CREATING A LAND USE/LAND COVER DICTIONARY BASED ON MULTIPLE PAIRS OF OSM AND REFERENCE DATASETS

ShuZhu Wang [1], Qi Zhou [1*], YaoMing Liu [1]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan, P.R. China - (wangshuzhu, zhouqi, 20171001314)@cug.edu.cn

Commission IV, WG IV/4

KEY WORDS: Land Use/Land cover, Mapping, OpenStreetMap, Dictionary, Reference dataset.

ABSTRACT:

OpenStreetMap (OSM) can supply useful information to improve land use/land cover (LULC) mapping. A dictionary is needed to convert each OSM tag into a LULC class. However, such a dictionary was mostly created subjectively or with only one pair of OSM and reference datasets. As a result, the existing dictionaries may not be applicable to other study areas. This study designed four measures: sample count, average area percentage, sample ratio and average maximum percentage; and used multiple pair of OSM and reference datasets to create a dictionary. 50 pan-European metropolitans were involved for testing and 1409 different OSM tags were found. We further found that: (1) Only a small proportion of OSM tags play a decisive role for LULC mapping. (2) An OSM tag may correspond to multiple different LULC classes, but the issue that which and how different LULC classes correspond to each OSM tag can be determined. Moreover, not only the proposed dictionary is useful for various applications, e.g., producing LULC maps and assessing the quality of an OSM dataset, but also the approach can be applicable to different study areas and/or LULC datasets.

1. INTRODUCTION

Land use (LU) and land cover (LC) refers to the classification of human activities and natural elements on the surface of the earth. LU refers to how people use the land, e.g., residential and commercial are common LU types. LC refers to how much a region is covered by natural features, e.g., forests, agriculture, artificial surfaces and bodies of water. Mapping land use/land cover (LULC) has been widely applied to natural resource management (Liang, 2017; Chen et al. 2019), LULC change modelling (Rimal et al. 2018; Verburg et al. 2019), urban and transportation planning (Cai et al. 2019; Long et al. 2022), and urban sprawls monitoring (Zeng et al. 2015; Xia et al. 2020).

Various data sources have been used for LULC mapping. The technology of remote sensing may be the most useful one (Chen et al. 2015), because it has benefits in detecting physical objects (e.g., roads, buildings and rivers) on the surface of the earth. However, median- and low- resolution remote sensing data are only useful for mapping LC rather than LU, because various LU types (e.g., residential, commercial and industrial lands) can hardly be sensed from the data (Fritz et al. 2017). Although high-resolution remote sensing data are useful for LU mapping, it is still expensive to purchase the data. As an alternative, various geographic data edited or provided by volunteers, called volunteered geographic information or VGI (Goodchild 2007), has also been used for LULC mapping. OpenStreetMap (OSM), as one of the most successful VGI projects in the world, has recently received increasing attention. The advantage of using OSM data includes: firstly, the data are freely available; secondly, they cover almost all the countries or regions in the world; and thirdly, the data are being updated on a minute by minute basis. Although many concerns have paid attention to the quality of OSM data (Haklay 2010; Zhou 2018; Zhang et al. 2022), existing studies have verified that an OSM-based LULC dataset can have a substantial classification accuracy compared with a corresponding reference dataset. For instance, Estima and Painho (2015) investigated the potential of using OSM data for LULC production, of which the OA was 76.7%. Arsanjani and Vaz (2015) assessed the OA of using OSM data for LU classification in seven European metropolitans, and they found that six out of the seven metropolitans had an OA higher than 75%. Dorn et al. (2015) reported a substantial agreement (kappa coefficient=0.61) between OSM and reference datasets. Moreover, OSM data can be combined with other data sources (e.g. remote sensing, POI, street view images) for LULC mapping (Johnson and Iizuka 2016; Liu and Long 2016; Schultz et al. 2017; Srivastava et al. 2018).

Commonly, OSM data consist of the three map elements, e.g., nodes, ways and relations. Nodes are used to describe point features; ways are used to describe linear features and area boundaries; and relations are used to explain how different map elements work together (e.g., an island is in a lake). In addition, tags are used to describe specific features of map elements. Each tag consists of two items, i.e., a key and a value. There is a need to convert the value of an OSM tag into an LULC class. This is because: in an OSM dataset, there are hundreds of different OSM tags or values; but in a reference dataset, there are much few LULC classes. Thus, it is needed to create an OSM-LULC dictionary, in order to establish the correspondence between each pair of OSM tag (denoted as the value of an OSM tag) and LULC class. However, such a dictionary was subjectively established in most studies (Estima and Painho 2015; Dorn et al. 2015; Fonte and Martinho 2017). Although Zhou et al. (2019) proposed an automated approach, they only used one pair of OSM and reference datasets for creating an OSM-LULC dictionary. As a result, the created dictionary based on one study area may not always be applicable to others. This is because that the number of different OSM tags varies in different study areas. For instance, it has been reported by Zhou et al. (2019) that the number of different OSM tags was 481 for London, but this number was only 178 for Sheffield. Moreover, an OSM tag may correspond to different LULC classes within a study area and the most

* Corresponding author

appropriate LULC class for each OSM tag may also vary in different study areas. As found in our study, the OSM tag *nature_reserve* may correspond to multiple different LULC classes, e.g., Forests and Water. Furthermore, to the best of our knowledge, very few studies have investigated that: (1) As there usually are hundreds of different OSM tags in a dataset, must all them be considered for LULC mapping? If not, which OSM tags play a decisive role (e.g., more common or have a relatively larger land area)? (2) Which and how OSM tags correspond to different LULC classes? Which LULC classes correspond to each OSM tag? An understanding of these questions is significant to use OSM data for LULC mapping.

