International Journal of Geo-Information

MDPI

Article

Refined Urban Functional Zone Mapping by Integrating Open-Source Data

Yue Deng [1,2] and Rixing He [1,2,*]

[1] College of Resources Environment and Tourism, Capital Normal University, Beijing 100048, China; dengyue@cnu.edu.cn
[2] Key Laboratory of Three-Dimensional Information Acquisition and Application, Ministry of Education, Capital Normal University, Beijing 100048, China
* Correspondence: herixing@cnu.edu.cn

Abstract: The determination of a reasonable spatial analysis unit is an essential step in urban functional zone (UFZ) division, which significantly affects the results. However, most studies on the division of functional zones are based on excessively large spatial units, such as blocks or traffic analysis zones (TAZs), which easily overlook the detailed characteristics of urban regions and introduce bias to the research conclusion. To address this issue, a refined zone segmentation method, namely, the Voronoi diagram for the polygon method, was proposed to generate refined spatial analysis units. Afterward, the functional topics of the spatial analysis unit were classified by a multiclass support vector machine (SVM) to produce the final UFZ map, where the functional topics of each spatial unit were obtained by coupling latent Dirichlet allocation (LDA). To verify the effectiveness of the proposed method, experiments were conducted in Beijing, China. The results indicated that the proposed segmentation method can generate fine-scale spatial units and provide fine-grained and higher accuracy UFZs (overall accuracy = 84%; kappa = 0.82).

Keywords: urban functional zone division; Voronoi diagram; LDA topic model; multiclass SVM classifier

check for updates

Citation: Deng, Y.; He, R. Refined Urban Functional Zone Mapping by Integrating Open-Source Data. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 421. https://doi.org/10.3390/ijgi11080421

Academic Editors: Wolfgang Kainz and Mingshu Wang

Received: 5 May 2022
Accepted: 23 July 2022
Published: 25 July 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 development promotes the gradual formation of functional zones such as residential zones, commercial zones and industrial zones within it to meet people's diverse social and economic demands [1]. As an extremely important geospatial feature of a city [2], the urban functional zone (UFZ) not only describes the distribution of the natural-physical environment (e.g., buildings and public facilities) of the city but also describes the socioeconomic activities (e.g., city vitality and nighttime economy) of humans [3,4]. The UFZ distribution pattern changes with city development, and UFZ formation can be affected by government planning, social production, and the lives of residents [1]. Fine-scale, accurate maps of intraurban functional zones play an important role in many fields, such as urban planning and construction, urban renewal, urban management, environmental monitoring, and epidemic investigation and prevention [5]. However, many countries lack recent and detailed UFZ distribution maps due to rapid urban development and the lack of effective urban geographic data [1]. Particularly, in the later stages of the urbanization process, urban functional zones generally do not undergo large-scale renewal, and urban planners do not replan functional zones on a whole city or block-by-block basis. Instead, local functional construction and adjustment are often carried out with buildings as the main body. Therefore, this research proposes a refined functional zone mapping method with buildings as the basic analysis unit by integrating open-source data such as buildings and POIs.

The remainder of this paper is organized as follows. Section 2 introduces the related works about urban function classification and zone segmentation. Section 3 describes the study area and data source. Section 4 introduces the workflow and the key components

of the proposed framework. Section 5 describes the experimental results, including zone segmentation results, topic modeling results, and classification results. Finally, we discuss several important parameters for UFZ extraction and draw conclusions in Section 6.

2. Related Works

2.1. Urban Function Classification

Since a UFZ is a heterogeneous urban zone composed of multiple objects, UFZ classification is different from image object classification. Land cover semantics and spatial patterns are basic cues to identify functional-zone categories. For example, a residential zone may contain residential buildings, public facilities, and roads. Therefore, a "bottom-up classification" method is needed to infer high-level functional zone using low-level land cover. Traditionally, remote sensing has been extensively applied to extract and analyze urban functional zones, which consider two-dimensional image features (such as spectral, geometrical, and textural attributes) [6], but the UFZ distribution pattern is strongly related to human socioeconomic activities [3,4,7]. Therefore, it is difficult to identify urban function use based on typical thematic features using pure remote sensing classification models [8]. To address this gap, geodemographic classification methods, which are derived from the social demographic data of urban regions, have also been developed for functional zone identification [9,10]. This method shows advantages in specific functional zones (such as workplace zones and residential zones) in small regions. However, data collection is labor intensive and costly and cannot meet the needs of rapid urbanization processes and development. A possible solution is jointly using multisource geospatial data such as OpenStreetMap (OSM) data, points of interest (POIs), smart card data, and taxi trajectories to provide rich social semantic features as a supplement to the natural properties of ground objects [11,12]. The existing research results indicate that the application accuracy of open-access social data to infer urban functions can reach approximately 70% [13], which is mainly attributed to the lack of reliable models [14]. It is difficult to quantify the relationship between the spatial distribution of open-access social data and urban function types. In addition, open-access social data are usually unbalanced and biased in regard to the category and spatial distribution, presenting challenges in data processing and interpretation.

Researchers have applied natural language processing (NLP) technology to the extraction of urban functions [15,16]. Yuan et al. inferred the functions within each region in a city with a topic-based inference model (discovers regions of different functions or DRoF) for the first time [17]. Gao et al. applied the latent Dirichlet allocation (LDA) topic model and incorporated check-in data acquired from social networks to determine urban functions [3]. Liu et al. integrated remote sensing images and social network data for land use delineation with a probabilistic topic model (LDA and probabilistic latent semantic analysis (PLSA)) [8]. The majority of the above modeling methods help to explore hidden semantics contained in POI data, but the model is typically based on the bag-of-words (BW) method, which ignores the spatial context of words [4]. To compensate for this deficiency, scholars have proposed neural network embedding models, such as word2vec [5], place2vec [15], and doc2vec [4]. The core idea of these models is to construct POI sequences with a spatial context through the shortest distance, sliding window or random point sampling method, thereby adopting the first law of geography as a criterion. In the modeling process, the frequency of data points and the spatial relationship between data points are considered. However, the above model suffers certain limitations in POI sequence construction. For example, the method based on the shortest path and random sampling only sorts POIs according to the Euclidean distance but does not consider the actual POI distribution in 3D space. Moreover, the sliding window method may lead to model overfitting due to oversampling.

2.2. Zone Segmentation

Previous analyses of UFZs mainly focused on functional classification but ignored zone segmentation. This approach is impractical because only when a reasonable spatial analysis unit is selected can multiple features derived from remote sensing images or social spatial data be linked to interpreting urban functions. Figure 1 shows several zone segmentation methods commonly adopted by researchers. As shown in Figure 1a), the grid-based segmentation method [14,18] does not consider the actual conditions of the ground structure distribution and function in the study area, and one surface feature may be divided into two grid cells. In addition, the grid-based zone segmentation methods are impacted by the scale effect, which means that different grid scales may generate different functional zone division results. In general, research on urban functions should be based on relatively stable basic spatial units over the corresponding historical period rather than randomly generated grids. As shown in Figure 1b), road-based segmentation methods [8,19] are easily affected by the completeness and quality of road data, and the selection process of different levels of road networks can also face the modifiable areal unit problem (MAUP) [20]. As shown in Figure 1c), large-scale block-based segmentation methods [21] blur internal differences in geographic functions. The update process of vector data in blocks experiences certain difficulties, which may cause conflicts between road networks and image data [22]. Moreover, large-scale regional segmentation methods (traffic analysis zones (TAZs) and blocks) can weaken or mask certain features represented by POIs, e.g., commercial buildings in residential areas. Therefore, in the study of urban functional area division, the option of a reasonable spatial analysis unit is a key scientific issue.

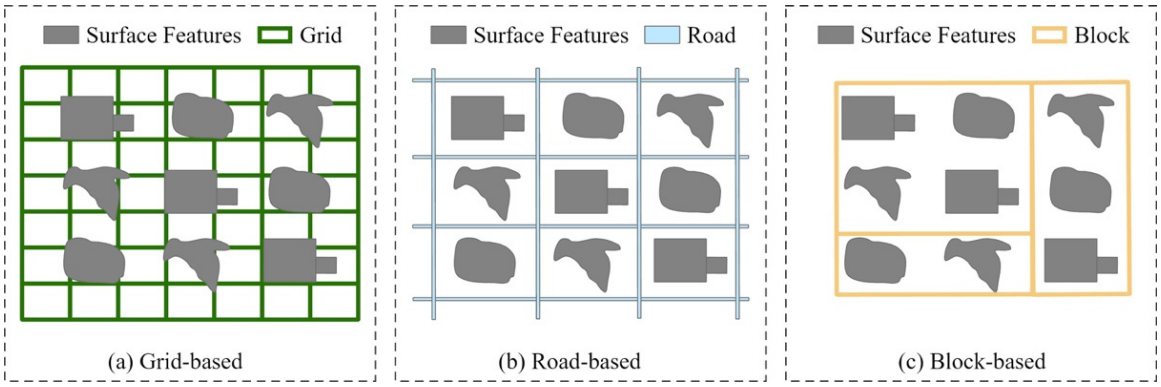


Figure 1. Examples of different zone segmentation methods.

