

Article

Using Multiple Sources of Data and “Voting Mechanisms” for Urban Land-Use Mapping

Kang Zheng ^{1,2,†}, Huiyi Zhang ^{1,2,†}, Haiying Wang ^{1,2,3,4,*} , Fen Qin ^{1,2,3,4}, Zhe Wang ⁵  and Jinyi Zhao ^{1,2}

¹ The College of Geography and Environment Science, Henan University, Kaifeng 475004, China

² Key Laboratory of Geospatial Technology for Middle and Lower Yellow River Regions, Ministry of Education, Kaifeng 475004, China

³ Henan Industrial Technology Academy of Spatio-Temporal Big Data, Henan University, Kaifeng 475004, China

⁴ Henan Technology Innovation Center of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China

⁵ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China

* Correspondence: whyhdgis@henu.edu.cn

† These authors contributed equally to this work.

Abstract: High-quality urban land-use maps are essential for grasping the dynamics and scale of urban land use, predicting future environmental trends and changes, and allocating national land resources. This paper proposes a multisample “voting mechanism” based on multisource data and random forests to achieve fine mapping of urban land use. First, Zhengzhou City was selected as the study area. Based on full integration of multisource features, random forests were used to perform the preliminary classification of multiple samples. Finally, the preliminary classification results were filtered according to the “voting mechanism” to achieve high-precision urban land-use classification mapping. The results showed that the overall classification accuracy of Level I features increased by 5.66% and 14.32% and that the overall classification accuracy of Level II features increased by 9.02% and 12.46%, respectively, compared with the classification results of other strategies. Therefore, this method can significantly reduce the influence of mixed distribution of land types and improve the accuracy of urban land-use classification at a fine scale.

Keywords: multisource data; random forests; urban land use; voting mechanisms



Citation: Zheng, K.; Zhang, H.; Wang, H.; Qin, F.; Wang, Z.; Zhao, J. Using Multiple Sources of Data and “Voting Mechanisms” for Urban Land-Use Mapping. *Land* **2022**, *11*, 2209. <https://doi.org/10.3390/land11122209>

Academic Editors: Linda See, Myroslava Lesiv and Dmitry Schepaschenko

Received: 15 November 2022

Accepted: 1 December 2022

Published: 5 December 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Urban dwellers account for about 55% of the global population, and this proportion is predicted to reach nearly 70% in 2059, with urban economic productivity accounting for more than 80% of the global GDP [1]. Land use has been an important research component for global ecological changes and sustainable social development [2–5]. As a major carrier of basic human life and daily activities, land merits accurate and timely understanding of urban land-use information, which is extremely important for tapping the inherent development potential of cities, improving urban spatial governance, and ensuring high-quality urban development [6–9].

Early urban land-use classification mapping started with field surveys and visual interpretation of remote-sensing images, which had higher classification accuracy [10]. However, this approach required interpreters to have profound knowledge and was characterized by poor timeliness and high cost [11]. Along with the development of remote sensing and computer technology, image-element-based and object-oriented methods have been gradually proposed, which have improved the ability to classify land-use features in remote-sensing images to a certain extent [12,13]. However, the heterogeneity of land-use features in remote-sensing images and the confusion of image elements usually lead to a decrease in urban land-use classification accuracy. For example, the spectral characteristics of commercial and residential areas in remote-sensing images are very similar, and it is

difficult to further improve the accuracy of land-use classification while relying only on remote-sensing images [14,15].

In recent years, due to the gradual availability of open-source data and the diversification of data sources, many other data sources in addition to remote-sensing images have become important sources of data for urban land-use classification, including mobile phone data, Sina Weibo social media data, Twitter check-in record data, OpenStreetMap (OSM) road network data, point-of-interest (POI) data, etc. Mobile phone data can be used to indicate the social function of land use [16]. Sina Weibo social media data can be used to estimate temporal trend patterns in land-use types [17]. The spatio-temporal characteristics of geo-tagged tweets in Twitter can reflect dynamic population distribution characteristics and further facilitate land-use classification [18]. OSM can be used to create a subdivision of urban land-use parcel units, and such parcel unit data can represent socio-economic functions [19]. However, when subdividing feature units, roads of different widths and administrative levels cannot be divided directly using road centerlines, resulting in a lack of parcel unit segmentation and further reducing the characterization of land-use features within the unit [20]. POI data can capture the spatial and temporal characteristics of human activities and provide new opportunities for classifying parcels of different uses in cities [21]. However, users of POI data must consider its intrinsic spatial dependencies and not only information on the frequency of POI occurrence [22].

Based on multisource Big Data, artificial intelligence algorithms can fully mine the spatial and temporal characteristics of urban land-use features to obtain richer feature information and achieve better urban land-use classification mapping [23]. For example, Gong et al. realized basic urban land-use type mapping in China in 2018 based on the random forest classification method, medium- and high-resolution remote-sensing imagery, and spatially and temporally assisted Big Data [24]. Zong et al. performed urban land-use classification in Lanzhou City by comparing urban classification accuracy under different feature data combinations using remote-sensing images, network data, and other data sources and selecting the optimal feature combination [25]. Tu et al. proposed an EULUC-seg framework based on the random forest model, which used high-resolution remote-sensing images and socioeconomic data to further refine the original urban classification results [26]. These studies have improved the accuracy of urban land-use classification to a certain extent. However, most of the sampling strategies are still single-sample and multiple-sample methods. Single-sample means that a point with certain rules or a random point in a parcel unit is selected as a model-training sample point [27]. These methods have a higher probability of misclassification of land-use features because a single point is hardly representative of parcel type. Multiple samples are expanded from a single sample by selecting multiple points in a parcel unit as model-training sample points [28,29]. This can most likely reduce the accuracy of land-use classification, as incorrect sample points in a parcel may interfere with the classification results. Both approaches fail to take into account the problem of land-use confusion in a parcel unit. Hence, the classification results may not be sufficient to meet the needs of other functions, such as urban planning and environmental monitoring, due to the insufficient processing of multisource data and the continued confounding of urban land-use classification results.

To reduce the influence of feature confusion on land-use classification and, thus, further improve land-use classification-mapping accuracy, this paper proposes an urban land-use classification method that uses a multisample “voting mechanism” while fully processing multiple sources of data. Using Zhengzhou City as the study area, road buffers are generated according to reasonable thresholds and combined with impervious surface data to generate high-quality basic land-mapping units. Multisource data, such as Sentinel-2 and POI, are preprocessed to construct a multitemporal high-dimensional feature library. To improve classification accuracy, multiple samples are collected for a single parcel as a sample dataset. A random forest model is then used for training and testing to obtain preliminary urban land-use classification results. Finally, the multisample classification results are filtered according to the “voting mechanism” to determine the types of parcels,

to achieve high-precision urban land-use classification mapping, and to provide decision support for sustainable urban development governance.

2. Study Area and Data

2.1. Study Area

Zhengzhou City, the capital of Henan Province, is located in the lower reaches of the Yellow River in China (Figure 1). Zhengzhou not only has a long history and a well-established culture, but is also an important transportation hub and national central city in China. It is 166 km wide from east to west and 75 km long from north to south. The total area of the city is 7446 km².

