

Article

# Understanding Completeness and Diversity Patterns of OSM-Based Land-Use and Land-Cover Dataset in China

ShuZhu Wang, Qi Zhou \*  and YuanJian Tian

School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; wangsz@cug.edu.cn (S.W.); tianyuanjian@cug.edu.cn (Y.T.)

\* Correspondence: zhouqi@cug.edu.cn

Received: 2 July 2020; Accepted: 31 August 2020; Published: 4 September 2020



**Abstract:** OpenStreetMap (OSM) data are considered essential for land-use and land-cover (LULC) mapping despite their lack of quality. Most relevant studies have employed an LULC reference dataset for quality assessment, but such a reference dataset is not freely available for most countries and regions. Thus, this study conducts an intrinsic quality assessment of the OSM-based LULC dataset (i.e., without using a reference LULC dataset) by examining the patterns of both its completeness and diversity. With China chosen as the study area, an OSM-based LULC dataset of the country was first generated and validated by using various accuracy measures. Both its completeness and diversity patterns were then mapped and analyzed in terms of each prefecture-level division of the country. The results showed the following: (1) While the overall accuracy was as high as 82.2%, most complete regions of China were not mapped well owing to a lack of diverse LULC classes. (2) In terms of socioeconomic factors and the number of contributors, higher correlations were noted for diversity patterns than completeness patterns; thus, the diversity pattern is a better reflection of socioeconomic factors and the spatial patterns of contributors. (3) Both the completeness and the diversity patterns can be combined to better understand an OSM-based LULC dataset. These results indicate that it is useful to consider diversity as a supplement for intrinsically assessing the quality of an OSM-based LULC dataset. This analytical method can also be applied to other countries and regions.

**Keywords:** OpenStreetMap; LULC mapping; data quality; intrinsic quality assessment; completeness; diversity

## 1. Introduction

Land-use (LU) and land-cover (LC) maps represent spatial information on different classes of natural and/or human-made geographical features on the Earth. These maps can be applied to natural resource management [1,2], urban and transportation planning [3,4], and the monitoring and modeling of urban sprawl [5–7]. Much focus has been placed on land-use and land-cover (LULC) mapping using different data sources. For instance, remote sensing data have been widely used to produce a global LC map [8,9] because the technology has benefits in terms of detecting physical objects (e.g., roads, buildings, rivers, and lakes) on the surface of the Earth. However, the use of remote sensing data has been criticized in the context of LU mapping because it is difficult to sense thematic attributes (e.g., residential, commercial, or industrial) of an object using only these data [10]. Other data sources, e.g., points of interest (POI) [11], street view images [12], and mobile phone data [13], have also been used for LU mapping, but most of them are often available only for studying a certain area of a city, rather than areas of a country or larger region. Another choice is geographic information provided by volunteers, the so-called “volunteered geographic information” (VGI) [14]. As one of

the most successful VGI projects, OpenStreetMap (OSM) has been used for LULC mapping [15–17]. OSM data are beneficial in this context because they are available for free, and provide global coverage and almost real-time updates (e.g., OSM data can theoretically be updated on a minute-by-minute basis). Nevertheless, many concerns have arisen regarding the quality of OSM datasets because most are provided by volunteers from different countries and/or backgrounds, embodying different ages, occupations, and incomes [18]. Therefore, much research has focused on assessing the quality of OSM data in terms of not only the LULC feature [15,19,20], but also other geographical features, e.g., roads [21–23] and buildings [24,25].

Most relevant studies have used an LULC reference dataset for quality assessment. For instance, Arsanjani and Vaz [19] compared the OSM-based LU datasets of seven of Europe's metropolises with the corresponding LU reference datasets from the Global Monitoring for Environment and Security Urban Atlas (GMESUA). They found that six out of the seven metropolises had approximately above 75% agreement between each pair of OSM and LU reference datasets; but the completeness (i.e., the coverage of an OSM dataset in a region) of the seven metropolises varied, from 42% for Budapest to 100% for Bucharest. Estima and Painho [26] used OSM-POI data to generate an LULC dataset of Portugal, the accuracy (i.e., the consistency between a pair of OSM and LULC reference datasets) of which was 76.7%. The corresponding LULC reference dataset was obtained from the Coordination of Information on the Environment (CORINE). Dorn et al. [27] assessed the quality of OSM-based LULC data in southern Germany and found that the LULC class (i.e., forest) had both a high accuracy (95.1%) and high completeness (97.6%), but another class (i.e., farmland) had a low completeness value (45.9%). Viana et al. [20] used OSM historical data for multi-temporal LULC mapping and compared the results with CORINE datasets. They also concluded that the accuracy values were substantially high (ranging from 77.3% to 91.9%), although the completeness value was low.

However, an LULC reference dataset is not always freely available, especially for a study area outside Europe. This means that for most parts of the world, we need to assess the quality of an OSM-based LULC dataset without any reference dataset (called intrinsic quality assessment). Several studies have been conducted on intrinsic quality assessment. For instance, See et al. [28] found that the quality of data contributed by volunteers increased when they indicated higher confidence. Comber et al. [29] used a number of control points to determine the accuracy of LC classes produced by a number of volunteers. However, they paid more attention to accuracy than completeness. It is also desirable to assess the completeness of an OSM dataset to determine whether free data are available. Some studies have proposed the use of proxy indicators (e.g., OSM building density and OSM street block density) to quantitatively estimate the completeness of an OSM dataset [24,30], but these studies have focused on assessing road and building data rather than LULC data in OSM.

Theoretically, it is necessary to employ an LULC reference dataset while assessing only accuracy and not completeness—the latter can be measured as the ratio of the covered areas of an OSM-based LULC dataset to the total area of a region. However, due to the availability of LULC reference datasets, most studies have assessed the quality of OSM-based LULC datasets in European cities, but few have paid attention to areas outside Europe. Thus, it is useful to investigate whether the quality of data (especially accuracy) of a different region is comparable to results reported for European study areas. Moreover, few studies have carried out an intrinsic quality assessment of OSM-based LULC datasets, which is especially desirable for a study area without any reference dataset. More importantly, in addition to completeness, is there any other useful measure(s) for an intrinsic quality assessment?

Therefore, this study conducts an intrinsic quality assessment of the OSM-based LULC dataset. The tenet of our approach is to understand both the completeness and diversity patterns of such a dataset. More precisely, an OSM-based LULC dataset is first generated and validated, and both its completeness and diversity patterns are mapped and analyzed. Diversity can be regarded as a quantitative measure that reflects the number of LULC classes in a dataset. While diversity has been widely applied to social science [31] and landscape analysis [32,33], to the best of our knowledge, this measure has rarely been employed for assessing the quality of an OSM-based LULC dataset.

The study makes two contributions:

- (1) The intrinsic quality assessment of an OSM-based LULC dataset is carried out. By contrast, most past studies have used an LULC reference dataset for quality assessment. Our analytical method can be applied to other regions, especially those for which a free LULC reference dataset is unavailable.
- (2) Both the completeness and the diversity patterns of an entire country (China) were mapped and analyzed, and the results indicate that the diversity measure may be used as a supplement for an intrinsic quality assessment.

The remainder of this study is structured as follows: Section 2 introduces the study area and data. Section 3 presents the methods used not only to generate the OSM-based LULC dataset, but also to analyze its completeness and diversity patterns. Section 4 reports and analyzes the results, and Section 5 analyzes the combined patterns of completeness and diversity. Sections 6 and 7 contain a discussion and the conclusions of this study, respectively.

