remote sensing

MDPI

Article

Mapping Essential Urban Land Use Categories in Nanjing by Integrating Multi-Source Big Data

Jing Sun [1], Hong Wang [1,*], Zhenglin Song [1], Jinbo Lu [1], Pengyu Meng [1] and Shuhong Qin [2]

[1] College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China; sunjing@hhu.edu.cn (J.S.); zlin_song@hhu.edu.cn (Z.S.); jinbo@hhu.edu.cn (J.L.); mengpengyu@hhu.edu.cn (P.M.)

[2] School of Earth Science and Engineering, Hohai University, Nanjing 211100, China; qinshuhong@hhu.edu.cn

* Correspondence: hongwang@hhu.edu.cn

Received: 29 June 2020; Accepted: 22 July 2020; Published: 24 July 2020

Abstract: High-spatial-resolution (HSR) urban land use maps are very important for urban planning, traffic management, and environmental monitoring. The rapid urbanization in China has led to dramatic urban land use changes, however, so far, there are no such HSR urban land use maps based on unified classification frameworks. To fill this gap, the mapping of 2018 essential urban land use categories in China (EULUC-China) was jointly accomplished by a group of universities and research institutes. However, the relatively lower classification accuracy may not sufficiently meet the application demands for specific cities. Addressing these challenges, this study took Nanjing city as the case study to further improve the mapping practice of essential urban land use categories, by refining the generation of urban parcels, resolving the problem of unbalanced distribution of point of interest (POI) data, integrating the spatial dependency of POI data, and evaluating the size of training samples on the classification accuracy. The results revealed that (1) the POI features played the most important roles in classification performance, especially in identifying administrative, medical, sport, and cultural land use categories, (2) compared with the EULUC-China, the overall accuracy for Level I and Level II in EULUC-Nanjing has increased by 11.1% and 5%, to 86.1% and 80% respectively, and (3) the classification accuracy of Level I and Level II would be stable when the number of training samples was up to 350. The methods and findings in this study are expected to better inform the regional to continental mappings of urban land uses.

Keywords: urban land use; classification; geospatial big data; POIs; Nanjing

1. Introduction

Urban areas accommodate about 55% of the global population, and the proportion is expected to reach almost 70% by 2050. Urban areas also generate over 80% of the global GDP [1]. Besides, according to UN-Habitat, cities consume 78% of the world's energy and produce more than 60% of greenhouse gas emissions. Yet, they account for less than 2% of the Earth's surface [2]. For better environmental monitoring, urban planning, and government management, identifying accurate and detailed urban land use patterns is important [3]. Remote sensing spectral information have the ability to distinguish the physical characteristics of land surface, such as vegetation, water, and buildings [4,5], and pixel-based image classification methods are often used to derive urban land cover and land use information. However, the heterogeneity of land use distribution and the existence of mixing pixels often lead to a lower accuracy of urban land use classification [6]. In recent years, zoning-based [7] and object-based [8] classification methods have aroused increasing popularity among remote sensing communities. Because of the open-source accessibility, OpenStreetMap (OSM) road network data are frequently used to segment urban land parcels [9,10]. The land parcels segmented by

the road network represent comparatively homogeneous socioeconomic functions, which can be used as the basic unit of land use classification [11]. Nevertheless, due to the roads at different administrative levels having different widths [12], urban parcels should not be directly segmented by the road center line. Previous studies simply adopted a unified threshold to generate a road buffer for whole study areas [10], which resulted in under- or over-segmentation of urban parcels, and thus led to reducing the purity of land uses within parcels.

Given the fact that the nature of urban land use categories is caused by differences in human social and economic activities, it is difficult to distinguish urban land use types only relying on remote sensing imagery [13]. For example, the spectral features of commercial, business, and residential land uses are similar, and we cannot differentiate them only using remote sensing images [14]. The urban light density recorded by nighttime light (NTL) images is closely related to the intensity of human activities [15] and can provide a different perspective for extracting urban land use information. Besides, various kinds of open social big data, such as mobile phone positioning data [14], point of interest (POI) data [16], Tencent Mobile phone locating-request (MPL) data [17–19], and taxi trajectory data [20], also provide new opportunities for identifying different urban land uses, due to its intrinsic capability in capturing the spatiotemporal rhythm of the human activities. Among these available social big data, POI data that contain geographic location and attribute information have been most widely used [10,21]. However, most of the previous studies only used the frequency information of POIs without considering their inner spatial dependency [22]. Secondly, since POIs were generated by people's annotation of interested places, there was often an unbalanced distribution in space and in different land use categories. For example, the number of commercial POIs was significantly larger than other land use categories—the POIs density in the central business district was always higher than the other areas. The imbalance of the original POIs cannot reflect the actual distribution of urban land uses, which requires to be preprocessed and regenerated before being used appropriately [16].

Previous studies have demonstrated that the comprehensive utilization of remote sensing data and POI data [9], mobile phone positioning data [14], or other geographical data can effectively improve the urban land use classification accuracy [23]. However, due to the lack in the data availability and a uniform standard on the urban land use classification practices, the mapping of urban land use in China and other countries was still limited [12]. Based on the urban areas delineated from 30 m resolution impervious surface data [24] and a 10 m resolution global land cover map (FROM-GLC10) [25], Gong et al. [12] has generated preliminary mapping results of essential urban land use categories in China (EULUC-China) using both remote sensing images (Sentinel-2, Luojia-01) and social big data (OSM, MPL, POIs). The classification accuracy for 27 validation cities ranges from 40.4% to 82.9% for Level I, and from 34% to 80% for Level II. However, the relatively lower accuracy of EULUC-China may not sufficiently meet the growing demands for urban planning, environmental monitoring, and other aspects at local to regional scopes [26], which can be attributed to the following aspects. Firstly, the buffer thresholds of roads for generating urban parcels were simply divided into major and minor categories without details assigned. Secondly, there were no pre-processing steps to resolve the unbalanced distribution of POI data. Thirdly, the number of parcel samples was still not sufficient (there were 440,798 parcels in China, but for training samples there were only 1795 and for validation samples there were 869). Fourthly, there were potentials to further explore the features extracted from multi-source big data [12]. For example, the building height information [27] and the texture features [28] extracted from remote sensing images have proven to help distinguish different urban land use types.

To address the above issues, this study aims to take Nanjing as the case study, comprehensively utilize multi-source remote sensing images and social big data, refine the generation of urban parcels, resolve the problem of unbalanced distribution between POIs, explore the geographic data features, and analyze the impact of sample size on the classification accuracy in order to improve the mapping result of EULUC-Nanjing. The structure of this article is as follows: Section 1 describes the background of this research and reviews related work on urban land use classification. The data

sources, methodology, and techniques are described in detail in Section 2. The experimental results are illustrated in Section 3. Section 4 provides a meaningful discussion of the results and analyzes the future research directions. Finally, this article is summarized in Section 5.

2. Materials and Methods

2.1. Study Area

Nanjing, the capital of Jiangsu province, is situated on the lower reaches of the Yangtze River in eastern China (Figure 1). As an ancient city, Nanjing enjoys a worldwide reputation not only for its history and culture, but also for its economy and politics. In recent years, with the rapid urbanization, the urban areas of Nanjing have expanded dramatically. As of 2018, the built-up area of Nanjing is approximately 1624 km^2 and the urban population is about 8.44 million, with an urbanization rate of 82.5% [29]. In this study, all built-up areas were regarded as our research area (marked with the red polygon in Figure 1c).