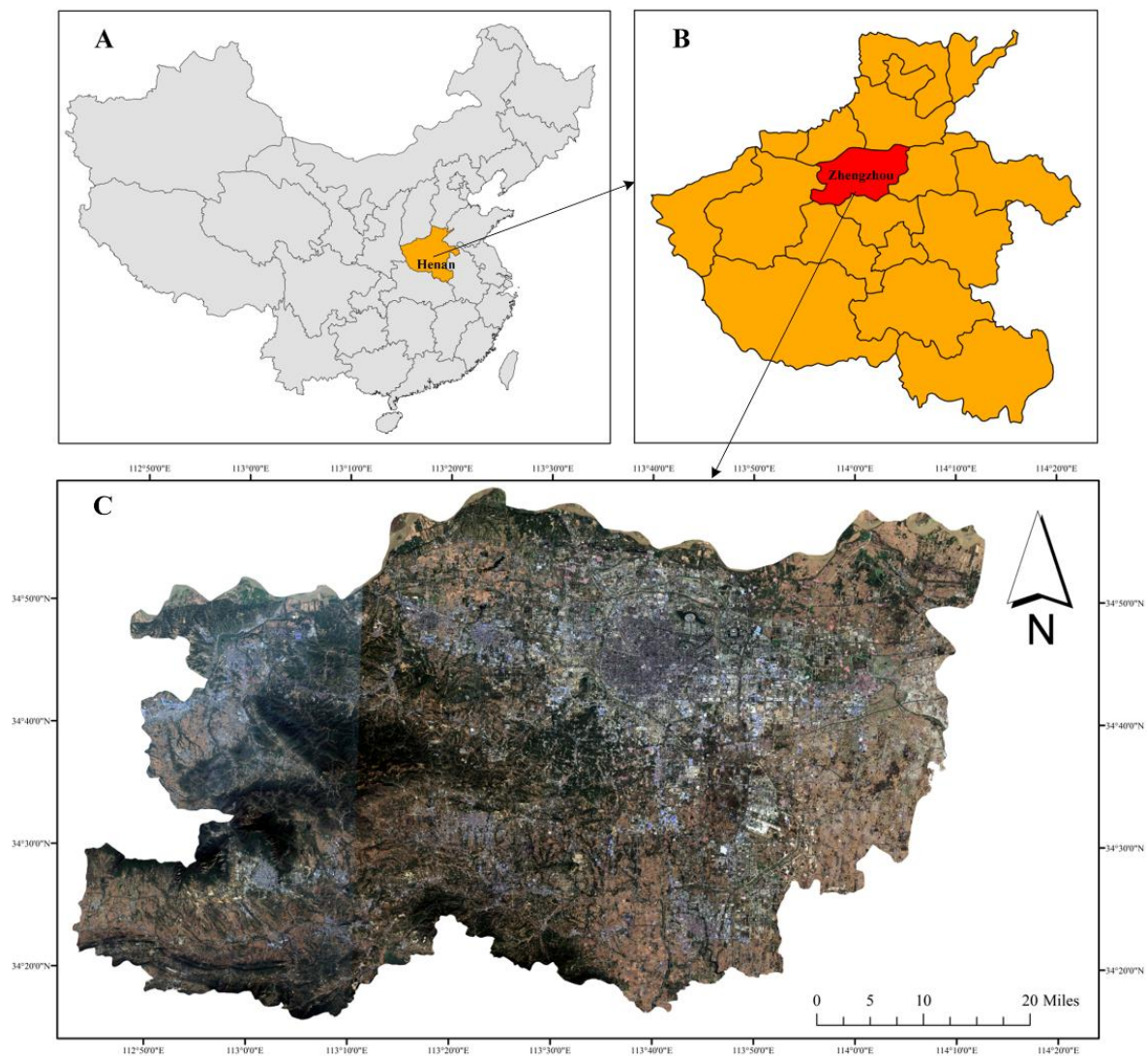


Figure 1. Location of the study area (A) and (B) Henan Province, China. (C) Sentinel-2 image of Zhengzhou, Henan Province.

2.2. Data and Sources

2.2.1. Sentinel-2

Sentinel-2 (<https://eros.usgs.gov/sentinel-2>, accessed on 1 September 2020) is a high-resolution multispectral-imaging satellite, which can be divided into the two satellites of 2A and 2B, with a revisit cycle of 10 days for one satellite and 5 days for two complementary satellites. Sentinel-2 also carries a multispectral imager (MSI) that can cover 13 spectral bands with ground resolutions of 10 m, 20 m, and 60 m, providing images of vegetation,

soil and water cover, inland waterways, and coastal areas for land monitoring, emergency relief services, and other purposes. In this paper, Sentinel-2 remote-sensing images with less than 5% cloudiness in 2018 were used, and remote-sensing images for February, June, September, and November were selected. The specific data sources are shown in Table 1.

Table 1. Remote-sensing image data sources.

Name	Time
S2A_MSIL1C_20180222T030731_N0206_R075_T49SFU_20180222T082850.SAFE	22 February 2018
S2A_MSIL1C_20180222T030731_N0206_R075_T49SGU_20180222T082850.SAFE	22 February 2018
S2A_MSIL1C_20180513T030541_N0206_R075_T49SFU_20180513T055623.SAFE	13 May 2018
S2B_MSIL1C_20180607T030539_N0206_R075_T49SGU_20180607T064334.SAFE	7 June 2018
S2A_MSIL1C_20180930T030541_N0206_R075_T49SGU_20180930T060706.SAFE	30 September 2018
S2B_MSIL1C_20181005T030549_N0206_R075_T49SFU_20181005T055242.SAFE	5 October 2018
S2B_MSIL1C_20181109T030931_N0207_R075_T49SFU_20181109T060447.SAFE	9 November 2018
S2A_MSIL1C_20181119T031011_N0207_R075_T49SGU_20181119T061056.SAFE	19 November 2018

2.2.2. Open-Source Data

Along with the rapid development of Internet technology, Big Data based on network media are gradually being enriched, and more open-data-sharing platforms are appearing. Data users can quickly access massive free geographic data on these platforms, such as OpenStreetMap data, POI data, and others, which provide new data sources for performing urban land-use classification.

OSM

OSM (<http://www.openstreetmap.org>, accessed on 1 September 2020) is a kind of open-source geographic database collected, processed, and produced by many volunteers, where all users can upload and edit data on its platform. OSM has become one of the more successful spontaneous geographic information platforms with the continuous application and popularity of electronic maps, and its data are widely used due to its fast updates, low cost, free access, and other advantages. It is widely used to delineate urban parcels and urban neighborhoods [30].

OSM vector data include point, line, and surface elements. Among them, point elements are divided into six categories: transport, natural, pofw (place of worship), traffic, places, and POIs (places of interest). Line elements are divided into three categories: waterways, roads, and railways. Finally, surface elements are divided into natural, traffic, water, land use, POIs, pofw, buildings, places, and transportation.

POI

POI, as one of the emerging Big Data sources, has a large volume of data, easy access, and fast update speed and describes spatial entities and their attributes related to urban spatial activity patterns, as well as visualizes current urban land-use situations. With the rapid development of modern electronic maps and geolocation technologies, location-based POI data form an important part of geospatial Big Data, which are divided into several specific classifications from the perspective of elements: residential, educational, medical, entertainment, financial, etc. [31]. In addition, POI data have the advantages of fast update speed, large data volume, a fine classification system, and easy access and are widely used for urban economic development research, urban boundary extraction, population studies, and urban spatial structure identification [32,33].

2.2.3. Ground Truth Data

Gong et al. proposed the concept of essential urban land-use categories (EULUCs) to describe the current urban development situation in China [24]. Based on the EULUC classification system and the current urban land-use situation in Zhengzhou, this paper divided urban land use into four Level I land-use categories and ten Level II land-use categories.

To obtain objective and representative sample units, this study generated several random points on a basic mapping unit and screened out parcel units where the random points were located. The screened parcel units were visually interpreted and field-surveyed to obtain land-use information. Some parcel units with larger areas and more complex land-use types were eliminated to reduce the degree of land-use mixture in the units. Finally, the sample parcel units were labeled with land-use types, and 413 real and valid sample parcel units were obtained. The samples and the numbers in each category are shown in Table 2 and Figure 2.

Table 2. Total sample size by feature type.

Level I	Count	Level II	Count
Residential	81	Residential	81
Commercial	84	Business office	34
		Commercial service	50
Industrial	64	Industrial	30
		Warehouse	34
Public	184	Administrative	34
		Educational	50
		Medical	34
		Sports and culture	33
		Parks and green space	33



Figure 2. Example of a sample of features: (a) educational land, (b) residential land, (c) institutional land, and (d) commercial land.

3. Methodology

This paper proposed a high-precision urban land-use classification method based on multisource data and a multisample “voting mechanism” (Figure 3). First, the OSM road network and impervious surface data [34] were used to create basic mapping units, and the final sample plots were verified by visual interpretation and field survey. Meanwhile, the multisource data were processed to build a high-dimensional feature library. The training sample data and high-dimensional feature data were input into random forests for training and were tested using validation sample data to obtain preliminary classification results. Finally, the preliminary classification results were filtered based on a multisample “voting mechanism” to obtain the final urban land-use classification map.