## 2. Study Area and Data

### 2.1. Study Area

(Mainland) China was chosen as the study area for several reasons. First, an LULC reference dataset for China, produced by either mapping agencies or commercial companies, was not available for free. However, the OSM dataset may be used as an alternative. Second, most relevant studies have focused on an analysis of OSM-based LULC datasets in Europe, and not in other regions of the world. China is the third-largest country in the world in terms of the land area. It is thus useful to investigate the regions of China that have been mapped/unmapped well, and whether the quality of OSM data on China is comparable to that for Europe. Third, there are 363 prefecture-level divisions in China. It might be useful to be able to analyze the completeness and diversity patterns of such a large number of divisions, as well as the factors that may significantly influence such patterns.

### 2.2. Data

The most important data in our study were those from the OSM dataset on China. This dataset was acquired for free from <http://download.geofabrik.de/index.html> in January 2019. The OSM dataset had a number of features, including POIs, in terms of point features, and data on roads, railways, and waterways, in terms of line features. The OSM dataset also covered the LULC, water, and buildings, in terms of polygon features. Both line and polygon features were used to produce an LULC dataset for all of China because such features can be characterized by either length or area. Furthermore, each object in an OSM dataset normally has a tag (e.g., `landuse = forest`) to describe its attribute, based on which (e.g., `forest`) we can classify each object into an LULC class. In addition to the OSM dataset, the administrative division dataset (consisting of data on 363 divisions in total) was acquired for free from the National Catalogue Service for Geographic Information (<http://www.webmap.cn/main.do?method=index>). This dataset was used to divide the OSM dataset of China into different sub-divisions.

## 3. Methods

The tenet of our approach is to understand both the completeness and the diversity patterns of an OSM-based LULC dataset. The OSM-based LULC dataset on China was first generated and then validated by calculating various accuracy measures. Moreover, this dataset was mapped and analyzed in terms of both completeness and diversity measures.

### 3.1. Production and Validation of OSM-Based LULC Dataset

#### 3.1.1. Production

Extensive efforts have been made to produce an OSM-based LULC dataset [15–17], most of which feature three typical steps. First, convert line features into polygon features; second, classify OSM objects into the corresponding reference classes according to their tags (e.g., landuse = forest); and third, merge multiple OSM features (or layers) into a single layer. This study follows these three steps to produce an OSM-based LULC dataset.

- Step 1: Convert line features into polygon features. According to Zhou et al. [17], it is feasible to convert a line feature into a polygon feature through buffering, i.e., to create a buffer region around the line feature, after which, the buffer region can be viewed as a polygon feature. The challenge here is to determine appropriate buffer radii for different OSM types because OSM line objects may be tagged with different attribute values (e.g., highways = primary, highways = secondary, and highways = residential). Such an appropriate radius was determined by Zhou et al. [17] through comparison with a corresponding reference LULC dataset (GMESUA). However, such a reference dataset was not available for our study area, and thus different buffer radii (ranging from 4.5–10 m) for different OSM types were manually determined by referring to the Technical Standard of Highway Engineering of China and the corresponding images on Google Earth (Table 1). The buffer radius was generally positively correlated with the importance of an OSM road type.
- Step 2: Classify OSM objects into corresponding reference classes. Owing to a lack of the LULC reference product, we manually classified all OSM objects (according to their tags) into 12 LULC classes: Agriculture, orchard, forest, grass, commercial, industrial, residential, public use, special use, transportation, water, and other lands (Table 2). All these LULC classes were obtained from the first level of the National Land Use Classification Standards of China.
- Step 3: Merge multiple LULC classes (or layers) into a single layer. This is a necessary step because some polygon objects in OSM may overlap but correspond to different LULC classes; it may therefore be difficult to determine a unique LULC class for the same geographical region. The solution is to make different (12) LULC classes overlap according to their average area, from small to large [17]. To be specific, the feature or class with the smallest average area was placed on the top and that with the largest average area was placed at the bottom. After this process, all LULC classes (or layers) were further merged into a single layer.

**Table 1.** Buffer radii for typical OpenStreetMap (OSM) road types.

OSM Road Type	Buffer Radius (m)
trunk	8
motorway	7.5
primary	10
secondary	7
tertiary	7
residential <sup>1</sup>	6
service	5
unclassified	4.5

<sup>1</sup>“residential<sup>1</sup>” denotes a road type (i.e., highways = residential).

**Table 2.** Correspondence between some typical OSM tags and land-use and land-cover (LULC) reference classes.

No.	Typical OSM Tags	LULC Reference Classes
1	allotments; farm	agriculture
2	orchard; vineyard	orchard
3	forest; scrub	forest
4	grass; meadow	grass
5	ATM; bank; bar; cafe; cinema; commercial; computer shop; fast food; furniture shop; giftshop; hostel; mall; nightclub; restaurant; retail; supermarket; tower	commercial
6	industrial; quarry	industrial
7	residential <sup>2</sup>	residential
8	artwork; college; hospital; library; museum; police; post office; school; sports center; stadium; toilet; university; zoo	public use
9	attraction; Christian; memorial; military; monument; Muslim; prison; shelter	special use
10	airport; bus station; parking; primary; rail; secondary; tertiary; trunk; unclassified	transportation
11	beach; dam; dock; reservoir; river; water; waterfall; wetland	water
12	cliff; heath	others

<sup>2</sup>“residential<sup>2</sup>” denotes a land-use type (i.e., landuse = residential).

### 3.1.2. Validation

Accuracy measures how closely the OSM-based LULC dataset matches a corresponding LULC reference product. This study employed a stratified sampling strategy. A number of sample points were first randomly chosen for each LULC class and then positively correlated with the area percentage of each class in the OSM dataset. The actual LULC class of each sample point was manually and independently marked by two people by referring to Google Earth. When any point was marked differently by two people, a third person joined in the analysis and a final decision was made by voting. All sample points were used as reference for assessing the accuracy of the OSM-based LULC dataset. Based on all points (3464 in total), several accuracy measures were calculated by comparing each pair of LULC classes in the OSM-based dataset with those in the reference. Three common measures—overall accuracy (OA), user accuracy (UA), and producer accuracy (PA)—were calculated:

$$OA = \frac{\sum_{i=1}^n P_i}{N} \quad (1)$$

$$UA_i = \frac{\sum_{i=1}^n P_i}{P_i(osm)} \quad (2)$$

$$PA_i = \frac{\sum_{i=1}^n P_i}{P_i(ref)} \quad (3)$$

where  $N$  denotes the number of points in total,  $n$  denotes the number of LULC classes ( $n = 12$ ),  $P_i(osm)$  and  $P_i(ref)$  denote the number of points classified as belonging to LULC class- $i$  in the OSM-based dataset and in the reference, respectively, and  $P_i$  denotes the number of points classified as belonging to LULC class- $i$  in both datasets.

### 3.2. Mapping and Analysis of Completeness and Diversity Patterns

#### 3.2.1. Completeness and Diversity Measures

Completeness has been widely used for assessing the quality of OSM-based LULC datasets [19,20,26,27]. Completeness is defined as the ratio of the area of land of a geographical unit covered by OSM data to the total land area of the unit.

An LULC map/dataset may also be characterized by a number LULC classes (e.g., 12). Different definitions of diversity have been offered [34], among which the Shannon diversity index or Shannon's entropy [35] is the mostly widely used in the literature. The Shannon diversity index can be used to measure a variety of LULC classes in an area, and thus was employed in our study. That is,

$$S = -\sum_{i=1}^n P_i \log P_i \quad (4)$$

where  $S$  denotes the Shannon diversity index of an OSM-based LULC dataset,  $P_i$  denotes the area percentage of LULC class  $i$  in a geographical unit, and  $n$  denotes the number of LULC classes. In our study,  $n = 12$ , and thus  $S$  could vary from zero (meaning only one LULC class) to 2.48 (meaning all 12 classes had the same percentage, i.e., 8.33%).

