# Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal

**Jacinto Estima and Marco Painho**

**Abstract** In the last decade, volunteers have been contributing massively to what we know nowadays as Volunteered Geographic Information (VGI). Through the research that has been conducted recently, it has become clear that this huge amount of information might hide interesting and rich geographical information. The OpenStreetMap (OSM) project is one of the most well-known and studied VGI initiatives. It has been studied to identify its potential for different applications. In the field of Land Use/Cover, an earlier study by the authors explored the use of OSM for Land Use/Cover (LULC) validation. Using the COoRdination of INformation on the Environment (CORINE) Land Cover (CLC) database as the Land Use reference data, they analyzed the OSM coverage and classification accuracy, finding an interesting global accuracy value of 76.7 % for level 1 land classes, for the study area of continental Portugal, despite a very small coverage value of approximately 3.27 %. In this chapter we review the existing literature on using OSM data for LULC database production and move this research forwards by exploring the suitability of the OSM Points of Interest dataset. We conclude that OSM can give very interesting contributions and that the OSM Points of Interest dataset is suitable for those classified as CLC class 1 which represents artificial surfaces.

**Keywords** Volunteered geographic information (VGI) · OpenStreetMap (OSM) · Land use · Land cover

J. Estima (✉) · M. Painho
ISEGI, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: jacinto.estima@gmail.com

M. Painho
e-mail: painho@novaims.unl.pt

© Springer International Publishing Switzerland 2015

J. Jokar Arsanjani et al. (eds.), *OpenStreetMap in GIScience*,
Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-319-14280-7_14

# 1 Introduction

## *1.1 Land Use and Land Cover Production*

Land Use/Land Cover (LULC) databases, as they characterize land, are fundamental inputs for a variety of applications such as LULC change monitoring, climate change and biodiversity monitoring, among others (Caetano et al. 2006; Ellis 2013; Fritz et al. 2009). While Land Cover (LC) is more related to natural environments characterizing biophysical features, Land Use (LU) represents human-related environments attempting to describe the human interaction with these natural features (Baulies and Szejwach 1997). The production of LC and LU maps is usually undertaken by highly trained and skilled people interpreting and classifying remote sensing data, and involves a complex and long process of four main steps: acquisition of remote sensing data used as the basis for the classification process, pre-processing data into a proper format to extract information, analysis/classification including quality assessment, and product generation and documentation (Cihlar 2000). Aerial photography or in situ data are required in some of these phases, either to clarify the interpretation in areas of uncertainty or for quality assessment and validation (Caetano et al. 2006), thus increasing the time and cost of production. These constraints mean that the focus of LULC mapping is on themes and areas that are considered more important and for use within multiple applications.

This also has a negative impact on the update strategies, which therefore happen less frequently. Consequently, these databases become outdated very quickly in some areas (Goodchild 2008b).

According to Cihlar (2000, p 1108), "*The research agenda needs to address the best ways of taking advantage of the new capabilities and, importantly, the ways of resolving problems identified during the production of the land cover maps over large areas*". This statement drives us to think that VGI data, and in this particular case OSM data, have to be investigated and exploited in order to be used for LULC database production.

## *1.2 Volunteered Geographic Information and OpenStreetMap*

The term Volunteered Geographic Information (VGI) was coined by Michael Goodchild in 2007 to describe "*the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies*" (Goodchild 2007, p 212). Other related terms were also introduced, in different years by different authors, such as Neogeography (Turner 2006) or Crowdsourcing geospatial data (Hudson-Smith et al. 2009).

There are a range of initiatives to which volunteers might participate and contribute and both the initiatives and the volunteers have been growing over the years according to the inventory made by Elwood et al. in 2009 where 99 VGI initiatives were identified (Elwood et al. 2012). Initiatives such as Wikimapia (2014), Google MyMaps (2014), GMapCreator (2014), London Profiler (2014), Map Tube (2014) and Flickr (2014) are just a few examples from this comprehensive list. According to the same authors, OpenStreetMap (OSM), also part of the list, is one of the most important and studied VGI initiatives. Started in August 2004 by Steve Coast and developed by the OpenStreetMap Foundation since 2006, this initiative is a worldwide mapping effort that includes more than a million volunteers around the globe, the number reached during 2013, and aims at providing free geographic data to anyone. Users collect data, including topographic data, mostly with GPS or GPS-enabled equipment. The collected data are then uploaded to the OSM database, along with descriptions, names and other attributes, through the OSM web page or established editors. The data then become available to anyone in the form of rendered maps and other services, including the possibility to download the data in vector format or embed it in websites using their Application Programming Interface (API).