The tenet of our approach is to use multiple pairs of OSM and reference datasets for creating an OSM-LULC dictionary. For each dataset, the most appropriate LULC class for each OSM tag was first determined. Then, an OSM-LULC dictionary was created by calculating four measures (i.e., sample count, average area percentage, sample ratio and average maximum percentage) and based on multiple pairs of OSM and reference datasets.

This study has three main contributions. Firstly, an approach to creating an OSM-LULC dictionary based on multiple pairs of OSM and reference datasets is proposed. This approach can also be applicable to different study areas. Secondly, an OSM-LULC dictionary is created based on 50 different pan-European metropolitans, and it includes 1409 different OSM tags. This dictionary may be used for the determination of an appropriate LULC class for these OSM tags in a different study area. Thirdly, from an analysis of the created dictionary, we found that a small proportion of OSM tags play a decisive role for LULC mapping. Moreover, the issue that which and how different LULC classes correspond to each OSM tag can be determined. These findings may be useful for not only producing LULC maps, but also obtaining training and/or validation samples.

This study is structured as follows: Section 2 introduces the approach to creating an OSM-LULC dictionary; section 3 presents study area and data, and also experimental steps; section 4 reports an experimental analysis of the created dictionary; sections 5 highlights various applications of this dictionary; section 6 discusses the limitations of this study and future work; and section 7 is the conclusion.

2. APPROACH TO CREATING AN OSM-LULC DICTIONARY

The tenet of our approach is to use multiple pairs of OSM and reference datasets for creating an OSM-LULC dictionary. In each pair of datasets, an OSM tag may correspond to different LULC classes, it is therefore necessary to determine which is the most appropriate LULC class for each OSM tag. As existing studies have reported that a substantial OA can be achieved comparing between OSM and reference datasets (Arsanjani et al. 2013; Estima and Painho 2015; Arsanjani and Vaz 2015; Zhou et al. 2019), we assumed that most OSM tags have been tagged by volunteers correctly. Following this assumption, the way to determine the most appropriate LULC class for each OSM tag includes two steps. Firstly, all objects of an OSM tag are intersected with those of different LULC classes, respectively. After that, the LULC class with the maximum intersecting area is viewed as the most appropriate one for this OSM tag. As an example, Figure 1 shows three pairs of schematic OSM and reference datasets. In the first pair of datasets (I), there are three objects tagged with 'a', 'b' and 'c' in the OSM dataset (Figure 1a) and three objects tagged with the LULC classes 'A', 'B' and 'C' in the reference dataset (Figure 1b). According to our approach, the OSM tag 'a' should correspond to the LULC class 'A' (Figure 1c) because their intersecting area is the maximum (40%). Similarity, the OSM tags 'b' and 'c' should correspond to the LULC classes 'B' and 'C', respectively.

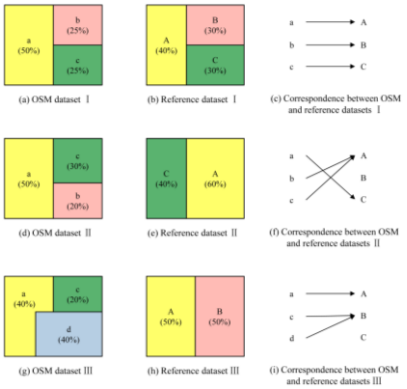


Figure 1. The approach to establishing a correspondence between each pair of OSM and reference datasets.

Theoretically, it is feasible to use only one pair of OSM and reference datasets for creating the OSM-LULC dictionary. But, we considered multiple pairs of datasets for two reasons. On the one hand, the most appropriate LULC class for each OSM tag may vary in different datasets. For instance, the most appropriate LULC class for the OSM tag 'a' is 'A' in the dataset I (Figure 1a-1c), but the most appropriate LULC class for this tag is 'C' in the dataset II (Figure 1d-1f). On the other hand, the number of different OSM tags may also vary in different datasets. For instance, in both the datasets I and II, the three OSM tags are 'a', 'b' and 'c'. But in the dataset III, the three OSM tags are 'a', 'c' and 'd'. Commonly, the number of different OSM tags may be positively correlated with the number of datasets analyzed. Furthermore, the three pairs of OSM and reference datasets (in Figure 1) are used to illustrate how to create an OSM-LULC dictionary. As there are three pairs of datasets, the total number of samples or datasets is three (denoted as Total samples=3).

Four attributes and four measures are designed to describe an OSM-LULC dictionary (Table 1). They are: Tag ID, Tag Name, Class ID and Class Name in terms of attributes; and Sample Count, Average Area Percentage, Sample Ratio and Average Maximum Percentage in terms of measures. First of all, the four attributes are introduced as follows.

(1) ***Tag ID:*** denotes the ID of an OSM tag. In Table 1, there are a total of four different OSM tags ('a', 'b', 'c' and 'd'). Thus, the Tag ID begins from '1' to '4'. Each OSM tag has a unique Tag ID.

(2) ***Tag name:*** denotes the name of an OSM tag. In Table 1, the four OSM tags are named as 'a', 'b', 'c' and 'd', respectively.

(3) ***Class ID:*** denotes the ID of an LULC class. In Table 1, there are a total of three different LULC classes ('A', 'B' and 'C'). Thus, the Class ID begins from '1' to '3'. Each LULC class has n unique Class ID.