#### 3.2.2. Mapping and Analysis

In contrast with accuracy, both completeness and diversity can be mapped without the need for an LULC reference dataset. Our analytical method therefore calculated the completeness and diversity values of each prefecture-level division and mapped them for the entire study area. With respect to differences in LULC characteristics between built-up areas and non-built-up areas, there were a relatively large number of artificial classes (e.g., commercial, industrial, and residential) in the built-up areas and a relatively large number of natural classes (e.g., forest, grass, and water) in the non-built-up areas. These two scenarios (I and II) were analyzed. For scenario I, both the completeness and the diversity values were calculated for the entire prefecture-level division, and for scenario II, these values were calculated only for the built-up areas of each prefecture-level division.

Furthermore, the completeness and diversity patterns of China were compared and analyzed using two methods: Visual analysis and quantitative assessment.

- Both the completeness and the diversity patterns were visually analyzed. A number of questions were considered. For instance, which areas had relatively high or low completeness and diversity values? Was there any difference between the completeness and the diversity patterns, in terms of scenarios I and II? Was there any correlation between the completeness and the diversity patterns of the same scenario?
- For quantitative assessment, a number of factors were employed to identify factors that could have influenced the completeness and diversity patterns. Three socioeconomic factors [the size of the built-up areas, their population, and gross domestic product (GDP)] were first considered. The corresponding data in 2019 were acquired from the National Bureau of Statistics of China (<http://www.stats.gov.cn>). These factors were chosen because studies have shown that the completeness of OSM data tends to be high in municipalities with high population density [27]. Completeness has also been positively correlated with the GDP [25]. Thus, it is useful to investigate whether these factors could still be positively correlated with the completeness and diversity patterns of the OSM-based LULC dataset of China. Moreover, the number of contributors (who had edited the OSM data) was calculated based on an analysis of OSM history data (<https://planet.openstreetmap.org/planet/full-history/>, accessed in January 2019). This number was calculated in terms of each prefecture-level division (for scenario I) and the built-up areas of each prefecture-level division (for scenario II) to determine whether the number of contributors had a positive correlation with the completeness and/or diversity patterns.

### 4. Results and Analysis

#### 4.1. Production and Validation of the OSM-Based LULC Dataset

The OSM-based LULC map of China is shown in Figure 1. First, most regions were either not covered by the OSM data or had not been mapped by the OSM volunteers (marked as “no data”). Thus, the dataset was far from complete. Second, both the LULC classes, forest (dark green) and water (light blue), were easy to observe because they occupied large areas compared with the other classes (e.g., residential in Figure 1b). Third, the OSM data were distributed unevenly across the country, e.g., there were relatively more data for the east, center, and northeast of China—regions that were mostly mapped with forests (Figure 1c).

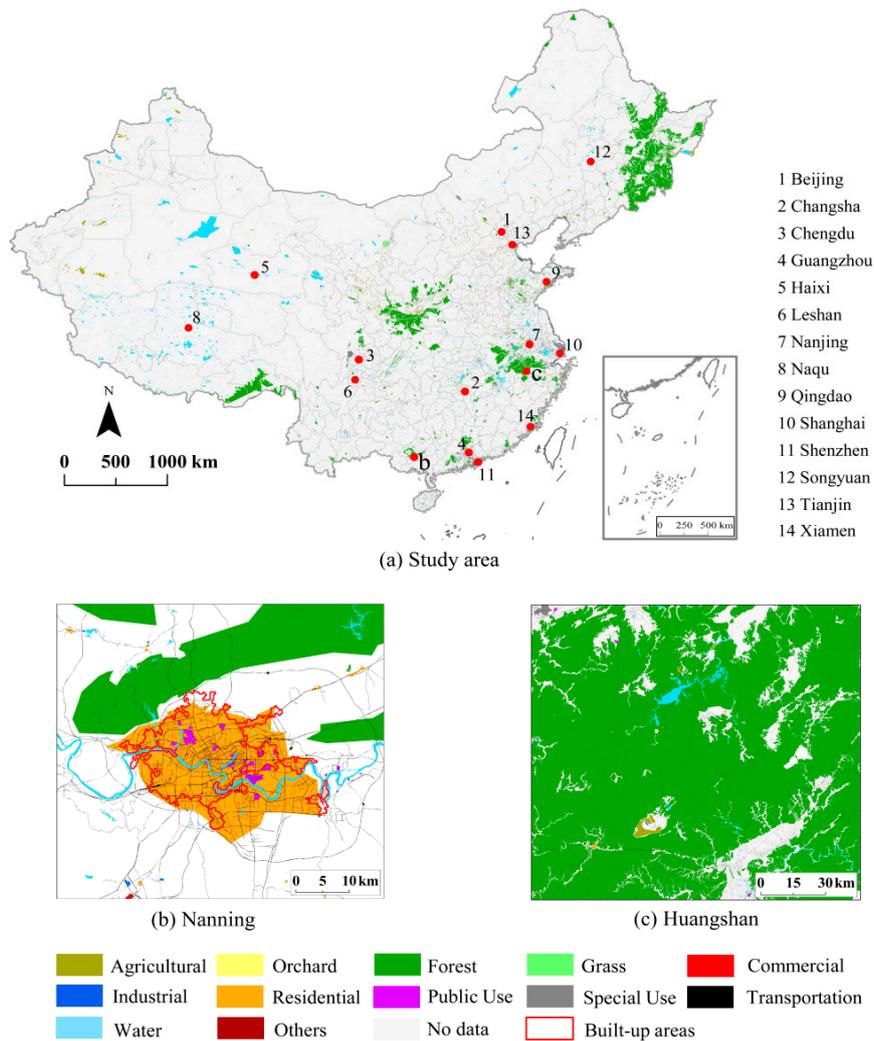


Figure 1. OSM-based LULC map of China.