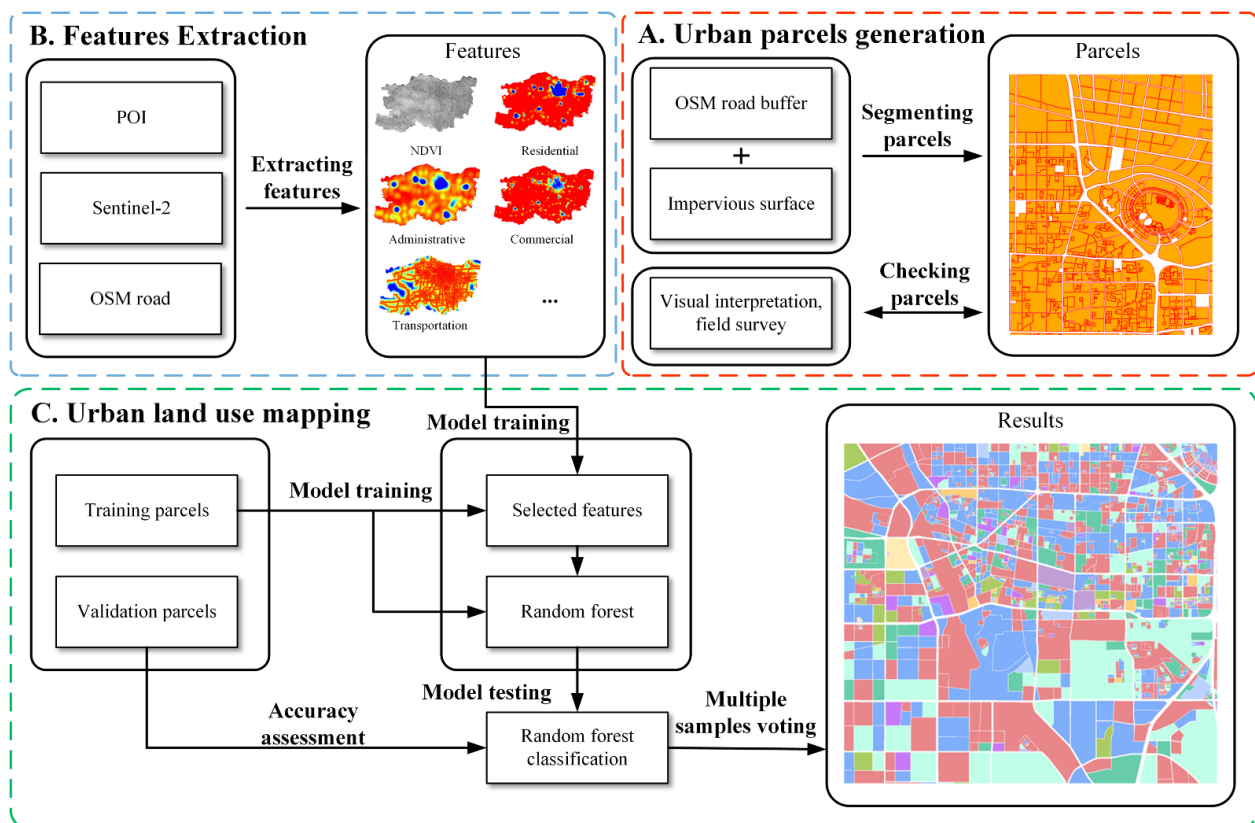


Figure 3. Main flowchart: (A) urban sample plot production; (B) multisource data feature extraction; (C) urban land-use map generation using random forests and multisample voting.

3.1. Road Network Data Processing

In this study, the acquired OSM road network data were integrated into seven categories, including highways, urban expressways, primary trunk roads, secondary roads, tertiary roads, minor roads, and special types of road. The road network was generated by sampling different classes of roads to establish buffer zones according to the actual situation in the city. The specific road sampling distances are shown in Table 3.

Table 3. Sampling distances of different levels of roads.

Category	Sampling Distance (m)	Road Width (m)
Expressway	20	40
Urban expressway	17.5	35
Primary main road	15	30
Secondary roads	10	20
Tertiary roads	6	12
Minor roads	5	10
Special types of road	4	8

3.2. Feature Extraction

Previous studies have shown that, when multitemporal remote-sensing image spectral features are used, the classification effect is much better than with single-temporal remote-sensing images [7]. Therefore, in this paper, remote-sensing images for four time periods (months) in Zhengzhou City in 2018 were selected, thus extracting a set of high-dimensional spatial and temporal features.

First, the average normalized vegetation index and the average normalized building index over the four time periods were calculated based on the bands, and two remote-sensing index features were obtained. Second, the mean, contrast, variance, synergy, correlation, dissimilarity, entropy, and second-order moments of the eight bands were calculated based on a gray-level co-occurrence matrix (GLCM) to obtain $4 \times 8 \times 8$ texture features. Finally, the POI data were combined into five classes, and kernel density analysis and Euclidean distance analysis were performed to obtain five features. Therefore, 2 spectral features, 256 texture features, and 5 POI features were selected in this study, for a total of 263 high-dimensional spatial and temporal features, which are detailed in Table 4.

Table 4. Features extracted from Sentinel-2 and POI data.

Type	Name	Variable	Count
Index feature	NDVI	Mean_NDVI	1
	NDBI	Mean_NDBI	1
Texture characteristic	Mean	Mean_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Variance	Variance_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Contrast	Contrast_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Homogeneity	Homogeneity_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Dissimilarity	Dissimilarity_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Correlation	Correlation_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	Entropy	Entropy_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
	SecondMoment	SecondMoment_b2, b3, b4, b5, b6, b7, b8, b8A	$4 * 8$
POI feature	Administrative center	Kernel_xzzx	1
	Residential community	Kernel_zzxq	1
	Public service	Kernel_ggfw	1
	Commercial center	Kernel_sfyl	1
	Transportation	EucDistance_jtd	1

3.2.1. Texture Features

The Sentinel-2 remote-sensing image bands of 1, 9, and 10 were removed, and bands 2, 3, 4, 5, 6, 7, 8, and 8A were resampled to unify the spatial resolution to 10 m, as shown in Table 5.

Table 5. Sentinel-2 parameters.

Band	Resolution (m)	Wavelength (nm)	Spectral Width (nm)
Band 2-Blue	10	490	65
Band 3-Green	10	560	35
Band 4-Red	10	665	30
Band 5-Vegetation Red Edge 1	20	705	15
Band 6-Vegetation Red Edge 2	20	740	15
Band 7-Vegetation Red Edge 3	20	783	20
Band 8-NIR	10	842	115
Band 8A-Vegetation Red Edge 4	20	865	20

The gray-level co-occurrence matrix method was introduced to extract attribute information, such as adjacent interval, intensity, and spatial arrangement of the gray-scale distribution of remote-sensing images in geographic space, and to improve the spatial distribution characteristics of the images.

3.2.2. Index Features

The normalized difference vegetation index (NDVI) and the normalized difference building index (NDBI) were introduced to enhance the characteristics of land-use classification within cities.

NDVI

This index is normally used to monitor and reflect vegetation cover status and is the most widely used vegetation index. It quantifies vegetation by measuring the difference between the near-infrared band (strong reflection of vegetation) and the red band (absorption of vegetation) and effectively reflects the spatial distribution and temporal variation of vegetation, weakens the influence of anomalous values brought about by atmospheric and other factors, and has better spatial and temporal adaptability [35]. The NDVI is calculated using Equation (1):

$$\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}} \quad (1)$$

where R refers to the band reflectance in the red band, and NIR is the band reflectance in the near-infrared band. The NDVI takes on values between -1 and 1 and is generally 0 when the surface cover is bare soil or rock; when the surface is covered with vegetation, the value is positive and is positively correlated with the vegetation cover. The greater the vegetation cover, the higher the value of the NDVI.

NDBI

The normalized building index is proposed to reflect information on building land use and the intensity of urbanization [36], as shown in Equation (2):

$$\text{NDBI} = \frac{\text{SWIR} - \text{NIR}}{\text{SWIR} + \text{NIR}} \quad (2)$$

where NIR is the band reflectance in the near-infrared band, and SWIR refers to the band reflectance in the mid-infrared band. It takes on values between -1 and 1 and is positively correlated with the likelihood of an image element being a building site.

3.2.3. POI Features

In this study, a total of 72,061 POI data points were obtained from the Baidu Map API interface. According to the EULUC classification system adopted in this paper, relevant POI data were reclassified by attribute according to the classification rules shown in Table 6.

Table 6. POI reclassification rules.

Classes	Description
Administrative center	Restaurants, hotels, retail stores, supermarkets, shopping malls, etc.
Residential community	Residential communities, amenities, etc.
Public service	Street offices, government agencies, police stations, etc.
Commercial center	Research institutes, schools, educational institutions, etc.
Transportation	Railroads, highways, main roads, secondary roads, minor roads, special roads, etc.

The processed residential communities, public services, administrative centers, and commercial centers were analyzed separately for kernel density and normalized to unify the POI data of each category to the same order of magnitude for easy analysis and comparison. A Euclidean distance analysis was performed for transportation.