The interest in exploring some of these initiatives has been increasing for areas such as navigation (Holone et al. 2007), emergency response (Goodchild and Glennon 2010; Zook et al. 2010), vernacular geography (Hollenstein and Purves 2010) and LULC production (Estima and Painho 2013b; Fritz et al. 2009). In this chapter we provide the current status on the use of data from OSM for LULC database production. We then move forward in our continuing research in this matter to explore a point-based dataset from the OSM database. We also discuss possible and interesting contributions OSM can give for LU database production, also considering some of the main issues related with its use and approaches to overcome them.

The chapter is structured as follows: After a brief introduction some related work is presented with an emphasis on studies using OSM data for LULC database production. We then describe the data and methods used for the practical part followed by the results and discussion. The paper ends with some conclusions, summarizing possible contributions of OSM for LULC database production purposes, and future research directions.

## 2 Volunteered Geographic Information and OpenStreetMap for Land Use/Land Cover Production

In this chapter we review studies already conducted on using VGI, with particular emphasis on OSM, for LULC database production. We start by summarizing the research on using VGI followed by the development of a more extended overview on the particular case of OSM.

### 2.1 Volunteered Geographic Information for Land Use/Land Cover Production

As stated before, VGI has been increasingly used to research novel applications for different areas, including LULC database production. In this particular domain two different approaches have been used so far: (1) asking volunteers to actively contribute to a specific project such as the validation of global land cover datasets (Fritz et al. 2009; Perger et al. 2012), and (2) using data contributed for other purposes/ projects to extract valuable information and develop new ways to use it in this domain (Estima and Painho 2013a, b, 2014). Geo-Wiki.Org (Fritz et al. 2009) is a project that fits in the first approach, described as a global network of volunteers who wish to help improve the quality of global land cover maps. “GLC-2000”, “MODIS”, and “GlobCover” global land cover databases are overlaid on a platform based on Google Earth (GE) and their areas of divergence highlighted. Then, a network of registered volunteers helps to solve these discrepancies using their local knowledge along with available GE satellite imagery and other ancillary data coming from other VGI projects such as pictures from Panoramico (http://www.panoramio.com/) and Degrees of Confluence Project (http://www.confluence.org/). Another example is the Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) initiative based on GE high-resolution imagery to collect LULC reference data (Clark and Aide 2011). It was tested with a small group of selected users acting as volunteers and not yet in a real crowdsourcing environment. Nevertheless they found important issues with using GE and its satellite imagery, e.g. the legal restrictions in the free use of the Google Maps/Earth APIs and that some classes that cannot be discriminated with the available imagery (e.g. different annual crops). In these examples, volunteers need to be available to contribute to these specific projects and they also need to have some familiarity with these tools, which might be discouraging for some groups of participants. To overcome this difficulty, some projects occasionally use contests and a mechanism of rewards to increase contributions and participation (Fritz et al. 2012; Perger et al. 2012).

Using the second aforementioned approach, some experiments were conducted by Leung and Newsam (2010) to derive maps of what-is-where from large collections of georeferenced photos in an automated way. In this initial work the authors derived LC classifications from georeferenced image collections for locations where ground-truth was available. The aim was to evaluate the quality of the results obtained from the automatic classification by comparing them with the available ground truth. They achieved a classification accuracy of approximately 75 %. Another interesting work was conducted by Estima and Painho (2013b) to explore the possibility of using Flickr photos as a source of ground-truth data to help in the accuracy assessment phase of LULC production. Using continental Portugal as the study area and COoRdination of INformation on the Environment (CORINE) Land Cover (CLC) as a reference LULC database, the authors explored all the publically available and geotagged Flickr photos in terms of their temporal and spatial distributions and their distribution over the different CLC classes.

The number of photos and their temporal resolution were the most positive aspects whereas their asymmetry and irregular distribution over different CLC classes the most negative. They concluded stating that this could be a valuable source of ground truth data if combined with other sources but could not be used alone. Foody and Boyd (2013) used two sources of volunteered data to illustrate the potential of amateur or neogeographical activity in map validation. They used photographs acquired from an internet-based collaborative project and interpreted by other volunteers to evaluate the Globcover map's representation of tropical forests in West Africa. They confirmed the potential value of VGI projects, such as the Degrees of Confluence project, for the provision of useful, spatially extensive, data to support map evaluation.