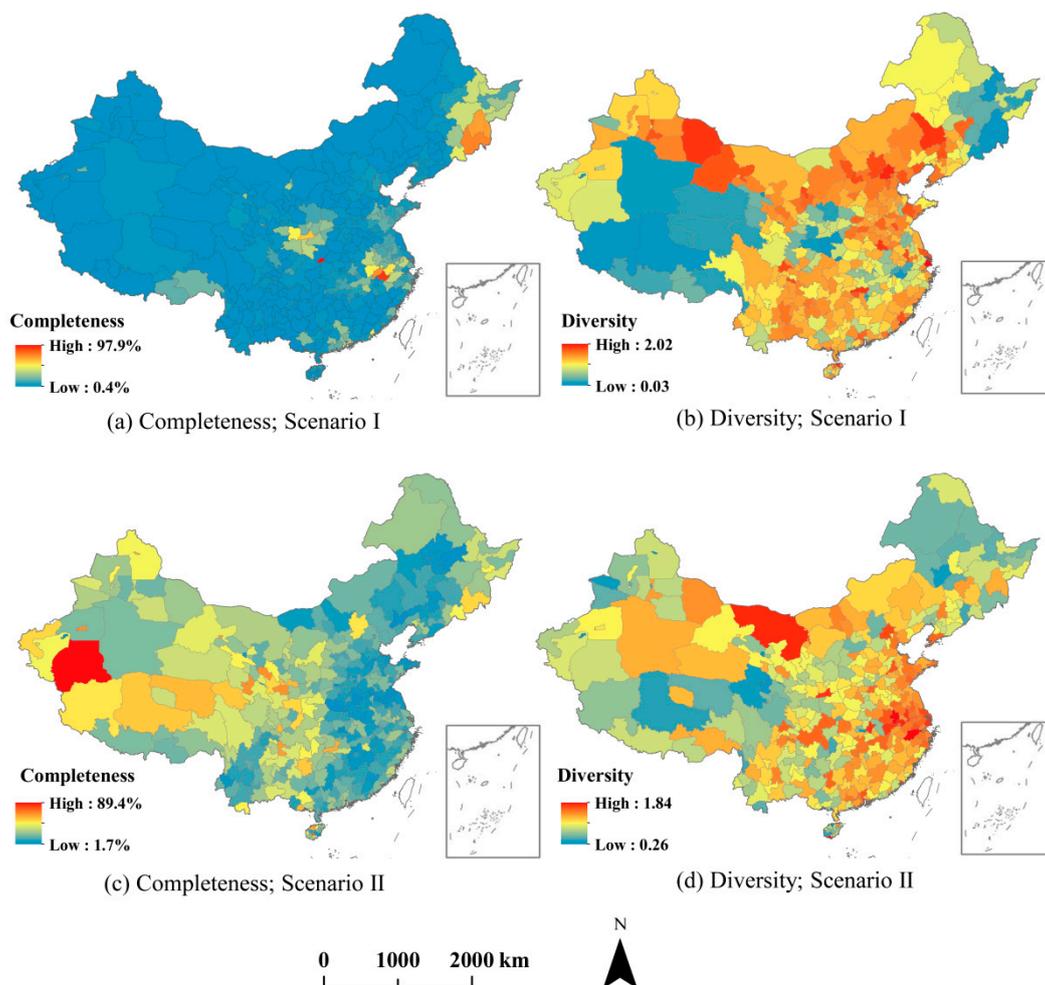
The confusion matrix and relevant accuracies of the OSM-based LULC dataset are listed in Table 3. Substantial overall accuracy (i.e., OA = 82.2%) was achieved. Moreover, eight out of the 12 LULC classes were close to or higher than 70% in terms of both the UA and the PA; two of them (forest and water) were close to or higher than 90%. These results verify the effectiveness of the produced LULC dataset of China. Furthermore, the accuracy values were verified as comparable to those reported in European study areas [19].

**Table 3.** Confusion matrix and relevant accuracies of the OSM-based LULC dataset for China.

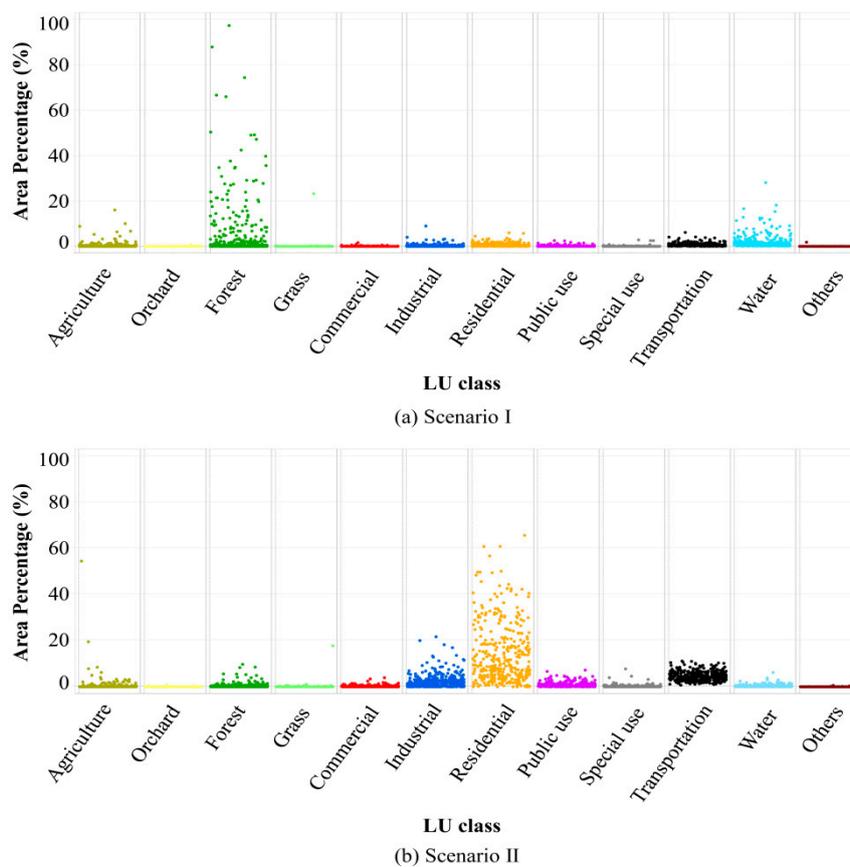
	Agricultural	Orchard	Forest	Grass	Commercial	Industrial	Residential	Public Use	Special Use	Transportation	Water	Others	PA (%)
Agricultural	260	6	32	8	4	3	6	0	4	3	8	14	74.7
Orchard	1	81	1	0	1	0	1	0	1	0	0	5	89.0
Forest	8	1	921	5	4	5	4	12	58	0	5	10	89.2
Grass	4	0	17	57	3	0	1	2	1	0	2	3	63.3
Commercial	0	0	0	6	112	1	2	5	0	4	0	0	86.2
Industrial	5	2	4	1	2	166	11	1	0	0	0	2	85.6
Residential	5	2	4	7	6	1	275	3	1	14	2	1	85.7
Public Use	0	3	6	36	6	3	19	138	2	13	7	2	58.7
Special Use	0	1	0	4	1	0	1	0	86	0	0	0	92.5
Transportation	10	1	8	30	2	7	8	0	3	325	2	1	81.9
Water	2	1	3	4	0	0	1	6	1	0	364	3	94.5
Others	5	2	11	5	6	14	5	3	23	1	10	62	42.2
UA (%)	86.7	81.0	91.5	35.0	76.2	83.0	82.3	81.2	47.8	90.3	91.0	60.2	OA (%) = 82.2

#### 4.2. Mapping and Analysis of Completeness and Diversity Patterns

The completeness values varied drastically with respect to the prefecture-level division (Figure 2a,c), i.e., from 0.4% to 97.9% for scenario I, and from 1.7% to 89.4% for scenario II. This illustrates a heterogeneous distribution of the OSM data for China. In scenario I, most of the prefecture-level divisions with high completeness values (e.g., values ranging from 50% to 97.9%) were located in the east, center, and northeast of China because a large area of these regions is occupied by forests (Figure 3a). However, the completeness pattern of scenario II showed a different distribution. Most divisions with high completeness values were located in the northwest, southwest, and northeast of China because there were more residential lands than forests in the built-up areas (Figure 3b). Moreover, most built-up areas with high completeness values had been mapped by (OSM) volunteers as a few large polygons, and tagged as belonging to the residential class. Inside these built-up areas were a few other LULC classes (e.g., commercial, industrial, and public use). As an example, Figure 1b shows the built-up areas of Nanning (a capital city in Southern China), mostly covered with a single residential land. This indicates that these built-up areas were not mapped well despite their relatively high completeness values.



**Figure 2.** Completeness and diversity patterns of the OSM-based LULC dataset of China, showing scenarios I (a,b) and II (c,d).



**Figure 3.** The area percentages of the 12 LULC classes of different prefecture-level divisions. For each LULC class (x-axis), each node denotes the area percentage (y-axis) of the class in a prefecture-level division. (a) scenario I and (b) scenario II.

The diversity patterns (Figure 2b,d) were different from the completeness patterns (Figure 2a,c). Most highly diverse divisions had relatively low completeness values—these divisions were located on the east coast of China (especially for scenario II), but their completeness values were low (marked in blue in Figure 2c). The difference between the completeness and diversity patterns is indicated in Figure 3. If the area percentage of a certain LULC class (e.g., forest or residential) was particularly high, the area percentages of the other LULC classes were low, and so were the diversity values of all classes.

