Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7245407, 13 pages
https://doi.org/10.1155/2022/7245407

Research Article

Urban Functional Area Recognition Based on Unbalanced Clustering

Junjie Wu,[1] Jian Zhang,[1] and Huixia Zhang[2]

[1] *Computer Department, Taiyuan Normal University, Yuci, Jinzhong 030619, China*
[2] *Institute of Geographical Science, Taiyuan Normal University, Yuci, Jinzhong 030619, China*

Correspondence should be addressed to Huixia Zhang; zhanghx@tynu.edu.cn

Received 31 July 2022; Accepted 26 August 2022; Published 22 September 2022

Academic Editor: Zaoli Yang

Urban functional area recognition refers to refining the main functions of building coverage areas. At present, multisource data analysis is prone to data imbalance, and types with large data volume are more likely to affect data analysis results. Therefore, this study took the main urban area of Taiyuan as the research object and used the Synthetic Minority Oversampling Technology (SMOTE) method to reduce the impact of data imbalance. In this study, the SOMTE method was used to incrementally process the microblog check-in data in the main urban area of Taiyuan, which reduced the phenomenon of data imbalance and further improved the recognition accuracy. The Point of Interest (POI) data were clustered through K-nearest neighbor, and microblog check-in data were semantically analyzed by Linear Discriminant Analysis (LDA). Then, the eigenvalues of the two kinds of data results were obtained by frequency density analysis. Finally, feature fusion was carried out by means of weighted average. The fused data were divided into single and mixed functional areas according to the difference of frequency density, which was rendered and displayed on the ArcGIS platform, so as to realize the visual identification and division of urban functional areas, and the results were compared with Gaode Map. The experimental results showed that this method can effectively identify urban functional areas with a recognition accuracy of 85%, which provided reference value for the planning and research of urban functional areas in the future.

1. Introduction

In the early research of urban structure, there is less research on urban functional areas, while there is more research on land use [1, 2]. With the gradual improvement of the current urbanization level, urban functional areas had also been refined accordingly. For example, the areas covered by buildings can be divided into industrial areas, residential areas, commercial areas, and other functional areas. The classification of urban functional areas had become a new research hotspot beyond land use and surface coverage. The identification of urban functional areas is to find the main functions of each region in the city [3]. At present, most of the research was based on POI single source data, but the methods were different. Researchers analyze the data of POI through frequency density analysis [4–7] and calculate the feature vector to identify and divide the functional area. In addition, the text prediction analysis of POI data through semantic analysis can get the urban functional area. On the basis of word2vec model, the potential semantic information of POI is deeply mined and text prediction is carried out to realize the recognition of urban functional areas [8–11]. In addition, spatial co-location patterns are used to mine the potential context of POI, extract the spatial distribution information of data, and extract feature vectors for clustering, so as to realize the recognition of urban functional areas [12, 13]. In view of the fact that this study used microblog check-in data for text recognition, it was decided to use LDA topic model to extract research topics and identify hot topics [14].

However, the above research and analysis were based on single source data, and a single data source cannot verify the authenticity and reliability of the data. Therefore, social media data [15, 16], GNSS trajectory data [17–20], mobile phone data [21, 22], and remote sensing images [23] are proposed for multi-source data analysis. Multi-source data improved the authenticity and reliability of the data, but the data were distributed in different fields, and the amount of data collected will vary, which was a phenomenon of data imbalance, resulting in data research results being biased towards the category of more data. Therefore, the purpose of this study was not only to identify urban functional areas but also to reduce data imbalance.

1.1. Data Sources. The study area was the main urban area of Taiyuan City, including Jiancaoping District, Xinghualing District, Wanbailin District, Yingze District, Xiaodian District, and Jinyuan District. The four main roads of North Central Street, South Central Street, West Central Road, and East Central Road formed a closed study area (see Figure 1).

Traditionally, urban functional area maps are mainly made by field surveys, which are time-consuming and difficult to update in time [24]. Free accessibility makes the OSM a promising segmentation data source for urban functional areas and has been successfully used for the mapping of urban functional areas [25]. Land-use patterns are closely related to human social and economic activities [26]. Open social data can yield a better understanding of urban functioning with more human maps [27]. The emergence of open social data (such as OSM (open street map), social media data, mobile phone data, POI, and GNSS trajectory) enables the social attributes of urban areas to be used to identify urban functional area [28–30]. The emergence of open social data enabled the social attributes of urban areas to identify urban functions. The data of this study included the OSM road network data of the main urban area of Taiyuan exported from the official website of OpenStreetMap, 33056 POI data, and 5627 microblog check-in data from November to December 2021. The POI data were obtained through the free open interface programming provided by Gaode Map, which mainly included four aspects of information: name, category, coordinates, and classification. Comprehensive POI information was the necessary information for navigation maps, and it also provided an important source of data for functional area identification. Microblog check-in data were obtained through the crawler microblog. It was a special kind of microblog data. Users used the intelligent terminal with GPS function to record the location at a certain time and write down their feelings at that time, so as to generate data with space-time information and text content. Microblog check-in data recorded users' interests and hobbies, reflected people' life trajectory, and had high research value. In recent years, it has also become an important source of urban functional area recognition.

1.2. Data Preprocessing. According to the parameters required in this study, firstly, the OSM road network map was topologically processed, and the unit grid was constructed according to the road network to obtain the unit map of the study area (see Figure 2). After deduplication and coordinate correction of POI data, 33056 pieces of data were finally obtained. The fields of the POI dataset include attribute information such as POI name, longitude, latitude, type, and region. POI types were divided into 6 categories and 12 medium categories. According to the type of statistics, POI data were divided into public management and public service land (medical clinics, sports, science and education culture, and government agencies), commercial service facility land (shopping related, catering, finance and insurance, hotels, and hospitality), square green space (tourist attractions), industrial land (companies and enterprises), residential land (commercial residence), and transportation facility land (business housing and transportation facilities). According to the API interface provided by Gaode Map, the longitude and latitude of the crawled microblog check-in data were converted. The microblog check-in dataset fields included user name, check-in content, check-in longitude, check-in latitude, and check-in place. By referring to public awareness [31], according to the actual floor area of buildings and public awareness. This study gave weights fordifferent functional area types: commercial service facility land (20), residential land (35), industrial land (40), public management and public service land (50), square green space (80), and transportation facility land (100). This method was to increase the number of different data types through weight. For example, there was one scenic spot in a region, and the number of scenic spots in the region was 80 by weight.

2. Research Methods

The identification model of urban functional area based on unbalanced clustering aims to reduce the imbalance of multi-source data and improve the proportion of the impact of minority data characteristics on the analysis results. This study adopted two data sources: POI and microblog check-in. The POI data included a feature point set containing spatial attribute information such as name, address, type, latitude, and longitude, which contained rich human and economic characteristics and natural characteristics and can reveal the use function of urban land [32]. The microblog check-in data refer to the content released by users during microblog check-in, with strong subjective color. These data were considered the key research source of hotspot function area identification. This study combined the two source data, extracted the geographic information and semantic information of the data, respectively, for feature fusion, and finally divided the urban functional area through the difference of frequency density. The specific steps were as follows. (1) The POI and microblog check-in data were non-equilibrium clustered, and the microblog check-in data were incrementally processed by the SMOTE method. (2) The K-nearest neighbor algorithm was used to count the POI in each research area, calculated the frequency density cf_i of the each POI type in research area, and constructed the POI feature vector through cfi. (3) The semantic analysis of microblog check-in data was carried out by using LDA topic model, the frequency density analysis of the analysis results was carried out, and the feature vector of microblog check-in data was constructed. (4) The multisource feature vectors were fused by weighted average. Finally, the single and mixed functional areas were divided by frequency density difference and rendered and displayed on ArcGIS platform, so as to realize the visual identification and division of urban functional areas.