In summary, the existing studies face two limitations. First, fine-grained UFZ mapping requires fine-mapping units, while TAZs and blocks cannot meet this demand well. Second, similar to how the context of words is considered in NLP when describing urban functions through the NLP technique, it is necessary to consider the spatial relationship of POIs rather than only the frequency. Most existing methods for POI spatial sequence construction are based only on the Euclidean distance but do not consider the actual POI distribution in space. Considering the challenges and issues of the current functional zone division methods, this paper aims to (1) propose a zone segmentation method to generate basic spatial units for fine-grained (building-level) UFZ mapping. (2) Construct categories of POI sequences based on zone segmentation. Moreover, the topic distribution is extracted for each zone with the LDA topic model. (3) The extracted topics are then input as semantic features into a multiclass support vector machine (SVM) classifier to obtain the final classification results.

3. Study Area and Data

3.1. Overview of the Study Area

The study area is located in Beijing. Since the implementation of the Reform and Opening-Up Policy in 1978, Beijing has experienced rapid urbanization with a growing gross domestic product (GDP) and population [23], thereby facing the problems of environmental pollution, housing shortages, traffic congestion, and unbalanced regional development. As the capital city of China and one of the largest cities worldwide, Beijing hosts a permanent population of 21.893 million people (as of November 2020) and covers an area of 16,400 km^2. Mountains cover an area of 10,072 km^2, 61% of the total area of Beijing. Most of the population is concentrated in the urban built-up areas occurring within the Fifth Ring Road, while the mountain areas are inaccessible. As the political and cultural center of China, Beijing must monitor the dynamic changes and development of its urban spatial structure and functional zones to accomplish smart and adaptive urban management. Therefore, we selected areas within the Fifth Ring Road as our case study area (within the red lines, as shown in Figure 2).

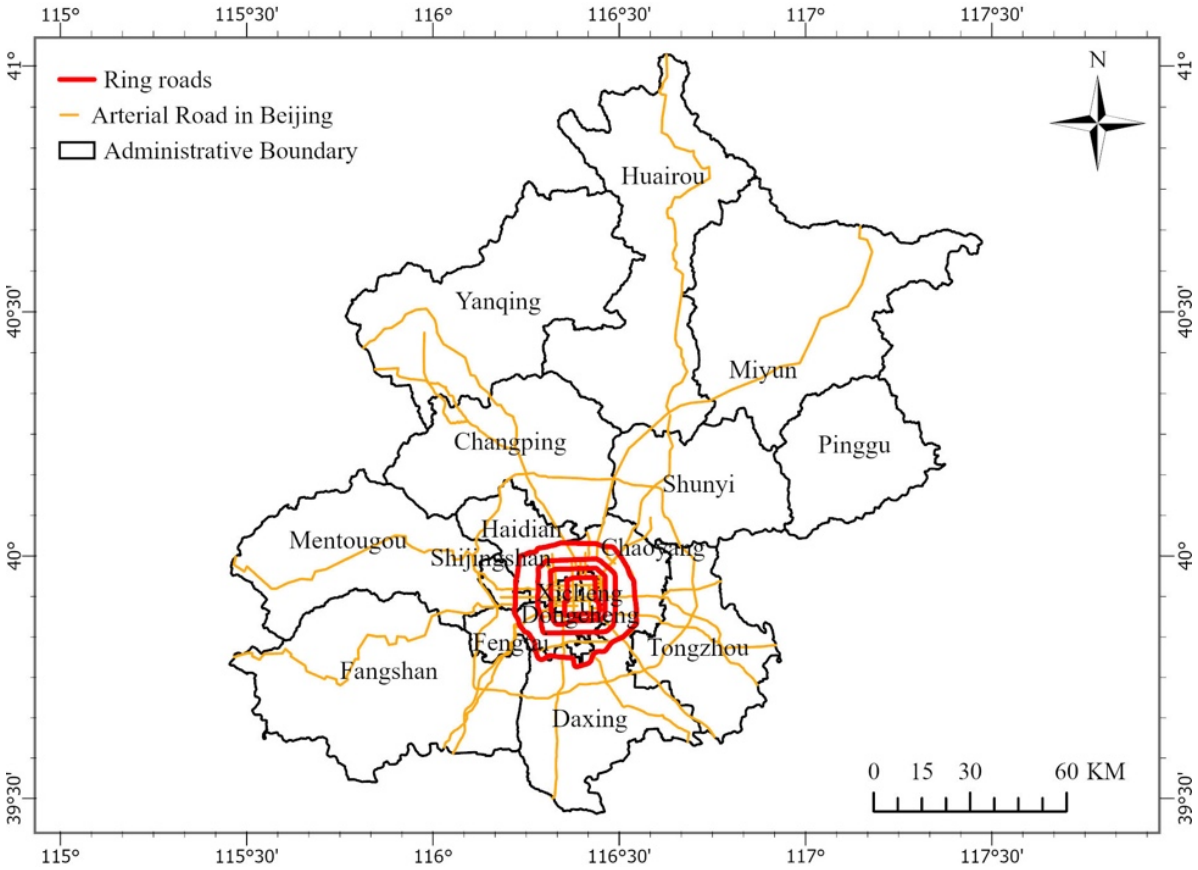


Figure 2. Case study area: Beijing (within the red lines).

3.2. Data

In the urban environment, POI data comprises a collection of location points such as supermarkets, cinemas, hospitals, hotels, and bus stations [4]. POI data can make socioeconomic features of UFZs available, define first-hand data on human activities and building uses, and provide researchers with the opportunity to explore the distribution of urban functions from bottom to top [3,24]. In contrast, urban building data can delineate the stable physical characteristics (such as building distribution features and 3D features) of a given city in great detail during a specific period. These two data sources can be complementary in urban land use representation and can realize finer and more accurate UFZ division.

3.2.1. POI Data

The POI dataset utilized in this research was derived from Gaode Maps (https://www.amap.com/, accessed on 20 July 2022), one of the largest online electronic map service websites in China (similar to Google Maps). This dataset was updated in April 2020, comprising 782,817 POIs within the Fifth Ring Road in Beijing. Each POI record exhibits many useful attributes, including name, coordinates (longitude and latitude), category, address detail (administrative zone and street name), and telephone number. Among these attributes, category attributes are closely related to urban functional areas, with a three-tier hierarchy, including 23 first-level categories, 267 second-level categories, and 903 third-level categories. This study excluded specific first-level categories (incidents and events, pass facilities, indoor facilities, and place name and address) unrelated to urban functions, yielding a total of 19 first-level categories. Figure 3 shows the relative frequency distribution of the total POIs. The relative frequency distribution of POIs follows Zipf's law, which is an important law in mathematical language. This reveals that a few POI categories account for a large proportion of all the POIs in Beijing. The five most frequent POI categories include shopping centers, daily life services, enterprises, food and beverages, and commercial houses.

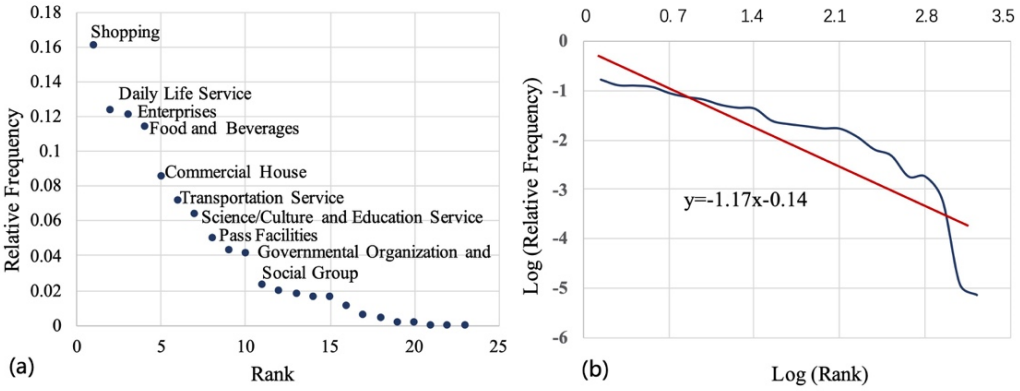


Figure 3. Rank-frequency plot for POIs. (**a**) Relative frequency plot. (**b**) Relative frequency log plot.

3.2.2. Building Data

The building dataset was created from Gaode Maps through open API (https://lbs.amap.com/demo/javascript-api/example/layers/buildings, accessed on 20 July 2022). This dataset includes 207,052 buildings within the Fifth Ring Road area of Beijing, of which the largest building covers an area of 195,448.77 m^2; the Beijing south railway station, and the smallest building covers an area of 34.48 m^2; located in Nan Luo Gu Alley (Figure 4). A total of 197,211 building surfaces were obtained by aggregating building surfaces and merging main and ancillary buildings within an adjacent distance smaller than 0.001 m. The building base map data mainly include physical attributes, such as the location, number of floors, and area, but these data lack social attributes such as function and use (offices, restaurants, etc.).