Table 4 lists the correlations among multiple factors (size of built-up area, population, GDP, and number of contributors) and the four patterns in Figure 2. The correlations between each of the three socioeconomic factors and number of contributors are also listed.

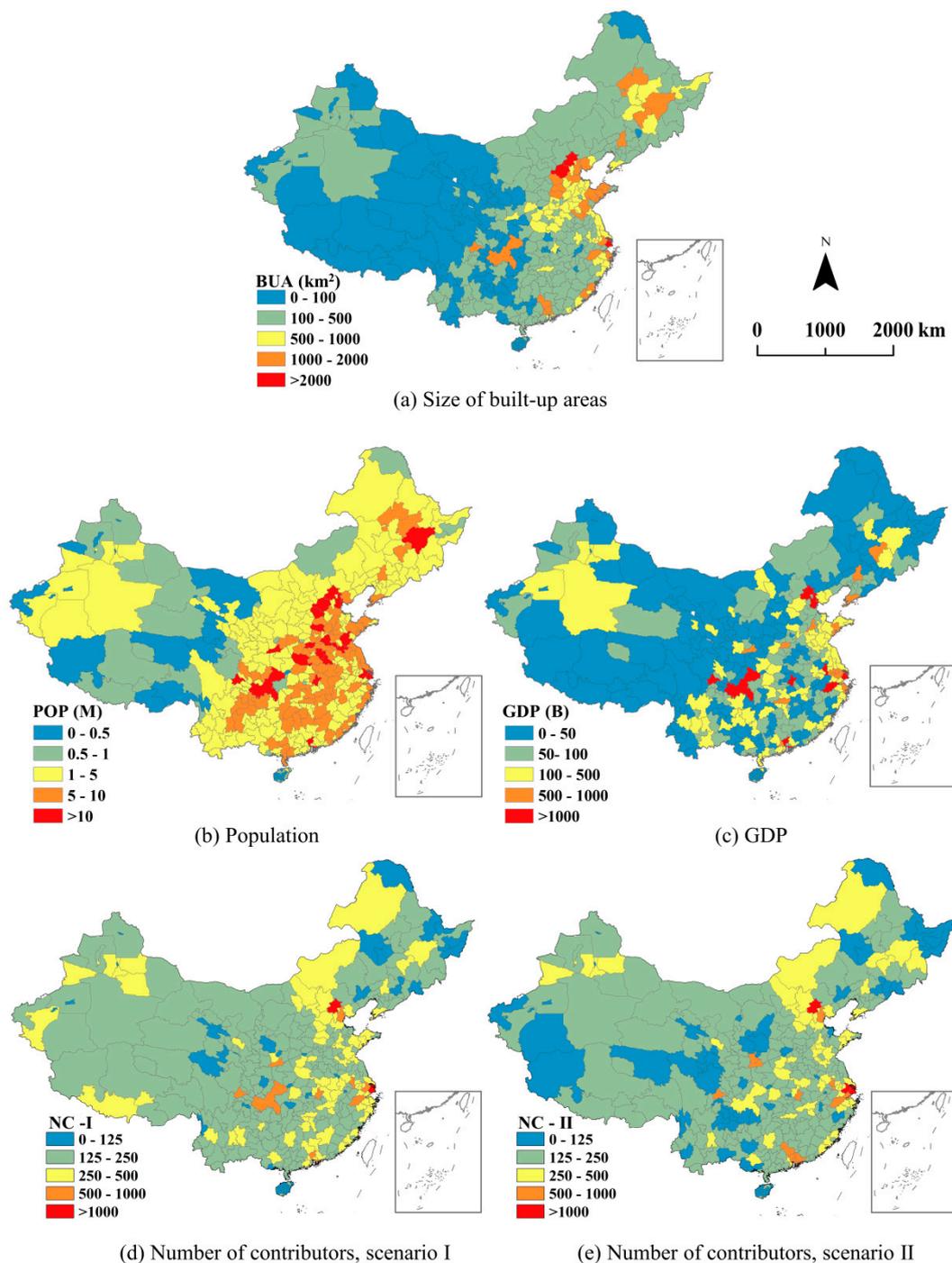
Significant correlations were observed between the completeness patterns and most factors (Table 4). Positive correlations were noted between Completeness-I, and BUA, GDP, and NC-I, and negative correlations were noted between Completeness-II, and BUA, POP, and NC-II. These results appear not to be fully consistent with those of past studies [25,27], which have reported the completeness of OSM data as being positively correlated with socioeconomic factors because there are more OSM contributors and associated mapping activities in dense populations and/or developed regions. The inconsistent results (for scenario II only) probably arose for two reasons: On the one hand, the sizes of the built-up areas in the west of China were relatively small (Figure 4a), and such areas might have been easily mapped by the volunteers. On the other hand, most built-up areas with high completeness values were mapped only with a few residential lands with few other LULC classes (e.g., Figure 1b). This was probably because the west of China has a relatively low population (Figure 4b), and thus few contributors (Figure 4d,e).

**Table 4.** Correlations between multiple factors and different patterns.

Pattern	Size of Built-Up Areas (BUA)	Population (POP)	Gross Domestic Product (GDP)	Number of Contributors, Scenario I (NC-I)	Number of Contributors, Scenario II (NC-II)
Completeness-I (completeness; scenario I)	0.115 *	−0.006	0.157 **	0.116 *	/
Diversity-I (diversity; scenario I)	0.285 **	0.282 **	0.248 **	0.249 **	/
Completeness-II (completeness; scenario II)	−0.262 **	−0.245 **	−0.029	/	−0.113 *
Diversity-II (diversity; scenario II)	0.367 **	0.375 **	0.412 **	/	0.429 **
Number of contributors, scenario I	0.681 **	0.659 **	0.855 **	1.000 **	1.000 **
Number of contributors, scenario II	0.722 **	0.645 **	0.799 **	1.000 **	1.000 **

\*\*  $p$ -value < 0.01; \*  $p$ -value < 0.05. “/” denotes “not applicable”.

Significant positive correlations were obtained between each of the diversity patterns (Diversity-I and Diversity-II) and all factors listed in Table 4. Moreover, their correlation coefficients were higher than those for the completeness patterns (Completeness-I and Completeness-II). This is because most highly diverse prefecture-level divisions of China were located in regions with a relatively high population (Figure 4b and GDP Figure 4c). More importantly, studies have report that socioeconomic factors are positively correlated with both the number of contributors and the density of their mapping data [18,36]. In this study, a moderate or high correlation was also found between each socioeconomic factor and the number of contributors. Therefore, diversity appeared to be better than completeness in reflecting the spatial pattern of contributors and their mapping activities. Furthermore, the correlation coefficients were higher for Diversity-II, probably because there were more contributors in built-up areas than in non-built-up areas.



**Figure 4.** Spatial patterns of various factors of China: (a) Size of built-up areas; (b) population; (c) GDP; (d) number of contributors (scenario I); and (e) number of contributors (scenario II).