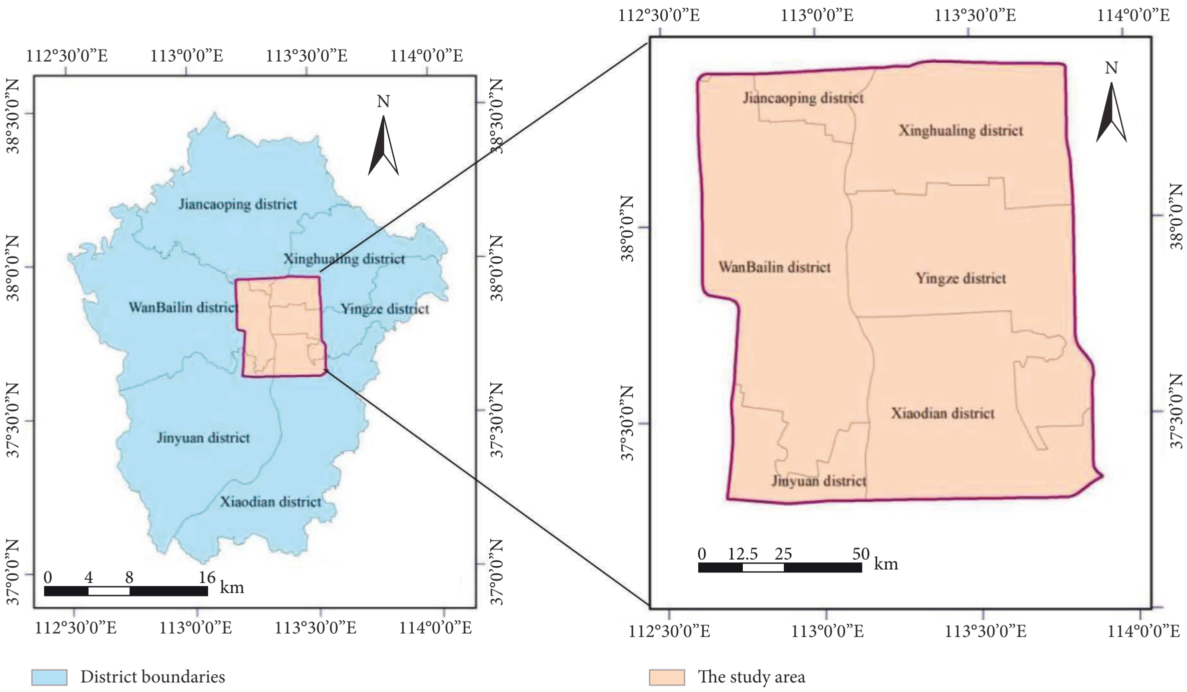


Figure 1: Geographical location of the study area. The study area was the main urban area of Taiyuan.

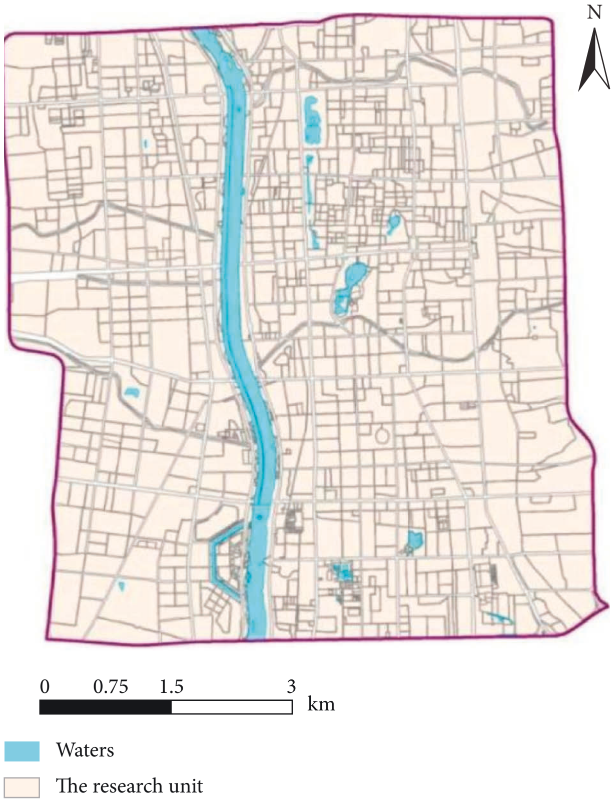


Figure 2: Unit diagram of study area. OSM road network map of main urban area of Taiyuan.

2.1. SMOTE Incremental Processing. Orriols-Puig et al. [33] studied the effects of imbalance on different LCS (Michigan Style Learning Classifier System) components, and then the learning classification system was used to solve the problem of data imbalance classification; Krawczyk and McInnes [34] proposed a scheme of local ensemble learning to deal with high-dimensional and multi-class unbalanced data. Zhu et al. [35] reviewed the main research progress of unbalanced data mining at home and abroad in recent years and made an in-depth analysis and comprehensive comparison of various existing technologies and methods to deal with unbalanced data mining from the data level and algorithm level. Cai et al. [15] used unbalanced clustering to fuse GPS and microblog check-in data to mine hotspots. In this study, the minority data (the microblog check-in data) in each research area were incrementally processed through the data layer. SMOTE in unbalanced clustering is a classic minority oversampling technology [36], and it is an improved scheme based on random oversampling algorithm. The basic flow of smote algorithm is as follows:

(1) For each sample in a minority class x_i, the distance from the sample to other samples in the minority sample set is calculated by Euclidean distance, and the K-nearest neighbor of k units is obtained.

(2) Set the sampling ratio N according to the unbalanced sample proportion calculation, take each minority sample x_i as the sample point, and randomly select several minority samples x_n through K-nearest neighbor.

$$N = \text{round}\left(\frac{\sum_{r=1}^{m} VC_r}{\sum_{r=1}^{m} VQ_r}\right), \tag{1}$$

where r is each regional unit; m is the total number of areas; VC_r represents the data amount of POI dataset

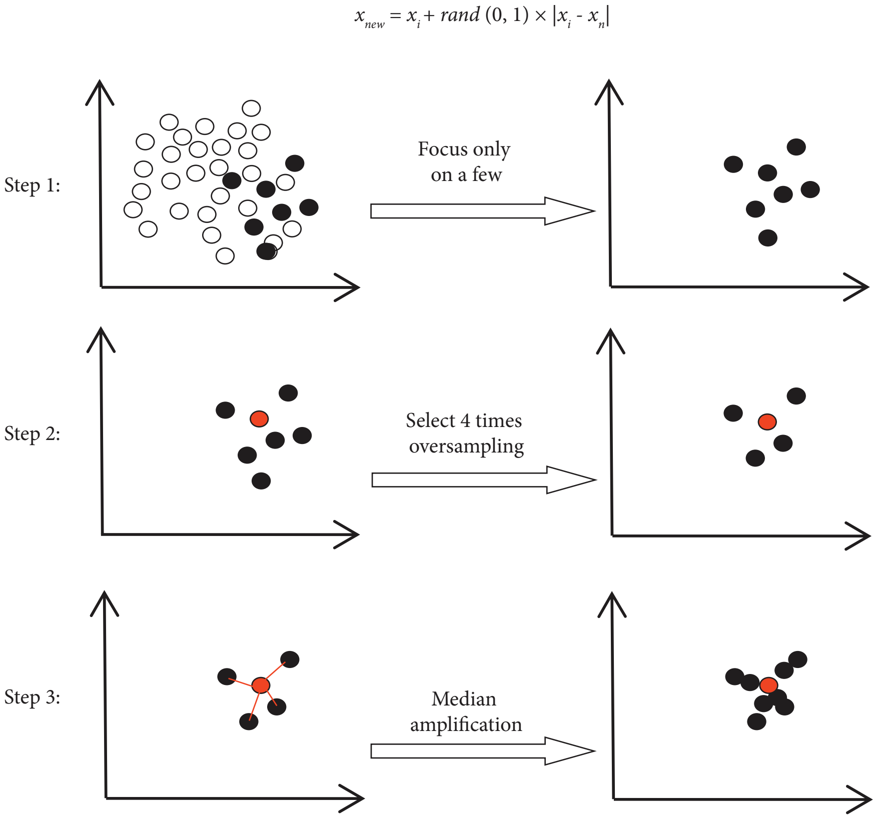

$x_{new} = x_i + rand\,(0, 1) \times |x_i - x_n|$
Step 1:
Focus only on a few
Step 2:
Select 4 times oversampling
Step 3:
Median amplification

Figure 3: Oversampling process. Incremental processing of data through 3 steps.