3.2.3. Urban Basic Geographic Data

Basic urban geographic data mostly includes Beijing green space data, water area data, and road network data. Relevant research has indicated [25,26] that the main land use type in the Fifth Ring Road area of Beijing is construction land, and large areas of green spaces and water bodies are also distributed. By the end of 2020, the per capita public green space area in Beijing reached 16.5 m^2. Therefore, green spaces and water areas should be included to improve the division accuracy while employing building data for zone division. The 2020 land use classification data considered in this study were downloaded from the Resource and Environmental Science and Data Center of the Chinese Academy of Sciences

(https://www.resdc.cn/, accessed on 20 July 2022). Green space and water area layers within the Fifth Ring Road area of Beijing were extracted.

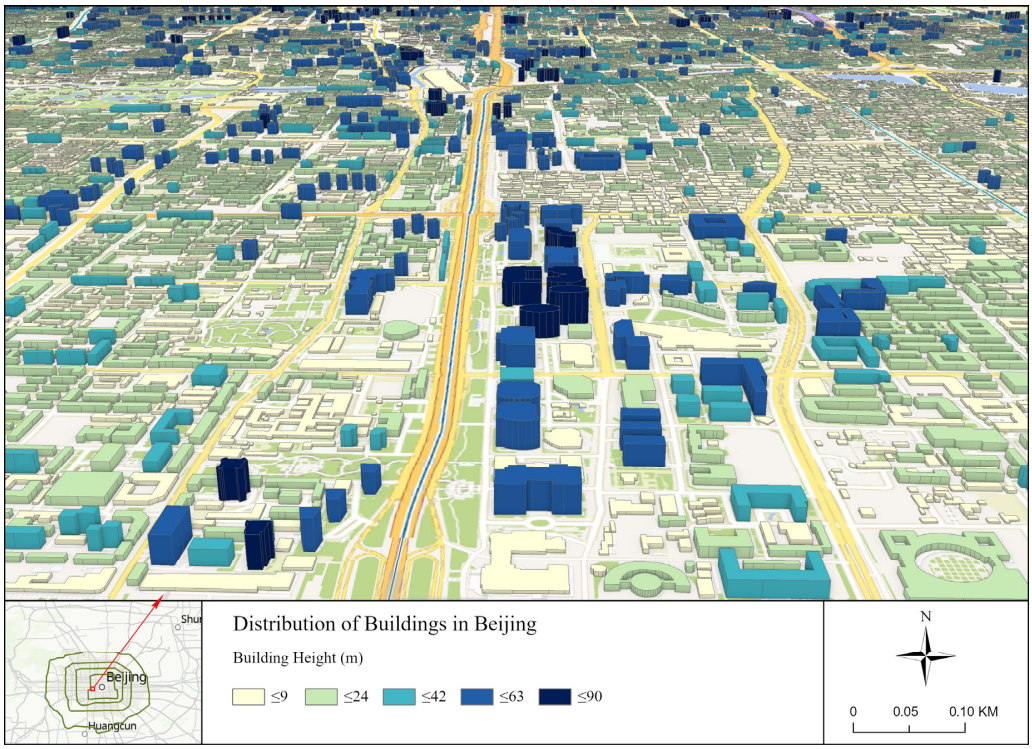


Figure 4. Example of building distribution in Beijing (the enlarged image shows Xi Dan).

4. Methodology

In this study, a method of integrating building and POI data was proposed for refined UFZ mapping. Generally, the urban functional zone is often naturally influenced by actual land use. The determination of the functional zone boundary and identification of its category are two critical issues in urban functional zone division. The former is to partition the functional zone on the spatial domain, and the latter is to classify the functional zone on the attribute domain. Specifically, the workflow of the proposed method comprises four steps (Figure 5):

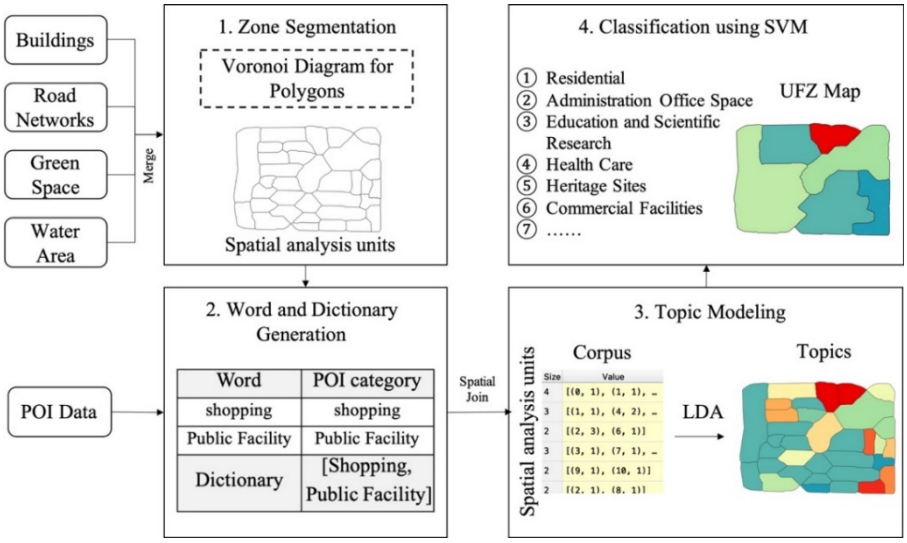


Figure 5. Methodology flowchart.

(1) Zone segmentation. The Voronoi diagram for the polygon method is employed to generate spatial analysis units by integrating building, green space, water area, and road network data.

(2) Word and dictionary generation. Each POI category is regarded as a single word. The dataset of all POI categories contained in a certain spatial analysis unit is regarded as a dictionary.

(3) Topic feature extraction. Each spatial analysis unit is regarded as a document, and the semantic features of each spatial analysis unit are mined with the LDA topic model to better classify UFZs.

(4) UFZ classification. The topics extracted from the LDA topic model are applied as semantic features. A multiclass SVM classifier is trained and applied to classify spatial analysis units into UFZs.

4.1. Zone Segmentation

In UFZ mapping, spatial analysis units should be determined first. The TAZ (traffic analysis zone) has been adopted as the spatial analysis unit in numerous studies involving UFZ division, urban land use classification, and scene classification [5,8,11,27]. However, highly mixed urban functions do exist in reality in each TAZ, and even a single building can contain different functions [28]. Buildings are a basic spatial component of urban areas, and their spatial structure creates a link between buildings and social functions. In theory, POIs within the same building exhibits a stronger spatial context than do POIs in different buildings. By dividing buildings into units, we can effectively capture the relationship between buildings in space. More accurately, space division assigns each building the meaning of functional influence on its surrounding spatial environment. Therefore, we propose a spatial segmentation method with buildings as basic anchors based on the Voronoi diagram to generate a more fine-grained spatial analysis unit.

4.1.1. Voronoi Diagram for Points

The division of spatial analysis units divides the research area according to specific rules. The core of the proposed zone segmentation method is the Voronoi diagram (Thiessen polygons), which is a space partitioning method that generates a series of polygons from a set of seeds (Figure 6). It's named after the famous Russian mathematician Georgy Voronoi. Any point in each polygon is the closest to the seed point in the polygon, so each polygon can be regarded as the influence area of the seed. Therefore, the zone segmentation method based on Thiessen polygons provides a more practical geographical significance than does the grid-based division method.

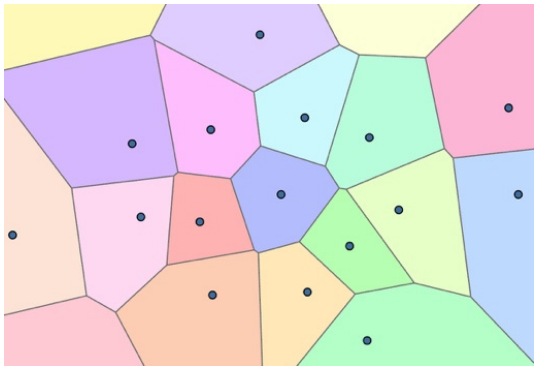


Figure 6. Thiessen polygons based on randomized points.

4.1.2. Voronoi Diagram for Polygons

The main advantage of the Voronoi diagram for the polygon method is that researchers can objectively determine the spatial partitions of urban buildings according to stable urban buildings during a certain period rather than through subjective interpretation. Conse-

quently, the Voronoi diagram for the polygon method is realized through the following four steps (Figure 7):

(1) Spatial feature merging. This step refers to the combination process in two cases. One case involves the combination of layers, which refers to the combination of various spatial feature layers (such as road, building, water area, and green space layers) to produce a complete feature layer of the urban space. The second case refers to the merging operation within a single layer, which entails the merging of similar adjacent features with an internal distance smaller than 0.001 m. It is proposed that similar adjacent features within a distance smaller than 0.001 m exhibit the same social functions. For example, the main and auxiliary buildings of a building complex (as shown in Figure 7a), B1 and B2, respectively, are combined into B, as shown in Figure 7b).

(2) Discretization of spatial features into points (Figure 7c). As the Voronoi diagram can only be generated from points, the spatial feature layer should be discretized into a series of points at certain intervals along its boundary.

(3) Generation of Voronoi diagrams around points (Figure 7d). Each diagram is associated with the original spatial features through the ID.

(4) Extraction of the segmentation boundary based on the ID of the original spatial features (Figure 7e).