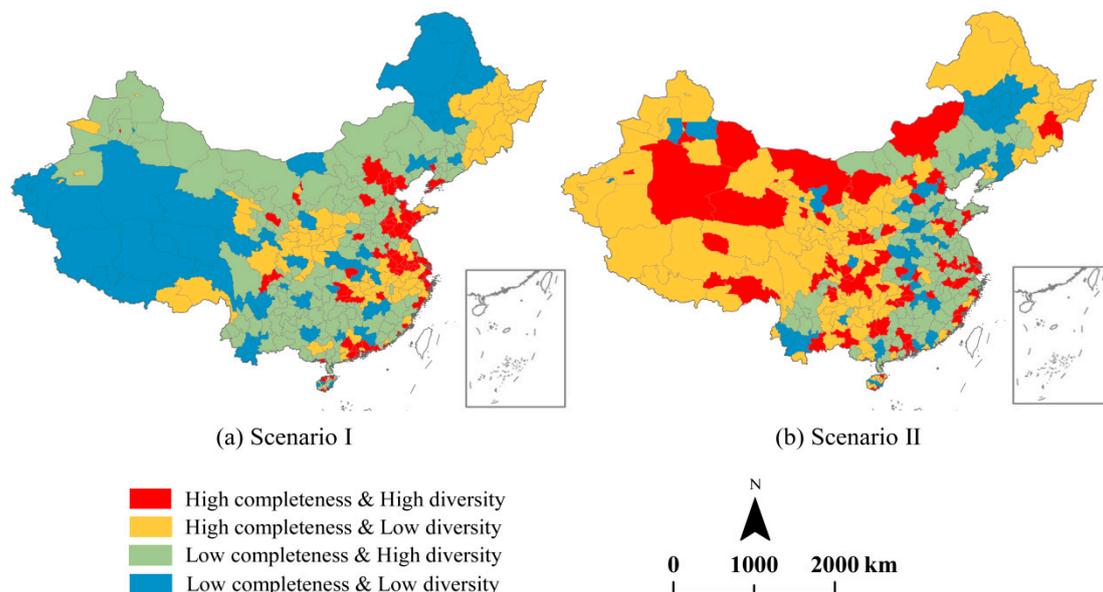
## 5. Combined Completeness and Diversity Patterns

While opposite trends were also observed (Figure 2), the correlations between completeness and diversity patterns were not very high, i.e., the correlation coefficients were  $-0.441$  and  $-0.330$  for scenarios I and II, respectively. It is therefore helpful to combine the completeness and the diversity patterns. The combined pattern has been plotted by referring to Zhang et al. [23]: The data distribution of each measure was checked, and if the data did not follow a normal distribution, they were subjected to a log-transformation. Subsequently, all the prefecture-level divisions of China were divided into four different groups:

- Group I (High completeness and high diversity): The completeness was higher than a certain threshold ( $A_c$ ), as was the diversity ( $A_d$ ).
- Group II (High completeness and low diversity): The completeness was higher than  $A_c$ , but the diversity was lower than  $A_d$ .
- Group III (Low completeness and high diversity): The completeness was lower than  $A_c$ , but the diversity was higher than  $A_d$ .
- Group IV (Low completeness and low diversity): The completeness was lower than  $A_c$ , and the diversity was lower than  $A_d$ .

The thresholds  $A_c$  and  $A_d$  denote the average completeness and the average diversity of all prefecture-level divisions, respectively.

In our study, only the completeness values were transformed with a log-transformation because they did not follow a normal distribution. We plotted the combined patterns for different scenarios (Figure 5), where the thresholds  $A_c$  and  $A_d$  were 5.09% and 1.16% for scenario I, and 18.84% and 1.10% for scenario II, respectively. The thresholds when using the average (completeness and diversity) were similar to those when using the median, and thus only the averages were used in the analysis. Moreover, the completeness values were low for most prefecture-level divisions of China.



**Figure 5.** Combined completeness and diversity patterns. (a) Scenario I and (b) scenario II.

The four groups for each scenario are shown in Figure 5:

- Group I: For both scenarios I and II, most prefecture-level divisions were municipalities, e.g., Beijing (Figure 6a), Shanghai, and Tianjin, capital cities, e.g., Guangzhou, Nanjing, Chengdu, and Changsha, and the relatively developed cities (Shenzhen, Qingdao and Xiamen), and regions on the east coast. These divisions probably received more attention from the volunteers, and thus both their completeness and diversity values were relatively high.
- Group II: The prefecture-level divisions of this group varied across scenario. In scenario I, most divisions were located in the east, center, and northeast of China owing to a large area percentage of forest. In scenario II, they were located in the southwest, northwest, and northeast of China owing to a large area percentage of residential land. This was nearly consistent with what is shown in Figure 2a,c.
- Groups III and IV: The prefecture-level divisions of these groups had low completeness values for several reasons: Some divisions (e.g., Haixi and Naqu) were characterized by a large land area, and thus the volunteers would have needed more time and effort to map these divisions

well. In addition, some divisions (e.g., Leshan (Figure 6b) and Songyuan (Figure 6c)) were less well known, especially compared with those in Group I, and thus would probably have received less attention by volunteers. Furthermore, most divisions showed a relatively high diversity value, which indicates that in most cases was no dominant LULC class. However, some divisions (e.g., Figure 6c) featured a relatively large area percentage of water (79.3% for the left graph in Figure 6c) or residential lands (85.6% for the right graph in Figure 6c), which resulted in a low diversity value.