C in the r-th area; and VQ_r represents the amount of data of check-in data Q in the r-th area.

(3) x_i is connected with each randomly selected nearest neighbor x_n into a line, in which a point is randomly selected as a new sample point x_{new}:

$$x_{\text{new}} = x_i + r\,\text{and}\,(0, 1) \times \left|x_i - x_n\right|. \quad (2)$$

The basic idea of the SMOTE algorithm was to analyze the minority samples and synthesize new samples according to the minority samples and add them to the dataset (see Figure 3). Through this algorithm, virtual sample points are obtained without any attributes. In the process of using SMOTE incremental processing, the attribute of the real microblog check-in data in this area was added to the incremental data, and the amplified data were used in subsequent analysis and processing.

2.2. Constructing POI Dataset with K-Nearest Neighbor. K-nearest neighbor algorithm, given a training dataset for a new input instance, found the k instances closest to the instance in the training dataset. Most of the k instances belonged to a class, and then classify the input instance into this class. In this study, the POI type was selected according to each research area to construct the dataset, and each kind of data adjacent to the POI was found through the K-nearest neighbor algorithm. The obtained data types form a training team. The Euclidean distance calculation formula is as follows:

$$L\left(x_i, x_j\right) = \left(\sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{1/2}, \quad (3)$$

where x_i,$x_j \in x$, x refers to all POI data in one area, x_i is the cluster of sample points, x_j is the data cluster around the sample point, $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}), x_j = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(n)})$, and n is the number of nearest neighbors selected.

We analyzed the frequency density of each type of POI in each functional area in the obtained dataset to obtain the feature vector about the type of POI ($cf = [cf_1, cf_2, \ldots, cf_n]$) to reflect the real attributes of each functional area (see Figure 4), and the dominant functions of different functional areas were obtained. The frequency density formula [4] is as follows:

$$cf_i = \frac{n_i/s_i}{\sum_{i=1}^{6} n_i/s_i}, \quad (4)$$

where i indicates the type of functional area; n_i represents the number of type i in the study area; s_i represents the number of type i in all study areas; and cf_i represents the proportion of frequency density of type i in all types in the study area.

Figure 4 shows the feature vector established for a research area through frequency density, that is, the frequency density proportion of various functional area types in the area. Through this figure, we can more intuitively observe

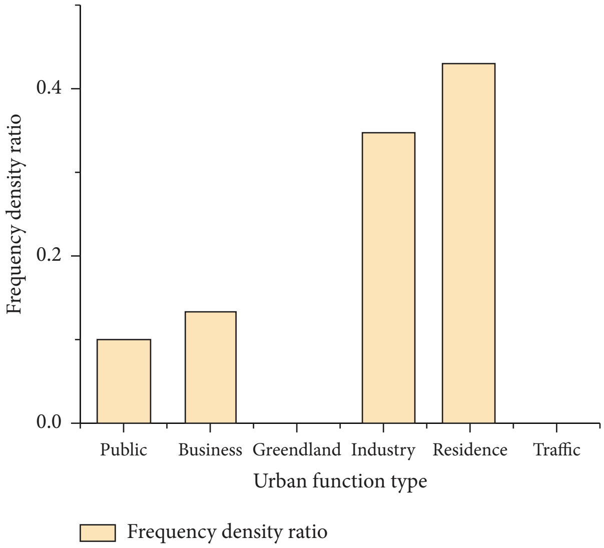


FIGURE 4: Proportion of frequency density in a certain area. Obtained the feature vector of the functional area type of the region through the density proportion.

what functional area the area belonged to and also provide data support for subsequent multi-source data feature fusion.

2.3. Construction of LDA Topic Model. LDA [37] was a popular method to divide urban functional areas based on topic modeling. This study judged the main functions of each region by extracting and analyzing the topic model of microblog check-in data. At present, there were two mature criteria to judge whether an LDA model was reasonable. One was consistency, and the other was perplexity. Perplexity was used to measure the quality of a probability distribution or probability model to predict the sample. As the number of topics increases, the perplexity of the model will decrease, so the prediction effect will be better. However, when there are too many topics, the trained model will be overfitted. We found that when the number of topics was 10 (see Figure 5), the perplexity of the model was the lowest. Therefore, this study used 10 topics for analysis and carried out LDA visual display of microblog check-in data (see Figure 6).

The function area was composed of 10 topics, each of which belonged to one of 12 categories, and was determined by the frequency of a large number of words. Topic 1 and Topic 3 occupy the majority of campuses and college, they belong to the type of science and education culture and should be divided into public management and public service land. In Topic 2, there are many railway stations, bus stations, and subway stations, which should be divided into land for transportation facilities. In Topic 4, there are medical land such as Shanxi University Hospital and Shanxi Provincial People's Hospital, which should be divided into land for public management and public service. Topic 5 includes Yingze Park, Longtan Park, and other tourist attractions, which should be divided into square green space; the names of innovation bases, companies, groups, and other enterprises in Topic 6 account for a large proportion and should be classified as industrial land. Topic 7: there are many shopping malls, such as Xintiandi, Taiyuan Wanxiang, and Wanda mall, which should be divided into land for commercial service facilities; Residential areas account for a large proportion in Topic 8 and should be divided into residential land. Topic 9 contains Taiyuan municipal government, management center, and other government departments, which should be divided into public management and public service land. There are catering related terms such as barbecue, pizza, and gourmet food in Topic 10, which should be divided into commercial service land.

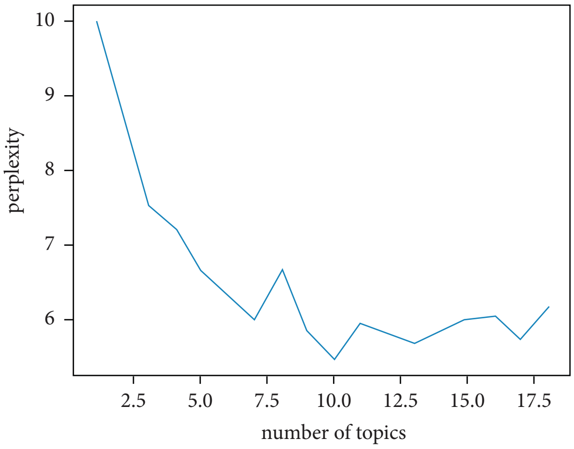


FIGURE 5: Line graph of perplexity relative to the number of topics. Selected the lowest perplexity value through 1 to 18 topics.

The accuracy of the results of LDA recognition of microblog check-in data can be divided into two types: the first type has a recognition accuracy of 88.3% for hot spots and the second type has a recognition accuracy of 66.7% for non-hot spots. The experimental research proves that the microblog check-in data are suitable for studying urban hot spots. The experiment showed that most of the microblog check-in data content was related to science and education, shopping, housing, medical treatment, catering, and green space, which proved that the microblog check-in data are suitable for urban hot spots with close human behavior activities and had a high research value for urban hot spots in the future.

LDA topic model is a unsupervised learning topic probability generation model, which can learn hidden layer topics from unstructured text. The data analyzed by LDA model also uses formula (4) for frequency density analysis to construct the feature vector of microblog check-in data ($W = [W_1, W_2, \ldots, W_n]$). Next, the feature vectors of POI and Weibo check-in data will be fused in this study.