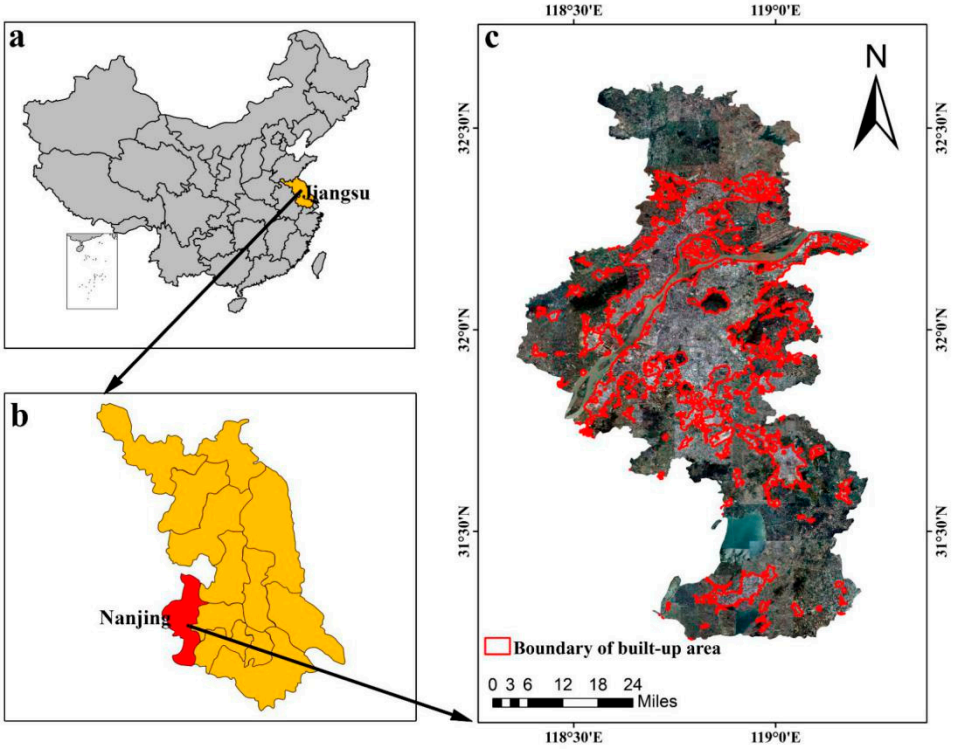

Figure 1. Location of the study area, (**a**) and (**b**) Jiangsu province, China, (**c**) the red polygons are the built-up area of Nanjing [12,30] with Sentinel-2A composite imagery as the background.

2.2. Data and Sources

2.2.1. Remotely Sensed Data

Sentinel-2A/B Composite Imagery

All preprocessed Sentinel-2A/B images from 1 January to 31 December 2018 were first collected from the Copernicus Open Access Hub (https://scihub.copernicus.eu). Then, the normalized difference vegetation index (NDVI) of each pixel was calculated. The pixel-based maximum NDVI values were finally used as a quality index to merge the whole-year images as the greenest composite imagery [12].

Luojia-1 Nighttime Light Image

LJ1-01, launched in June 2018, is the first remote sensing satellite focusing on nighttime light in China. It can obtain high-precision nighttime light imagery with a swath of 250 km and a spatial resolution of 130 m [15]. In this study, we used the Luojia-1 nighttime light images acquired from December 2018 as an indicator of human activities. The data can be downloaded freely at the High-Resolution Earth Observation System of the Hubei Data and Application Center (http://59.175.109.173:8888/app/login.html).

Google Earth High-Resolution Image

Google Earth high-resolution image [31] of Nanjing obtained in May 2018 was downloaded from BIGEMAP (http://www.bigemap.com), with a resolution of 2 m, including three bands of blue, green, and red. In this study, we used it as an auxiliary data to help preprocess POIs.

2.2.2. Social Media Data

Point of Interest

We purchased the POI data obtained in December 2018 from Gaode Map Services (https://lbs.amap.com/), which is one of the most popular and largest web map service providers in China. The data have a total of 760,000 records with 19 categories in the top level and 400 categories in the second level.

Mobile Phone Locating-Request Big Data

Tencent mobile phone locating-request big data records the real-time location of Tencent active users such as QQ (an instant-messaging software, 800 million users in China), WeChat (a chatting software, 350 million users in China), Tencent games (an online game community, 200 million users in China), and others, which can reflect the spatial distribution of the population in the study area [18]. The hourly data for an entire week from 12 to 18 August 2019 (12–16 is the workday, 17–18 is weekend), with a spatial resolution of 25 m were collected from the Tencent Location Big Data platform (http://heat.qq.com). The data are in the format of TXT, including four fields: count, longitude, latitude, and time. The count field carries the heat information of the population.

2.2.3. Other Data

The Impervious Surface

The impervious surface is defined as the ground surface, such as roof, asphalt, or cement road, which is the key index to characterize the degree of urbanization and evaluate the quality of the urban ecological environment. In this study, we used the 2018 impervious surface [12] extracted from Landsat images by the "exclusion and inclusion" framework [30] to represent the urban area.

OSM Road Network

OSM road network data collected in December 2018 were downloaded from OpenStreetMap (https://www.openstreetmap.org) [32]. The data are in vector format and contain a series of information, such as road names, road grades, etc.

Building Footprint Dataset

The building footprint data collected in 2018 were downloaded from BIGEMAP (http://www.bigemap.com), which are in vector format, including information about the area and number of floors for each building. The building polygons are mainly extracted from the Google Earth high-resolution image, and the building height is inverted from its shadow length.

FROM-GLC10

FROM-GLC10 [25] is the first 10 m resolution global land-cover map in the world, which can be freely downloaded from the website: http://data.ess.tsinghua.edu.cn/. In this research, we used FROM-GLC10 [25] to determine the land cover types in the non-built-up area in Nanjing.

2.3. Methods

The study included the following four procedures (Figure 2). Firstly, urban parcels were generated by an overlay analysis of the road buffer, the water, and the impervious surface data. Secondly, multiple features were extracted from Sentinel-2 images, Luojia-1 nighttime light image, POI data, MPL data, and Nanjing building footprint data. Thirdly, training and validation samples were collected by both the visual interpretation and field investigation. Fourthly, the mapping of EULUC-Nanjing and assessment of classification accuracy were performed.

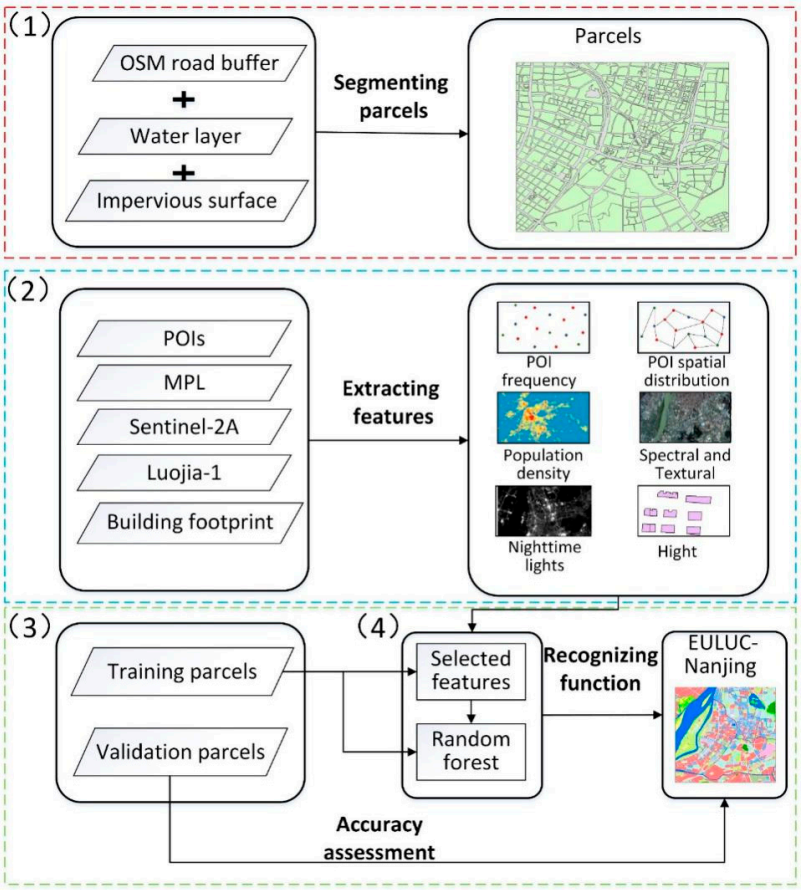

Figure 2. Procedures of mapping EULUC-Nanjing [12].