3.3. Multisample Voting

A limited number of selected sample units and mixed land use within the parcels results in less room for improving land-use classification accuracy [37]. Therefore, a research concept of multipoint-result-voting selection was designed (Figure 4). First, by deploying

multiple random points inside a basic mapping unit, the sample size was effectively increased, and the randomness of the model was reduced. After the multiple points were trained by random forests, preliminary land-use classification results were obtained. The predicted results for the multiple points were then filtered by “voting”, and the main land-use type of the site was selected as the land-use type of the basic mapping unit to further determine and refine the classification results. This approach could reduce the influence of mixed features to a certain extent, screen out erroneous samples, and improve land-use classification accuracy.

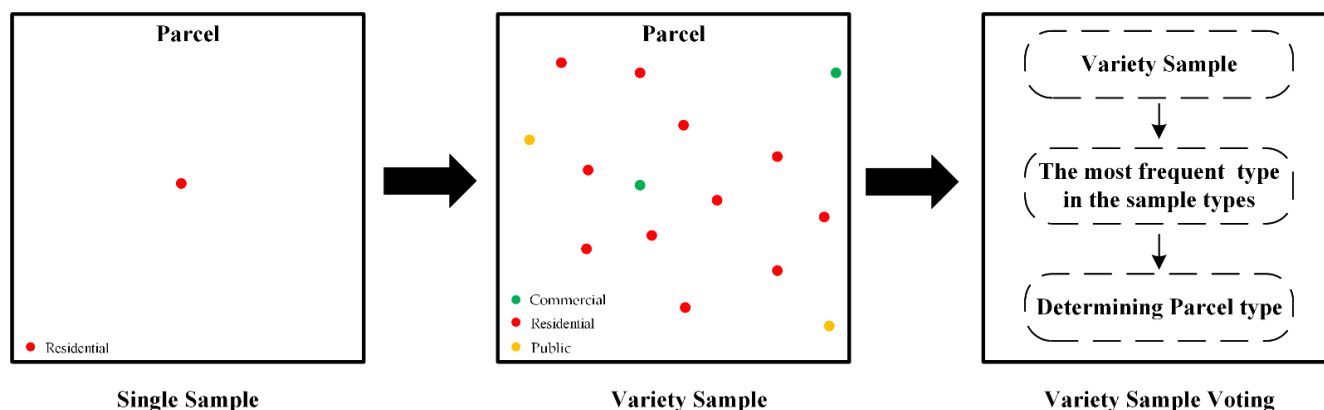


Figure 4. Multisample result voting selection process.

3.4. Random Forest

Random forest is an integrated learning algorithm for data mining with decision-tree-based classifiers [38] that is now widely used for land-use classification [39]. It is composed of a number of weak decision tree classifiers, and multiple decision trees are combined to form a forest. Each decision tree learns features and predictions independently with its own rules so that different decision trees do not have the same results. The results of all the decision-makers are aggregated, and the final classification result is decided by “voting”. The model not only can effectively and quickly process large amounts of data with high classification accuracy, but also has strong anti-noise and anti-interference ability when facing a large amount of redundant data and can effectively avoid overfitting. Studies using the random forest model for classification have proved the superiority of the model over other classification methods [40].

From the algorithmic perspective of the model, the classification process of random forest is simple. First, a training set is generated by random sampling of the original dataset; second, individual subdecision trees are constructed. The training data are placed on the subdecision trees, and each subdecision tree produces one result; finally, the results of the subdecision trees are voted on, and the one with the highest number of votes is the final classification result.

Regarding the parameters of random forest, the decision tree classifier includes both *mtry* and *ntree* parameters, where *mtry* refers to the number of variables used in a single decision tree node, and *ntree* refers to the number of decision trees in the random forest model. Moreover, the random forest model evaluates model generalization performance by the out-of-bag (OOB) score, where OOB is the ratio of the number of misclassified samples to the total number of samples in the original data that were not sampled.

3.5. Evaluation Methodology

In this study, overall accuracy (OA), producer’s accuracy (PA), and user’s accuracy (UA) were used for accuracy evaluation.

OA is the proportion of the number of correctly classified samples to the total number of samples. The mathematical expression is as follows:

$$OA = \frac{\sum_{i=1}^p X_{ii}}{N} \quad (3)$$

where p is the number of categories, X_{ii} represents the number of correct classifications of a particular category, and N represents the total number of samples.

PA is the proportion of correctly classified samples to the predicted samples. Its mathematical expression is as follows:

$$PA_i = \frac{X_{ii}}{X_{+i}} \quad (4)$$

where X_{ii} represents the number of correctly classified samples in a category, and X_{+i} represents the total number of samples in a category.

UA is the proportion of correctly classified samples to the real samples. Its mathematical expression is as follows:

$$UA_i = \frac{X_{ii}}{X_{+i}} \quad (5)$$

where X_{ii} represents the number of correctly classified samples in a category, and X_{+i} represents the total number of real features.

4. Results

In this study, the random forest model was selected to classify urban land use. The model was built using R language, with the parameters of *mtry* set to 16 and *ntree* set to 600.

4.1. Urban Parcels

Currently, the grid-cell-based approach is commonly used in many studies to segment urban sites. However, this segmentation ignores the geospatial correlation of features and destroys the spatial connection of urban sites in terms of function and structure [41]. Therefore, it is necessary to start from urban spatial structure and functional units to better combine remote-sensing image data and auxiliary data to obtain good classification results. In this study, the OSM road network was used to segment the impervious surfaces of Zhengzhou City based on a basic mapping unit. First, the road network was generated based on the sampling distance and OSM road data, and the network topology was checked and modified. The road network was then used to segment the urban impervious surfaces of the study area to obtain preliminary basic mapping units. Then, parcels with small and fragmented areas after segmentation were integrated, thus reducing the fragmentation of the basic mapping units, and finally, 12,466 basic mapping units were obtained (Figure 5).

4.2. Sample Size

To investigate the relationship between the number of training samples and the performance of the random forest and to filter out a more suitable number of datasets, the sample set was divided, and different percentages of training and validation samples were set. The samples were divided into five groups according to different percentages, where the percentages of samples in the validation set in order were 10%, 20%, 30%, 40%, and 50%. Urban land-use classification was performed by a random forest model with multiple points and validated by overall accuracy and OOB scores. The classification results for each percentage are shown in Table 7.

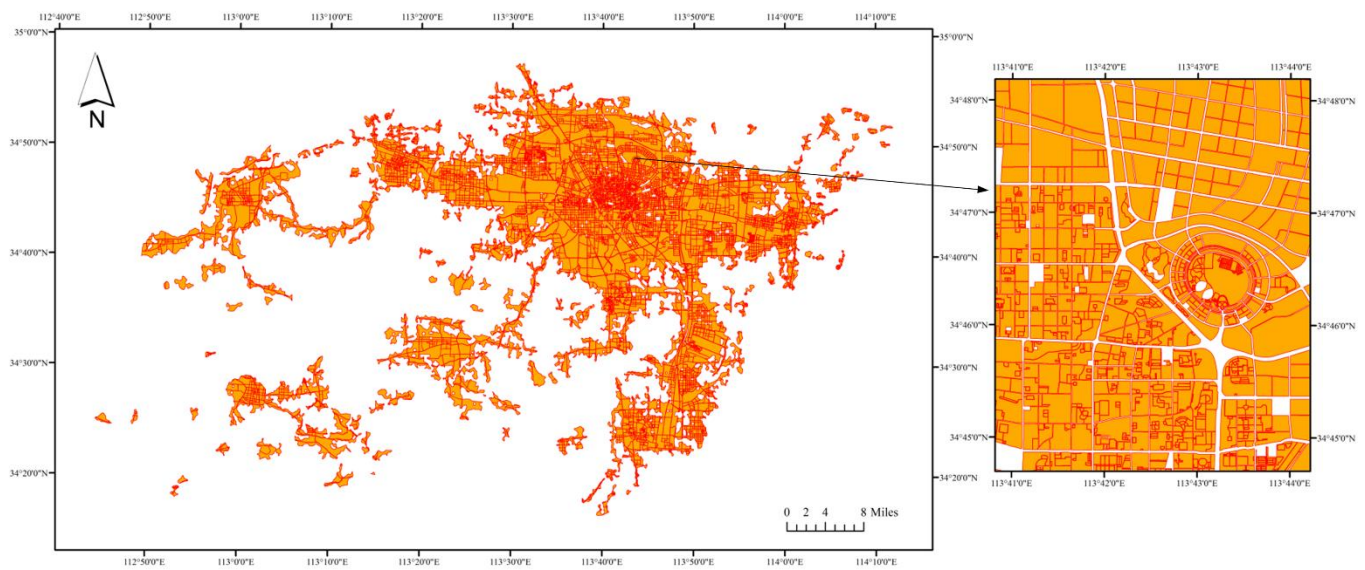


Figure 5. Urban parcels.