2.4. Data Feature Fusion and Analysis. Feature fusion was a hot research direction of multi-source data analysis. It integrated data from different sources and contained the same features, extracted sample features for data fusion, and finally identified and divided the fused data through classification algorithm, which can better reflect the characteristics of data features. This study takes the research

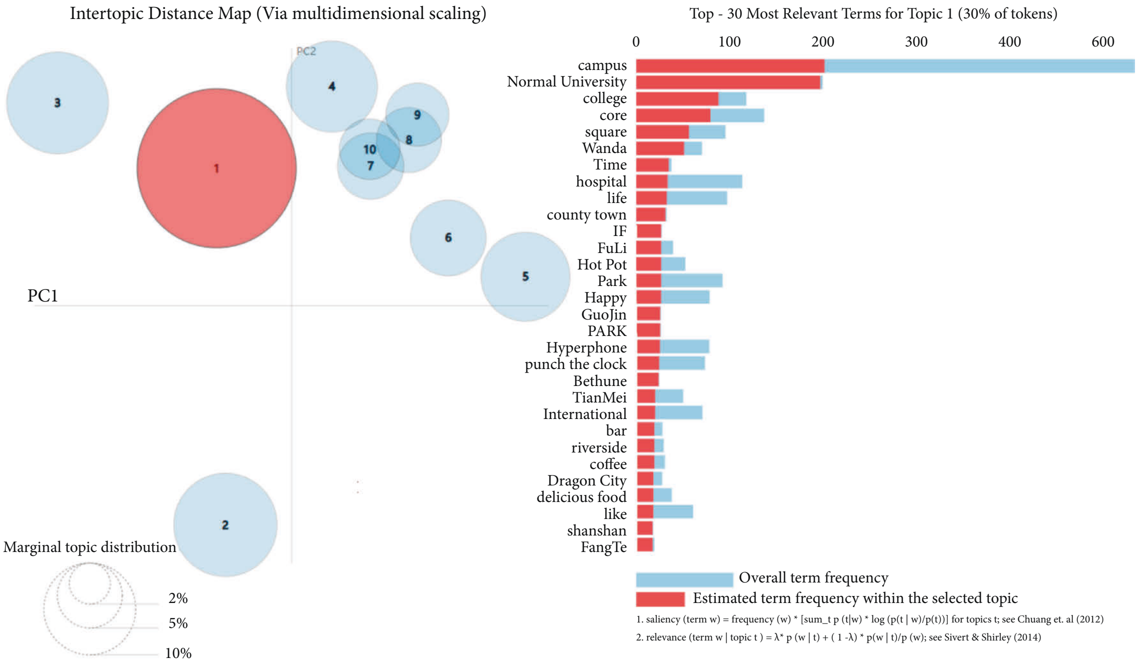


FIGURE 6: LDA visualization. Through LDA semantic analysis, 10 topic distribution probabilities were predicted.

area as the data support and fuses the features by means of weighted average. The formula is as follows:

$$R_i = \frac{cf_i + w_i}{n}, \tag{5}$$

where cf_i, W_i represent the eigenvector value of class i of POI and microblog check-in; n represents the number of eigenvector values of each type, i.e., 2; and R_i represents the class i eigenvector value of a certain region, that is, the frequency density value after weighted average.

According to the calculated frequency density of each area, the existing density analysis mostly takes 50% as the threshold. If the maximum value was greater than 50%, the area will be divided into single functional area; otherwise, it will be divided into mixed functional area. This judgment was not completely reasonable. For example (see Figures 7 and 8), the maximum frequency density value was 50%, while the second frequency density value was 46%, and the last was 4%. If 50% is taken as the threshold, the area will be divided into public management and public service land. According to the actual judgment, it was more reasonable to divide the area into public-residence (public management and public services-residential land).

Therefore, in this study, the frequency density difference will be used as the basis for the division of functional areas [5]. By sorting the numerical results of each research area, obtain the first two relatively large density values (the maximum value type is a, followed by b). If the difference between a and b is more than 20%,, the functional area will be identified as a single functional area (a); otherwise, the two types will be mixed as a mixed functional area (a-b).

2.5. Accuracy Evaluation Standard. Aiming at the imbalance of multi-source data fusion, this paper compared three experimental schemes from the data level and clustering algorithm level, namely, single source data recognition functional area, multi-source data with unbalanced clustering, and multi-source data without unbalanced clustering (see Table 1).

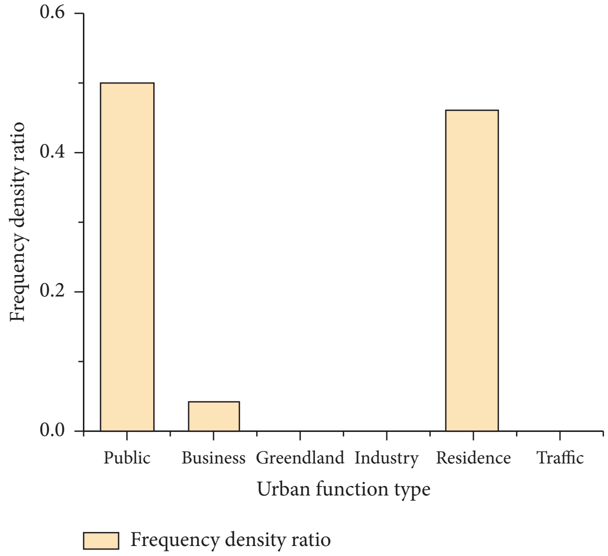
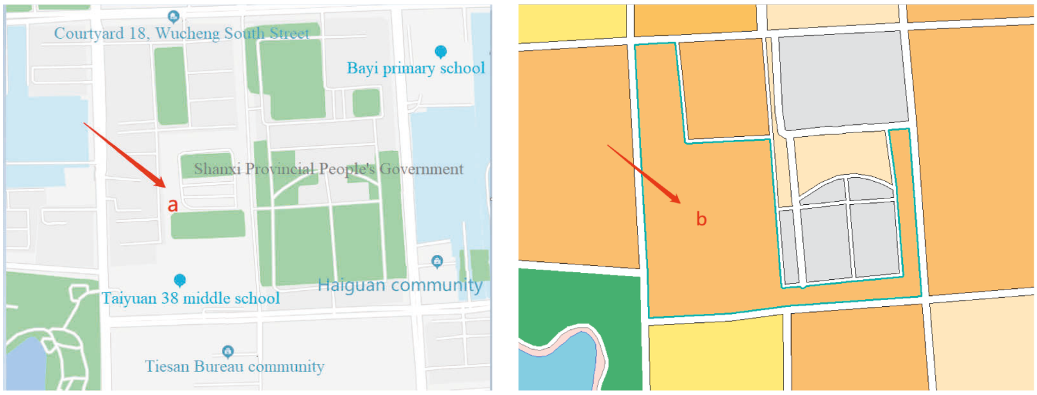


FIGURE 7: Frequency density ratio after fusion in a certain region. The feature vector of the functional type of the region was obtained by weighted average.

In order to verify the accuracy of the identification results of this study, 20 blocks were randomly selected from 1650 study areas and compared with the land use of real blocks. The compliance evaluation was divided into four levels, with a full score of 3, that is, it was completely consistent, 2 was relatively consistent, 1 was relatively inconsistent, and 0 was completely inconsistent. The



(a)

(b)

Figure 8: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

Table 1: Overview of experimental methods.

Programme	Characteristic	Method description
1	Single source data identification function area [4]	Use POI to identify urban functional areas through frequency density analysis.
2	Multi-source data without unbalanced clustering	POI and microblog check-in data are used for machine learning and natural language processing, respectively. Finally, data fusion is used to identify urban functional areas.
3	Multi-source data with unbalanced clustering	Firstly, unbalanced clustering is used for data multiplication, machine learning, and natural language processing, respectively, and finally data fusion is used to identify urban functional areas.

formula for calculating the overall recognition accuracy [38] is

$$a = \frac{\sum_{i=1}^{n} x_i}{kn} \times 100\%, \tag{6}$$

where k is the full fraction, i.e., $k = 3$; x_i is the actual score of the compliance of block i; and n is the number of sample study areas.