2.3.1. Classification System

According to the EULUC classification system [12], urban land use types are divided into five classes in Level I (01 Residential, 02 Commercial, 03 Industrial, 04 Transportation, 05 Public management and service) and twelve classes in Level II (0101 Residential, 0201 Business office, 0202 Commercial service, 0301 Industrial, 0401 Road, 0402 Transportation station, 0403 Airport, 0501 Administrative, 0502 Educational, 0503 Medical, 0504 Sport and cultural, 0505 Park and greenspace). For the non-built-up area, there are four land cover types in Nanjing, including cropland, forest, grassland, and water, by referring to the FROM-GLC10 data [25].

2.3.2. Parcels Generation

In this study, according to the attributes of OSM data (Figure 3), the roads were divided into seven levels as the primary, secondary, tertiary, residential, motorway, trunk road, and railway. In order to set buffer thresholds for different road levels, the 604 samples were randomly selected from the above seven road levels and their width were measured using high-spatial-resolution imagery in the BIGEMAP software. Here, the size of samples from each road level were more than 30 percent of its total numbers. Then, the upper quartile of the roads' width statistics of each level was adopted as the buffer thresholds (Table 1). By overlaying the road buffer and the water layer from FROM-GLC10 [25] and the impervious surface data, the built-up area of Nanjing was segmented into 8209 land parcels (Figure 4).

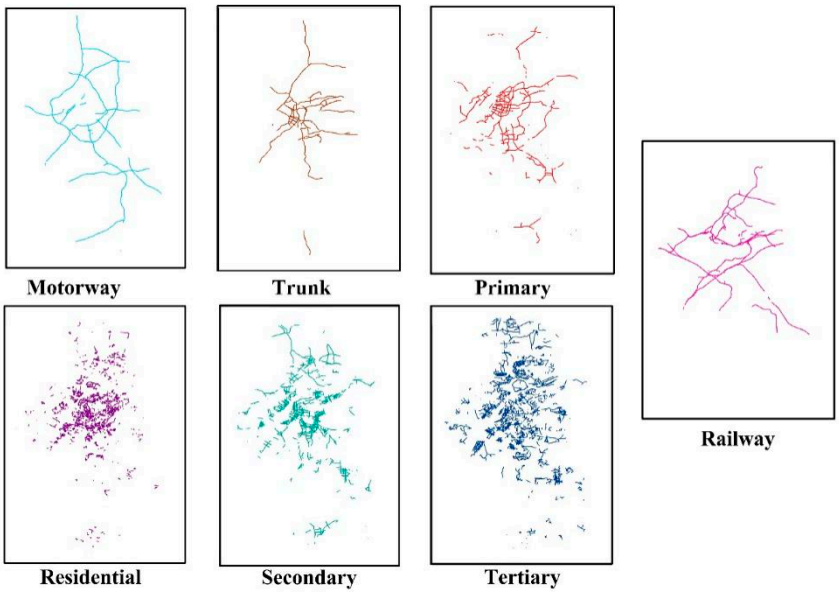

Figure 3. Distribution of different types of roads.

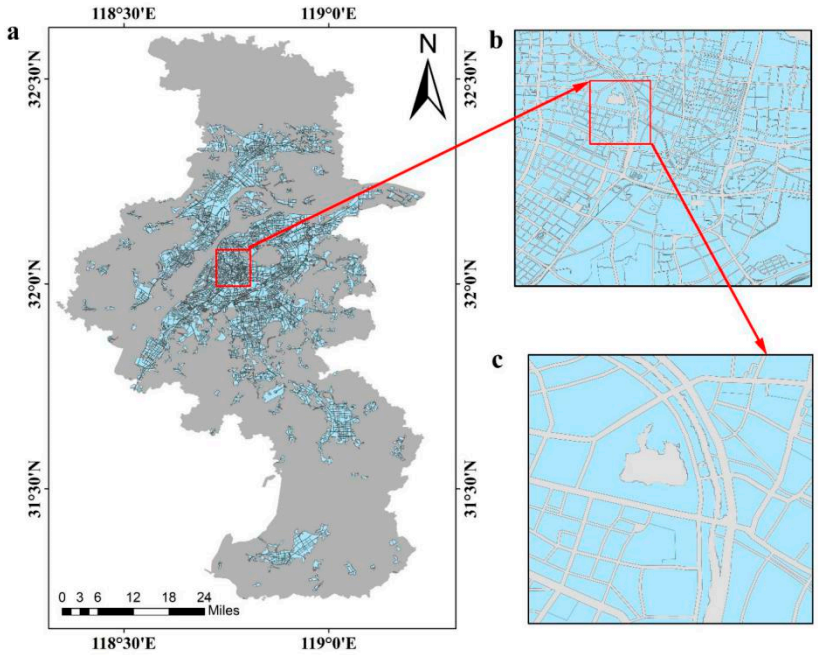

Figure 4. Urban land parcels of Nanjing. (**a**) Overall pattern, (**b**) and (**c**) are detailed zoomed-in views of urban land parcels.

Table 1. Buffer thresholds for different types of roads.

Level	Primary	Secondary	Tertiary	Residential	Motorway	Trunk	Railway
Buffer thresholds	44 m	34.8 m	30.4 m	21.5 m	42 m	60.5 m	7.7 m

2.3.3. POIs Processing

By overlaying a total record of 760,000 POI data with impervious surface polygons, the 220,000 POI points within the built-up area were retained. According to the EULUC classification system [12], the POI data were reclassified into four classes at Level I and nine classes at Level II. It should be noted that the transportation category was excluded in this study [12]. Statistically, the commercial, residential, public, and industrial POIs accounted for 52%, 27%, 19.8%, and 1.2% of the total numbers, respectively. Additionally, the POIs density in the downtown business district is higher than that in the surrounding areas. In order to solve the unbalanced distribution problem, these POI data were regenerated by the following three steps: (1) the commercial POI points with a distance less than 10 m were deleted, (2) the industrial and residential POIs were added according to the method proposed by Zhang et al. [16]. Firstly, each building abstracted from the building footprint data was regarded as the basic unit. For each unit, the geometrical, spectral, and textural features were extracted from the Google Earth high-resolution image. Next, the buildings overlapped with the POIs were labeled by the corresponding POI's Level I category. The labeled buildings were employed to train a random forest (RF) model for recognizing other non-labeled buildings. Finally, the centers of the building polygons identified as the residential and industrial categories were converted to POI points of the corresponding classes. (3) For the public category, POI points were added on the buildings by visual interpretation due to the relatively lower classification accuracy via the above RF classification method.