Table 7. Sample size and feature classification results for each experimental group. The bolded content is the best performed.

Groups	Level I Accuracy (%)	OOB Score	Level II Accuracy (%)	OOB Score
1	72.57	78.14	61.98	73.22
2	72.21	79.16	62.21	74.72
3	73.72	80.80	64.73	75.90
4	71.72	79.28	60.90	71.94
5	70.66	78.73	59.19	70.13

Table 7 shows that the classification accuracy values of primary features, and the OOB scores reached the highest values in the third group with increasing numbers of samples, but the differences in accuracy from neighboring groups were not very large. Regarding the classification accuracy of secondary features, it also showed an overall improvement trend with increasing numbers of samples, and the accuracy of group 3 was significantly higher than the accuracies of other groups. Therefore, in this paper, 70% of the samples were split into training samples, and the remaining samples were used as validation samples.

Based on the above discussion, to investigate the influence of the number of sample points on the final classification results in the process of multiple-sample-point voting, six sets of control experiments were performed in which the sample size was increased from 5 to 20. The experiments were validated with an overall accuracy assessment to select the appropriate sample size for voting. The specific results are shown in Figure 6.

As the number of voting samples for a single parcel increased, the classification accuracy increased accordingly at first. However, when the sample size was 17, the classification accuracy started to decrease. The classification accuracy reached its highest value when the sample size was 14, with 85.37% for primary features and 78.05% for secondary features. Therefore, the number of samples randomly taken in the multiple-sample-voting process was set to 14.

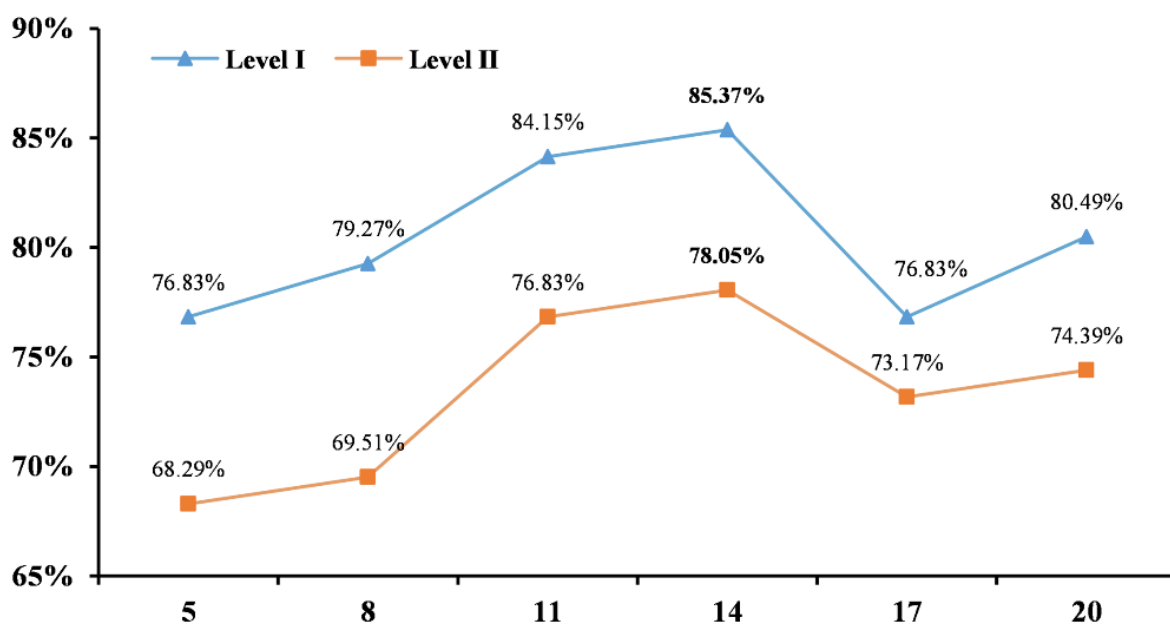


Figure 6. Effect of sample size on multiple-sample-voting strategy.

4.3. Classification Results of Sample Selection Strategies

At present, the ways to select sample points in a basic mapping unit include both single sample points and multiple sample points. These methods have improved land-use classification accuracy to a certain extent, but the confusion of features present in plots has not been resolved.

In this study, a multisample voting mechanism was used to solve the problem of feature confusion in a single plot. Figure 7 shows the results for OA accuracy after training different sample selection strategies with segmentation in random forests according to nine segmentation ratios. Figure 7 clearly shows that the feature accuracy with multiple samples was not as good as that with a single sample. This was the case because of the purity of the parcels; the parcel itself had the problem of feature confusion, and the single-sample procedure had a certain probability of obtaining the wrong sample after uniform expansion, leading to lower accuracy. However, the distribution interval of the classification results was smaller, and the classification process was more stable; therefore, multiple samples could improve model robustness to a certain extent. In addition, the results of screening by multisample voting were significantly better than those of other strategies, regardless of Level I or Level II precision. This occurred because the screening process cleaned up some of the erroneous samples, thus improving feature classification accuracy. Therefore, the multisample voting mechanism proposed in this paper was favorable for land-use classification.

4.4. Urban Land-Use Classification Results

The land-use classification was completed for Zhengzhou City according to the sample split ratio; the classification results for each level of feature are shown in Table 8. The OA values of Level I and Level II features were 85.37% and 78.05% respectively. In addition, among the Level I features, the PA of the residential category was the highest at 0.91, followed by those of the industrial and public categories. The public and residential categories had the best UA performances. On the whole, the primary features were not difficult to distinguish. Among the Level II features, the commercial services and educational categories had lower PA accuracies that were lower than the OA values. The category of business office had a relatively lower UA of 64.29%. However, the medical category had the highest PA accuracy of 93.33%, and the residential category achieved the highest UA of 93.05%.

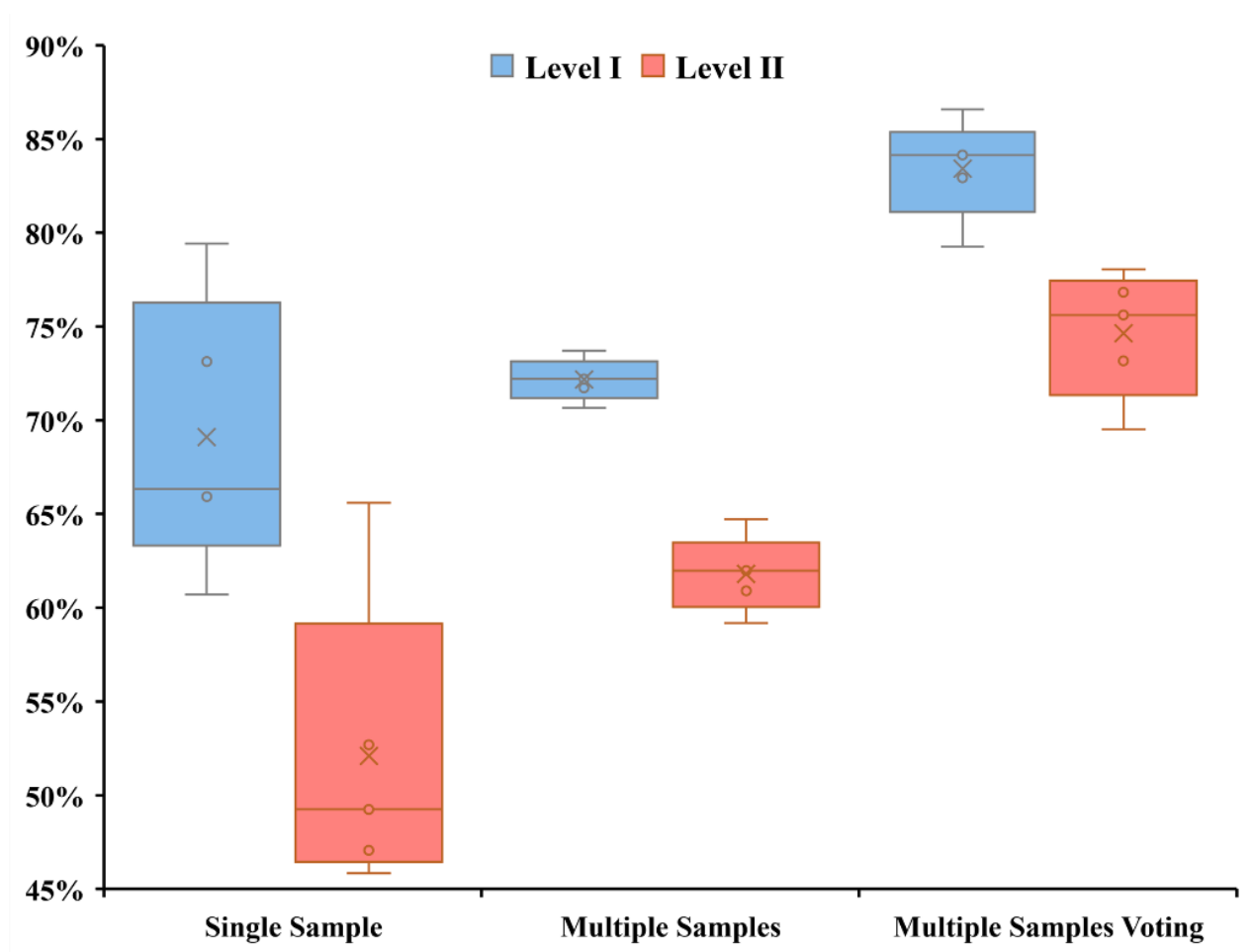


Figure 7. Accuracy distribution of different sample selection strategies.