3. Results

According to programme 3 of this study, the distribution map of functional areas in the main urban area of Taiyuan (see Figure 9) was obtained. The distribution map included 1650 research areas. A total of 18 functional areas were displayed by using ArcGIS platform, including 6 single functional areas and 12 mixed functional areas. Recognition accuracy reaches 85%.

3.1. Comparison with Gaode Map

3.1. Comparison with Gaode Map. In order to test the accuracy of the experimental results, this study selected several hot areas for comparative analysis with Gaode Map. The area in Figure 10 is located in Yingze District, Taiyuan City, and included Wenying Park, the monument to righteous deeds, Chunyang Palace, Shanxi Provincial Art Museum, and Qifeng Street Primary School. It was an area related to public management and public service land-square green space, which was in line with the identification results. The area in Figure 11 is located in Xiaodian District, Taiyuan, below Shanxi National Fitness Center. The area was located in a prosperous business district, included phase 2 Maoye Tiandi shopping mall, with high population density. It was an area related to land for commercial service facilities, which was in line with the identification results. The area in Figure 12 is located near Longtan Park, Xinghualing District, Taiyuan. This area included land for public management and public services such as Taiyuan Water Supply and Water Saving Management Center, Taiyuan Grain and Material Reserve Bureau, Taiyuan Forestry Bureau, and Taiyuan Planning Commission, as well as land for commercial service facilities such as China Merchants Bank, People's Insurance Group of China, and Agricultural Bank of China. Therefore, this area was an area related to public management and land for public services-land for commercial service facilities, which was in line with the identification results. The area in Figure 13 is located in Jiancaoping District of Taiyuan City and included Taiyuan City College, Shengliqiao Campus of Shanxi Water Conservancy Vocational and Technical College, and Shanxi Academy of Environmental Sciences. The school belonged to science and education culture, and the research institute belonged to government organs. These two middle classes belonged to public management and public service land. Therefore, this area was related to public management and public service land, which was in line with the identification results. The area in Figure 14 is located in Wanbailin District, Taiyuan City, and included two residential areas: Zijing Holiday and Xiangyue Lanxi. Therefore, this area was related to residential land, which was in line with the identification results. The area in Figure 15 is

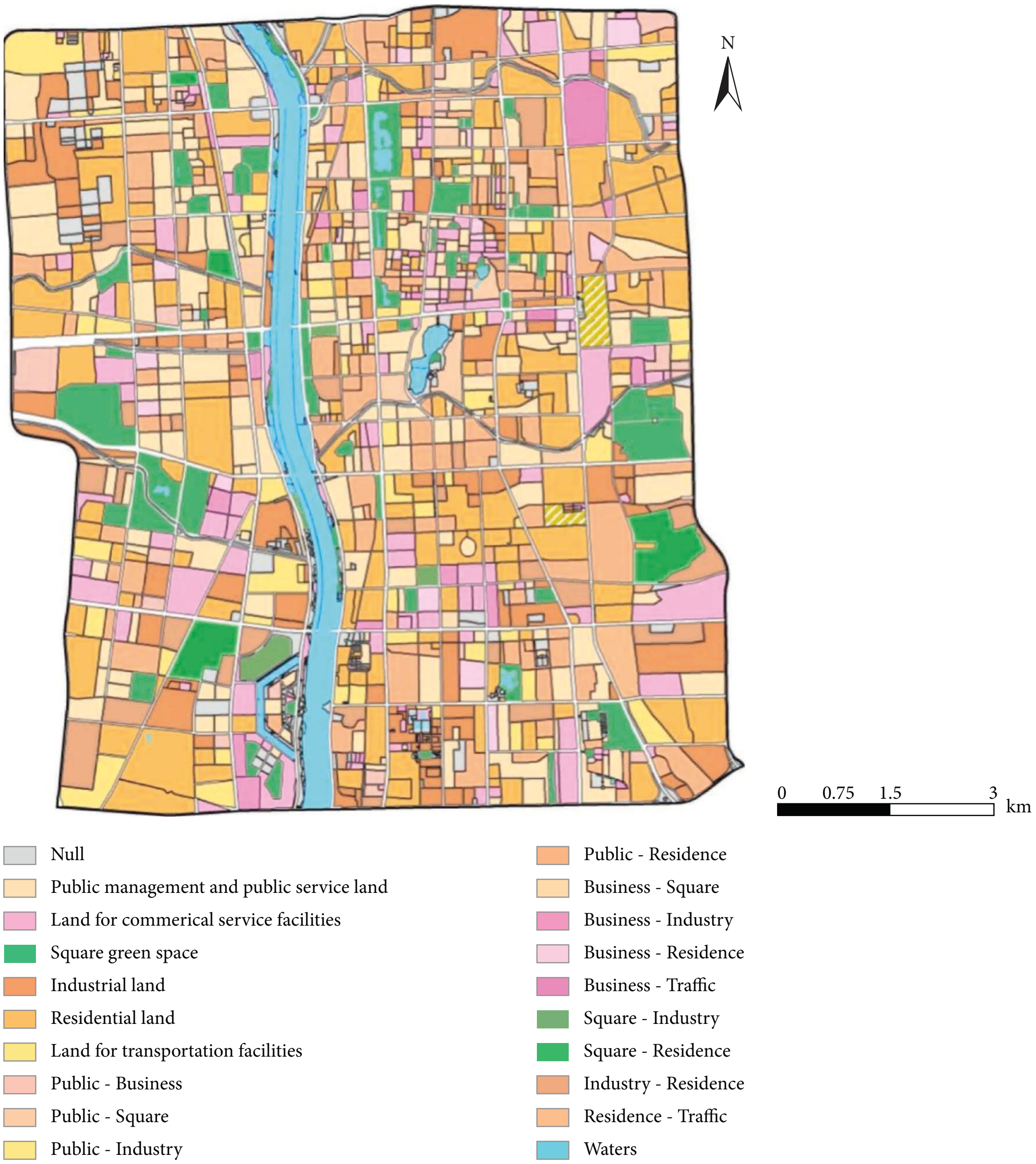
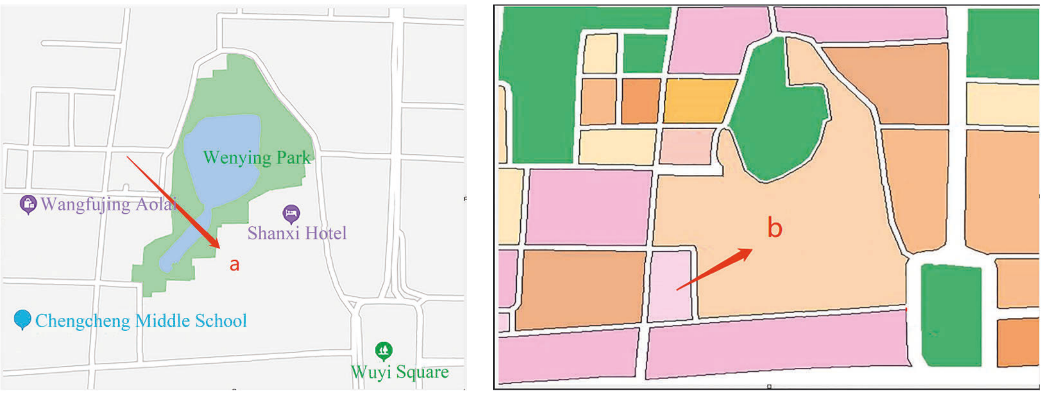


Figure 9: Distribution of functional areas in the main urban area of Taiyuan. Urban functional areas included 6 single functional areas and 12 mixed functional areas.

located in Jinyuan District, Taiyuan City, and included the two residential areas of Houwanshan and Qianwanshan, and Wanbailin Experimental Primary School belonged to public management and public service land. Therefore, this area was public management and public service land-residential land, which was in line with the identification results. This study also selected two suburbs for comparative analysis. The area in Figure 16 is located in Wanbailin District, which included Taiyuan Jinxi Electromechanical Equipment Manufacturing Plant and covers the whole area. Therefore, this area was related to industrial land, which was in line with the identification results. The area in Figure 17 is located in Yingze District, which included Jiaxin Rolling Gate Factory, Shanxi Huajian Electric Power Construction Co., Ltd., and other companies. In addition, there was a Dongcheng 100 community. Therefore, this area was industrial land-residential land, which was in line with the identification results.

