more accurate characterization of the required features.

In Figures 10, Class 1 has the lowest price and transaction volume, indicating that the residential value in this area is not recognized by the market and lacks liquidity, making it a key planning area for urban renewal. The price and transaction volume of Class 2 take second place, mainly due to the early development of commercial housing after 1999. Due to the age of the housing being around 20 years, the living quality is relatively low, and it can be mainly renovated to improve the living quality and meet the housing needs of low-income groups. Class 3 is mainly in the demand area, while Class 6 shows a high cost-effectiveness and is widely accepted by the market.Class 4 belongs to the improvement area, with more active transactions and higher prices, as well as good public implementation support, especially in the school district. Class 7 can be defined as a top-level area, which is Hankou Binjiang Business Distric. It also boasts high-quality, excellent school districts, and river views, possessing the top resources of the city.

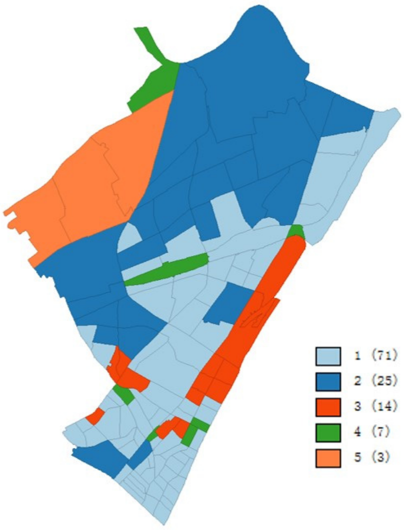

Figure 8 Double factor clustering with k=5

Table 3 Cluster value with k=5

CL	Cluster Center		Within cluster S.S.
	Price	Volume	
C1	-0.15705	-0.45573	13.5305
C2	-0.239736	1.07884	9.80291
C3	2.17597	0.0205741	22.7591
C4	-1.87807	-0.957552	1.38051e-030
C5	-0.057699	3.93358	1.04504
between-cluster S.S./total S.S.= 0.861943			

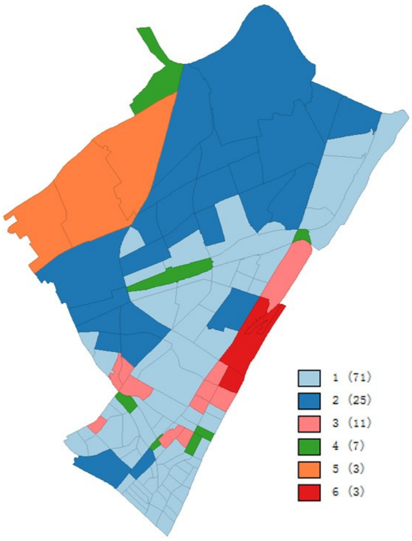

Figure 9 Double factor clustering with k=6

Table 4 Cluster value with k=6

CL	Cluster Center		Within cluster S.S.
	Price	Volume	
C1	-0.15705	-0.45573	13.5305
C2	-0.239736	1.07884	9.80291
C3	1.64151	0.014351	4.59762
C4	-1.87807	-0.957552	1.38051e-030
C5	-0.057699	3.93358	1.04504
C6	4.13565	0.043392	3.49628
between-cluster S.S./total S.S.= 0.901943			

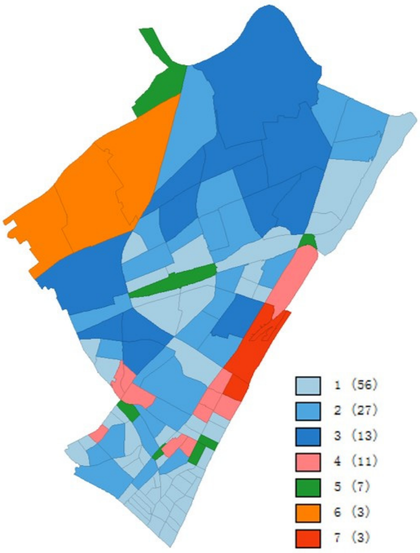

Figure 10 Double factor clustering with k=7

Table 5 Cluster value with k=7

CL	Cluster Center		Within cluster S.S.
	Price	Volume	
C1	-0.145351	-0.588283	7.92378
C2	-0.238291	0.291853	4.20974
C3	-0.197728	1.51368	3.54046
C4	1.64151	0.014351	4.59762
C5	-1.87807	-0.957552	1.38051e-030
C6	-0.057699	3.93358	1.04504
C7	4.13565	0.043392	3.49628
between-cluster S.S./total S.S.= 0.957424			

4 CONCLUSION

By mining real estate registration data, the value of the data can be fully utilized and the potential of the data can be stimulated. The research results indicate that urban real estate exhibits agglomeration in space, and corresponding second-hand housing transactions also exhibit agglomeration effects. Further exploration of local clustering features reveals spatial differentiation between hot spots and clustering patterns. The housing structure exhibits both homogenization and polarization, with steep price gradients. Urban residential planning needs to balance low-income groups, those in urgent need, and those who need improvement. Taking Jiang'an District as an example, the value of the ring road and location is vague, and the traditional core area of the inner ring road has outdated housing, making it difficult for the living quality to meet people's needs for a high-quality life. The overall supply of high-quality housing is insufficient, and sporadic urban renewal is difficult to drive the development of the entire area.

From the analysis results, it can be seen that the development of Hankou Binjiang Business District has a significant pilot effect, which can provide reference for subsequent urban renewal.

REFERENCES

[1] ZHANG Tiehong, YIN Junbo, et al. Discussion on the Optimization Path of Real Estate Registration Data Integration Technology[J]. GEOSPATIAL INFORMATION, 2022, 20(9):28-31,71.

[2] LIU Wei1, LIU Songxian. Research on Optimization of Key Technologies for Real Estate Registration Data Integration[J]. GEOMATICS & SPATIAL INFORMATION TECHNOLOGY, 2022, 45(9):205-207.

[3] Li Linhui, Wang Yu, Liu Yueyan, et al. A Fast Fusion Model for Multi-Source Heterogeneous Data Of Real Estate Based on Feature Similarity [J]. Geomatics and Information Science of Wuhan University, 2023, DOI: 10.13203/j. whugis20220742.

[4] Efthymiou, D. Antoniou,C. Investigating the impact of recession on transportation cost capitalization: a spatial analysis[J]. Journal of Transport Geography, 2015, 42:1-9.

[5] James D. Shilling, C.F. Sirmans, Barrett A. Slade. Spatial Correlation in Expected Returns in Commercial Real Estate Markets and the Role of Core Markets[J]. The Journal of Real Estate Finance and Economics, 2017, 54(3):297-337.

[6] Shengwen Li, Xinyue Ye, Jay Lee, et al. Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective[J]. Applied Spatial Analysis and Policy, 2017, 10(3):421-433.