## *2.2 OpenStreetMap for Land Use/Land Cover Production*

The exploration of data from the OpenStreetMap (OSM) project is also recent. In 2013, Estima and Painho (2013a) explored the use of OSM data for LULC database production, particularly for validation purposes. The authors explored three OSM datasets—buildings, land use, and natural areas—using continental Portugal as the study area and CLC as the reference LULC database. They analyzed the spatial coverage and distribution, established a correspondence between OSM and CLC nomenclatures, and explored the coverage and accuracy when compared with CLC level 1 classes (CLC nomenclature can be found in Table 1). They found that the coverage area is not homogeneous among all the CLC level 1 classes, with values of approximately 75, 20, 2, 0.8, and 0.2 % for classes water bodies, artificial surfaces, forest and semi natural areas, agricultural areas, and wetlands, respectively. A table summarizing discrepancies between OSM and CLC classification nomenclatures was also reported, showing that multiple correspondences increase when we move from CLC level 1 to CLC level 3, given the increasing level of detail. To give just a brief example, the class "Farm" from the OSM Landuse dataset corresponds to the class 2 from the CLC level 1, to classes permanent crops, pastures, and heterogeneous agricultural areas from the CLC level 2, and to classes fruit trees and berry plantations, pastures, annual crops associated with permanent crops, or complex cultivation patterns from the CLC level 3. In terms of classification accuracy, the values reported were very promising, showing around 99.5, 84.3, 83.5, 46.6, and 1.2 % for classes water bodies, artificial surfaces, forest and semi-natural areas, agricultural areas, and wetlands, respectively, of the CLC level 1 nomenclature, and a global value of 76.7 %. They conclude that OSM might be very useful for LULC classification for classes with good coverage and accuracy such as classes' artificial surfaces and water bodies.

The possibility of using VGI data to replace training data acquired from in-site visits in the process of LULC classification was also investigated by Jokar Arsanjani et al. (2013a). Using the city of Koblenz, Germany, as the study area, they applied a supervised classification approach to classify data from the RapidEye

**Table 1** Corine land cover nomenclature

| Level 1 | Level 2 | Level 3 |
|---|---|---|
| 1 Artificial surfaces | 11 Urban fabric | 111 Continuous urban fabric |
| | | 112 Discontinuous urban fabric |
| | 12 Industrial, commercial and transport units | 121 Industrial or commercial units |
| | | 122 Road and rail networks and associated land |
| | | 123 Port areas |
| | | 124 Airports |
| | 13 Mine, dump and construction sites | 131 Mineral extraction sites |
| | | 132 Dump sites |
| | | 133 Construction sites |
| | 14 Artificial, non-agricultural vegetated areas | 141 Green urban areas |
| | | 142 Sport and leisure facilities |
| 2 Agricultural areas | 21 Arable land | 211 Non-irrigated arable land |
| | | 212 Permanently irrigated land |
| | | 213 Rice fields |
| | 22 Permanent crops | 221 Vineyards |
| | | 222 Fruit trees and berry plantations |
| | | 223 Olive groves |
| | 23 Pastures | 231 Pastures |
| | 24 Heterogeneous agricultural areas | 241 Annual crops associated with permanent crops |
| | | 242 Complex cultivation patterns |
| | | 243 Land principally occupied by agriculture |
| | | 244 Agro-forestry areas |
| 3 Forest and semi natural areas | 31 Forests | 311 Broad-leaved forest |
| | | 312 Coniferous forest |
| | | 313 Mixed forest |
| | 32 Scrub and/or herbaceous vegetation associations | 321 Natural grasslands |
| | | 322 Moors and heathland |
| | | 323 Sclerophyllous vegetation |
| | | 324 Transitional woodland-shrub |
| | 33 Open spaces with little or no vegetation | 331 Beaches, dunes, sands |
| | | 332 Bare rocks |
| | | 333 Sparsely vegetated areas |
| | | 334 Burnt areas |
| | | 335 Glaciers and perpetual snow |

(continued)