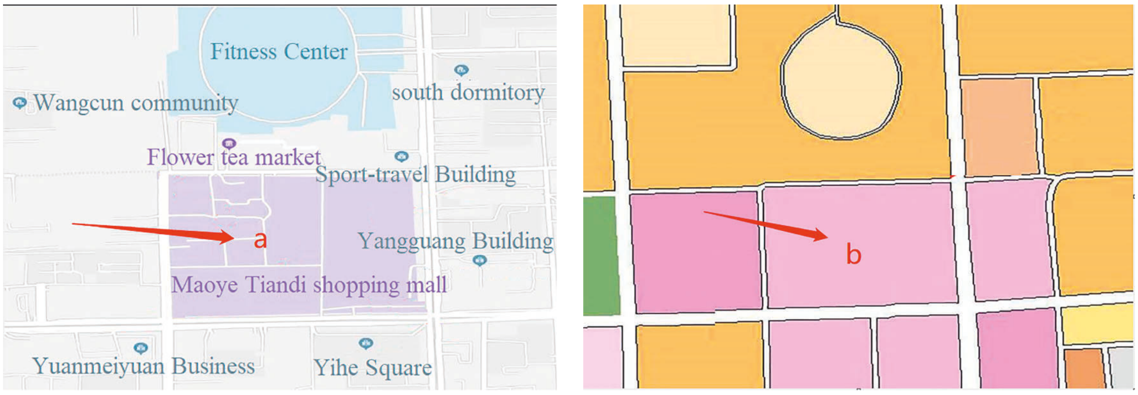
4. Discussion

4.1. Comparison with Accuracy Evaluation. The identification accuracy was used to verify the effectiveness of the identification scheme in this study by comparing the distribution results of functional areas identified by the three



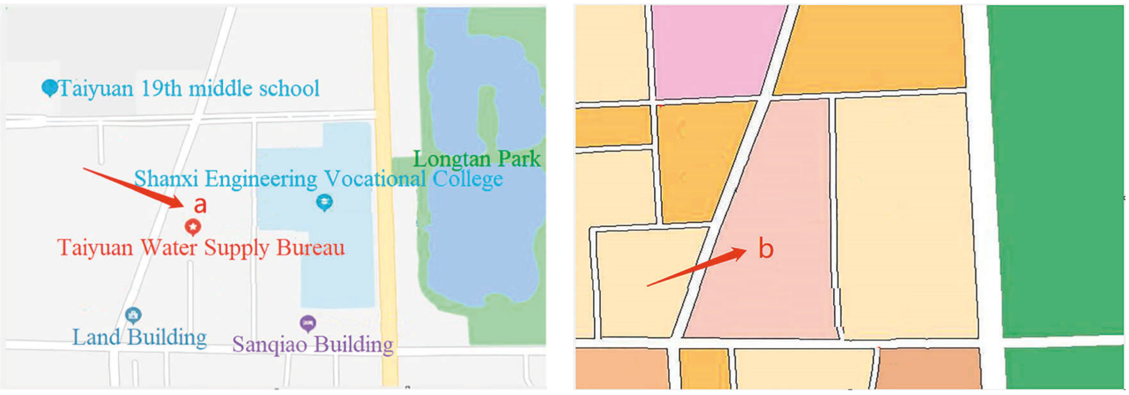
(a) (b)

Figure 10: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.



(a) (b)

Figure 11: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

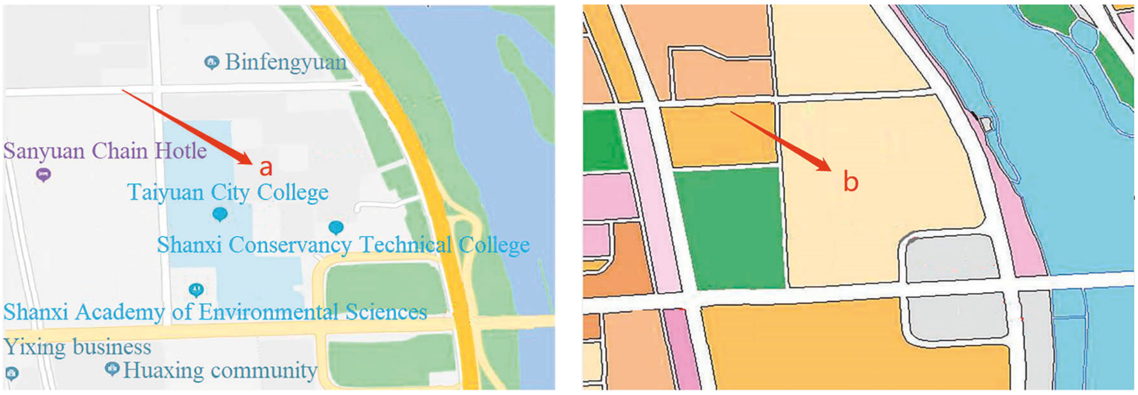

Taiyuan 19th middle school
Longtan Park
Shanxi Engineering Vocational College
a
Taiyuan Water Supply Bureau
Land Building
Sanqiao Building
b

(a) (b)

Figure 12: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

schemes (see Table 2). The identification accuracy of the three schemes is shown in Table 2.

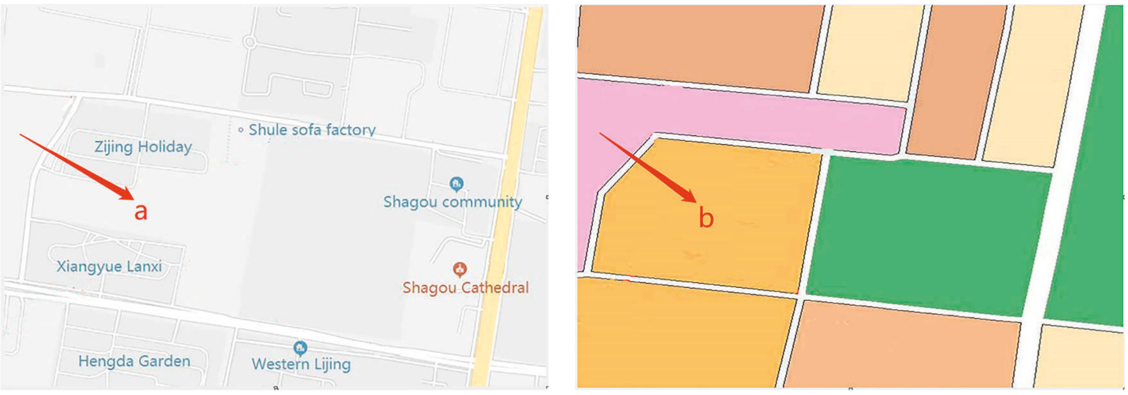
On the whole, the accuracy of single source recognition functional area was the lowest, 73.33%, which proved that multi-source data can improve the recognition accuracy of urban functional area, reaching 81.67%, while most multi-source data had the problem of data imbalance. This study multiplied a small amount of data through unbalanced



(a)

(b)