2.3.4. Features Extraction

From Remote Sensing Data

Eighteen spectral features and 26 texture features were extracted from Sentinel-2 composite imagery (Table 2) for each urban parcel. Spectral features included the mean and standard deviations of blue, green, red, near-infrared, short-wave-infrared bands, NDVI, normalized difference built-up index (NDBI), and normalized difference water index (NDWI). Texture features included the entropy, contrast, correlation of blue, green, red, near-infrared bands, and the entropy of normalized difference vegetation index (NDVI). For the Luojia-1 nighttime light image, we calculated the mean of digital number (DN) values for each urban parcel and applied it as a feature to urban land use classification (Table 2). The names of features are shown in Supplementary Table S1.

Table 2. Features extracted from multi-source big data. The figure in brackets is the number of features.

Data	Features
Sentinel-2 (44)	Spectral features (18): Mean and standard deviation of blue, green, red, near-infrared, short-wave-infrared bands (band11,12), NDVI, NDBI, and NDWI.
	Textural features (26): Entropy, contrast and correlation of blue, green, red, and near-infrared bands. Entropy of NDVI.
Luojia-1 (1)	Mean of digital number values.
POIs (29)	Frequency features (19): Total number of all POIs, total number and proportion of each type of POIs within each parcel.
	Spatial features (10): Spatial distribution of POIs.
MPL (48)	Mean of hourly active population during weekdays and weekends.
Building footprint data (12)	Total number of buildings, proportion of different grade buildings, average height of buildings in each parcel, floor area ratio, and height density index.

From POIs

There was a total of 19 frequency features and 10 spatial features extracted from the regenerated POIs (Table 2). The frequency features included the total number of all POIs, and the number and proportion of each POI class within each parcel. The spatial features were abstracted by the Google word2vec model, which is a deep-learning tool for transforming natural language words into high-dimensional spatial vectors [33]. In this study, we regarded the Nanjing built-up-area as a corpus, 8209 parcels as the documents, POI categories as the basic words, and the POIs spatial distribution within the parcels as the word sequences in the documents. Thus, the spatial distribution features of POIs can be quantified by the Word2vec model. The calculation of spatial features included three steps [20]: (1) the parcels and POIs were used to build a Parcel-POI corpus. In order to obtain an adequate number of words, we divided the generated POIs into 44 categories in Level III on the basis of Level II (Supplementary Table S2), (2) all POI category vectors were obtained using the continuous bag of words (CBOW) model in Word2Vec and the dimension value of 10 was set by the trial and error, and (3) the parcel vectors were calculated by averaging inside POI vectors with weightings. The 10-dimensional parcel vectors were the final spatial features we would use in the EULUC-Nanjing mapping. The results of the Parcel-POI corpus, POI category vectors, and the parcel vectors can be downloaded from https://pan.baidu.com/s/1jb9lOpqFcqNlVQZ4UrC8rA.

From the Building Footprint Data

Based on the number of floors, the buildings were divided into eight levels (Figure 5). For each parcel, the total number, each level's proportion, and the average height and floor area ratio (FAR) of all buildings were calculated. Considering that the business office building is higher but occupies a small area, and the commercial service building is lower but covers a large area, this study proposed a height density index (total height of buildings in parcel/parcel area), which would differentiate the difference between the commercial and business land uses. Twelve features were finally extracted from Nanjing building footprint data (Table 2).

From Mobile Phone Locating-Request Big Data

Given that people's daily activities generally change periodically on a weekly basis, and there is a certain difference in the population distribution between work and non-work days [19], the weekly Tencent MPL data were divided into two groups: working day and rest day. Kernel density analysis is commonly used in urban hot spot exploration [34]. Since the original data was a series of spatial points, the kernel density analysis was used to make the MPL data into population heat maps. Furthermore, to better reflect the law of population distribution and reduce the data error, the results of kernel density analysis at the same time on working days and rest days were averaged. Finally, the mean of the hourly population heat map value of working days and rest days was calculated from each parcel and 48 features were obtained (Table 2).

2.3.5. Urban Land Use Classification

The samples were collected by both visual interpretation and field investigation from the Google Earth [31] high-resolution image. The selected samples should be typical and stable with a low mixing of land uses. However, there are few samples with high purity, such as the medical and administrative categories, which are usually mixed with other land uses. Therefore, we edited these samples manually to ensure that the purity of each sample was more than 90%. For example, as shown in Figure 6, the parcel contained two land use types (educational and medical) before editing (Figure 6a), while after editing, it was divided into two parcels with high-purity (Figure 6b). Finally, a total of 680 samples were collected, according to a ratio about 3:1 [35], 500 samples (from 26 to 98 in Level II) were used for training and the remaining 180 samples (from 15 to 25 in Level II) were used for validation. The samples can be downloaded from https://pan.baidu.com/s/1jb9lOpqFcqNlVQZ4UrC8rA.

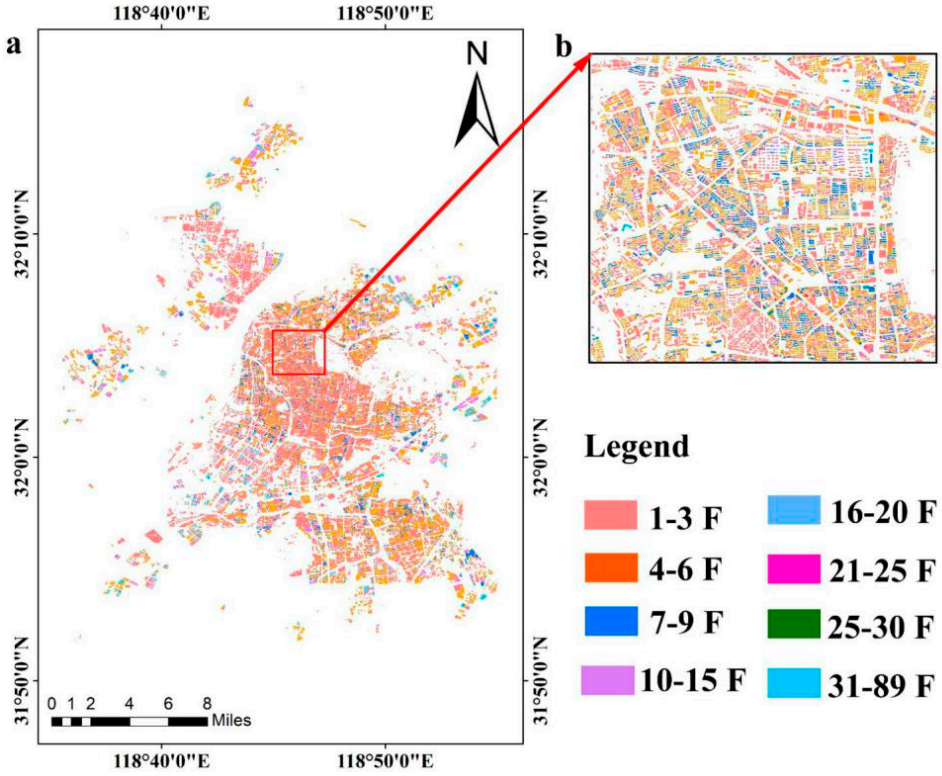

Figure 5. (**a**) The distribution of buildings with different height grades in Nanjing (the "F" in legend means floor). (**b**) A zoomed-in figure of the area in the red frame.