**Table 1** (continued)

| Level 1 | Level 2 | Level 3 |
|---|---|---|
| 4 Wetlands | 41 Inland wetlands | 411 Inland marshes |
| | | 412 Peat bogs |
| | 42 Maritime wetlands | 421 Salt marshes |
| | | 422 Salines |
| | | 423 Intertidal flats |
| 5 Water bodies | 51 Inland waters | 511 Water courses |
| | | 512 Water bodies |
| | 52 Marine waters | 521 Coastal lagoons |
| | | 522 Estuaries |
| | | 523 Sea and ocean |

*Source* Corine Land Cover Nomenclature 2011

sensor, and they used data downloaded from the OSM project as field measurements to select the most optimal training sites. They performed a comparison of the resultant LU map with the Global Monitoring for Environment and Security Urban Atlas (GMESUA) map achieving a Kappa index of 89 %, which proves that OSM data is suitable to use as a source for training site definition. They also stress that the quality of VGI is heterogeneous and location-dependent, and they recommend checking the amount of contributions and also considering other VGI data, such as Flickr photos.

Another study investigated a new approach to generating land-use patterns from VGI without applying remote-sensing techniques and/or engaging official data (Jokar Arsanjani et al. b). Using OSM datasets and Vienna, Austria, as the study area, the authors applied a Hierarchical GIS-based decision tree approach to classify and segment parcels. The results were evaluated by conducting a texture-variability analysis of the LU maps generated using each dataset, and producing a confusion matrix to compare each LU class in the two datasets. Results of the texture analysis showed that the LU patterns derived from OSM data are richer than those derived from GMESUA. The confusion matrix showed a high level of agreement between the two classifications but this decreased when we move from level 1 towards the more detailed level 3. Although they conclude that VGI can be a potential data source for mapping LU patterns, they only used one source of VGI, OSM, and they did not test any other sources. Nevertheless, they pointed out as advantages of such an approach that no inputs from remote-sensing or any other administrative data were used, no financial cost exists as the OSM data is freely available and no field work was required, a number of incorrectly labeled features in the GMESUA were identified when OSM was incorporated, and the process of updating LU maps is facilitated due to the updating rate of OSM while GMESUA requires time and high financial costs to be updated by authorities.

A different approach was previously proposed by Hagenauer and Helbich (2012). They applied Artificial Neural Networks (ANNs) and Genetic Algorithms

(GAs) as a machine learning methodology to delineate continuous urban areas using all the information diversity of OSM, where a large set of potential OSM attributes was derived for inductive learning. Using OSM and GMESUA data, they applied this methodology to 42 randomly selected GMESUA urban regions and analyzed the significance of the attributes used and the performance of the model. The model performed comparatively well for most regions, with a few remarkable exceptions. The study shows that if enough OSM data for reasoning is present, urban patterns can be predicted to a large extent. This approach could be very useful to help map continuous fabric classes, from OSM data, for LULC databases.

The representation of natural features in OSM was also explored by Mooney et al. (2010), who examined the level of detail present in the representation of such polygon features. They tried to verify if there was enough detail in the representation of those features to provide a high-quality spatial representation. They used data for Austria, Estonia, Switzerland, Bretagne, Lower Saxony, Iceland, Ireland, and Scotland to calculate the statistical distribution of the mean distance between connected vertices of polygons. They found that many of the features are under-represented, with a small number of vertices used to delineate them, while some of them might be considered over-represented (e.g. small urban green spaces and golf courses). Some OSM data collection characteristics, such as the different GIS skill levels of OSM volunteers or the differences in accuracy of equipment and methods used, influence the under-representation of some features. These under-represented features have a serious impact on using OSM data in certain Earth science applications, mainly those that use OSM as ground-truth data. They recommend that the quality of the OSM representation of "natural" polygons and other features should be established against a recognized ground-truth dataset.

In this sense, other authors have been exploring the quality of OSM data that are of interest for LULC database production. Barron et al. (2014) developed a comprehensive framework for intrinsic OSM quality analysis that included the logical consistency of "natural" and "landuse" polygons. They developed a tool to generate information about OSM data quality for a selectable area without a reference dataset but using only OSM's data history. This tool intends to help users to assess the OSM data quality of a given area for a specific application. As an example, for map applications such as LULC database production, the tool automatically identifies erroneously overlapping land use polygons and analyzes not only the equidistance between the polygons' adjacent vertices, which is a good way to determine the quality of those polygons, but also the evolution of their equidistance over time.