(4) ***Class name:*** denotes the name of an LULC class. In Table 1, the three LULC classes are named as 'A', 'B', and 'C', respectively. The above four attributes are used to describe the ID and Name of an OSM tag or LULC class. The blew four measures are used to describe the area and count characteristics of an OSM tag and/or its correspondence with each LULC class.

(5) ***Sample count (SC):*** denotes how frequent an OSM tag 't' is appeared in different study areas or datasets (denoted as SC(t)). As an example, the SC for the OSM tag 'a' is 3 because the OSM tag 'a' was appeared in all the three datasets (I, II and III). A large SC indicates that an OSM tag is common among different study areas or datasets.

Class ID				1		2		3	
Class Name				A		B		C	
Tag ID	***Tag Name***	***SC***	***AAP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***
1	a	3	50	67	90			33	80
2	b	2	22.5	50	100	50	100		
3	c	3	25	33	100	33	100	33	100
4	d	1	30			100	75		

Table 1. A schematic OSM-LULC dictionary (Total samples=3).

(6) ***Average area percentage (AAP):*** denotes the average of the area percentages of an OSM tag ('t') in multiple different OSM datasets (denoted as $AAP(t)$).

$$AAP(t) = \frac{\sum_{i=0}^{SC(t)} AP(t,i)}{SC(t)} \quad (1)$$

Where, $AP(t,i)$ denotes the area percentage (*AP*) of an OSM tag 't' in the i^{i} OSM dataset. In Figure 1, the *AP* for the OSM tag 'c' is 25%, 30% and 20% respectively in the datasets I, II and III. Thus, the average of these area percentages (*AAP*) is 25%= (25%+30%+20%)/3. A large *AAP* indicates that the *AP* for an OSM tag is relatively high in an OSM dataset.

(7) ***Sample ratio (SR):*** denotes the percentage of study areas or datasets that an OSM tag 't' corresponds to an LULC class 'c' (denoted as $SR(t,c)$).

$$SR(t,c) = \frac{SC(t,c)}{SC(t)} \quad (2)$$

Where, $SR(t,\ c)$ denotes the number of study areas or datasets that an OSM tag 't' corresponds to an LULC class 'c'. In Figure 1, the OSM tag 'a' corresponds to the LULC class 'A' in both the datasets I and III (Figure 1c and 1i), but the OSM tag 'a' corresponds to the class 'C' in the dataset II. Thus, the *SR* is approximately 67% (2/3) for the pair of OSM tag 'a' and LULC class 'A'; and 33% (1/3) for the pair of OSM tag 'a' and LULC class 'C'. A relatively large *SR* indicates that an OSM tag corresponds to an LULC class in most cases.

(8) Average maximum percentage (AMP): Within a pair of OSM and reference datasets, the objects for an OSM tag may intersect with those for multiple LULC classes. For each OSM tag, there is at least one pair of OSM tag and LULC class, whose objects are intersected with the maximum area. The maximum percentage (denoted as $MP(t,c)$) denotes the ratio of the maximum area for a pair of OSM tag 't' and LULC class 'C' to the total area for the OSM tag 't'. The *AMP* (t,c) denotes the average of all the $MP(t,c)$ in different study areas or datasets.

$$AMP(t,c) = \frac{\sum_{i=0}^{SC(t)} MP(t,c)}{SC(t,c)} \quad (3)$$

In Figure 1, the OSM tag 'a' corresponds to the LULC class 'A' in both the datasets I and III. In the dataset I, the $MP(a,A)$ is 80%=40%/50% (Figure 1a and 1b); and in the dataset III, the $MP(a,A)$ is 100%=40%/40% (Figure 1g and 1h). Thus, the $AMP(a,A)$ is 90%=(80%+100%)/2. A large *AMP* indicates that within a pair of OSM and reference datasets, an OSM tag is highly consistent with the most appropriate LULC class.

3. DESIGN OF EXPERIMENTS

3.1 Study area and data

A total of 50 metropolitans in pan-European region were chosen as study areas (Figure 2). First of all, such a large number of samples were chosen, in order to minimize the subjective of only using a few study areas to create the OSM-LULC dictionary. Then, all the reference datasets for these 50 metropolitans were freely available, and they are required for creating an OSM-LULC dictionary. More important, multiple pairs of OSM and reference datasets were involved for investigating: Which OSM tags are more common or have relatively larger land areas? And which is/are the most appropriate LULC class (es) corresponded to each OSM tag? To be specific, two categories of datasets were required for creating an OSM-LULC dictionary.

- Category I: OSM datasets of the 50 metropolitans. These datasets were acquired from Geofabrik's free download server: http://download.geofabrik.de/index.html (accessed in June 2020). This server provided not only a list of countries for users to freely download OSM data of interest, but also standard data formats (shapefiles) to analyze OSM data in the GIS (Geographic Information System) software.
- Category II: Corresponding reference datasets of the 50 metropolitans. These datasets (called urban atlas or UA) were produced by European Environment Agency in 2012 and were freely available from the website https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012/# (accessed in June 2020).

Figure 2. Locations of the 50 pan-European metropolitans.

3.2 Experimental steps

The approach to creating an OSM-LULC dictionary was implemented in a commercial GIS software-ArcGIS (version 10.6). The general steps are listed as follows.