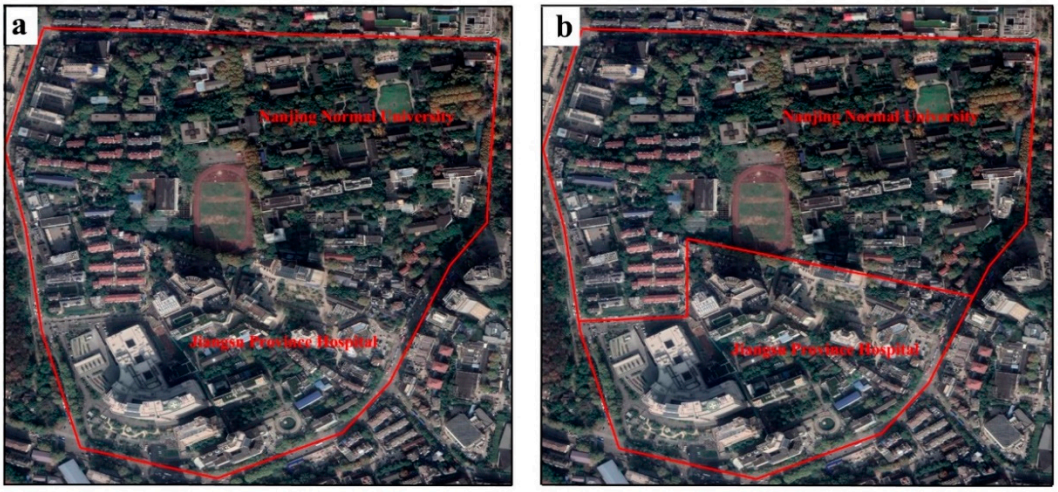

Figure 6. (**a**) The parcel before editing. (**b**) The parcel after editing.

The random forest (RF) model is an ensemble classifier that consists of multiple decision trees and has been extensively used in urban land use classification [36]. It has a higher advantage in processing high-dimensional data compared with other algorithms [37]. In this study, in order to improve the performance of the RF model, the optimal features were selected from the total of 134 features by using the recursive feature elimination algorithm with the help of the Caret package [38] in R. The selected features and 500 training samples were employed to train a RF model for identifying the land use types of the remaining parcels. The parameters of 'ntree' and 'mtry' for the RF model were set to 500 and 3, respectively. Finally, the 180 independent validation samples were adopted

to establish the confusion matrices for Level I and Level II categories, and the overall accuracy (OA) and kappa coefficient [39–41] were used to evaluate the classification result.

As for the built-up area, the mapping of Level I and Level II categories was directly used from the classification results. As for the non-built-up area, the FROM-GLC10 land cover map of Nanjing City was used [25]. Finally, the built-up and non-built-up areas were combined to produce a complete land use map.

3. Results

3.1. The Results of POI Regeneration

Compared with the original POIs, the number of regenerated POIs substantially increased (Table 3), especially for industrial, residential, and public POI points. Furthermore, the skewness distribution of the original POIs (Figure 7a) has been rectified to a relatively uniform distribution after being regenerated (Figure 7b).

Table 3. A comparison of original POIs and the regenerated POIs in number.

Level I	Level II	Original	Regenerated
01 Residential	0101 Residential	60,341	82,770
02 Commercial	0201 Business office	25,802	25,802
	0202 Commercial service	91,038	81,702
03 Industrial	0301 Industrial	2594	13,961
05 Public	0501 Administrative	10,524	17,142
	0502 Educational	9591	14,805
	0503 Medical	8049	11,498
	0504 Sport and cultural	5827	8434
	0505 Park and greenspace	2554	7685

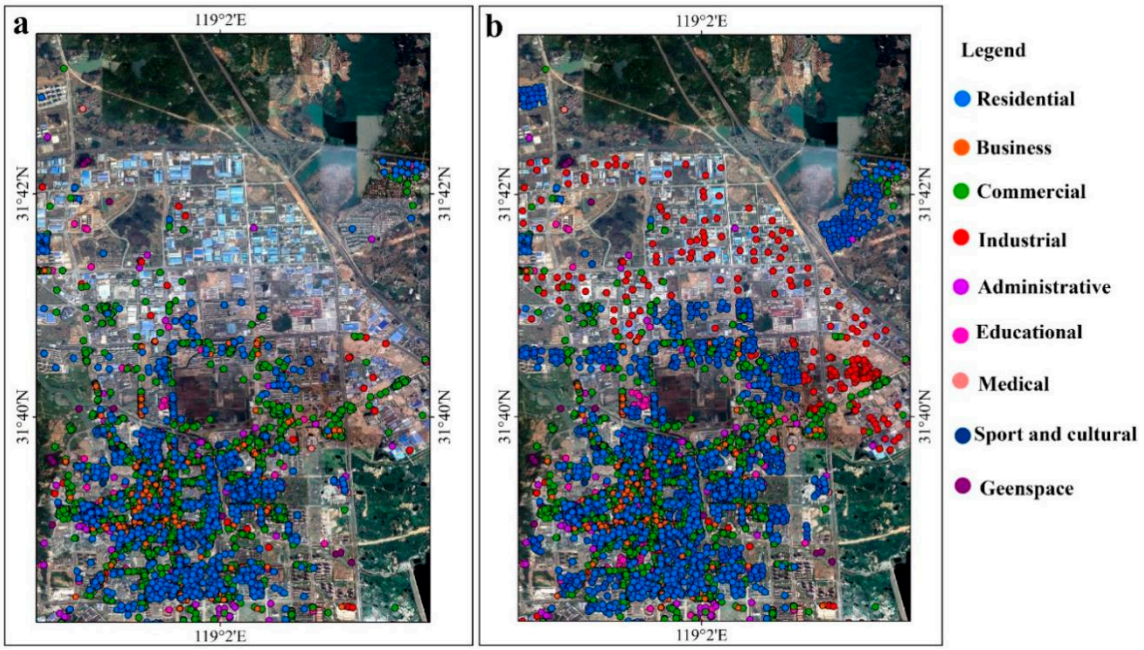

Figure 7. A spatial distribution of (**a**) the original POI points and (**b**) the regenerated POI points.

3.2. Feature Selection

The accuracy assessment based on OOB error indicated that the number of features, 68 in Level I and 61 in Level II, achieved the highest OA (Figure 8). The specific features selected are shown in Supplementary Tables S3 and S4. Table 4 shows that using the above 68 and 61 features, the classification OA (obtained through 180 independent validation samples) increased by 2.6% and 4% for Level I and Level II respectively, and the kappa coefficients both increased by 0.04.

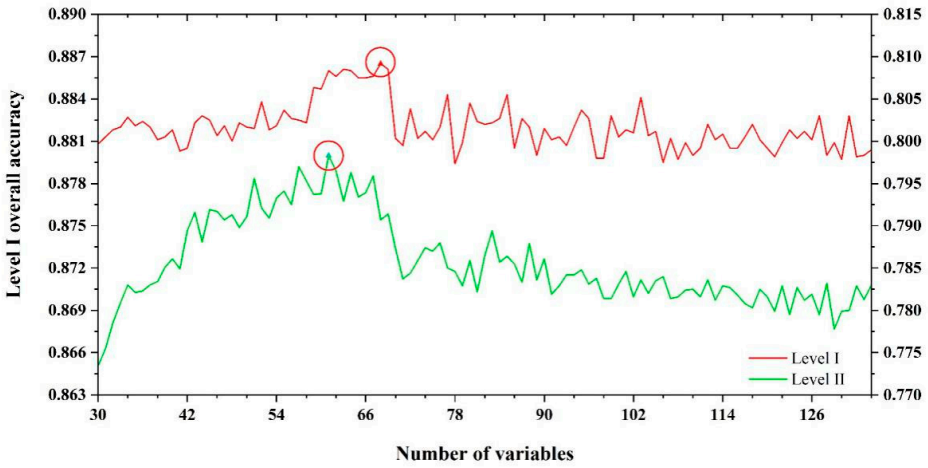

Figure 8. The classification overall accuracy with the varying number of features.

Table 4. The classification accuracy before and after feature selection.