Methods to analyze the completeness of building footprints over space and time were described and analyzed by Hecht et al. (2013) for the German states of North Rhine-Westphalia and Saxony. They used unit-based and object-based methods to analyze the level of completeness of building footprints contained within OSM by always comparing them with a reference dataset regarded as complete. They conclude that unit-based methods require less computation but have limitations in their level of detail when compared with object-based methods. Their results in applying these methods to the mentioned areas of Germany showed that OSM building footprints, as of November 2012, are characterized by a low degree of completeness, below 30 %, and a strong geometrical heterogeneity, and the level of completeness is higher in urban than in rural areas.

A similar study for the German city of Munich was developed by Fan et al. (2014). In this study the authors developed a quality assessment of building footprint data, after they found that the number of buildings in OSM was over 77 million on 5 May 2013. Building footprints were assessed using four criteria: (1) completeness, (2) semantic accuracy, (3) position accuracy, and (4) shape accuracy, where OSM data were compared with the reference data from the German Amtliches Topographisch-kartographisches Informatiosystem—Authorative Topographic-Cartographic Information System (ATKIS) to perform a quantitative assessment. They concluded that, for the case study of Munich (Germany), a high level of completeness was found but OSM building footprints still lack important attributes such as name, type, and height, among others. They found, however, more than 1,200 newly constructed buildings which were not documented in the ATKIS data. On the other side, although OSM building footprints are very similar in terms of shape, they have on average a 4 m offset to their corresponding ones in ATKIS in terms of position accuracy. Building footprints might be an important source of information to help in the classification or validation of urban areas, and these results are a very good indicator. Jokar Arsanjani and Vaz (2015) analyzed the completeness and thematic accuracy of seven European metropolises and thanks to the promising accuracy values concluded that these parameters greatly vary from location to location, which confirms the heterogeneity of contributions.

## 3 Materials and Methods

In this chapter we introduce the practical part of this study, which is an advance on our previous studies in exploring the suitability of OSM data for LULC purposes (Estima and Painho 2013a). In this case, the Points of Interest dataset was explored. We start by presenting and describing the study area and data used for this study, and explaining the methodology used to accomplish our objective.

### *3.1 Study Area and Data*

The defined study site is continental Portugal, located on the southwestern side of Europe covering a total area of 8,908,220.16 Ha. The land cover is mainly composed of agricultural and forest areas covering around 95 % of the country.

The OSM database under analysis covers the area of continental Portugal and was downloaded from the Geofabrik website (Geofabrik 2014). This database is current as of 23 July 2013, and is divided into six datasets: places, points, railways, roads, waterways, buildings, landuse, and natural areas. Places and points are represented by point geometries; railways, roads and waterways by line geometries;

and buildings, landuse and natural areas by polygon geometries. As already mentioned, moving further in our research, the points dataset was used in this study.

The CLC database is composed of version 16 (04/2012) of Corine Land Cover (CLC) for the CLC2006 inventory, downloaded from the European Environment Agency (EEA 2014). This dataset, in vector format, was developed using the European Terrestrial Reference System 1989 (ETRS89) with the Lambert Azimuthal Equal Area, also known as ETRS89-LAEA. Using a Minimum Mapping Unit (MMU) of 25 Ha, the land cover is classified according to the CLC nomenclature, which is hierarchically divided into three levels of classes, as shown in Table 1. For the purpose of this investigation we used the five classes from level one: (1) artificial surfaces (AS), (2) agricultural areas (AA), (3) forests and semi-natural areas (F), (4) wetlands (W), (5) water bodies (WB).

### *3.2 Methods*

The methodology adopted to conduct this analysis was as follows:

1. We explored the point dataset defined in the previous section in terms of content and coverage;
2. We established a relationship between each point type and the CLC classes, based on their description documented on the OSM Map Features website (OpenStreetMap Map Features 2014);
3. For each point location, we compared the classification given in the previous step with the respective class extracted from the CLC database, using a confusion matrix approach. We also analyzed the classification accuracy for each OSM point type.

## 4 Results and Discussion

In this chapter we present and discuss the results obtained by applying the methodology described in the previous section.