Table 8. Classification accuracy of Level I and Level II land-use features.

Level I	PA	UA	OA	Level II	PA	UA	OA
Residential	91.13%	81.50%	85.37%	Residential	77.00%	93.05%	78.05%
Commercial	79.44%	77.67%		Business office	92.31%	64.29%	
				Commercial service	67.09%	84.57%	
Industrial	84.94%	64.98%		Industrial	84.00%	72.81%	
				Warehouse	76.92%	76.19%	
Public	85.10%	94.48%		Administrative	85.51%	72.84%	
				Educational	70.94%	89.44%	
				Medical	93.33%	72.00%	
				Sports and culture	82.50%	79.29%	
			Parks and green space	72.90%	92.86%		

Figures 8 and 9 show the results of land-use classification for Level I and Level II features in Zhengzhou City, with a total classified area of 1993.22 km². Among the primary features, the public category had the highest percentage among all the features at 67.11%, followed by the residential category, with only 18.27%. Among the secondary features, the category of residential accounted for 40.12%, which was 19.35% higher than that of parks and green space, which accounted for the second-highest percentage.

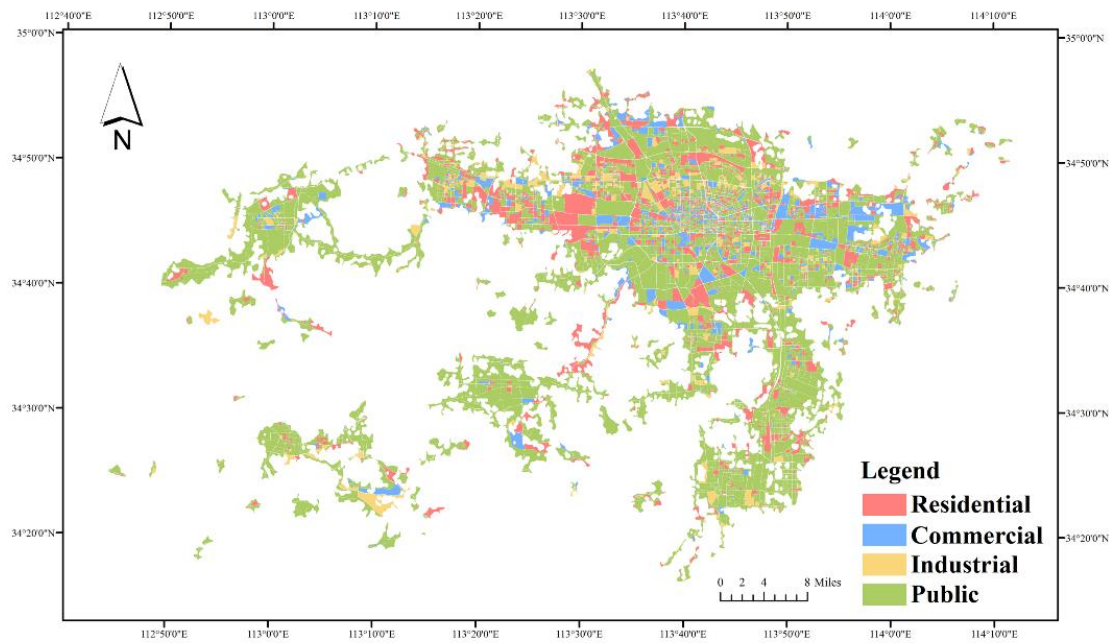


Figure 8. Classification map of Level I features in Zhengzhou.

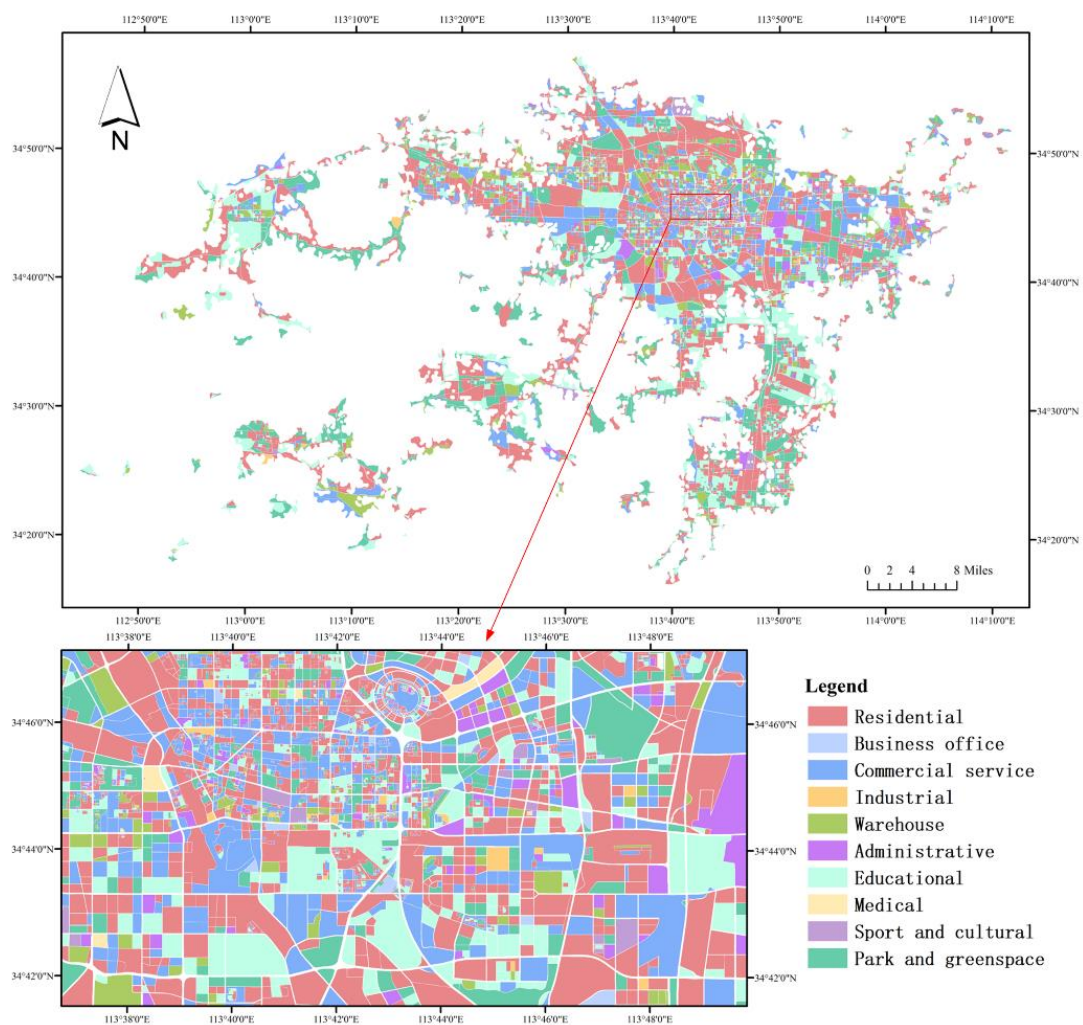


Figure 9. Classification map of Level II features in Zhengzhou.

5. Discussion

In this paper, the “voting” mechanism of multiple point prediction for basic mapping units was combined with a random forest model to extract land-use classification information within the city. The overall classification accuracy of Level I features was 85.07%, and the overall classification accuracy of Level II features was 74.62%, both of which were higher than the results of other methods, thus realizing fine-scale urban land-use mapping. The research results show that the use of spatial and temporal feature information from remote-sensing images and POI data, as well as the use of the multisample voting classification method for land parcels, could reduce the influence of confusion of land-use features in land parcels on the classification results to a certain extent and could achieve the purpose of promoting the fine-scale mapping of urban land-use.