No. of Features	Level I		Level II	
	OA	Kappa Coefficient	OA	Kappa Coefficient
134	83.5%	0.74	76%	0.73
68	86.1%	0.78	-	-
61	-	-	80%	0.77

3.3. Performance of EULUC-Nanjing

The confusion matrices of Level I (Table 5) and Level II (Table 6) showed that the OA and kappa coefficient were 86.1% and 0.78 respectively, in Level I, and 80% and 0.77 respectively, in Level II. The residential, industrial, and public land use in Level I achieved a higher accuracy with both producer's accuracy (PA) and user's accuracy (UA) of more than 80%, while the commercial land use had a relatively lower user's accuracy of 68%. As for the Level II category, the industrial land use had the highest accuracy with both producer's accuracy (PA) and user's accuracy (UA) of 92%, while the business land use had the lowest accuracy with UA of 60% and PA of 64%.

Table 5. Confusion matrix for Level I categories in EULUC-Nanjing.

			OA: 86.1%		Kappa Coefficient: 0.78		
Level I	**Residential**	**Commercial**	**Industrial**	**Public**	**Total**	**UA**	**PA**
Residential	20	1	0	4	25	80%	83%
Commercial	3	27	1	9	40	68%	87%
Industrial	0	0	22	3	25	88%	96%
Public	1	3	0	86	90	96%	84%
Total	24	31	23	102	180		

The feature importance plots indicated that seven POI spatial features (POIspa_2, POIspa_8, POIspa_6, POIspa_1, POIspa_9, POIspa_7, and POIspa_5), three POI frequency features (POIp201,

POIp101, and POI101), three texture features (b2cormean, b2entstd, and b3entstd), and two spectral features (ndvimean, b8mean) were identified as the top 15 important features (Figure 9a) in the Level I class. In the Level II class, six POI spatial features (POIspa_8, POIspa_1, POIspa_2, POIspa_7, POIspa_6, and POIspa_9), four POI frequency features (POIp502, POIp501, POI503, and POIp201), three texture features (b2entstd, b2cormean, and b3entmean), and two spectral features (ndvimean, b4mean) were selected in the top 15 (Figure 9b). Compared with spectral and texture features extracted from Sentinel-2 imagery, POI features, especially for the POI spatial features, were more important in the EULUC-Nanjing mapping.

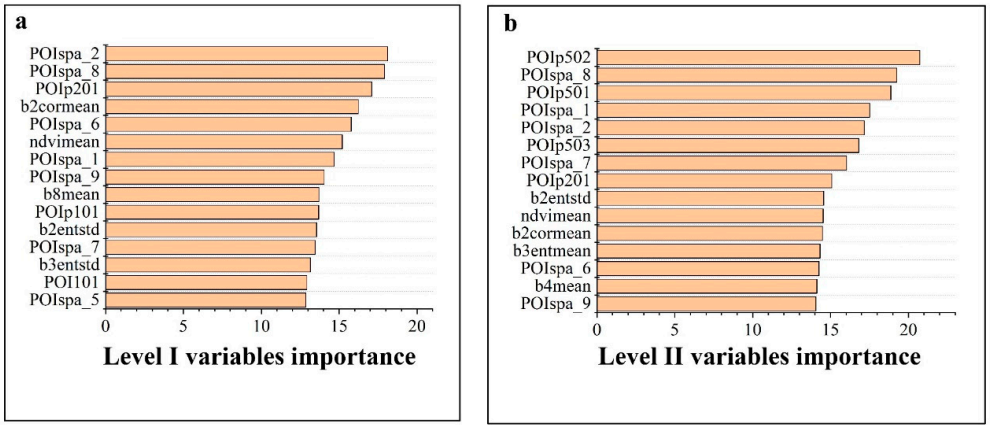

Figure 9. (**a**) The feature importance rank in the Level I classification framework. (**b**) The feature importance rank in the Level II classification framework.

Figure 10 shows the mapping results of EULUC-Nanjing. Within the 1624 km^2 built-up area of Nanjing, the public service, residential, and industrial land accounted for 37.6% (610.9 km^2), 23.2% (378 km^2), and 22.7% (370.1 km^2) respectively, while the transportation and commercial land only covered 13.3% (216.7 km^2) and 2.9% (48.6 km^2).

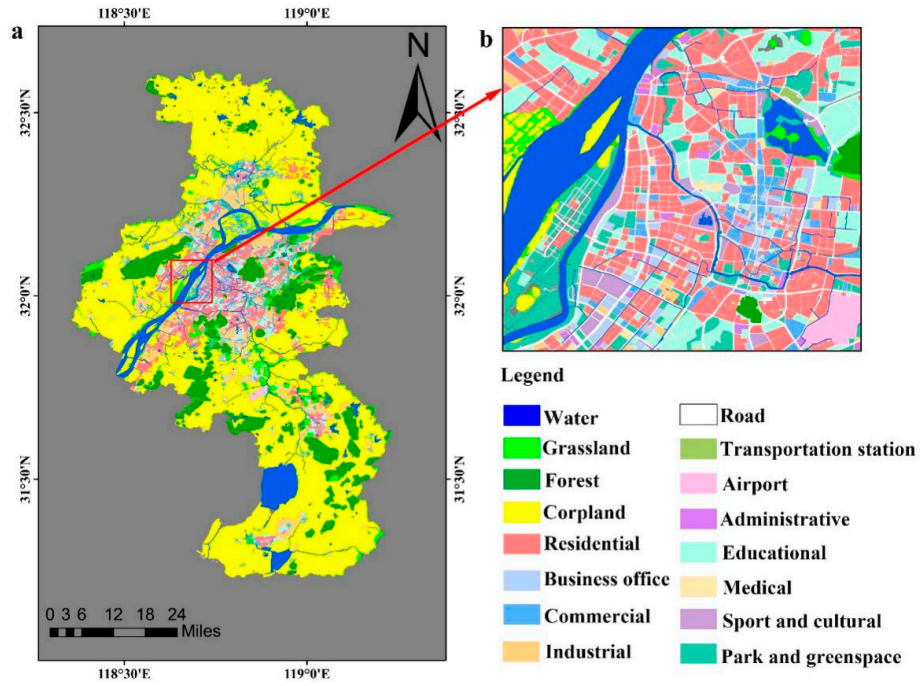

Figure 10. (**a**) The map of EULUC-Nanjing in 2018. (**b**) Zoomed-in area of the red frame.

Table 6. Confusion matrix for Level II categories in EULUC-Nanjing.

					OA: 80%		**Kappa Coefficient: 0.77**					
	Residential	**Business**	**Commercial**	**Industrial**	**Administrative**	**Educational**	**Medical**	**Sport**	**Greenspace**	**Total**	**UA**	**PA**
Residential	22	1	1	0	0	1	0	0	0	25	88%	81%
Business	1	9	3	1	1	0	0	0	0	15	60%	64%
Commercial	0	2	20	0	1	1	0	0	1	25	80%	71%
Industrial	1	0	0	23	0	1	0	0	0	25	92%	92%
Administrative	0	0	2	0	12	0	0	1	0	15	80%	75%
Educational	1	0	0	1	0	20	1	1	1	25	80%	76%
Medical	0	1	2	0	1	0	11	0	0	15	73%	91%
Sport	1	1	0	0	0	0	0	12	1	15	80%	86%
Greenspace	1	0	0	0	1	3	0	0	15	20	75%	83%
Total	27	14	28	25	16	26	12	14	18	180		

4. Discussion

4.1. Contribution of Different Features

In order to investigate the role of different features in urban land use classification, the features from five data sources in Table 2 were removed in turn and the features from the remaining four data sources were used as input variables in the RF model. From the changes of the OA, the influence degree of the features from different data sources on Level II categories can be found. The result (Table 7) showed that POI data had a great impact on the business, commercial, educational, administrative, medical, and sports culture categories. In particular, the influences on the latter three land uses were the most obvious. The features derived from Sentinel-2 imagery had more influences on residential, industrial, and greenspace land. The height features had an influence on the business and commercial land. In addition, the mean DN values from Luojia-1 data were also helpful to distinguish commercial and business land from other land uses. This may be because the night light intensity in the commercial districts was relatively higher than that in other places.