### *4.1 Analysis of the OSM Dataset*

In this first step we explored the point dataset in terms of content and coverage. This data are composed of a collection of 49,861 Points of Interest (PoI) within the study area, classified according to type of PoI. A list of predefined types is available for use when a new point is registered (OpenStreetMap Map Features 2014), but each user can also define new types. Although this possibility gives a lot of flexibility in the mapping and classification process, it creates additional difficulties to perform further analysis, mainly related to the lack of proper descriptions but also to the possibility of introducing spelling errors.

Table 2 shows a list of POI types found within the collection of points. A closer look shows some types that are not of interest for the purpose of our study, mainly because they do not represent any type of LULC or related feature, or the relation is not clear (e.g. "attraction", "heritage", "no", "yes"). Different spelling for the same type were also found (e.g. "community_centre", "comunity centre", and "Comunity_centre"), a typical error related to the possibility of users creating their own types. Taking into account the description available for each feature type, and only for those types available in the wiki list, the types marked with asterisk (*) in Table 2 were considered attributable to a CLC class and selected for further analysis. This represents a total of 26,290, corresponding to around 52 % of the total number of initial points.

Figure 1 shows the spatial distribution of the selected PoIs over the study area. It is possible to observe the concentration of points over the coast, where touristic places and larger cities are represented, as well as along some of the main roads.

## *4.2 Correspondence Between OSM Point Types and CLC Classes*

After selecting the types of PoI to use in the previous task, a CLC equivalent class was attributed to each type according to their description in the wiki website. Only two CLC classes were used: classes 1 and 5, representing AS and WB, respectively. This was already expected due to the higher probability of more volunteers visiting places fitting in these classes. There were some special cases where we also took into account our knowledge of the feature type class versus their surroundings. The case of the "bridge" feature type, which would apparently be classified as Artificial Surfaces, was classified as Water Bodies since bridges are usually over water bodies and are not represented in LULC databases due to their size. Table 3 shows the list of PoI types for each given CLC level 1 class.

## *4.3 Classification Accuracy Analysis*

After assigning a CLC level 1 class to each PoI type, the evaluation of the classification was the next step. In this task we first filled the PoI dataset with the CLC class, based on the correspondence defined in the previous step. We then intersected it with the CLC database to have, for each point location, the classification defined by the PoI description and the classification taken from the CLC database. A new attribute was created to identify agreements/disagreements between the two

**Table 2** List of types of OSM PoIs

| | | | | | |
|---|---|---|---|---|---|
| arts_centre* | charging station* | flagpole | marketplace* | reservoir* | tertiary |
| adit | charging_station* | food_court* | mast | reservoir_covere* | tertiary_link |
| alpine_hut | chimney | footway | measurement_stat | residential* | theatre* |
| animal_shelter | cinema* | ford* | megalith | resort | theme_park* |
| antenna | city_gate* | forester's lodge | memorial | rest_area* | toilets |
| archaeological_s | clinic* | fort | milestone | restaurant* | tower |
| artwork | clinica fisiote | forte de sao jo | mineshaft | road* | townhall* |
| ashtray | clock | fountain* | mini_roundabout* | ruins | track |
| atm* | college* | fuel* | moinho do cuco | satellite_centre | traffic_signals |
| attraction | communications_t | gasometer* | monument* | school* | traffic-signs |
| baby_hatch | community_centre* | gate | motel* | scout_hut | trail_riding_sta |
| bank* | comunity centre* | give_way | motorcycle_parki | secondary | tram_stop* |
| bar* | comunity_centre* | grave_yard | motorway_junctio* | seguranca socia | trunk_junction |
| battlefield | conference_centr | guest_house | museum* | service | turning_circle* |
| bbq | construction | halt | newspaper* | services* | turntable* |
| beacon | convent | health | newstand | shelter | undefined |
| beauty | courthouse* | health_centre | nightclub* | shop* | university* |
| bed & breakfast | coutada | healthcare | no | shower* | user defined |
| bench | crane* | heritage | nursing_home* | silo* | vending_machine |
| biblias e casa | critpy | horses | oil_tank | snack_bar | veterinary* |
| bicycle_parking | cross | hospital* | old_cafe | social_centre* | viewpoint |
| bicycle_rental | crossing | hostel* | optical | social_facility* | waste_basket |
| biergarten | dentist* | hotel* | park* | solicitor | waste_deposal |
| boundary_stone | disused | hunting_stand | parking* | souvenirshop | waste_disposal |
| bridge* | diving_center | ice_cream* | parking_entrance* | spa | waste_dispostal |