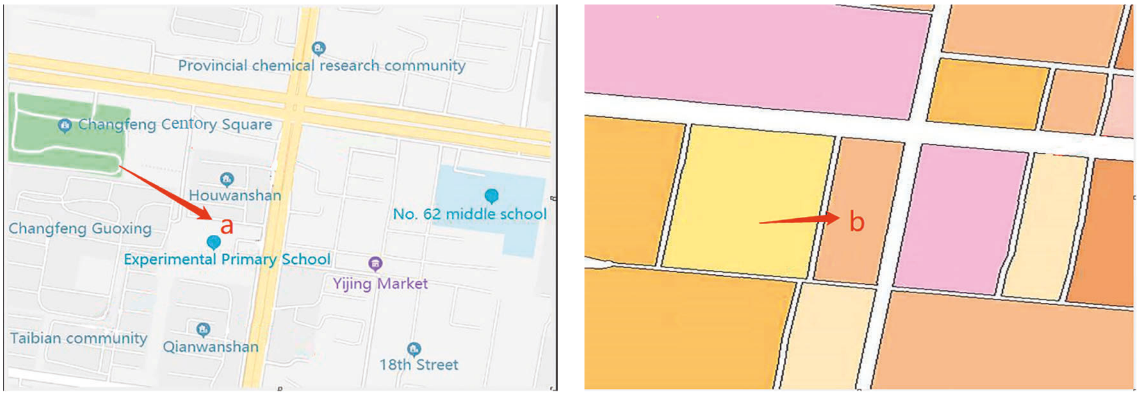
Figure 13: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.



(a)

(b)

Figure 14: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.



(a)

(b)

Figure 15: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

clustering method, so as to reduce the phenomenon of data imbalance and improve the recognition accuracy of urban functional area, which reached 85%.

4.2. Comparison with Existing Studies. At present, the earliest research on urban functional area recognition was based on a single data source (see scheme 1 in Tables 1 and 2);

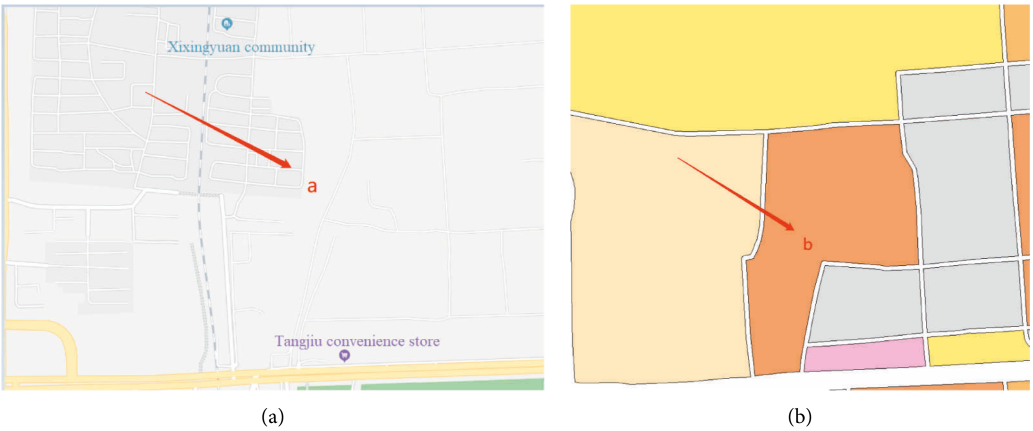


Figure 16: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

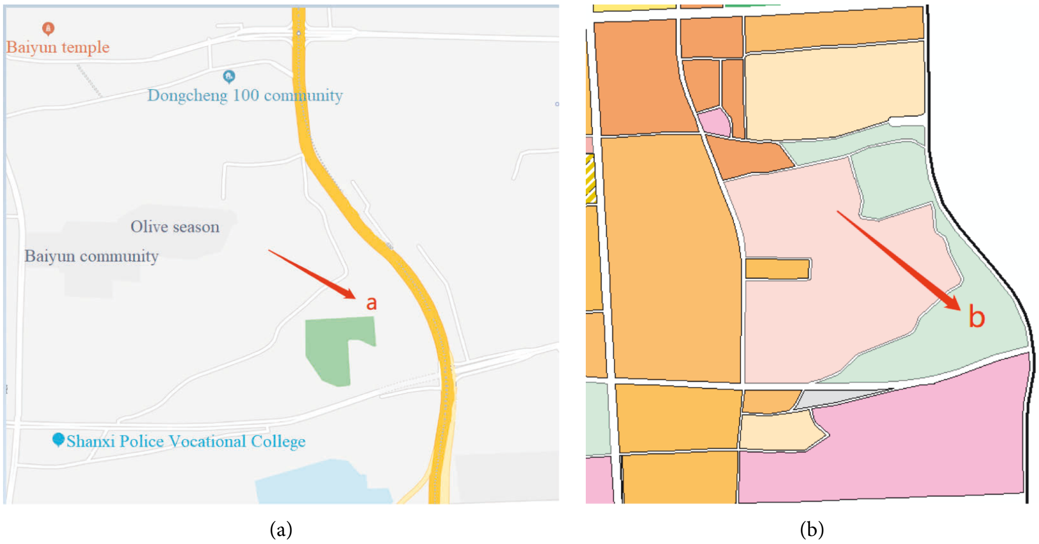


Figure 17: Comparison of identification results between Gaode Map and functional area. (a) Gaode Map. (b) Functional area recognition results.

Table 2: Functional area recognition accuracy.

Programme	Characteristic	Accuracy (%)
1	Single source data identification function area [4]	73.33
2	Multi-source data without unbalanced clustering	81.67
3	Multi-source data with unbalanced clustering	85

based on the data of interest points, Chi et al. [4] divided it into single function area and mixed function area through frequency density analysis. Guo et al. [5] assigned weight to POI and improved the recognition accuracy of functional areas through frequency density difference. The method of urban functional area identification and analysis based on frequency density and POI function type ratio was proposed [6]. The area identified by single source data was not accurate and authoritative, which resulted in the proposal for multi-source data. For example, Li et al. [18] had preliminarily realized more accurate quantitative identification and evaluation of the mixing degree of urban functions on a fine scale based on POI and GPS data, which can provide support and basis for the improvement of urban comprehensive functions. Yang et al. [22] integrated the mobile signaling data and POI data, used the day and night difference of intensity and the degree of internal function mixing, and completed the judgment of regional dominant function type and the evaluation of function mixing degree. Chen et al. [23] proposed an urban functional area recognition framework integrating multi-source geographic data. The overall accuracy ranged from 73.3% to 84.8%.

However, the above research did not take into account the problem of data imbalance. The output categories of many models were based on the threshold. For example, in logistic regression, those less than 0.5 were negative examples, and those greater than 0.5 were positive examples. The data features of a few classes were easy to be covered by the data features of most classes, and the former often contained valuable feature information. This data imbalance will lead to the identification of functional areas biased towards the data features of most classes. In this study, unbalanced clustering was used to reduce the phenomenon of data imbalance, and this method improved the eigenvalue of a small number of data and further improved the recognition accuracy. In summary, the proposed method outperforms existing methods.

5. Conclusions

The research direction of this study was to reduce the imbalance of multi-source data fusion by using unbalanced clustering. This study identifies the functional areas of the main urban areas of Taiyuan. Through the feature fusion of POI data and microblog check-in data, 6 single functional areas and 12 mixed functional areas are obtained. Through comparative analysis with Gaode Map, the identification results of this method were in line with the actual situation of public cognition; at the same time, different schemes were compared and analyzed through the recognition accuracy of functional areas. The recognition research of urban functional areas based on unbalanced clustering has high accuracy, and the accuracy was 85%, which verified the feasibility of this scheme.

The results of this study showed that POI data and microblog check-in data had obvious research and application value in functional area recognition and were of great significance for low-cost and efficient identification of functional areas. At the same time, the scheme used in this study was applicable to other location data for analysis, such as public comments, WeChat circle of friends, bus card data, and so on. Although this study had achieved some research results, there were still deficiencies. The microblog check-in data cannot be divided according to the release time period, which resulted in the problem of poor timeliness of data results. This will be the author's future research direction..