- Step 1: Download the OSM datasets of the 50 metropolitans from the Geofabrik's server. These datasets were organized into a number of layers, including buildings, landuse, natural, places, pofw, pois, traffic, transport, railways, roads, water and waterways. Most layers were represented by area features, although some were represented by point or line features (e.g. roads, railways and waterways). In this study, only the layers represented by area features were involved for the analysis because various LULC classes in the reference datasets were also represented by area features.
- Step 2: Calculate the two measures *SC* and *AAP* based on the 50 different metropolitans.
- Step 3: Download the corresponding reference datasets of these 50 metropolitans and determine 14 LULC classes for the analysis.

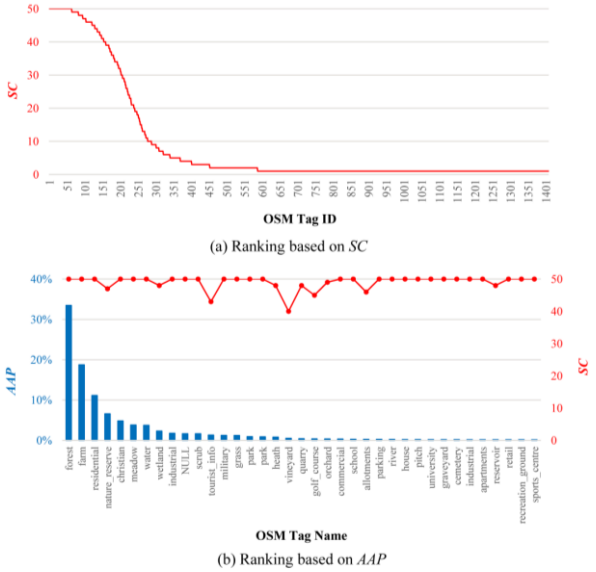
• Step 4: For each metropolitan, use the Tabulate Intersection tool to intersect each pair of OSM and reference datasets and determine the most appropriate LULC class for each OSM tag. The LULC class having the maximum intersecting area with this OSM tag was viewed as the most appropriate one.
• Step 5: Calculate the other two measures *SR* and *AMP* for each pair of OSM tag and LULC class.
• Step 6: Finally, create an OSM-LULC dictionary by referring to the Table 1.

4. RESULTS AND ANALYSIS

This section reports an analysis of the created OSM-LULC dictionary based on the four measures, i.e., *SC*, *AAP*, *SR* and *AMP*. The first two measures (*SC* and *AAP*) were used to investigate which OSM tags are common and/or have a relatively large land area. The other two measures (*SR* and *AMP*) were used to investigate whether and which different LULC classes correspond to each OSM tag.

4.1 Analysis of *SC* and *AAP*

A total of 1409 different OSM tags were found based on analyzing the 50 metropolitans. These OSM tags were ranked according to *SC* (Figure 3a) and *AAP* (Figure 3b), respectively.



(a) Ranking based on *SC*

(b) Ranking based on *AAP*

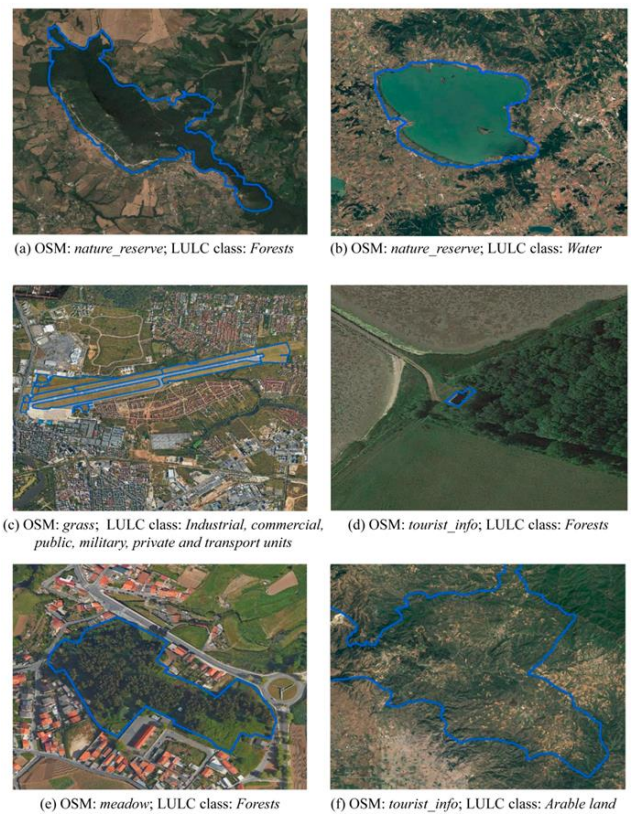
Figure 3. Ranking OSM tags according to *SC* (a) and *AAP* (b).

The ranking of the *SC* in Figure 3a approximately follows a long-tail distribution. That is, for most OSM tags, there was a relatively small *SC* value (e.g., <5). More precisely, for 1,000 of the 1,409 OSM tags, each was found in at most five different metropolitans (comparing with 50 metropolitans in total). In contrast, for less than 300 OSM tags, each was found in more than ten metropolitans. This indicates that a small proportion of OSM tags are common. In addition, for more than 800 OSM tags, the SC value is equal to one because these OSM tags can only be found in one metropolitan.
The ranking of the *AAP* in Figure 3b also follows a long-tail distribution. Only 37 of the 1409 OSM tags were plotted, in order to show their tag names. These OSM tags (37) all had a relatively large percentage of land areas (>0.1%). More precisely, the OSM tag *forest* has the maximum value (*AAP*=33%). But, for the fifth largest OSM tag *christian*, the *AAP* value is smaller than 5%; and for the 15th largest OSM tag *park*, this value is smaller than 1%. This illustrates that most OSM tags had a small percentage of land areas.