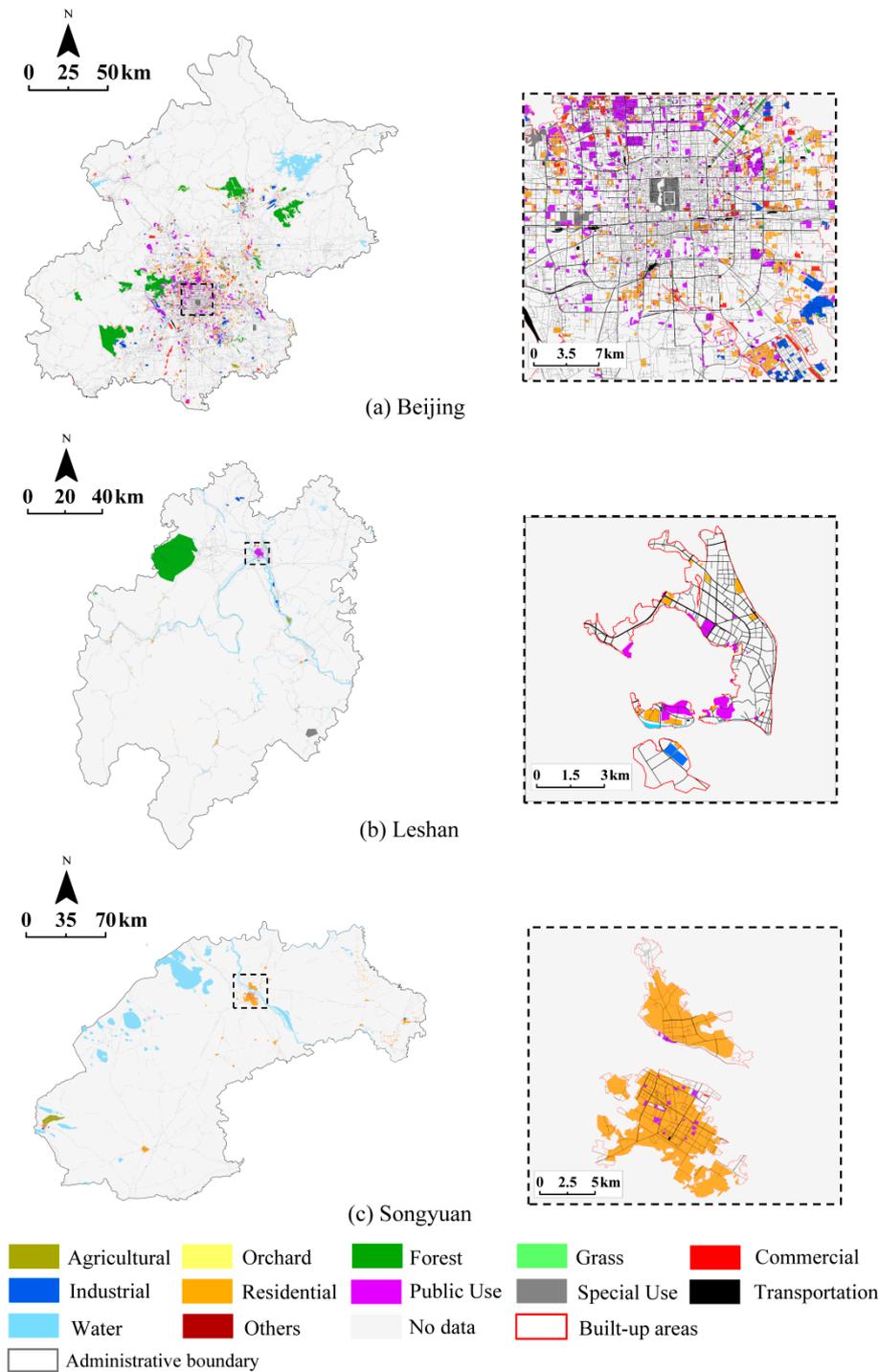


Figure 6. OSM-based LULC maps of some typical cities: (a) Beijing, (b) Leshan, and (c) Songyuan.

## 6. Discussion

### 6.1. Quality Measures

This study used three measures (accuracy, completeness, and diversity) for an intrinsic quality assessment of the OSM-based LULC dataset for China. By contrast, most past studies [19,20,27] not only employed an LULC reference dataset for comparison with an OSM-based dataset, but also considered two (accuracy and completeness) of these measures only. We think that there are several advantages of considering the diversity for an intrinsic quality assessment of an OSM-based LULC dataset.

First, the diversity patterns of OSM are significantly different from its completeness patterns (Figure 2). For instance, a highly complete prefecture-level division must not be high in terms of diversity because the diversity of LULC classes may vary across regions. As an example, the diversity value was commonly higher in built-up areas than in non-built-up areas. More importantly, the diversity of an OSM-based LULC dataset is also related to the contributors' mapping activities. For instance, we found that some built-up areas had been edited as a few large polygons and tagged with the OSM type "residential" (Figure 6c), but others had been edited with more details (e.g., as a greater number of small polygons) and tagged with different OSM types (residential, industrial, and public use, Figure 6a,b). Thus, it is useful to detect the diversity of OSM types by employing the diversity measure.

Second, the diversity patterns showed significantly positive correlations with all socioeconomic factors (size of built-up areas, population, and GDP in Table 4). By contrast, the completeness pattern (scenario II) showed a significantly negative correlation with both the size of the built-up areas and the population. Moreover, socioeconomic factors were positively correlated with the number of contributors (Table 4) and, probably, the density of their mapping data [18,36]. Some built-up areas with high completeness values for China were mapped with relatively large residential lands, but with few details from other classes, probably because such built-up areas had received less attention from contributors. However, most highly complete and highly diverse prefecture-level divisions of China probably received more attention from contributors. Thus, it is useful to consider diversity for understanding the spatial pattern of the contributors and their mapping activities.

Despite the above advantages, we suggest using the diversity measure as a supplement, rather than a way to simply conclude that diversity is positively correlated with the quality of a dataset. This is because the actual diversity value might have varied across regions. Moreover, an accuracy assessment may not be necessary because past studies have verified that the accuracy of OSM data is relatively high [19,26,27], which was also the case for our study area.

### 6.2. Applications

The proposed method offers potential for use in the service of understanding both the completeness and the diversity patterns of regions. For instance, an OSM user may not only use highly complete prefecture-level divisions to produce an OSM-based LULC map, but also highly diverse divisions (even if some of them are incomplete). This is because a diversity of LULC classes may be used as training and/or validation samples [37], which in turn can also be used for generating an LULC map, probably with another data source, e.g., remote sensing data [16,38], POIs [39], and/or street views [40]. An OSM contributor may not only use divisions that have low completeness values, but also those with low diversity values (even if some of them have relatively high completeness values). This is because highly complete regions might also not have been mapped well. Therefore, OSM users and contributors can use combined patterns to better understand OSM-based LULC datasets.

### 6.3. Limitations

This study is limited by a lack of available reference datasets, which applies to most countries and regions of the world. As a result, it was difficult to calculate both the accuracy and the actual diversity values for each prefecture-level division. While global LC products (e.g., GlobalLandcover30 and Climate Change Initiative Land Cover) are becoming increasingly available, they provide only

LC and not LU data; thus, various LU classes, such as commercial, industrial, public use, residential, and transportation, are not obtainable. The LULC maps and datasets provided by both mapping agencies and commercial companies are not available for free for China. Nevertheless, both the completeness and the diversity patterns of an OSM-based LULC dataset can be calculated without using any reference. Therefore, our analytical method for understanding spatial patterns can still be applied to other countries and regions, especially for those without LULC reference datasets available for free.

## 7. Conclusions

This study carried out an intrinsic quality assessment of OSM-based LULC datasets by understanding both the completeness and the diversity patterns of such datasets. An OSM-based LULC dataset was first generated and validated with various accuracy measures, and was then mapped and analyzed in terms of both completeness and diversity. China, which has rarely been studied before, and is characterized by a lack of a freely available LULC reference dataset, was chosen as the study area. Both the completeness and the diversity patterns of the generated dataset were analyzed for two scenarios. In scenario I, the entire land area of each prefecture-level division was analyzed; in scenario II, only the built-up areas of each prefecture-level division were analyzed. Moreover, the correlations between both patterns (completeness and diversity) and the three socioeconomic factors (built-up area size, population, and GDP), and the number of contributors were investigated. The results showed the following:

- (1) The OA of the OSM-based LULC dataset of China was as high as 82.2%, which illustrates that the generated LULC dataset for the country was effective, and is comparable to those for European study areas in past work.
- (2) Both the completeness and the diversity patterns varied with prefecture-level division. Moreover, the completeness patterns were significantly different from the corresponding diversity patterns. In particular at the scale of built-up areas, divisions with high completeness values might not have been mapped well owing to a low diversity value.
- (3) The correlations between diversity patterns and each of the three socioeconomic factors, and the number of contributors were not only higher than those considering for completeness patterns, but also significantly positive. Thus, the diversity pattern is a better reflection of socioeconomic factors and the spatial pattern of contributors.
- (4) Both the completeness and the diversity patterns can be combined into different groups (high completeness and high diversity, high completeness and low diversity, low completeness and high diversity, and low completeness and low diversity). The combined patterns benefit both OSM users and volunteers in that they provide a better understanding of OSM-based LULC datasets.

The above results indicate that it is useful to consider diversity as a supplement for the intrinsic quality assessment of an OSM-based LULC dataset. While only China was investigated here, the analytical method proposed in this study can be used for understanding the spatial patterns of an OSM-based LULC dataset in other countries and regions.

In future work, first, it is useful to employ an LULC reference dataset to further assess the quality of the OSM-based LULC dataset for China. Second, the Shannon diversity index alone was employed in our study, but some other diversity indices [34] can be considered. It should be informative to use the analytical method proposed here to intrinsically assess the quality of an OSM-based LULC dataset for other countries and regions.

**Author Contributions:** Conceptualization, Qi Zhou; Formal analysis, ShuZhu Wang and YuanJian Tian; Writing—original draft, Qi Zhou and ShuZhu Wang. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project was supported by the National Natural Science Foundation of China [grant number 41771428], the Fundamental Research Funds for the Central Universities, and the China University of Geosciences (Wuhan) [grant number CUGESI1801].

**Acknowledgments:** The authors thank all anonymous reviewers and the editor for their valuable comments and suggestions that have helped improve this paper substantially.