(continued)

**Table 2** (continued)

| | | | | | |
|---|---|---|---|---|---|
| brothel | doctor* | icon | parking_space* | spa | wastewater_plant* |
| buffer_stop | doctors* | incline | passing_place | speed_camera* | water_tank |
| buoy | drinking_water | incline_steep | path | sport clube leir | water_tower* |
| buoy | driving school | incline_up | pharmacy* | station* | water_well |
| bus_station* | driving_school | info | picnic_site | steps | water_works* |
| bus_stop | elevator | information | pier* | stop | waterfall |
| café* | embassy* | junction | pillar buoy | storage_tank | watering_place |
| cairn | emergency_access | kindergarten* | place_of_worship* | street_lamp | watermill |
| caixa geral de d | emergency_phone | laboratory | police* | studio* | wayside_cross |
| camp_site | escola superior | landmark | post_box | subway entrance* | wayside_shrine |
| camping park | ev_charging* | lavoir | post_office* | subway_entrance* | wifi |
| capela | farmacia | lawyer* | posto abastecime | survey point | wind_turbine |
| car_rental* | fast_food* | leisure_centre | primary_link | survey_pillar | windmill |
| car_wash* | ferry_terminal | level_crossing* | prison* | survey_point | works* |
| chalet | fitness_center | lookout_tower | register_office* | telephone | |
| caravan_site | fire_hydrant* | library* | pub* | swimming_pool* | yes |
| castle* | fire_station* | lift | public_building* | taxi | zoo* |
| cemiterio | first_aid | lighthouse | recycling | teahouse | |

*Legend* types marked with asterisk (*) were considered attributable to a CLC class and selected for further analysis

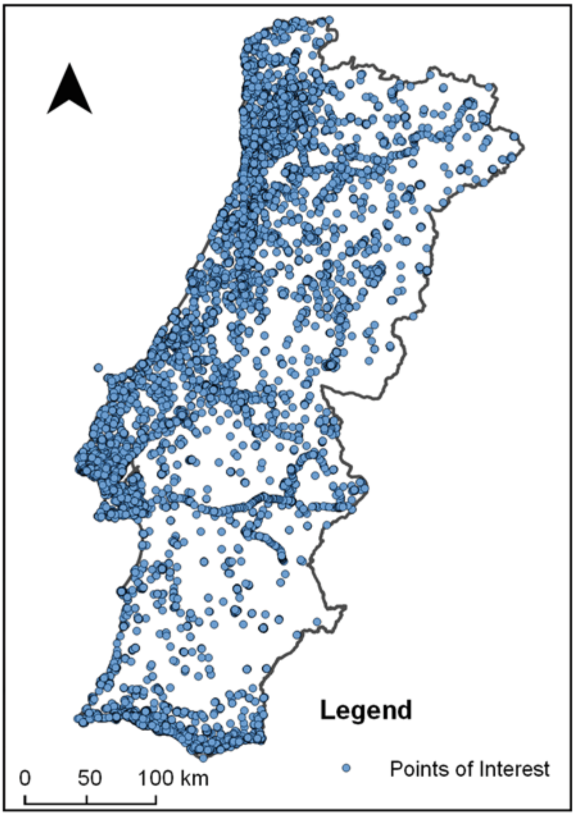

**Fig. 1** Spatial distribution of the points of interest over the study area

classifications. This agreement/disagreement is depicted, along with their spatial distribution, in Fig. 2. Red points represent locations where both classifications are not matching and green points represent locations where both classifications are equal.

Table 4 summarizes the classification of the OSM point accuracy. Points classified as and WB classes obtained 77.96 and 1.47 % correct classification, respectively, when compared with the CLC classification for the same locations. One of the reasons for the poor result of the WB class might be related with the MMU of 25 Ha of the CLC database. It is natural that body areas of small dimension do not represent the predominant class when using such a MMU value.

Finally we analyzed the classification accuracy for each OSM point type. In Table 5, each PoI type is classified according to its range of accuracy. This is important to understand the suitability of each OSM PoI type to use in LULC databases. The lower accuracy of some OSM point type might be also related with the MMU. A “rest_area”, for instance, might be located within a forest crossed by a motor way. In the same way, a “water_tower” might be located within an area where another class is predominant.