4.2 Analysis of *SR* and *AMP*

The OSM-LULC dictionary created based on the 50 metropolitans is listed in Table 2. Due to the limited space, only 20 of the 1409 OSM tags were listed, and they were also ranked on the top in terms of the *AAP*. It can be seen in Table 2 that:
(1) Most OSM tags can correspond to multiple different most appropriate LULC classes. For instance, the OSM tag *forest* can correspond to two LULC classes, i.e., *Forests* and *Herbaceous vegetation associations*; and the OSM tag *nature_reserve* can correspond to seven LULC classes, i.e., *Arable land*, *Artificial non-agricultural vegetated areas*, *Forests*, *Herbaceous vegetation associations*, *Pastures*, *Water* and *Wetland*.
(2) The *SR* varies with different pairs of OSM tag and LULC class. For instance, the *SR* is 96% for the pair of the OSM tag *forest* and the LULC class *Forests*. This means that the OSM tag *forest* was found corresponding to the LULC class *Forests* in 48 of the 50 metropolitans. However, for the pair of the OSM tag *nature_reserve* and different LULC classes, the maximum *SR* value is much lower (64%). It indicates that this OSM tag might correspond to different most appropriate LULC classes in different metropolitans.
(3) The *AMP* also varies with different pairs of OSM tag and LULC class. For instance, the *AMP* for the pair of the OSM tag *forest* and the LULC class *Forests* is 81%, which is also higher than that (60%) for the pair of the OSM tag *nature_reserve* and the LULC class *Forests*. This indicates that the consistency for the former pair is much higher than that for the latter one.

Moreover, the reasons for why an OSM tag may correspond to multiple different LULC classes were also investigated. As an example, Figure 4 shows some typical pairs of OSM tag and LULC class for the analysis, and the corresponding images in Google Earth were also given out as references.



(a) OSM: *nature_reserve*; LULC class: *Forests*

(b) OSM: *nature_reserve*; LULC class: *Water*

(c) OSM: *grass*; LULC class: *Industrial, commercial, public, military, private and transport units*

(d) OSM: *tourist_info*; LULC class: *Forests*

(e) OSM: *meadow*; LULC class: *Forests*

(f) OSM: *tourist_info*; LULC class: *Arable land*

Figure 4. Comparing some typical pairs of OSM tag and LULC class with corresponding images in Google Earth. Each OSM object was highlighted with a blue line frame.

Class ID					1		2		3		4		5		6		7		8		9		10		11		12		13		14	
Class Name					*Arable land*		*Artificial non-agricultural vegetated areas*		*Complex and mixed cultivation patterns*		*Forests*		*Herbaceous vegetation associations*		*Industrial, commercial, public, military, private and transport units*		*Mine, dump and construction sites*		*Orchards*		*Open spaces with little or no vegetation*		*Pastures*		*Permanent crops*		*Urban fabric*		*Water*		*Wetlands*	
Tag ID	***Layer***	***Tag Name***	***SC***	***AAP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***	***SR (%)***	***AMP (%)***
1	*landuse*	*forest*	50	33.46							96	81	4	49																		
2	*landuse*	*farm*	50	18.71	86	70			2	48													12	61								
3	*landuse*	*residential*	50	11.12																							100	71				
4	*landuse*	*nature_reserve*	47	6.55	9	52	2	30			64	60	11	50									4	34					4	65	6	47
5	*pofw*	*christian*	50	4.81	4	45	68	62			2	35	2	32	18	51							2	30			4	50				
6	*landuse*	*meadow*	50	3.79	30	46	2	97			2	30	2	78	4	50							60	58								
7	*water*	*water*	50	3.70																									100	83		
8	*water*	*wetland*	48	2.29	10	50					27	42	6	56									19	54					8	50	29	60
9	*landuse*	*industrial*	50	1.73											100	82																
10	*buildings*	*NULL*	50	1.62											20	51											80	61				
11	*landuse*	*scrub*	50	1.62	16	36	2	20			32	52	22	60							2	41	24	44			2	28				
12	*pois*	*tourist_info*	43	1.28	5	68	19	85			9	95			51	76	2	97					2	77			12	69				
13	*landuse*	*military*	50	1.21	8	42	2	54			28	63	4	46	54	62							4	53								
14	*landuse*	*grass*	50	1.18	10	30	2	38					2	36	52	50							30	47			4	45				
15	*landuse*	*park*	50	0.89			90	54			10	33																				
16	*pois*	*park*	50	0.85			90	55			10	33																				
17	*landuse*	*heath*	48	0.77	15	57	6	44			25	55	23	62	2	69					4	74	25	52								
18	*landuse*	*vineyard*	40	0.48	30	67			5	68	3	55			3	85							23	65	35	64	3	57				
19	*landuse*	*quarry*	48	0.37													96	71			2	32									2	46
20	*pois*	*golf_course*	45	0.34			98	83			2	52																				

*Both the SR and AMP were divided into four intervals (0-25%, 25-50%, 50-75% and 75-100%) and highlighted in different shades of red and green colours. respectively.

Table 2. The OSM-LULC dictionary created according to 14 different LULC classes (Total samples=50)*.

(1) Some OSM tags may have multiple meanings. For instance, the OSM tag *nature_reserve* was described in OpenStreetMap Wiki as "a protected area of importance for wildlife, fauna or features of geological or other special interest" (https://wiki.openstreetmap.org/wiki/Tag:leisure%3Dnature_reserve, accessed in June 2020), this OSM tag may corresponding to the LULC class *Forests* for a forests and protected area (Figure 4a); and it may also correspond to the class *Water* for a protected area of the lake (Figure 4b). Thus, for the OSM tag *nature_reserve*, the most appropriate LULC class depends on the category of a protected area.