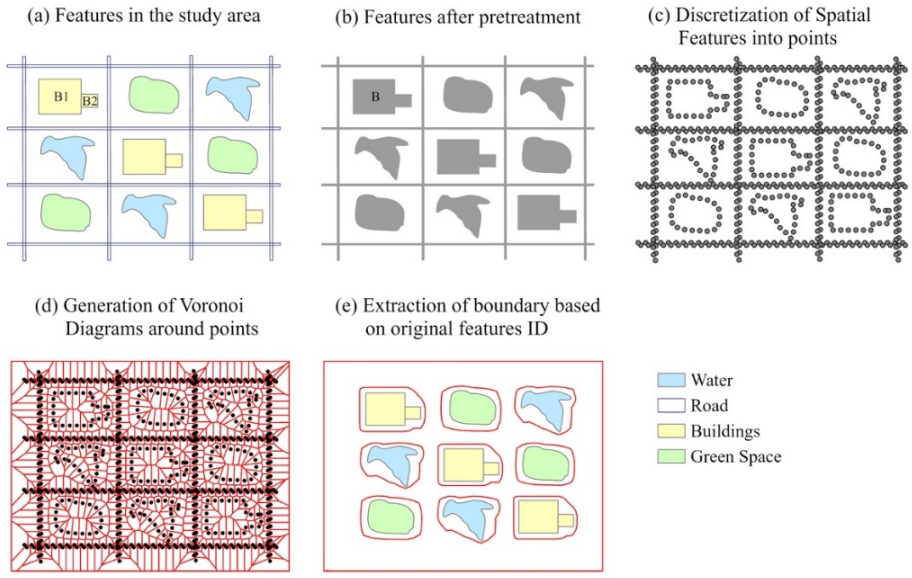


Figure 7. Voronoi diagram for the polygon method.

It is worth noting that the parameters that may have a significant effect on the results need to be evaluated in greater detail when using the Voronoi diagram for the polygon method for zone segmentation. More specifically, in the case of discretization of spatial features into points (Step 2), the selection of a discretization interval that is too large may produce unwanted "saw-like" geometry issues between adjacent buildings (Figure 8a). The most intuitive consequence of the phenomenon is that the line used to divide the zone will pass through spatial features such as buildings. As a result, the same building may be divided into multiple different areas, which further affects the division accuracy of UFZs. In contrast, the computational demand increases exponentially when a small discretization interval is selected. Due to the importance of correctly setting the parameter, Section 6.1 presents the test for determining the discretization interval. Based on this, the optimal parameter is adopted to generate the spatial analysis unit in later stages.

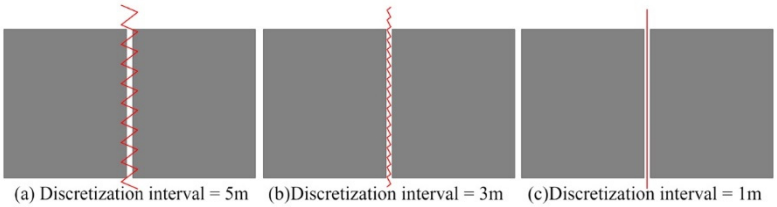


Figure 8. Illustration of the effect of the discretization interval causing "saw-like" geometry on the boundary between adjacent buildings (**a**) compared to the ideal combination (**b**,**c**).

4.2. Topic Feature Extraction

Similar to the distribution of words in natural language, the distribution of POI categories follows Zipf's law [15,16]. Therefore, the NLP-based LDA technique is suitable for urban function extraction [15,16]. The LDA model is a topic model employed to infer the topic distribution of documents. This model can obtain the topic of each document in the document set in the form of a probability distribution, and this model can conduct topic clustering or text classification according to the topic distribution [29]. In this study, the LDA model extracts the topic of each segmented unit by calculating the distribution characteristics of POIs, which provides a higher thematic feature for urban functional zones. Satisfactory results have been achieved in urban functional zone characterization [30]. In this method, each category of POI is regarded as a word, all POI categories comprise the dictionary, and each segmented spatial analysis unit is regarded as a document. The LDA variables in this study are defined in Table 1.

Table 1. Definition of the LDA variables in this study.

Variables	Definition
Word	POI category
Dictionary	All POI categories
Document	Set of POI categories in a spatial analysis unit
Topic	Function of the spatial analysis unit

The function distribution $\theta_i = \{\theta_{i,1}, \theta_{i,2}, \ldots \theta_{i,k}\}$ can be determined by integrating Equation (1) over θ and z.

$$\rho(\theta, z, w|\alpha, \beta) = \rho(\theta|\alpha) = \prod_{n=1}^{N} \rho(z_n|\theta)\rho(w_n|z_n, \beta), \quad (1)$$

where w denotes a set of categories of POIs, z denotes the inferred latent functions, and θ denotes the probability distribution of the functions in the spatial analysis units following Dirichlet allocation with parameter α. According to the LDA topic model, the function category of the spatial analysis unit and its probability distribution can be obtained.

4.3. UFZ Classification

There are many-to-many relationships between the topic distribution and the categories of urban functional zone for each segmentation unit. Manually determining the categories of urban functional zones to which each segmentation unit belongs leads to a large workload. Determining the categories of urban functional zones represented by each vector through a scientific classification algorithm is necessary. SVM is an algorithm that can be used to search hyperplanes in high-dimensional feature space to resolve linearly inseparable problems [31], which shows high precision and accuracy among classification algorithms. Therefore, after using LDA to obtain the topic distribution for each spatial analysis unit, an SVM classifier is applied to classify UFZs. The SVM classifier is implemented with the libsvm package (https://www.csie.ntu.edu.tw/~cjlin/libsvm/, accessed on 20 July 2022). The proportion of training and testing datasets is determined based on experiments [8,30]. Seventy percent of the selected UFZ samples were applied as the

training dataset, whose topics were input into a multiclass SVM classifier. The remaining 30% of the selected UFZ samples were employed as the testing dataset. The penalty C factor and kernel parameter γ are two finely tuned parameters in the SVM model. To determine the best parameter configuration, we applied 25% of the training dataset as a validation dataset, and we verified the accuracy through multiple combinations to obtain the optimal parameters. Then, we employed the training dataset to train the optimized SVM model and applied the optimized model to the test dataset for accuracy evaluation.

5. Experimental Results

The UFZ classification standard adopted in this study was formulated with reference to the code for the classification of urban land use and planning standards of developed land (GB 50137-2011, details of the standard can be found at https://sthjj.jiujiang.gov.cn/zwgk_215/zcjd/202006/P020200623535114893887.pdf, accessed on 20 July 2022). According to the land use characteristics of Beijing, combined with the classification system based on the primary and secondary categories of GB 50137-2011, we divided the UFZ into 5 primary categories and 10 secondary classifications. The classification system is listed in Table 2.

Table 2. The UFZ classification system.

Category 1	Category 2
Residential	Residential (Re [1])
Administration and Public Services	Administrative Office Space (Ad [1]) Education and Scientific Research (Ed [1]) Health Care (Hea [1]) Heritage (Her [1])
Commercial and Business Facilities	Commercial Facility (Co [1]) Business Facility (Bu [1]) Sport and Recreation Facility (Sp [1])
Road, Street and Transportation	Transportation Service (Tr [1])
Municipal Utilities	Public Facility (Pb [1])

[1] The abbreviation of Category 2.

5.1. Zone Segmentation Results

Based on the obtained building, green space, water area, and road network data, we segmented the study area into a series of spatial analysis units. A typical area (Peking University) was selected to demonstrate the effectiveness of the proposed zone segmentation method by comparison with the MT method [32] and road-based zone segmentation method [11] (Figure 9).

In the segmentation results, all of the buildings were completely segmented, thereby suitably retaining edges, especially in the districts with regularly distributed buildings (Figure 9a). Compared with the MT method that only uses buildings for zone segmentation, our method considers the distribution of roads, green spaces, and water to eliminate the phenomenon that the segmented zones may crossroads, green spaces, or water areas (Figure 9b A and B). Compared to the road-based segmented zones (Figure 9c), the spatial analysis units generated with our segmentation method are more appropriate for refined UFZ mapping, as a single spatial analysis unit may contain mixed functions due to the sparseness of roads, which is a major issue influencing the mapping accuracy [33]. Thus, the improved segmentation method performed better than the MT, and road-based segmentation methods did when considering roads, green spaces, and water areas. The segmented spatial analysis units in this study are expected to hold great potential for the task of refined urban functional mapping.