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

## References

1. Jones, D.A.; Hansen, A.J.; Bly, K.; Doherty, K.; Verschuyf, J.P.; Paugh, J.I.; Carle, R.; Story, S.J. Monitoring land use and cover around parks: A conceptual approach. *Remote Sens. Environ.* **2009**, *113*, 1346–1356. [\[CrossRef\]](#)
2. Liang, J.; Zhong, M.; Zeng, G.; Chen, G.; Hua, S.; Li, X.; Yuan, Y.; Wu, H.; Gao, X. Risk management for optimal land use planning integrating ecosystem services values: A case study in Changsha, Middle China. *Sci. Total Environ.* **2017**, *579*, 1675–1682. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Mosadeghi, R.; Warnken, J.; Tomlinson, R.; Mirfendersesk, H. Comparison of fuzzy-AHP and AHP in a spatial multi-criteria decision making model for urban land-use planning. *Comput. Environ. Urban Syst.* **2015**, *49*, 54–65. [\[CrossRef\]](#)
4. Reisi, M.; Aye, L.; Rajabifard, A.; Ngo, T. Land-use planning: Implications for transport sustainability. *Land Use Policy* **2016**, *50*, 252–261. [\[CrossRef\]](#)
5. Hegazy, I.R.; Kaloop, M.R. Monitoring urban growth and land use change detection with GIS and remote sensing techniques in Daqahlia governorate Egypt. *Int. J. Sustain. Built Environ.* **2015**, *44*, 117–124. [\[CrossRef\]](#)
6. Rimal, B.; Zhang, L.; Keshtkar, H.; Hacck, B.N.; Rijal, S.; Zhang, P. Land use/land cover dynamics and modeling of urban land expansion by the integration of cellular automata and Markov chain. *Int. J. Geo-Inf.* **2018**, *77*, 154. [\[CrossRef\]](#)
7. Zeng, C.; Liu, Y.; Stein, A.L.; Jiao, L. Characterization and spatial modeling of urban sprawl in the Wuhan metropolitan area, China. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 10–24. [\[CrossRef\]](#)
8. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X.; He, C.; Han, G.; Peng, S.; Lu, M.; et al. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 7–27. [\[CrossRef\]](#)
9. Grekousis, G.; Mountrakis, G.; Kavouras, M. An overview of 21 global and 43 regional land-cover mapping products. *Int. J. Remote Sens.* **2015**, *36*, 5309–5335. [\[CrossRef\]](#)
10. Fritz, S.; See, L.; Perger, C.; McCallum, I.; Schill, C.; Schepaschenko, D.; Duerauer, M.; Karner, M.; Dresel, C.; Laso-Bayas, J.C.; et al. A global dataset of crowdsourced land cover and land use reference data. *Sci. Data* **2017**, *4*, 170075. [\[CrossRef\]](#)
11. 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. [\[CrossRef\]](#)
12. Kang, J.; Körner, M.; Wang, Y.; Taubenböck, H.; Zhu, X.X. Building instance classification using street view images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 44–59. [\[CrossRef\]](#)
13. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007. [\[CrossRef\]](#)
14. Goodchild, M. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [\[CrossRef\]](#)
15. Arsanjani, J.J.; Helbich, M.; Bakillah, M.; Hagenauer, J. Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2264–2278. [\[CrossRef\]](#)
16. Schultz, M.; Voss, J.; Auer, M.; Carter, S.; Zipf, A. Open land cover from OpenStreetMap and remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *63*, 206–213. [\[CrossRef\]](#)
17. Zhou, Q.; Jia, X.; Lin, H. An approach for establishing correspondence between OpenStreetMap and reference datasets for land use and land cover mapping. *Trans. GIS* **2019**, *23*, 1177–1464. [\[CrossRef\]](#)
18. Shen, J.; Yang, S.; Zang, K.; Zhou, J.; Fan, H.; Mao, B. The opinion of crowd participatory for OpenStreetMap: A survey in China. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 1525–1531. [\[CrossRef\]](#)
19. Arsanjani, J.J.; Vaz, E. An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *35*, 329–337. [\[CrossRef\]](#)

20. Viana, C.M.; Encalada, L.; Rocha, J. The value of OpenStreetMap historical contributions as a source of sampling data for multi-temporal land use/cover maps. *Int. J. Geo-Inf.* **2019**, *8*, 116. [[CrossRef](#)]
21. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [[CrossRef](#)]
22. Girres, J.F.; Touya, G. Quality assessment of the French OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459. [[CrossRef](#)]
23. Zhang, Y.; Li, X.; Wang, A.; Bao, T.; Tian, S. Density and diversity of OpenStreetMap road networks in China. *J. Urban Manag.* **2015**, *44*, 135–146. [[CrossRef](#)]
24. Zhou, Q. Exploring the relationship between density and completeness of urban building data in OpenStreetMap for quality estimation. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 257–281. [[CrossRef](#)]
25. Tian, Y.; Zhou, Q.; Fu, X. An analysis of the evolution, completeness and spatial patterns of OpenStreetMap building data in China. *ISPRS Int. J. Geo-Inf.* **2019**, *88*, 35. [[CrossRef](#)]
26. Estima, J.; Painho, M. Investigating the potential of OpenStreetMap for land use/land cover production: A case study for continental Portugal. In *OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography*; Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015.
27. Dorn, H.; Törnros, T.; Zipf, A. Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. [[CrossRef](#)]
28. See, L.; Comber, A.; Salk, C.; Fritz, S.; Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; Obsersteiner, M. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE* **2013**, *88*, e69958. [[CrossRef](#)]
29. Comber, A.; See, L.; Fritz, S.; Velde, M.; Perger, C.; Foody, G. Using control data to determine the reliability of volunteered geographic information about land cover. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *23*, 37–48. [[CrossRef](#)]
30. Zhou, Q.; Tian, Y.J. The use of geometric indicators to estimate the quantitative completeness of street blocks in OpenStreetMap. *Trans. GIS* **2018**, *22*, 1550–1572. [[CrossRef](#)]
31. Hristova, D.; Williams, M.; Musolesi, M.; Panzarasa, P.; Mascolo, C. Measuring urban social diversity using interconnected geo-social networks. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016.
32. Vogiatzakis, L.N.; Manolaki, P. Investigating the diversity and variability of Eastern Mediterranean Landscapes. *Land* **2017**, *66*, 71. [[CrossRef](#)]
33. Velazquez, J.; Gutierrez, J.; Hernando, A.; Garcia-Abril, A.; Martin, M.A.; Irastorza, P. Measuring mosaic diversity based on land use map in the region of Madrid, Spain. *Land Use Policy* **2018**, *71*, 329–334. [[CrossRef](#)]
34. Nagendra, H. Opposite trends in response for the Shannon and Simpson indices of landscape diversity. *Appl. Geogr.* **2002**, *22*, 175–186. [[CrossRef](#)]
35. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Champaign, IL, USA, 1949.
36. Neis, P.; Zielstra, D.; Zipf, A. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet* **2013**, *5*, 282–300. [[CrossRef](#)]
37. Fonte, C.C.; Martinho, N. Assessing the applicability of OpenStreetMap data to assist the validation of land use/land cover maps. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2382–2400. [[CrossRef](#)]
38. Johnson, B.A.; Lizuka, K. Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna Bay area of the Philippines. *Appl. Geogr.* **2016**, *67*, 140–149. [[CrossRef](#)]
39. 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](#)]
40. Srivastava, S.; Lobry, S.; Tuia, D.; Vargas-Muñoz, J. Land-use characterisation using Google Street View pictures and OpenStreetMap. In Proceedings of the Association of Geographic Information Laboratories in Europe Conference (AGILE), Lund, Sweden, 12–15 June 2018.