(2) Some regions have been mapped by OSM volunteers with much more details than they presented in the reference dataset. As an example, the OSM tag *grass* mostly corresponded to the LULC class *Industrial, commercial, public, military, private and transport units*, because there was plenty of grass in the airport (Figure 4c). As another example, in Figure 4d, the OSM tag *tourist_info* corresponded to the LULC class *Forests* because an OSM object tagged as *tourist_info* was found inside the forest.

(3) Some OSM objects may also be incorrectly tagged by volunteers. As an example, the OSM tag *meadow* was sometimes found to correspond to the LULC class *Forests*. But in the corresponding image (Figure 4e), the OSM object was mostly covered with forest rather than meadow. That means this OSM object may be incorrectly tagged by volunteers. As another example, the OSM tag *tourist_info* was sometimes found to correspond to the LULC class *Arable land* (Figure 4f). However, this OSM tag was described in OpenStreetMap Wiki as "*an information source for tourists, travellers and visitors, e.g., tourist information centres and offices*" (https://wiki.openstreetmap.org/wiki/Tag:tourism%3Dinformation, accessed in June 2020). Thus, the example in Figure 4f may also be an incorrect tag.

5. APPLICATIONS

This study proposed an approach to creating an OSM-LULC dictionary and involved 50 pairs of OSM and reference datasets in pan-European metropolitans for the analysis. There are multiple applications of this dictionary.

5.1 LULC mapping

First of all, the dictionary can be used for producing an OSM-based LULC map, for which application it is needed to first establish the correspondence between each pair of OSM tag and LULC class. As an example, ten additional pan-European metropolitans were used for LULC mapping and validation, and the main steps are listed as follows.

Firstly, different OSM tags were picked up for LULC mapping. More precisely, three different scenarios were considered, in order to investigate: 1) Is it possible to use a few OSM tags (e.g. whose *AAP*>0.1%) for LULC mapping rather than all (commonly there are 100 or more OSM tags in a metropolitan)? 2) Can the performance of an LULC map be improved if only using OSM tags with relatively larger *SR* and *AMP*.

For each metropolitan, the following three scenarios were considered for picking up OSM tags.

- Scenario I: All the OSM tags that can be found in the dictionary (with 14 LULC classes) were picked up.
- Scenario II: All the OSM tags that can be found in the dictionary (with 14 LULC classes) and whose *AAP* is larger than 0.1% were picked up.
- Scenario III: All the OSM tags that can be found in the dictionary (with 14 LULC classes) and whose *AAP*, *SR* and *AMP* are respectively larger than 0.1%, 70% and 70% were picked up.

Secondly, determine the most appropriate LULC class for each selected OSM tag. That is, for each selected OSM tag, the LULC class with the maximum *SR* is determined as the most appropriate one.

Thirdly, rank all the classes from top to bottom according to their average land areas from small to large, and then merge all these classes into an LULC map. The purpose of this step was to avoid overlaps among classes and also to produce a detailed LULC map.

Fourthly, Compare the produced LULC map with the corresponding reference dataset and assess this map with two quality measures: overall accuracy (OA) and completeness. The OA denotes that in an LULC map, how many land areas have been correctly classified; and the completeness denotes that in a metropolitan, how many land areas have been covered with an LULC class (Arsanjani and Vaz 2015).

The experimental results are reported in Table 3. It can be seen from this table that both the OA and completeness are almost the same for the scenarios I and II, which verifies that it is feasible to use only a few OSM tags (e.g., *AAP*>0.1%) for LULC mapping. Moreover, all the OAs for the scenario III are larger than those for the scenarios I and II, which verifies that the performance of an LULC map can be improved by picking up OSM tags with relatively higher *SR* and *AMP*. Despite the advantage, it should be noted that all the completeness values for the scenario III are much lower because fewer OSM tags were involved for LULC mapping. Thus, in a practical application, we suggest to use the maximum *SR* to determine an appropriate LULC class for each OSM tag, while both the corresponding *SR* and *AMP* are relatively high (e.g., >70%). But, it may be better to consider multiple different LULC classes, while either the *SR* or *AMP* is relatively low. In this case, the most appropriate LULC class may be manually determined by referring to the images in Google Earth. Nevertheless, the dictionary provides an understanding of that: 1) which OSM tags can correspond to multiple different LULC classes and 2) which LULC class(es) is/are corresponded to each OSM tag.

5.2 Quality assessment

Last but not least, the dictionary may also be used for the quality assessment of an OSM dataset. This is because potentially incorrect tags may be detected using this dictionary (Figure 4e and 4f). The tenet of the approach for the quality assessment is to refer to the *SR* and/or *AMP* with a relatively low value. A low *SR* indicates that a pair of OSM tag and LULC class is not common. As an example, only in one of the 50 metropolitans (*SR*=2.0%), the OSM tag *meadow* corresponded to the LULC class *Forests*; But in 30 metropolitans (*SR*=60%), the OSM tag *meadow* corresponded to the LULC class *Pastures*, which was semantically a more reasonable pair. More important, a low *AMP* indicates a low consistency between a pair of OSM tag and LULC class. As another example, the *AMP* for the pair of the OSM tag *meadow* and the LULC class *Forests* was only 29.7% (Table 2), which illustrates that the intersecting area (or consistency) for this pair is rather low.