5.1. Contributions of Different Features

In previous studies, Tu et al. pointed out that the features influencing Sentinel-2 were the main factors affecting classification when compared to other features [26]. Chang et al. and Sun et al. each suggested that the spatial features of POI were beneficial for urban land-use classification [1,12]. To explore the contributions of different features to urban land-use classification, this paper discusses the contributions of all the features using random forests. As can be seen in Figure 10, in the mean decrease Gini contribution regarding Level I features, the most important ones were administrative centers and commercial centers, and among the top eight most important features, POI features were clearly more important than remote-sensing features. Regarding Level II features, the first four important features were all based on POI data and had higher values, followed by remote-sensing features. In mean decrease accuracy, regarding Level I features, both remote-sensing features and POI features accounted for larger proportions, and the difference between them was smaller, but POI features were still more important. Regarding Level II features, the top three most important features were among the POI features, and the contribution of the POI features was also greater than that of remote-sensing features.

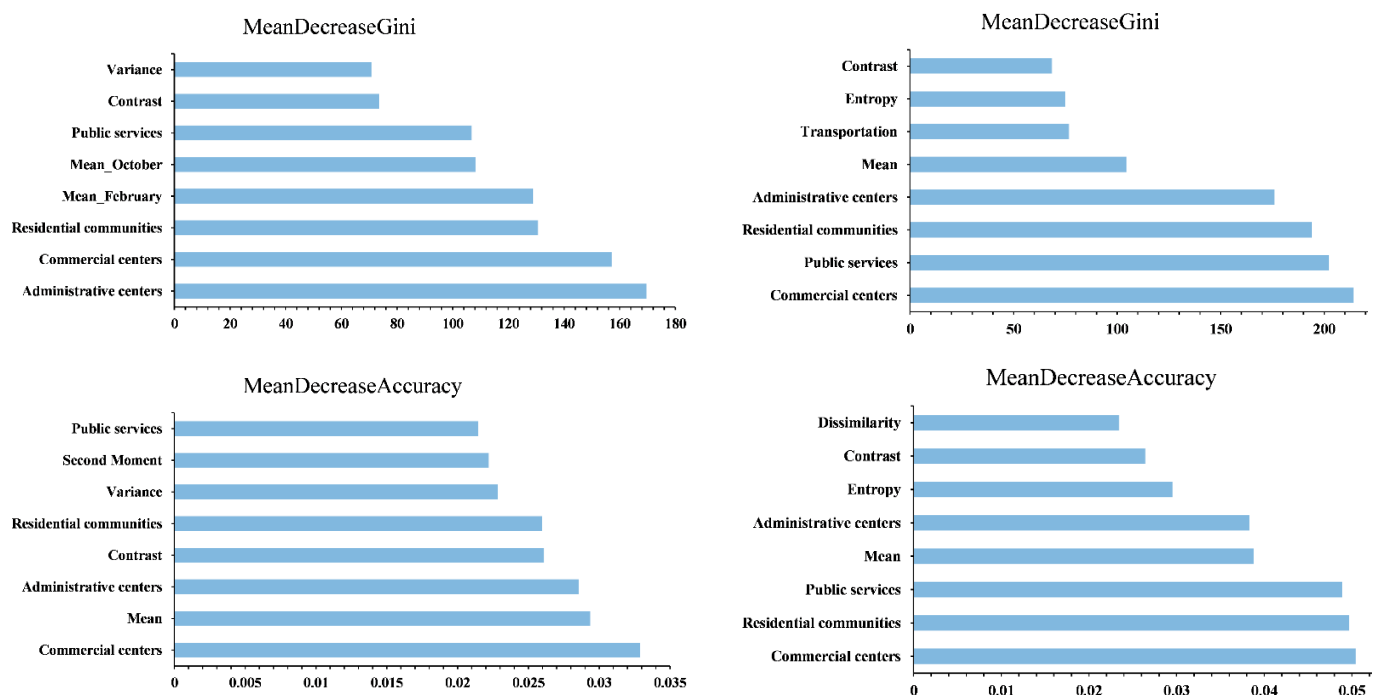


Figure 10. Contributions of different features. The (left) panel shows the feature contributions of Level I feature classification. The (right) figure shows the contributions of features for Level II feature classification.

In general, POI features appeared to be more important in land-use classification. This may be because the processed POI spatial features could distinguish different land-use features well and could effectively improve land-use classification accuracy.

5.2. Shortcomings

Although the method proposed in this paper achieved a good classification effect for Zhengzhou urban land use, it still had some shortcomings. First, under the EULUC classification system, the mixing of land uses within cities imposed severe constraints on classification accuracy. It is necessary to consider adjusting urban land-use classification and cognitive rules and to consider adopting more multidimensional urban Big Data features to realize fine-scale land-use mapping in cities. Second, the selection of feature data needs to be improved. Although a high-dimensional feature library was constructed in this study, it contained some features that could not contribute to distinguishing land-use classes, leading to data redundancy and low computational efficiency. Future studies should consider streamlining feature data, increasing the diversity and uniqueness of features, and improving feature quality.

6. Conclusions

This paper took Zhengzhou City as the study area, proposed the construction of a high-dimensional feature library using multisource data, adopted a multisample voting mechanism for urban land-use information extraction on the basis of random forest classification, and achieved 85.37% accuracy for primary feature classification mapping and 78.05% accuracy for secondary feature classification mapping. Among these approaches, the superiority of the multitemporal high-dimensional feature library in extracting urban land-use information maximized the extraction of effective information and enhanced the authenticity, reliability, and applicability of the classification results. In addition, the research idea of multipoint result voting could effectively reduce the influence caused by the mixing of multiple land uses within a parcel and improved urban land-use classification accuracy. Ultimately, this paper further improved the accuracy of fine-scale urban land-use mapping by combining multisource data and a multisample voting mechanism, as well as providing a reference basis for urban managers to conduct sustainable planning.

The feature selection and methods still need to be further discussed and improved. On the one hand, data such as population data and building height can be added. Moreover, the method can be optimized and improved in future studies by combining cutting-edge technologies, such as deep learning, to improve the generalization ability in different spatial and temporal scenarios.

Author Contributions: Conceptualization, H.W.; methodology, K.Z., H.Z. and H.W.; validation, K.Z. and H.Z.; formal analysis, K.Z. and H.Z.; investigation, K.Z. and H.Z.; resources, H.W. and F.Q.; data curation, Z.W. and H.Z.; writing—original draft preparation, H.Z. and K.Z.; writing—review and editing, H.W. and K.Z.; visualization, H.Z., K.Z. and J.Z.; supervision, H.W. and F.Q.; project administration, H.W. and F.Q.; funding acquisition, H.W. and F.Q. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the support from the National Major Project of High-Resolution Earth Observation System (Grant Number 80-Y50G19-9001-22/23), the Young Key Teacher Training Plan of Henan (Grant Number 2020GGJS028), the Key Scientific Research Project Plans of Higher Education Institutions of Henan (Grant Number 21A170008), and the Technology Development Plan Project of Kaifeng (Grant Number 2003009).

Data Availability Statement: Not applicable.