Data Availability

The data supporting the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Natural Science Foundation of Shanxi Province (201901D111291).

References

[1] E. Banzhaf and M. Netzband, “Monitoring urban land use changes with remote sensing techniques,” *Applied Urban Ecology*, vol. 15, pp. 18–32, 2011.

[2] F. Rodrigues, F. C. Pereira, A. Alves, S. Jiang, and J. Ferreira, “Automatic classification of points-of-interest for land-use analysis,” in *Proceedings of the fourth international conference on advanced geographic information systems, applications, and services (GEOProcessing)*, pp. 41–49, Valencia, Spain, 2012.

[3] Z. W. Gao, W. W. Sun, and P. G. Cheng, “Identify urban functional zones using multi feature latent semantic fused information of high-spatial resolution remote sensing image and POI data,” *Remote Sensing Technology and Application*, vol. 36, no. 3, pp. 618–626, 2021.

[4] J. Chi, L. M. Jiao, and T. Dong, “Quantitative identification and visualization of urban functional area based on POI data,” *Journal of Geomatics*, vol. 41, no. 2, pp. 68–73, 2016.

[5] Y. F. Guo, G. W. Lan, and D. L. Fan, “Identifying functional urban regions with POI data,” *Journal of Guilin University of Technology*, vol. 41, pp. 1–9, 2021.

[6] Y. Hu and Y. Han, “Identification of urban functional areas based on POI data: a case study of the Guangzhou economic and technological development zone,” *Sustainability*, vol. 11, no. 5, p. 1385, 2019.

[7] Z. Wang, D. Ma, D. Sun, and J. Zhang, “Identification and analysis of urban functional area in Hangzhou based on OSM and POI data,” *PLoS One*, vol. 16, no. 5, Article ID e0251988, 2021.

[8] B. Yan, K. Janowicz, and G. Mai, “From ITDL to Place2Vec - reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts,” in *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, vol. 35, pp. 1–10, CA, USA, November 2017.

[9] Y. Yao, X. Li, X. P. Liu et al., “Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model,” *International Journal of Geographical Information Science*, vol. 31, no. 4, pp. 825–848, 2017.

[10] W. Zhai, X. Y. Bai, Y. Shi, Y. Han, Z. R. Peng, and C. Gu, “Beyond Word2vec: an approach for urban functional region extraction and identification by combining Place2vec and POIs,” *Computers, Environment and Urban Systems*, vol. 74, pp. 1–12, 2019.

[11] Z. J. Zheng, R. B. Zheng, and J. Y. Xu, “Identification of urban functional regions based on poi data and place2vec model,” *Geography and Geo-Information Science*, vol. 36, no. 04, pp. 48–56, 2020.

[12] Z. L. Chen, L. L. Zhou, and W. H. Yu, “Identification of the urban functional regions considering the potential context of interest points,” *Acta Geodaetica et Cartographica Sinica*, vol. 49, no. 07, pp. 907–920, 2020.

[13] W. Huang, L. Cui, M. Chen, D. Zhang, and Y. Yao, “Estimating urban functional distributions with semantics preserved POI embedding,” *International Journal of Geographical Information Science*, vol. 36, pp. 1–26, 2022.

[14] J. P. Qiu and C. Shen, “Analysis of hot topics in domestic big data research based on LDA model,” *Journal of Modern Information*, vol. 41, no. 09, pp. 22–31, 2021.

[15] L. Cai, Y. Z. Li, and J. Fang, “Study on clustering mining of imbalanced data fusion towards urban hotspots,” *Computer Science*, vol. 46, no. 08, pp. 16–22, 2019.

[16] R. Miao, Y. Wang, and S. Li, "Analyzing urban spatial patterns and functional zones using sina Weibo POI data: a case study of Beijing," *Sustainability*, vol. 13, no. 2, p. 647, 2021.

[17] Y. Y. Gu, L. M. Jiao, and T. Dong, "Spatial distribution and interaction analysis of urban functional areas based on multi-source data," *Geomatics and Information Science of Wuhan University*, vol. 43, no. 7, pp. 1113–1121, 2018.

[18] M. Y. Li, Y. Ma, and X. M. Sun, "Application of spatial and temporal entropy based on multi-source data for measuring the mix degree of urban functions," *City Planning Review*, vol. 42, no. 02, pp. 97–103, 2018.

[19] B. Yu, Z. Wang, H. Mu, L. Sun, and F. Hu, "Identification of urban functional regions based on floating car track data and POI data," *Sustainability*, vol. 11, no. 23, p. 6541, 2019.

[20] j. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194, Beijing China, August 2012.

[21] H. Y. Han, X. Yu, and Y. Long, "Identifying urban functional zones using bus smart card data and points of interest in beijing," *City Planning Review*, vol. 40, no. 06, pp. 52–60, 2016.

[22] Z. S. Yang, J. H. Su, and H. Yang, "Exploring urban functional areas based on multi-source data: a case study of Beijing," *Geographical Research*, vol. 40, no. 02, pp. 477–494, 2021.

[23] S. Chen, H. Zhang, and H. Yang, "Urban functional zone recognition integrating multisource geographic data," *Remote Sensing*, vol. 13, no. 23, p. 4732, 2021.

[24] M. Herold, X. Liu, and K. C. Clarke, "Spatial metrics and image texture for mapping urban land use," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 991–1001, 2003.

[25] X. Liu, J. He, Y. Yao et al., "Classifying urban land use by integrating remote sensing and social media data," *International Journal of Geographical Information Science*, vol. 31, no. 8, pp. 1675–1696, 2017.

[26] A. Soliman, K. Soltani, J. Yin, A. Padmanabhan, and S. Wang, "Social sensing of urban land use based on analysis of twitter users' mobility patterns," *PLoS One*, vol. 12, no. 7, Article ID e0181657, 2017.

[27] S. Du, S. Du, B. Liu, X. Zhang, and Z. Zheng, "Large-scale urban functional zone mapping by integrating remote sensing images and open social data," *GIScience and Remote Sensing*, vol. 57, no. 3, pp. 411–430, 2020.

[28] A. Crooks, D. Pfoser, A. Jenkins et al., "Crowdsourcing urban form and function," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 720–741, 2015.

[29] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1988–2007, 2014.

[30] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, 2015.

[31] W. F. Zhao, Q. Q. Li, and B. J. Li, "National remote sensing bulletin," *National Remote Sensing Bulletin*, vol. 15, no. 5, pp. 973–988, 2011.

[32] S. Jiang, A. Alves, F. Rodrigues, J. Ferreira, and F. C. Pereira, "Mining point-of-interest data from social networks for urban land use classification and disaggregation," *Computers, Environment and Urban Systems*, vol. 53, pp. 36–46, 2015.

[33] A. Orriols-Puig, E. Bernado-Mansilla, D. E. Goldberg, K. Sastry, and P. Lanzi, "Facetwise analysis of XCS for problems with class imbalances," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 1093–1119, 2009.

[34] B. Krawczyk and B. T. McInnes, "Local ensemble learning from imbalanced and noisy data for word sense disambiguation," *Pattern Recognition*, vol. 78, pp. 103–119, 2018.

[35] Y. J. Zhu, Z. Wang, H. Y. Zha, and D Gao, "Boundary-eliminated pseudoinverse linear discriminant for imbalanced problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2581–2594, 2018.

[36] K. W. Li, W. R. Zhang, Q. H. Lu, and X Fang, "An improved SMOTE imbalanced data classification method based on support degree," *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, in *Proceedings of the 2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 34–38, Beijing China, October 2014.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[38] Y. H. Kang, Y. Y. Wang, and Z. J. Xia, "Identification and classification of wuhan urban districts based on POI," *Journal of Geomatics*, vol. 43, no. 1, pp. 81–85, 2018.