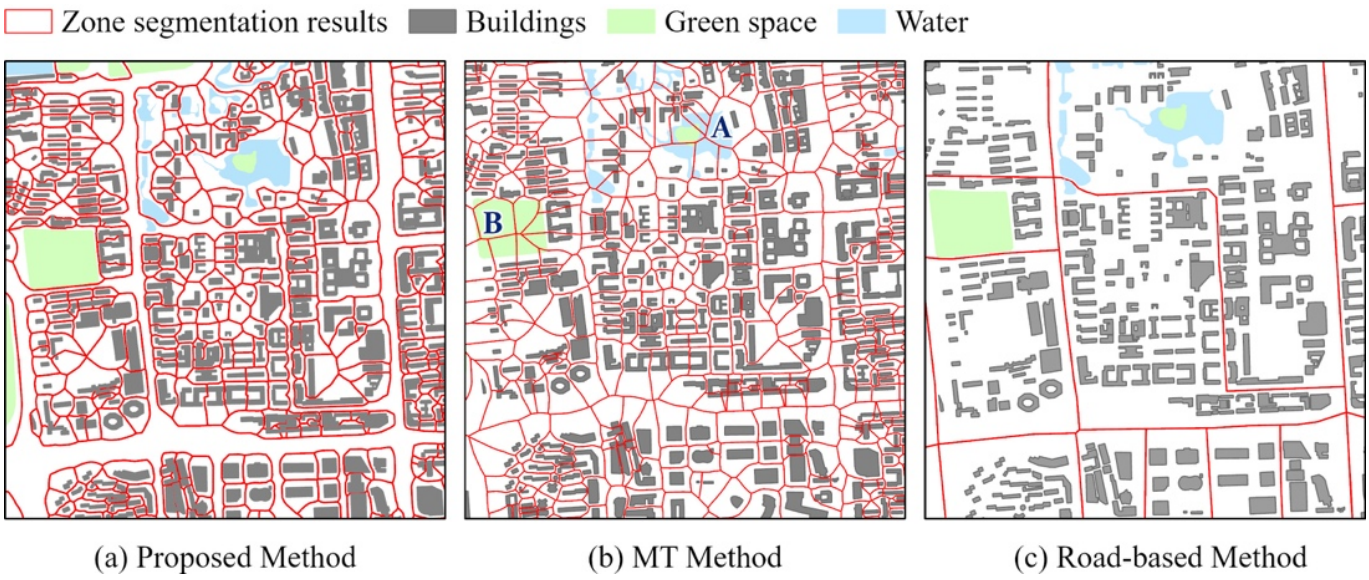


Figure 9. Zone segmentation results. (**a**) Proposed method. (**b**) MT method. (**c**) Road-based method.

5.2. Topic Modeling Results

Considering each spatial analysis unit as a document, we set the number of topics to 21 and the hyperparameter to 1.1 for topic modeling. Figure 10 shows the 21 topics and the word distribution probability of each topic. Here, each topic was represented by all POI categories considered. As shown in Figure 10, POI data and the LDA topic model can be employed to delineate the functional type of spatial analysis units. Our study area is located in Beijing, where highly mixed UFZs served as theme-mixed documents, which enabled the LDA model to yield accurate classification results.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Accommodation Service	0.73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Daily Life Service	0.11	0	0.01	0.03	0.06	0.03	0	0.91	0.14	0.05	0.02	0.01	0	0.01	0	0.14	0.25	0	0	0.28	0
Commercial House	0.07	0.23	0	0.02	0.03	0.08	0	0.01	0.08	0.04	0	0.01	0.04	0.21	1	0.01	0.1	0.01	0.32	0.31	0.01
Shopping	0.06	0	0.02	0.02	0.01	0	0	0.04	0.11	0.06	0.95	0.22	0	0	0	0.06	0.19	0	0	0.38	0
Food and Beverages	0.03	0	0.97	0	0	0	0	0.01	0.19	0.05	0.01	0.01	0	0	0	0.07	0.01	0	0	0	0
Transportation Service	0.01	0	0	0.01	0	0	0	0	0.1	0.13	0	0	0.03	0.78	0	0.02	0.2	0.04	0.24	0	0
Road Furniture	0	0.76	0	0	0	0.01	0.01	0.01	0.05	0.02	0	0	0.03	0	0	0	0.12	0.02	0	0	0
Enterprises	0	0	0	0.9	0	0	0	0.01	0.14	0.04	0.01	0.02	0	0	0	0.02	0.08	0	0.44	0.03	0
Medical Service	0	0	0	0	0.9	0	0	0	0.01	0	0	0	0	0	0	0	0.06	0	0	0	0
Science/Culture and Education Service	0	0	0	0.02	0	0.02	0	0.01	0.11	0	0	0	0.89	0	0	0.06	0	0	0	0	0
Sports and Recreation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.62	0	0	0	0	0
Governmental Organization and Social Group	0	0	0	0	0	0.87	0	0.08	0	0	0	0	0	0	0	0	0	0	0	0	0
Finance and Insurance Service	0	0	0	0	0	0	0	0	0	0.61	0	0	0	0	0	0	0	0	0	0	0
Auto Repair	0	0	0	0	0	0	0		0.01	0	0	0	0	0	0	0	0	0	0	0	0.31
Auto Service	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.63
Public Facility	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	0	0	0
Auto Dealers	0	0	0	0	0	0	0.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04
Motorcycle Service	0	0	0	0	0	0	0	0	0	0	0	0.72	0	0	0	0	0	0	0	0	0
Tourist Attraction	0	0	0	0	0	0	0.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10. Twenty-one topics and the word probability in each topic.

By analyzing the word probability distribution of the above 21 topics, we found noteworthy phenomena. (1) Most topics were dominated by a small number of words. For example, except for topics 9, 17, 19, and 20, every other topic was dominated by a word with a probability higher than 0.6, and the highest probability could reach 1. This may reflect the single function and use of buildings in Beijing. (2) Automobile dealer- and

motorcycle service-related POIs only occupied a very low probability distribution in topics 7, 12, and 21, and the probability distribution in the remaining topics reached 0, which may be related to the vehicle traffic restriction regulations implemented in Beijing, resulting in a downturn of automobile dealerships and the motorcycle service industry. (3) Sports and recreation-related POIs exhibited a probability of 0.62 in topic 10 and a probability of 0 in the remaining topics, indicating the lack of sports and recreation services in Beijing, which requires the attention of urban managers and planners because the lack of these services may lead to a decline in the happiness index of residents [34]. (4) Topic 15 attained a unique topic distribution, and the POI words "commercial houses" constituted the only words. The probability reached 1, which indicates that topic 15 truly relied on the POI words "commercial houses". This further reflected the lack of commercial, transportation, and other service facilities near commercial houses in Beijing. In general, the application of the LDA topic model to mine the topic distribution of the different categories of UFZs greatly contributed to UFZ identification.

5.3. Classification Results

The semantic feature of each spatial analysis unit was represented by a 21-dimensional vector. Table 3 lists the classification results obtained with the multiclass SVM classifier. Each topic was represented by the first three high-probability words. As indicated in Table 3 and Figure 9, topics 3, 8, and 11 were the top topics in commercial zones, and among these topics, POI words such as shopping, daily life services and food and beverages attained the highest probability. In business zones, topics 2, 13, and 19 included the POI word "enterprises", indicating that the POI word "enterprises" was significant for institutional zone recognition.

Table 3. Topics and top-three word distributions in the different UFZ categories.

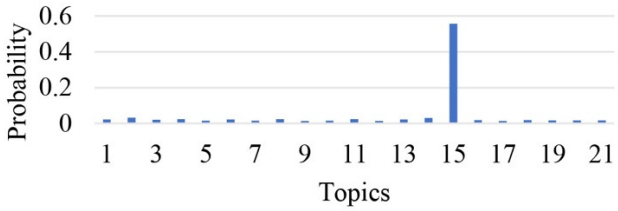
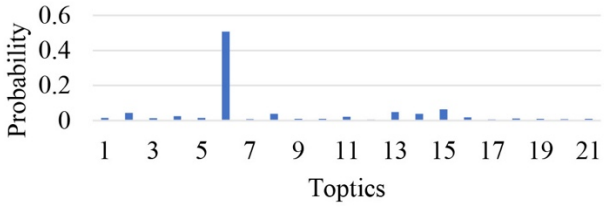
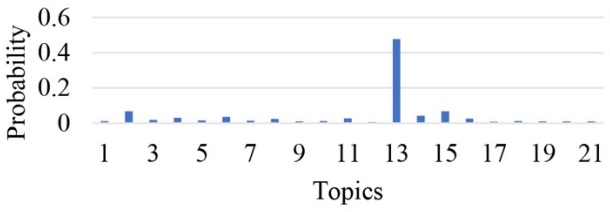
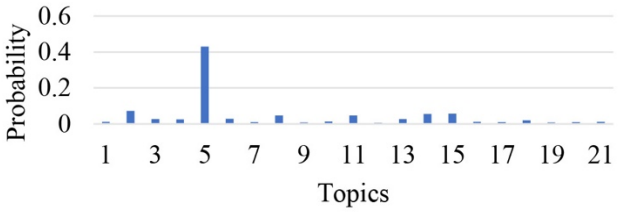
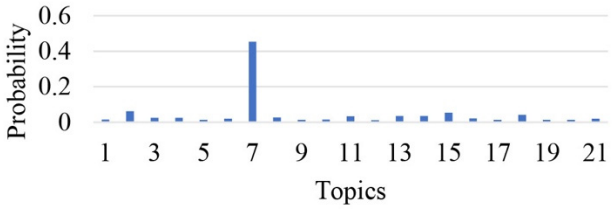
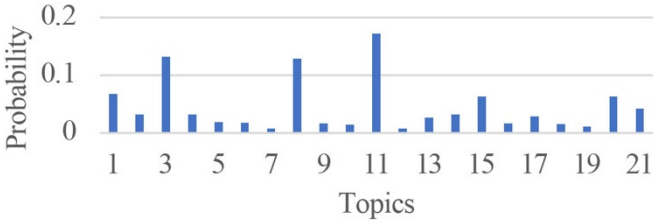
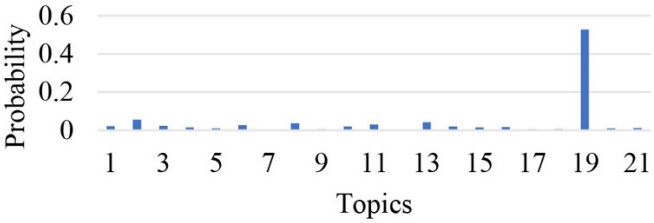
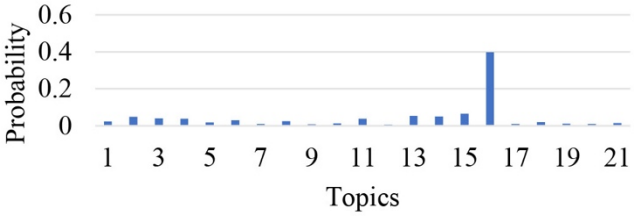
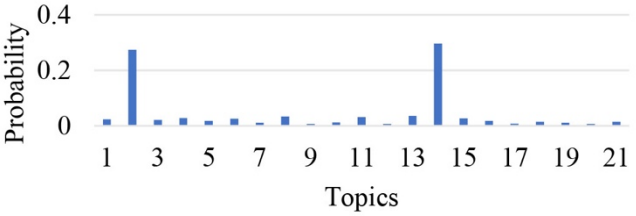
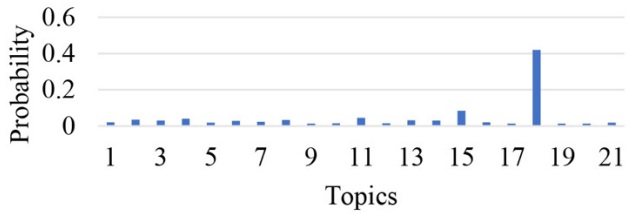
UFZ Category	Top Words in Each UFZ Category	Average Probability Distribution in Each UFZ Category
Residential	Commercial Houses Transportation Services Road Furniture	
Administrative Office Spaces	Governmental Organization and Social Group Commercial Houses Science/Culture and Education Services	
Education and Scientific Research	Science/Culture and Education Services Commercial Houses Road Furniture	