Table 7. Changes of the OA in Level II categories in EULUC-Nanjing after removing features of a certain type. (The values with large changes are shown in bold.)

	0101	**0201**	**0202**	**0301**	**0501**	**0502**	**0503**	**0504**	**0505**
All features (134)	88%	67%	72%	96%	62.5%	75%	73%	67%	75%
POI	88%	**46%**	**52%**	96%	**15%**	**58%**	**33%**	**35%**	70%
Building height	88%	**62%**	**58.9%**	96%	62.5%	75%	73%	67%	70%
MPL	88%	67%	76%	92%	68%	75%	73%	67%	75%
Luojia-1	88%	**55%**	**67%**	96%	62.5%	71%	67%	67%	75%
Sentinel-2	**80%**	64.9%	68%	**86%**	62.5%	71%	73%	67%	**60%**

In this study, it can be found that the POI features had the greatest contributions in the EULUC-Nanjing mapping (Table 7 and Figure 9). However, Tu et al. [26] pointed out that compared with POI data, the features obtained from Sentinel-2 imagery were the main factors affecting the classification performance. This may be because this study used the regenerated POI data by overcoming the problem of unbalanced distribution and extracted the POI spatial features. Therefore, the POI spatial features derived from the regenerated POIs are expected to better inform the regional to continental mappings of urban land uses.

4.2. The Impact of the Sample Sizes

In order to explore whether 500 samples can meet the classification requirement, the number of training samples was gradually increased by 10% each time and the variation of classification accuracy was observed. Figure 11 shows that when the number of samples reaches 70% (350), the accuracy for both the Level I and Level II categories tends to be stable, which is consistent with the stable classification concept proposed by Gong et al. [25]. Therefore, 500 training samples used in this study have met the number of samples required for the maximum classification accuracy.

Since there are some small water bodies within the impervious surface polygons, the water layer from FROM-GLC10 was used as a mask to remove these water areas. However, there were still water bodies within segmented urban land parcels. This is due to the limited spatial resolution of Landsat images being used to extract impervious surface, and Sentinel-2 images being applied for FROM-GLC10 generation. The water bodies of less than 100 m^2 or rivers with a width of less than 10 m cannot be removed. Therefore, high-spatial-resolution images will be considered to improve the parcels' purity in the future work. In addition, although we have tried to use high-purity samples to train the RF model, actually, the land use parcels are often mixed [42], and even a single building contains multiple functions. Liu et al. [43] have found that 18.82% of the buildings in Tianhe district of Guangzhou had mixed functions; for example, some buildings had residential and commercial

functions, while others had residential, leisure, and official functions. The existence of multi-functional buildings makes it hard to determine a single land use type in one urban land parcel. The results of EULUC-China showed that the Level II classification OA in Beijing, Hong Kong, and Shenzhen city were only about 43%, 44%, and 51%, respectively. The low accuracy may be due to the highly mixed land use functions within one building in these cities. Although the OAs of EULUC-Nanjing have been greatly improved (from 75.0% and 75.0% to 86.1% and 80% in Level I and Level II categories, respectively), there was still considerable confusion between the business and the commercial land use (Table 6). By overlapping the object parcels segmented from the high-resolution images with the POI data, Zhang et al. [16] found that the feature of the object categories contributed the most to the extraction of urban functional zones. This enlightens us to use the object parcels obtained from high-resolution images as the basic analysis unit in the future to reduce the possibility of mixed land use.

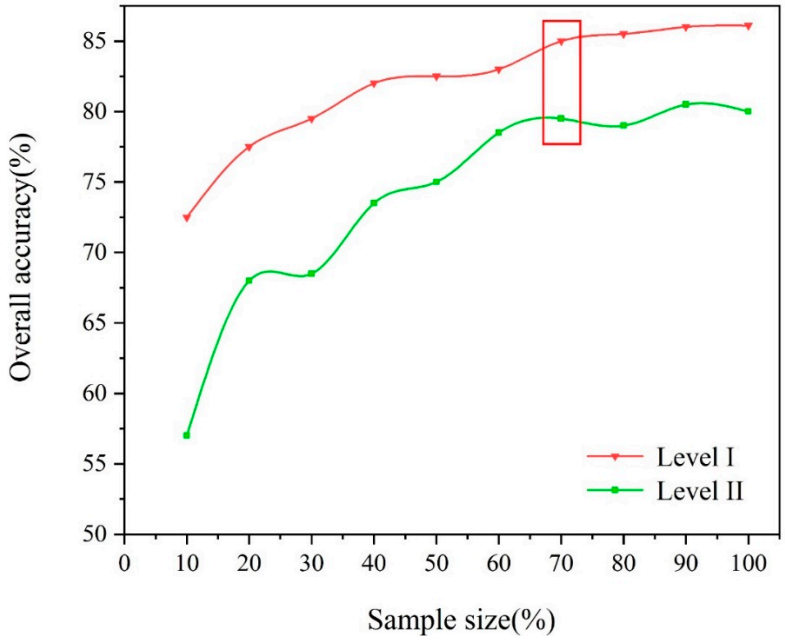

Figure 11. The variation of urban land use classification accuracy with the training sample size.

In this study, it can be found that the POI spatial features made the greatest contributions to the EULUC-Nanjing (Figure 9). Due to the large area of the urban parcels, there are different types of POI data in a parcel. The POI points have the coordinates but no geometric attribute information, such as the area, so it is difficult to determine the weight of the different POI types. At present, this study only adopted the same weight for different POI types in a parcel to calculate the POI vectors. If the segmented objects can be obtained, it could apply the POI category to mark the attributes of these objects and use the object area as the weight to construct a more accurate parcel vector.

Compared with previous studies [12], this study refined the spatial and temporal resolution of MPL data; however, the results showed that the contribution of this kind of data in classification was the lowest (Table 6). Chen et al. [44] used the MPL data and a dynamic time warping (DTW) distance-based k-medoids method to aggregate those buildings with similar social activities into functional areas with a clustering accuracy rate of up to 85%. Therefore, the dynamic change of MPL data in a parcel can be used as a feature in the regional or global urban land use classification. Furthermore, with the increasing of mobile phone users, mobile phone positioning data also can be used to obtain a large amount of the dynamic trajectory data. Limited to the density of the base station, mobile phone positioning data has a lower resolution of 300–500 m, while the MPL data has a relatively higher spatial resolution of 25 m. The number of the Tencent users is relatively smaller but the number

of mobile phone users is larger. Therefore, fusing the MPL data and mobile phone positioning data would be feasible and effective in the national or global urban land use mapping.