Furthermore, Figure 5 shows several typical OSM tags and their most appropriate LULC classes for the study area of London. These OSM tags were picked up because their AMP values are not higher than 60% in the dictionary (Table 2), even while their corresponding SR values reach to the maximum. In this figure, each object was divided into one of two cases.

- Case I: the OSM tag for this object was reclassified into its most appropriate LULC class (highlighted in blue);
- Case II: the OSM tag for this object was reclassified into other LULC classes (highlighted in red).

Study Metropolitan	Scenario I		Scenario II		Scenario III	
	OA	Completeness	OA	Completeness	OA	Completeness
Arhus	70.83%	41.75%	70.99%	41.39%	74.89%	36.13%
Brno	88.26%	91.65%	88.75%	88.12%	91.62%	80.16%
Linz	82.97%	54.98%	83.10%	54.98%	84.80%	53.13%
Munchen	75.85%	86.45%	75.95%	85.68%	87.17%	68.01%
Radom	45.39%	62.31%	45.37%	61.93%	73.34%	33.90%
Stavanger	64.83%	33.82%	64.82%	33.64%	70.53%	28.80%
Nis	87.93%	37.51%	87.98%	37.49%	89.17%	36.66%
Einhoven	49.52%	92.90%	49.39%	92.08%	60.88%	61.01%
Laussane	70.70%	93.63%	70.70%	93.17%	74.63%	77.52%
Plovdiv	63.90%	45.46%	63.94%	44.59%	68.21%	37.81%

Table3. LULC mapping considering different scenarios.

We can see from Figure 5 that: for each OSM tag, a number of objects were not reclassified consistently as their most appropriate LULC classes. This indicates that it is feasible to use our dictionary for picking up OSM tags with potential inconsistent errors.

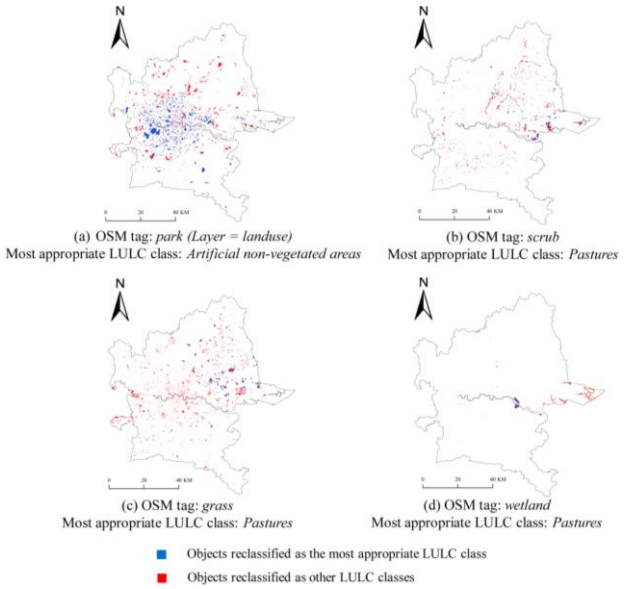


Figure 5. Detecting inconsistent errors for several OSM tags.

6. LIMITATION AND FUTURE WORK

Despite the above applications, there are several limitations of this study.

Firstly, as the most appropriate LULC class may vary with different study areas, it may not be possible to automatically determine the most appropriate LULC class for all OSM tags, while using our dictionary for LULC mapping. This is especially the case for some OSM tags whose *SR* values are relatively low (e.g., <60% or 70%). As an alternative, it is possible to use this dictionary to pick up candidate LULC classes for each OSM tag and then determine the most appropriate one by referring to other data sources (e.g. images in Google Earth).

Secondly, it is useful to detect potential inconsistent errors with our dictionary (Section 5.2). But it is hard to precisely determine which object(s) has/have been tagged by volunteers incorrectly, especially if a corresponding LULC reference dataset is not available. As an alternative, other data sources (e.g. images in Google Earth) may also be referred to for the analysis.

Last but not least, the OSM-LULC dictionary may vary in involving different metropolitans for the analysis, and it may also vary in involving different LULC classes for the analysis. In this study, 50 different pan-European metropolitans were chosen as study areas to create an LULC dictionary in order to minimize the subjective of using only a few study areas. Thus, the created dictionary may still be applied to other countries and regions in the world. Moreover, global LULC data products become more and more available (Chen et al. 2015; Grekousis et al. 2015; Fritz et al. 2017). Our proposed approach can further be applied to creating an OSM-LULC dictionary based on other reference datasets.

In the future work, firstly, it is worth to validate our proposed approach by creating an OSM-LULC dictionary based on different study areas and reference datasets. Secondly, it may also be interesting to add some other measures (e.g., the average area percentage (*AAP*) for each pair of OSM tag and LULC class) in order to improve the performances of this dictionary for various applications.

7. CONCLUSION

This study proposed an approach to creating an OSM-LULC dictionary. The tenet of this approach was to involve multiple pairs of OSM and reference datasets for the analysis. First of all, each pair of OSM and reference datasets were intersected and the most appropriate LULC class for each OSM tag was determined. Then, the four measures, i.e., sample count (*SC*), average area percentage (*AAP*), sample ratio (*SR*) and average maximum percentage (*AMP*), were designed and calculated based on multiple pairs of OSM and reference datasets.