Table 3. *Cont.*

UFZ Category	Top Words in Each UFZ Category	Average Probability Distribution in Each UFZ Category
Health Care	Medical Services Road Furniture Transportation Services	
Heritage Sites	Tourist Attractions Commercial Houses Road Furniture	
Commercial Facilities	Shopping Daily Life Services Food and Beverages	
Business Facilities	Enterprises Commercial Houses Science/Culture and Education Services	
Sport and Recreation Facilities	Sports and Recreation Science/Culture and Education Services Road Furniture	
Transportation Services	Road Furniture Transportation Services Commercial Houses	
Public Facilities	Public Facility Commercial Houses Shopping	

The obtained UFZ classification results are shown in Figure 11, and two typical regions were selected to demonstrate the accuracy of our method. Commercial zones covered more than one-fourth of the study area, and most commercial zones, mainly including food and beverages, were located along roads. Most educational zones were located in the northwest built-up areas, while heritage zones were mainly located within the Second Ring Road. Most recreational zones were located among green spaces and water areas, and most remaining zones were located within schools.

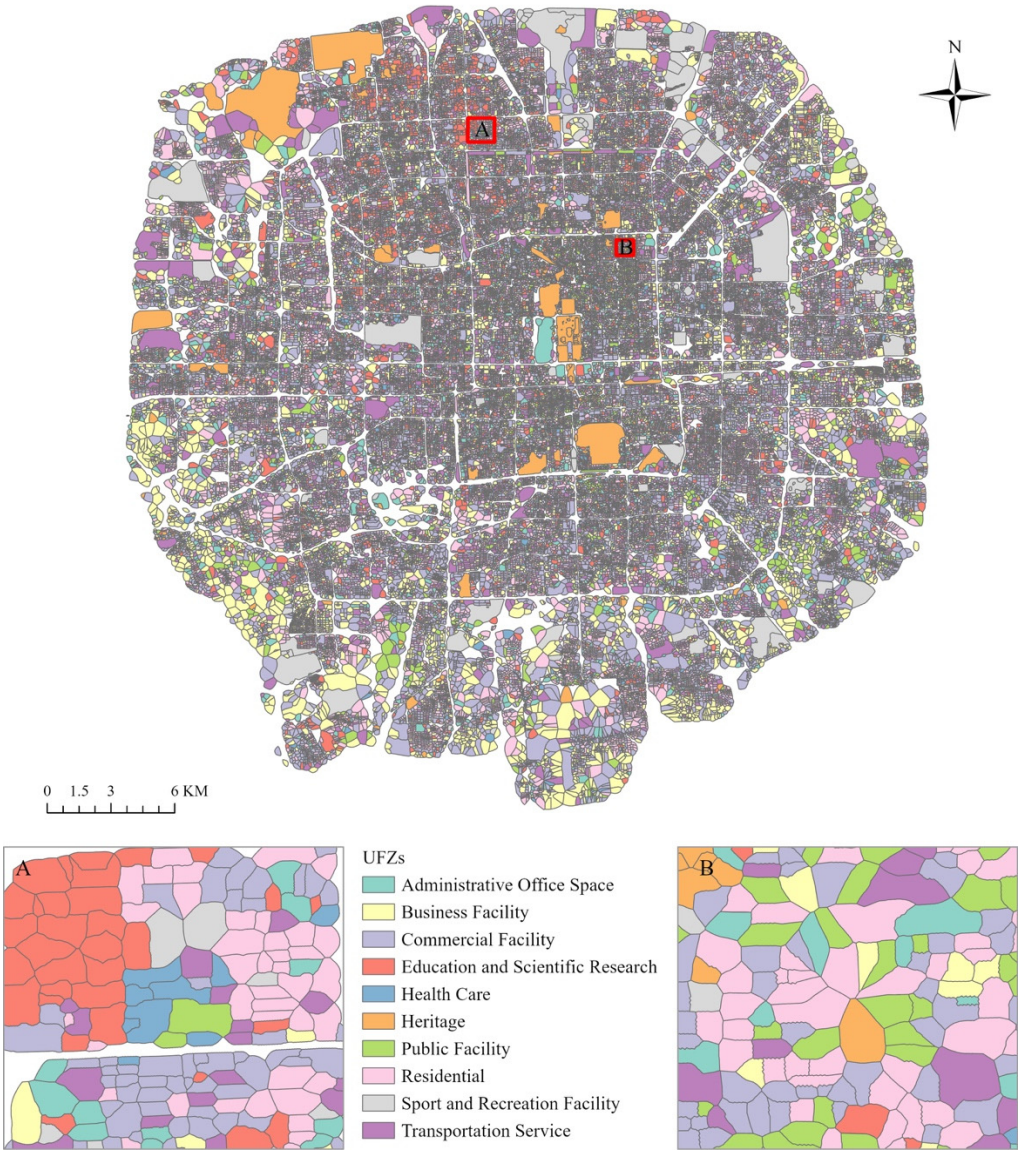


Figure 11. UFZ classification results (proposed method).

6. Discussion

6.1. Parameter Sensitivity Analysis of Zone Segmentation

To determine the optimal setting for the discretization interval (Section 4.1.2), a test was run on buildings within the Fifth Ring Road in Beijing. Several settings of the discretization interval from 0.01 to 1 m were considered, and a relatively large discretization interval was set without affecting the division result, as it directly impacted the memory demand. As shown in Figure 12, it appears that the discretization interval has an exponential effect on the number of generated points. For discretization values greater than 0.3 m (tail of the distribution), computational demand remained relatively stable, while for

higher values (head of the distribution), it grew sharply, more than doubling at each step. Discretization intervals ≥0.3 m were thus preferred as being more computationally effective. The discretization interval <0.3 m had no significant effect on the results. In contrast, a discretization interval >3 produced a "saw-like" geometry that accounted for 2.1%, which had a more obvious impact on the zone segmentation. Therefore, to balance the memory demand and the accuracy of the segmentation results, 2 m was selected as the optimal discretization interval to participate in the zone segmentation.

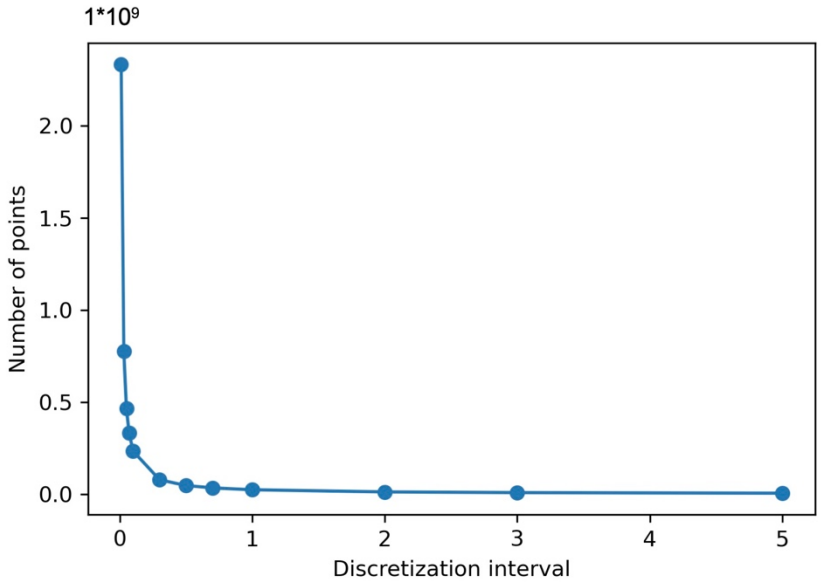


Figure 12. Relation of discretization interval and number of points generated.