5. Conclusions

Based on the concept of EULUC-China, this study proposed a more consolidated framework of urban land use mapping at a regional scale. We refined the generation of the urban parcels, resolved the problem of unbalanced distribution of POIs, extracted the POI spatial features, and mapped the EULUC-Nanjing. The results showed that: (1) compared with EULUC-China, the classification OA in Level I and Level II were improved by 11.1% and 5% respectively, (2) the spatial features from POI data were identified to be the most important variables, and (3) the classification accuracy tended to be stable when the number of training samples reached 350. This study recommended that the POI data should be preprocessed before they were used and the spatial features from POIs cannot be overlooked in the national or global urban land use mapping, especially in cities with a lot of tall buildings, and there are multiple land use functions in a single building, such as Hong Kong, Shen Zhen, etc. Also, the urban land use information provided in this study can be applied to help urban planners monitor urban land use changes, analyze urban structures [45], and make scientific and reasonable planning for the existing urban land resources, so as to promote the healthy development of the city. In addition, this study also had some limitations. For example, the extraction of urban parcels needed to be further refined, and the problem of mixed land use was not considered. Different types of POIs used the same weights when constructing parcel vectors, and the land use information of MPL data was not deeply mined. In the future work, for a multi-scale analysis unit, such as a single building, the objects or the parcels segmented from the high-spatial-resolution images will be considered in the urban land use. Attempts will be made to address the weighting problem of different types of POIs. Meanwhile, the MPL data and mobile phone positioning data can be fused to derive dynamic trajectory data with high spatial and temporal resolutions to better uncover land use functions in other cities.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-4292/12/15/2386/s1: Table S1: The name of all 134 features, Table S2: The Level III categories for POI data, Table S3: The selected features of Level I categories, Table S4: The selected features of Level II categories.

Author Contributions: Conceptualization, J.S.; methodology, J.S. and H.W.; formal analysis, J.S.; investigation, J.S., Z.S., J.L., P.M. and S.Q.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, H.W.; visualization, J.S.; supervision, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the National Natural Science Foundation of China (No. 41471419 and No. 31971579).

Acknowledgments: We express our appreciation to Peng Gong and Bin Chen for their constructive comments. The research was also supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Project Code: KYCX20_0520) and the Fundamental Research Funds for the Central Universities.

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

References

1. The World Bank. Available online: https://www.worldbank.org/en/topic/urbandevelopment/overview (accessed on 16 July 2020).
2. United Nations. Available online: https://www.un.org/en/climatechange/cities-pollution.shtml (accessed on 16 July 2020).
3. Arsanjani, J.J.; Helbich, M.; Bakillah, M.; Hagenauer, J.; Zipf, A. Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2264–2278. [CrossRef]
4. Fisher, P. The pixel: A snare and a delusion. *Int. J. Remote Sens.* **1997**, *18*, 679–685. [CrossRef]
5. Lu, D.; Weng, Q. Use of impervious surface in urban land-use classification. *Remote Sens. Environ.* **2006**, *102*, 146–160. [CrossRef]
6. Shaban, M.; Dikshit, O. Improvement of classification in urban areas by the use of textural features: The case study of Lucknow city, Uttar Pradesh. *Int. J. Remote Sens.* **2001**, *22*, 565–593. [CrossRef]

7. Wu, S.-S.; Qiu, X.; Usery, E.L.; Wang, L. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. *Ann. Assoc. Am. Geogr.* **2009**, *99*, 76–98. [CrossRef]
8. Voltersen, M.; Berger, C.; Hese, S.; Schmullius, C. Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level. *Remote Sens. Environ.* **2014**, *154*, 192–201. [CrossRef]
9. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [CrossRef]
10. Yuan, J.; Zheng, Y.; Xie, X. 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, Beijing, China, 12–16 August 2012; pp. 186–194.
11. 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]
12. 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]
13. Zhang, X.; Li, P.; Cai, C. Regional urban extent extraction using multi-sensor data and one-class classification. *Remote Sens.* **2015**, *7*, 7671–7694. [CrossRef]
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. Li, X.; Zhao, L.; Li, D.; Xu, H. Mapping Urban Extent Using Luojia 1-01 Nighttime Light Imagery. *Sensors* **2018**, *18*, 3665. [CrossRef] [PubMed]
16. Zhang, X.; Du, S.; Wang, Q. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRSJ. Photogramm. Remote Sens.* **2017**, *132*, 170–184. [CrossRef]
17. Chen, B.; Song, Y.; Huang, B.; Xu, B. A novel method to extract urban human settlements by integrating remote sensing and mobile phone locations. *Sci. Remote Sens.* **2020**, 100003. [CrossRef]
18. Chen, B.; Song, Y.; Jiang, T.; Chen, Z.; Huang, B.; Xu, B. Real-Time Estimation of Population Exposure to PM2.5 Using Mobile- and Station-Based Big Data. *Int. J. Env. Res. Public Health* **2018**, *15*, 573. [CrossRef]
19. Niu, N.; Liu, X.; Jin, H.; Ye, X.; Liu, Y.; Li, X.; Chen, Y.; Li, S. Integrating multi-source big data to infer building functions. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1871–1890. [CrossRef]
20. Tang, J.; Liu, F.; Wang, Y.; Wang, H. Uncovering urban human mobility from large scale taxi GPS data. *Phys. A Stat. Mech. Its Appl.* **2015**, *438*, 140–153. [CrossRef]
21. Hobel, H.; Abdalla, A.; Fogliaroni, P.; Frank, A.U. A semantic region growing algorithm: Extraction of urban settings. In *AGILE 2015*; Springer: Cham, Switzerland, 2015; pp. 19–33.
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.* **2016**, *31*, 825–848. [CrossRef]
23. Ge, P.; He, J.; Zhang, S.; Zhang, L.; She, J. An Integrated Framework Combining Multiple Human Activity Features for Land Use Classification. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 90. [CrossRef]
24. Gong, P.; Li, X.; Zhang, W. 40-Year (1978–2017) human settlement changes in China reflected by impervious surfaces from satellite remote sensing. *Sci. Bull.* **2019**, *64*, 756–763. [CrossRef]
25. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y.; et al. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef]
26. Tu, Y.; Chen, B.; Zhang, T. Regional Mapping of Essential Urban Land Use Categories in China: A Segmentation-Based Approach. *Remote Sens.* **2020**, *12*, 1058. [CrossRef]
27. Liu, S.; Qi, Z.; Li, X.; Yeh, A.G.-O. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and sar data. *Remote Sens.* **2019**, *11*, 690. [CrossRef]
28. Gong, P.; Marceau, D.J.; Howarth, P.J. A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. *Remote Sens. Environ.* **1992**, *40*, 137–151. [CrossRef]
29. Nanjing Municipal Bureau Statistics. Available online: http://221.226.86.104/file/2018/renkou/3-8.htm (accessed on 16 July 2020).
30. Lisle, R.J. Google Earth: A new geological resource. *Geol. Today* **2006**, *22*, 29–32. [CrossRef]

31. Li, X.; Gong, P. An "exclusion-inclusion" framework for extracting human settlements in rapidly developing regions of China from Landsat images. *Remote Sens. Environ.* **2016**, *186*, 286–296. [CrossRef]
32. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
33. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
34. Duan, Y.; Liu, Y.; Liu, X.; Wang, H. Identification of polycentric urban structure of central Chongqing using points of interest big data. *J. Nat. Resour.* **2018**, *33*, 70–82.
35. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI* **1995**, *14*, 1137–1145.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. 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]
38. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
39. Alberg, A.J.; Park, J.W.; Hager, B.W. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J. Gen. Intern. Med.* **2004**, *19*, 460–465. [CrossRef]
40. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
41. Kraemer, H.C. Kappa coefficient. *Wiley Statsref Stat. Ref. Online* **2014**, 1–4.
42. Wang, Y.P.; Wang, Y.; Wu, J. Urbanization and informal development in China: Urban villages in Shenzhen. *Int. J. Urban Reg. Res.* **2009**, *33*, 957–973. [CrossRef]
43. Liu, X.; Niu, N.; Liu, X. Characterizing mixed-use buildings based on multi-source big data. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 738–756. [CrossRef]
44. Chen, Y.; Liu, X.; Li, X.; Liu, X.; Yao, Y.; Hu, G.; Xu, X.; Pei, F. Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landsc. Urban Plan.* **2017**, *160*, 48–60. [CrossRef]
45. 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]

© 2020 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 (http://creativecommons.org/licenses/by/4.0/).