More precisely, a total of 50 pairs of OSM and reference datasets in pan-European metropolitans were chosen as study areas for creating an OSM-LULC dictionary. Finally, a number of 1409 different OSM tags were found and they were reclassified into five and 14 different LULC classes, respectively. Moreover, this dictionary was also analyzed with the four proposed measures. Results showed that:

Firstly, most OSM tags (>1,000) were only found in less than five study areas (*SC*<5). Moreover, only 37 of the 1409 OSM tags had a percentage of average area (*AAP*) larger than 0.1%. This indicates that a small proportion of OSM tags can play a decisive role.

Secondly, an OSM tag may correspond to multiple different LULC classes within a pair of OSM and reference datasets; The most appropriate LULC class for each OSM tag may also vary among different pairs of datasets. Thus, both the *SR* and *AMP* may also vary in different pairs of OSM tag and LULC class.

With the proposed dictionary, it is possible to understand the differences of different OSM tags and different pairs of OSM tag and LULC class. This is essential not only for producing LULC maps, but also for picking up training and/or validation data from an OSM dataset. Therefore, we concluded that it has benefits for creating an OSM-LULC dictionary based on multiple pairs of OSM and reference datasets.

REFERENCES

Arsanjani, J. J., Helbich, M., Bakillah, M., & Hagenauer, J. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, 27(12), 2264–2278.

Arsanjani, J. J., & Vaz, E. (2015). An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *International Journal of Applied Earth Observation and Geoinformation*, 35(2015), 329–337.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., and Yuan, J., (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. *Transportation Research Part A: Policy and Practice*, 127:71-85.

Chen, C., Park, T., Wang, X., Piao, S., Xu, B., Chaturvedi, R.K., et al. (2019). China and India lead in greening of the world through land-use management. *Nature Sustainability*, 2:122-129.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., et al. (2015). Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103(2015), 7–27.

Dorn, H., Törnros, T., & Zipf, A. (2015). Quality evaluation of VGI using authoritative data-A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information*, 2015(4), 1657–1671.

Estima, J., & Painho, M. (2015). Investigating the potential of OpenStreetMap for land use/land cover production: A case study for continental Portugal. In: Arsanjani, et al. *OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography*, Springer International Publishing, Switzerland.

Fonte, C. C., & Martinho, N. (2017). Assessing the applicability of OpenStreetMap data to assist the validation of land use/land cover maps. *International Journal of Geographical Information Science*, 31(12), 2382–2400.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., et al. (2017). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data*, 4, 170075.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(2007), 211–221.

Grekousis, G., Mountrakis, G., & Kavouras, M. (2015). An overview of 21 global and 43 regional land-cover mapping products. *International Journal of Remote Sensing*, 36(21), 5309–5335.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 2010(37), 682–703.

Johnson, B.A., & Lizuka, K. (2016). 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. *Applied Geography*, 67(2016), 140–149.

Liang, J., Zhong, M., Zeng, G., Chen G., Hua, S., Li, X., et al. (2017). Risk management for optimal land use planning integrating ecosystem services values: A case study in Changsha, Middle China. *Science of the Total Environment*, 579(1), 1675–1682.

Liu, X, & Long, Y. (2016). Automated identification and characterization of parcels with OpenStreetMap and points of interest. *Environment and Planning B: Planning and Design*, 43(2), 341–360.

Long, X., Chen, Y., & Zhang, Y. (2022). Visualizing green space accessibility for more than 4,000 cities across the globe. *Environment and Planning B: Urban Analytics and City Science*, 23998083221097110.

Rimal, B., Zhang, L., Keshtkar, H., Hacck, B. N., Rijal, S., & Zhang, P. (2018). Land use/land cover dynamics and modeling of urban land expansion by the integration of cellular automata and markov chain. *ISPRS International Journal of Geo-Information*, 7(4), 154.

Schultz, M., Voss, J., Auer, M., Carter, S., & Zipf, A. (2017). Open land cover from OpenStreetMap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 63(2017), 206–213.

Srivastava, S., Lobry, S., Tuia, D., & Vargas-Muñoz, J. (2018). Land-use characterisation using Google Street View pictures and OpenStreetMap. In: *Proceedings of the Association of Geographic Information Laboratories in Europe Conference (AGILE)*, 12–15 June, Lund.

Verburg, P.H., Alexander, P., Evans, T., Magliocca, N.R., Malek, Z., Rounsevell, M. et al. (2019). Beyond land cover change: towards a new generation of land use models. *Current Opinion in Environmental Sustainability*, 38:77-85.

Xia, C., Yeh, A.G., and Zhang, A., (2020). Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: A case study of five Chinese megacities. *Landscape and Urban Planning*, 193: 103669.

Zeng, C., Liu, Y., Stein, A. L, & Jiao, L. (2015). Characterization and spatial modeling of urban sprawl in the Wuhan metropolitan area, China. *International Journal of Applied Earth Observation and Geoinformation*, 34(2015), 10–24.

Zhang, Y., Zhou, Q., Brovelli, M. A., & Li, W. (2022). Assessing OSM building completeness using population data. *International Journal of Geographical Information Science*, 1-24.

Zhou, Q. (2018). Exploring the relationship between density and completeness of urban building data in OpenStreetMap for quality estimation. *International Journal of Geographical Information Science*, 32(2), 257–281.

Zhou, Q., Jia, X., & Lin, H. (2019). An approach for establishing correspondence between OpenStreetMap and reference datasets for land use and land cover mapping. *Transactions in GIS*, 23(6), 1177–1464.

APPENDIX

The data that support the findings of this study are openly available in figshare at https://figshare.com/s/df28d9020ad35f5220bf.