6.2. Parameter Sensitivity Analysis of LDA

In this section, we evaluate the two parameters of our proposed methods, including the number of topics and hyperparameter α applied in the LDA model. When analyzing the influence of the number of topics, α was set to 1.1, and when analyzing the influence of α, the number of topics was set to 21.

(1) Number of topics

As shown in Table 4, the number of topics impacted the classification results. A small number of topics reduced the discrimination ability of features in distinguishing different UFZs. When the number of topics reached 60, all 10 predefined function types were dominated by a certain POI category. When the number of topics was greater than 21, no new functions emerged. Therefore, we propose that when the number of topics is 21, all urban functions can be distinguished.

Table 4. Parameter influence analysis with the number of topics.

UFZ Categories	Topics = 26	Topics = 22	Topics = 21	Topics = 18	Topics = 14	Topics = 10	Topics = 6
Residential	1	2	1	2	1	1	1
Administrative Office Space	2	3	1	1	1	1	0
Education and Scientific Research	1	1	1	2	1	1	1
Health Care	1	1	1	1	1	1	0
Heritage	1	1	1	0	0	0	0
Commercial Facility	11	10	9	6	6	3	2
Business Facility	4	1	3	2	2	1	1
Sport and Recreation Facility	1	0	1	1	0	0	0
Transportation Service	3	2	2	2	1	1	0
Public Facility	1	1	1	1	1	1	1

(2) Hyperparameter

Previous studies have indicated that the hyperparameter α exerts a minor impact on the classification results. A larger α value could aggregate the topic distribution between different UFZs. As shown in Figure 13, when α ranged from 0.5 to 2.0, the overall accuracy ranged from 0.81 to 0.84. Moreover, when α was set to 1.1, the best result was achieved.

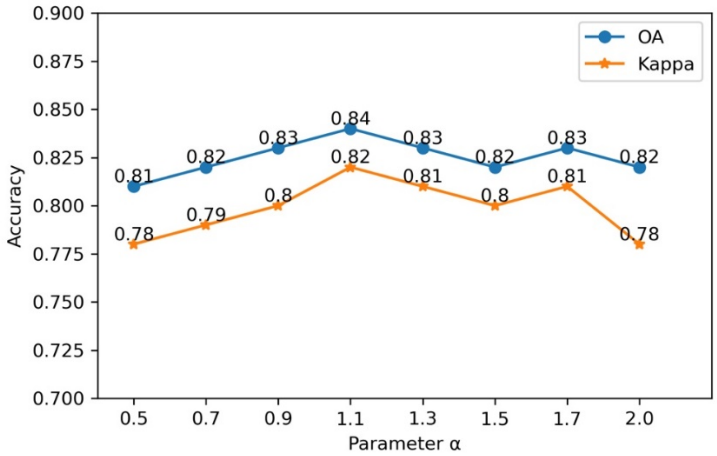


Figure 13. Parameter influence analysis with hyperparameter α.

6.3. Verification of the Classification Results

(1) UFZ Mapping Units Assessment

To examine the zoning effects, we compared the POI class richness and diversity between the proposed method and the road-based segmentation method. Among them, POI class richness refers to the number of classes present in a unit. The POI class diversity index was calculated by Shannon's diversity index, which represents the diversity of urban functional uses. Table 5 shows that small-scale units produced by the proposed method had a low diversity index and low richness index. Compared with small scales, the mean diversity and richness indices were greater in road-based segmentation units because the greater units included a larger extent with varying functions. However, this increased the difficulty of extracting urban functional use because the units included too much POI information.

Table 5. Results of richness and diversity indexes measured for two segmentation units.

Zoning Method	POI Class Richness			POI Class Diversity		
	Mean	St. Dev. [1]	Median	Mean	St. Dev.	Median
Proposed method	3.3	2.6	2	0.84	0.64	0.69
Road-based method	17.2	3.5	18	2.30	0.39	2.45

[1] Standard deviation.

In addition, Figure 14 shows the different classification results obtained by different segmentation methods in the selected region. As shown in Figure 14a–c, the functional zone classification results of the Keyu district in Z-Park show that mark No.1 is displayed as a purely residential zone in the results based on the road segmentation method. The classification results based on the method proposed in our study show that it has mixed functional uses, including residential, business facility, commercial facility, educational and scientific research, transportation service, and administrative office space. As shown in Figure 14d–f, the functional zone classification results of Wangfujing show that mark No. 2 is displayed as a pure commercial zone in the results based on the road segmentation method. The classification results based on the method proposed in our research show that although it is mainly a commercial zone, it also has a small number of

other mixed functional uses, including heritage, transportation service, residential, business facility, administrative office space, and public service. The results demonstrate that compared to the road-based segmentation method, the application of spatial analysis units obtained by the proposed method to identify UFZs was more reasonable and more effective on a fine-grained scale.

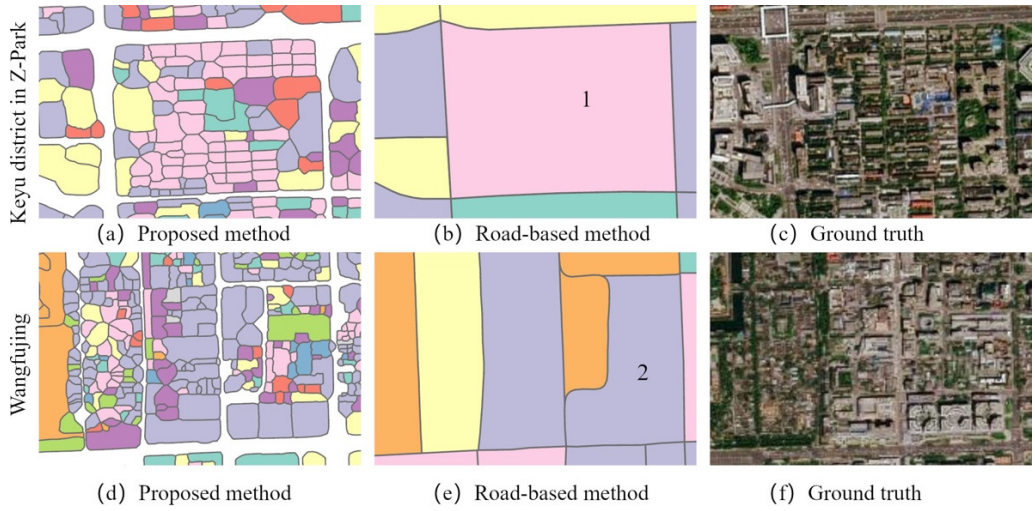


Figure 14. Illustration of UFZ mapping results with the different zone segmentation strategies. (**a**,**d**) Proposed method, (**b**,**e**) road-based method, (**c**,**f**) ground truth.

(2) Accuracy Assessment

As it is difficult to obtain the actual function of all UFZs, a total of 500 UFZs with different functions were manually identified as a reference, assisted by local knowledge, Gaode Maps, and field investigations. The associated confusion matrix is shown in Figure 15. In the proposed method, an overall accuracy of 0.84 was achieved with a kappa value of 0.82, and in the road-based method, an overall accuracy of 0.76 was achieved with a kappa value of 0.76. In general, the proposed zone segmentation strategies can significantly improve the classification results to obtain refined urban functional zone maps.

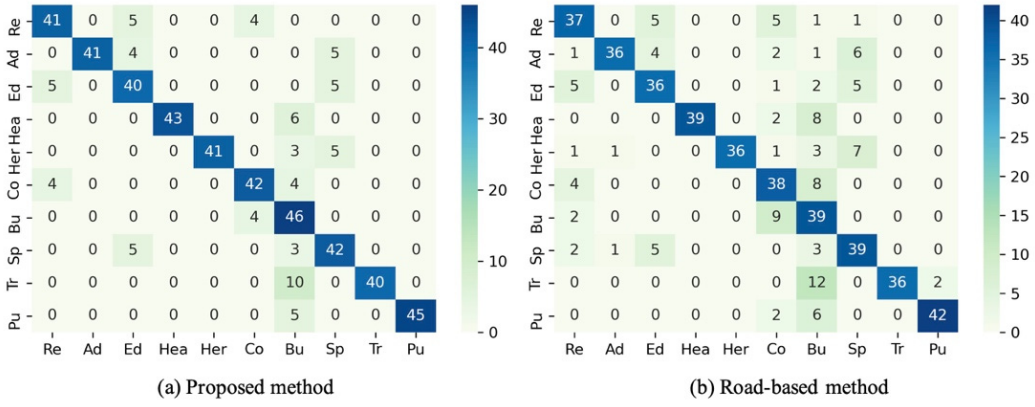


Figure 15. Confusion matrix of the classification results.