Acknowledgments: The research was also supported by the National Earth System Science Data-Sharing Infrastructure: National Science and Technology Infrastructure of China—Data Center of Lower Yellow River Regions (<http://henu.geodata.cn> accessed on 19 September 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, J.; Wang, H.; Song, Z.; Lu, J.; Meng, P.; Qin, S. Mapping Essential Urban Land Use Categories in Nanjing by Integrating Multi-Source Big Data. *Remote Sens.* **2020**, *12*, 2386. [\[CrossRef\]](#)
2. Adam, E.; Mutanga, O.; Odindi, J.; Abdel-Rahman, E.M. Land-Use/Cover Classification in a Heterogeneous Coastal Landscape Using RapidEye Imagery: Evaluating the Performance of Random Forest and Support Vector Machines Classifiers. *Int. J. Remote Sens.* **2014**, *35*, 3440–3458. [\[CrossRef\]](#)
3. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying Urban Land Use by Integrating Remote Sensing and Social Media Data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [\[CrossRef\]](#)
4. Guo, P.; Zhang, F.; Wang, H. The Response of Ecosystem Service Value to Land Use Change in the Middle and Lower Yellow River: A Case Study of the Henan Section. *Ecol. Indic.* **2022**, *140*, 109019. [\[CrossRef\]](#)
5. Liu, J.; Chen, L.; Yang, Z.; Zhao, Y.; Zhang, X. Unraveling the Spatio-Temporal Relationship between Ecosystem Services and Socioeconomic Development in Dabie Mountain Area over the Last 10 Years. *Remote Sens.* **2022**, *14*, 1059. [\[CrossRef\]](#)
6. Huang, B.; Zhao, B.; Song, Y. Urban Land-Use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [\[CrossRef\]](#)
7. Wang, Z.; Wang, H.; Qin, F.; Han, Z.; Miao, C. Mapping an Urban Boundary Based on Multi-Temporal Sentinel-2 and POI Data: A Case Study of Zhengzhou City. *Remote Sens.* **2020**, *12*, 4103. [\[CrossRef\]](#)
8. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An Object-Based Convolutional Neural Network (OCNN) for Urban Land Use Classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [\[CrossRef\]](#)
9. Zhao, Y.; Wu, Q.; Wei, P.; Zhao, H.; Zhang, X.; Pang, C. Explore the Mitigation Mechanism of Urban Thermal Environment by Integrating Geographic Detector and Standard Deviation Ellipse (SDE). *Remote Sens.* **2022**, *14*, 3411. [\[CrossRef\]](#)
10. Feng, M.; Li, X. Land Cover Mapping toward Finer Scales. *Sci. Bull.* **2020**, *65*, 1604–1606. [\[CrossRef\]](#)
11. Mao, W.; Lu, D.; Hou, L.; Liu, X.; Yue, W. Comparison of Machine-Learning Methods for Urban Land-Use Mapping in Hangzhou City, China. *Remote Sens.* **2020**, *12*, 2817. [\[CrossRef\]](#)
12. Chang, S.; Wang, Z.; Mao, D.; Guan, K.; Jia, M.; Chen, C. Mapping the Essential Urban Land Use in Changchun by Applying Random Forest and Multi-Source Geospatial Data. *Remote Sens.* **2020**, *12*, 2488. [\[CrossRef\]](#)
13. Jinghua, Z.; Zhiming, F.; Luguang, J. Progress on Studies of Land Use/Land Cover Classification Systems. *Resour. Sci.* **2011**, *33*, 1195–1203.
14. Jia, Y.; Ge, Y.; Ling, F.; Guo, X.; Wang, J.; Wang, L.; Chen, Y.; Li, X. Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data. *Remote Sens.* **2018**, *10*, 446. [\[CrossRef\]](#)
15. Zheng, K.; Wang, H.; Qin, F.; Han, Z. A Land Use Classification Model Based on Conditional Random Fields and Attention Mechanism Convolutional Networks. *Remote Sens.* **2022**, *14*, 2688. [\[CrossRef\]](#)
16. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.-L.; Li, T.; Zhou, C. A New Insight into Land Use Classification Based on Aggregated Mobile Phone Data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007. [\[CrossRef\]](#)
17. Wang, Y.; Wang, T.; Tsou, M.-H.; Li, H.; Jiang, W.; Guo, F. Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China. *Sustainability* **2016**, *8*, 1202. [\[CrossRef\]](#)
18. Chen, B.; Tu, Y.; Song, Y.; Theobald, D.M.; Zhang, T.; Ren, Z.; Li, X.; Yang, J.; Wang, J.; Wang, X.; et al. Mapping Essential Urban Land Use Categories with Open Big Data: Results for Five Metropolitan Areas in the United States of America. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 203–218. [\[CrossRef\]](#)
19. Liu, X.; Long, Y. Automated Identification and Characterization of Parcels with OpenStreetMap and Points of Interest. *Environ. Plan. B Plan. Des.* **2016**, *43*, 341–360. [\[CrossRef\]](#)
20. Yuan, J.; Zheng, Y.; Xie, X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *KDD'12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 12 August 2012; pp. 186–194.
21. Zhang, X.; Du, S.; Wang, Q. Hierarchical Semantic Cognition for Urban Functional Zones with VHR Satellite Images and POI Data. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 170–184. [\[CrossRef\]](#)
22. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [\[CrossRef\]](#)
23. Abdi, A.M. Land Cover and Land Use Classification Performance of Machine Learning Algorithms in a Boreal Landscape Using Sentinel-2 Data. *GISci. Remote Sens.* **2020**, *57*, 1–20. [\[CrossRef\]](#)
24. Gong, P.; Chen, B.; Li, X.; Liu, H.; Wang, J.; Bai, Y.; Chen, J.; Chen, X.; Fang, L.; Feng, S.; et al. Mapping Essential Urban Land Use Categories in China (EULUC-China): Preliminary Results for 2018. *Sci. Bull.* **2020**, *65*, 182–187. [\[CrossRef\]](#)
25. Zong, L.; He, S.; Lian, J.; Bie, Q.; Wang, X.; Dong, J.; Xie, Y. Detailed Mapping of Urban Land Use Based on Multi-Source Data: A Case Study of Lanzhou. *Remote Sens.* **2020**, *12*, 1987. [\[CrossRef\]](#)
26. Tu, Y.; Chen, B.; Zhang, T.; Xu, B. Regional Mapping of Essential Urban Land Use Categories in China: A Segmentation-Based Approach. *Remote Sens.* **2020**, *12*, 1058. [\[CrossRef\]](#)
27. Dai, Y.; Liu, Z.; Wang, Y.; Yang, Z. Urban land use classification based on big data: Case of Xining. *J. Beijing Norm. Univ. (Nat. Sci.)* **2021**, *57*, 399–410.

28. Hengkai, L.; Lijuan, W.; Songsong, X. Random forest classification of land use in hilly and mountainous areas of southern China using multi-source remote sensing data. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 244–251.
29. Wei-chun, Z.; Hong-bin, L.; Wei, W. Classification of Land Use in Low Mountain and Hilly Area Based on Random Forest and Sentinel-2 Satellite Data: A Case Study of Lishi Town, Jiangjin, Chongqing. *Resour. Environ. Yangtze Basin* **2019**, *28*, 1334–1343.
30. Ding, J.; Liu, Y.; Zhang, X.; Yang, W. Research on Updating Spatial Data Combined with Open Street Map. *Bull. Surv. Mapp.* **2016**, *6*, 94. [[CrossRef](#)]
31. Xin, L. Recognition of Urban Polycentric Structure Based on Spatial Aggregation Characteristics of POI Elements: A Case of Zhengzhou City. *Acta Sci. Nat. Univ. Pekin.* **2020**, *56*, 692–702. [[CrossRef](#)]
32. Cai, J.; Huang, B.; Song, Y. Using Multi-Source Geospatial Big Data to Identify the Structure of Polycentric Cities. *Remote Sens. Environ.* **2017**, *202*, 210–221. [[CrossRef](#)]
33. Wang, H.; Qin, F.; Xu, C.; Li, B.; Guo, L.; Wang, Z. Evaluating the Suitability of Urban Development Land with a Geodetector. *Ecol. Indic.* **2021**, *123*, 107339. [[CrossRef](#)]
34. Gong, P.; Li, X.; Wang, J. Annual Maps of Global Artificial Impervious Area (GAIA) between 1985 and 2018. *Remote Sens. Environ.* **2020**, *236*, 111510. [[CrossRef](#)]
35. Li, S.; Xu, L.; Jing, Y.; Yin, H.; Li, X.; Guan, X. High-Quality Vegetation Index Product Generation: A Review of NDVI Time Series Reconstruction Techniques. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102640. [[CrossRef](#)]
36. Guha, S.; Govil, H.; Gill, N.; Dey, A. A Long-Term Seasonal Analysis on the Relationship between LST and NDBI Using Landsat Data. *Quat. Int.* **2021**, *575–576*, 249–258. [[CrossRef](#)]
37. Shi, Y.; Qi, Z.; Liu, X.; Niu, N.; Zhang, H. Urban Land Use and Land Cover Classification Using Multisource Remote Sensing Images and Social Media Data. *Remote Sens.* **2019**, *11*, 2719. [[CrossRef](#)]
38. Belgiu, M.; Drăguț, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
39. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [[CrossRef](#)]
40. Pu, R.; Landry, S.; Yu, Q. Assessing the Potential of Multi-Seasonal High Resolution Pléiades Satellite Imagery for Mapping Urban Tree Species. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *71*, 144–158. [[CrossRef](#)]
41. Dong, X.; Xu, Y.; Huang, L.; Liu, Z.; Xu, Y.; Zhang, K.; Hu, Z.; Wu, G. Exploring Impact of Spatial Unit on Urban Land Use Mapping with Multisource Data. *Remote Sens.* **2020**, *12*, 3597. [[CrossRef](#)]