6.4. Efficacy and Limitations

As discussed above, the main purpose of this study was to generate refined UFZ mapping through the proposed Voronoi Diagram for the Polygons method. The main justification and scientific arguments of this research are that the Voronoi diagram can generate more geographically meaningful spatially segmented units (compared to TAZs). Specifically, this study objectively defined a unit of analysis that is able to capture the

smallest and arguably most fundamental level of spatial subdivision and developed a reliable and replicable method to generate and measure it. The main innovations of this paper are summarized as follows:

(1) The existing Voronoi diagram method generates map divisions based on points rather than polygons. Buildings are polygon data, which cannot be directly used for map segmentation. In the proposed method, building boundaries are discretized into points, spatial point tracking is performed, and refined map segmentation is generated through polygons. The Voronoi diagram method is thereby extended from one-dimensional spatial data dependencies to two-dimensional space.

(2) The Voronoi Diagram for the Polygons method is useful and has broad application prospects because it uniformly covers the totality of space within a set study area, which enables the topology of contiguous space to be captured at the plot level. Indeed, because all analysis units are determined by adjacency, the Voronoi Diagram for the Polygons method can be used for analyses that are based on topological distances (a set number of topological steps between cells) rather than geographic distances (a set metric distance around elements, or along the street network).

(3) In the modeling process using POIs, methods based on the shortest path and random sampling only sort POIs according to the Euclidean distance but do not consider the actual POI distribution in 3D space. The sliding window method may lead to model overfitting due to oversampling. In this study, taking buildings and their scope of influence as the basic analysis unit can avoid potential disasters caused by an uneven distribution of POI categories. For example, certain types of POIs are easily overlooked due to their small size, and some important functional areas (such as a post office next to a mall) are overlooked.

However, several limitations still exist. For example, because the map segmentation method considers each spatial feature as an individual input for division, spatial features should not consist of multiple parts. Second, the time consumption of constructing a Voronoi diagram in the process of map division has a considerable relationship with the number of buildings in the city and the interval of discrete points. The time complexity is $O(n \log n)$, and the space complexity is $O(n \log n)$. Parallel or block computing should be used for large-scale computing. Finally, there may exist no perfect building dataset for some developing countries, and remote sensing technologies [35,36] need to be used to extract buildings to obtain useful datasets for segmentation.

7. Conclusions

UFZs reflect the actual distribution characteristics of the urban spatial structure and social economy. UFZ distribution mapping on a fine scale is very important for urban decision-makers and planners. However, the rapid urbanization process leads to UFZ diversification and complications within cities, which poses a great challenge to accurate and effective UFZ mapping. This study established a framework to delineate UFZs at the building level considering semantic features extracted from POIs. First, the Voronoi diagram for the polygon segmentation method was proposed to generate spatial analysis units by integrating building, green space, water area, and road network data. Subsequently, reclassified Gaode-based POIs and the LDA model were introduced to extract semantic features. Finally, we input all semantic features into a multiclass SVM classifier applied to obtain the UFZ map. The results indicated that the proposed model could effectively segment the study area to obtain refined UFZs with higher classification accuracy (overall accuracy = 84%; kappa = 0.82).

In future work, other open data (such as remote sensing images and mobile phone positioning data) will be introduced to improve the UFZ division accuracy.

Author Contributions: Conceptualization, Yue Deng and Rixing He; methodology, Yue Deng; software and data curation, Yue Deng; writing—original draft preparation, Yue Deng; writing—review and editing, Yue Deng; visualization and supervision, Rixing He; project administration, Rixing He; funding acquisition, Rixing He. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Special Projects of the Ministry of Public Security in Strengthening Basic Police Work (Grant No. 2021JC35).

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by Gaode Maps. The data and results are only for research purposes, and do not involve commercial interests. We thank the editors and the anonymous reviewers for their insightful comments and suggestions.

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

References

1. Du, S.; Du, S.; Liu, B.; Zhang, X. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.* **2021**, *261*, 112480. [CrossRef]
2. Jokar Arsanjani, 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]
3. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
4. Niu, H.; Silva, E.A. Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. *Comput. Environ. Urban Syst.* **2021**, *88*, 101651. [CrossRef]
5. 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]
6. Huang, X.; Yang, J.; Li, J.; Wen, D. Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 403–415. [CrossRef]
7. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
8. 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]
9. Martin, D.; Cockings, S.; Harfoot, A. Development of a geographical framework for census workplace data. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2013**, *176*, 585–602. [CrossRef]
10. Longley, P. Geographical Information Systems: A renaissance of geodemographics for public service delivery. *Prog. Hum. Geogr.* **2005**, *29*, 57–63. [CrossRef]
11. Yuan, N.J.; Zheng, Y.; Xie, X.; Wang, Y.; Zheng, K.; Xiong, H. Discovering Urban Functional Zones Using Latent Activity Trajectories. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 712–725. [CrossRef]
12. Tu, W.; Cao, J.; Yue, Y.; Shaw, S.L.; Zhou, M.; Wang, Z.; Chang, X.; Xu, Y.; Li, Q. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2331–2358. [CrossRef]
13. Song, J.; Lin, T.; Li, X.; Prishchepov, A.V. Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest: A Case Study of Xiamen, China. *Remote Sens.* **2018**, *10*, 1737. [CrossRef]
14. Hu, Y.; Han, Y. Identification of Urban Functional Areas Based on POI Data: A Case Study of the Guangzhou Economic and Technological Development Zone. *Sustainability* **2019**, *11*, 1385. [CrossRef]
15. Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.-R.; Gu, C. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Comput. Environ. Urban Syst.* **2019**, *74*, 1–12. [CrossRef]
16. Li, W. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans. Inf. Theory* **1992**, *38*, 1842–1845. [CrossRef]
17. Yuan, N.J.; Zheng, Y.; Xie, X. Discovering Functional Zones in a City Using Human Movements and Points of Interest. In *Spatial Analysis and Location Modeling in Urban and Regional Systems*; Thill, J.-C., Ed.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 33–62.
18. Chi, J.; Jiao, L.; Dong, T.; Gu, Y.; Ma, Y. Quantitative identification and visualization of urban functional area based on POI data. *J. Geomat.* **2016**, *41*, 68–73. [CrossRef]
19. 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]
20. Openshaw, S. Ecological fallacies and the analysis of areal census data. *Environ. Plan. A* **1984**, *16*, 17–31. [CrossRef]
21. Wang, N.; Li, W.; Tao, R.; Du, Q. Graph-based block-level urban change detection using Sentinel-2 time series. *Remote Sens. Environ.* **2022**, *274*, 112993. [CrossRef]
22. Zhou, W.; Ming, D.; Lv, X.; Zhou, K.; Bao, H.; Hong, Z. SO–CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sens. Environ.* **2020**, *236*, 111458. [CrossRef]

23. Liu, Y.; Zhang, Y.; Jin, S.T.; Liu, Y. Spatial pattern of leisure activities among residents in Beijing, China: Exploring the impacts of urban environment. *Sustain. Cities Soc.* **2020**, *52*, 101806. [CrossRef]
24. Niu, H.; Silva, E.A. Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *J. Urban Plan. Dev.* **2020**, *146*, 04020007. [CrossRef]
25. Li, X.; Hu, T.; Gong, P.; Du, S.; Chen, B.; Li, X.; Dai, Q. Mapping Essential Urban Land Use Categories in Beijing with a Fast Area of Interest (AOI)-Based Method. *Remote Sens.* **2021**, *13*, 477. [CrossRef]
26. Sun, Z.; Li, X.; Fu, W.; Li, Y.; Tang, D. Long-term effects of land use/land cover change on surface runoff in urban areas of Beijing, China. *J. Appl. Remote Sens.* **2013**, *8*, 084596. [CrossRef]
27. Zhao, L.; Li, Y. Study on Urban Road Network Traffic District Division based on Clustering Analysis. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 3556–3560.
28. 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]
29. Blei, D.M.; Ng, A.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *30*, 993–1022. [CrossRef]
30. Du, S.; Du, S.; Liu, B.; Zhang, X.; Zheng, Z. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. *GISci. Remote Sens.* **2020**, *57*, 411–430. [CrossRef]
31. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
32. Fleischmann, M.; Feliciotti, A.; Romice, O.; Porta, S. Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Comput. Environ. Urban Syst.* **2020**, *80*, 101441. [CrossRef]
33. Zhang, Y.; Li, Q.; Huang, H.; Wu, W.; Du, X.; Wang, H. The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China. *Remote Sens.* **2017**, *9*, 865. [CrossRef]
34. Deng, Y.; Liu, J.; Luo, A.; Wang, Y.; Xu, S.; Ren, F.; Su, F. Spatial Mismatch between the Supply and Demand of Urban Leisure Services with Multisource Open Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 466. [CrossRef]
35. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
36. Chen, K.; Fu, K.; Gao, X.; Yan, M.; Sun, X.; Zhang, H. Building extraction from remote sensing images with deep learning in a supervised manner. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1672–1675.