**Table 3** CLC classes given to each PoI type

| Class agricultural areas (AA) | | | | | Class water bodies (WB) |
|---|---|---|---|---|---|
| arts_centre | crane | lawyer | post_office | subway entrance | bridge |
| atm | dentist | level_crossing | prison | subway_entrance | ford |
| bank | doctor | library | pub | swimming_pool | pier |
| bar | doctors | marketplace | public_building | theatre | reservoir |
| beauty | embassy | mini_roundabout | register_office | theme_park | |
| bus_station | ev_charging | monument | residential | townhall | |
| cafe | fast_food | motel | rest_area | tram_stop | |
| car_rental | fire_hydrant | motorway_junctio | restaurant | turning_circle | |
| car_wash | fire_station | museum | road | turntable | |
| castle | food_court | newspaper | school | university | |
| charging station | fort | nightclub | services | veterinary | |
| charging_station | fountain | nursing_home | shop | wastewater_plant | |
| cinema | fuel | park | shower | water_tower | |
| city_gate | gasometer | parking | silo | water_works | |
| clinic | hospital | parking_entrance | social_centre | works | |
| college | hostel | parking_space | social_facility | zoo | |
| community_centre | hotel | pharmacy | speed_camera | | |
| comunity centre | ice_cream | place_of_worship | station | | |
| courthouse | kindergarten | police | studio | | |

**Fig. 2** PoI type class versus CLC class

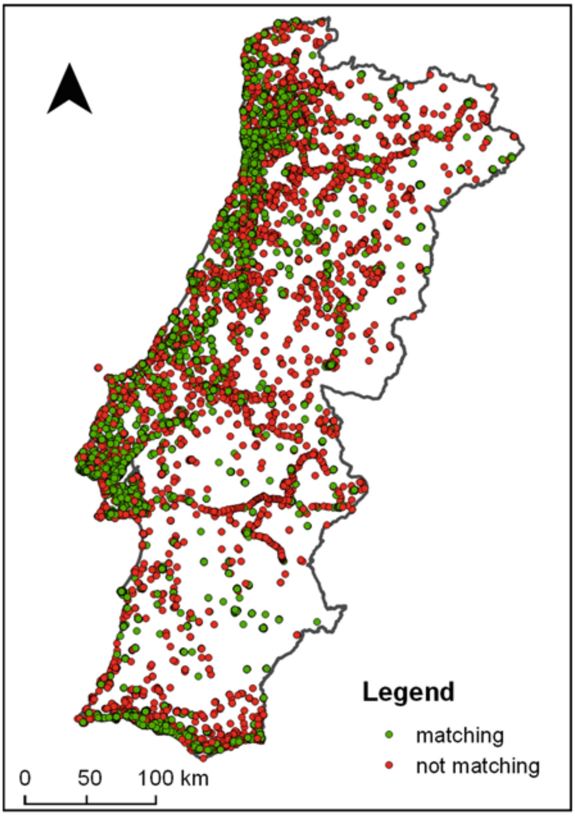

**Table 4** Classification of OSM points

| | | Classification based on OSM points | |
|---|---|---|---|
| | | AS | WB |
| CLC classes containing the point locations | AS | 20,421 | 1 |
| | AA | 4,110 | 46 |
| | F | 1,556 | 20 |
| | W | 19 | 0 |
| | WB | 85 | 1 |
| Total | | 26,191 | 68 |
| Correct (%) | | 77.96 | 1.47 |
| Wrong (%) | | 22.03 | 98.53 |

**Table 5** Classification accuracy by PoI type

| Accuracy classes (%) | | | | | | |
|---|---|---|---|---|---|---|
| 0–50 | 50–60 | 60–70 | 70–80 | 80–90 | 90–100 | 100 |
| water_tower | place_of_worship | works | clinic | fire_station | cinema | charging station |
| castle | social_facility | station | townhall | parking_space | bank | charging_station |
| rest_area | speed_camera | motel | hotel | car_wash | courthouse | comunity centre |
| motorway_junctio | water_works | city_gate | museum | hospital | university | doctor |
| zoo | silo | food_court | parking | bus_station | pharmacy | embassy |
| level_crossing | monument | | mini_roundabout | nightclub | veterinary | ev_charging |
| pier | fire_hydrant | | swimming_pool | arts_centre | theatre | fort |
| theme_park | residential | | turning_circle | kindergarten | police | ice_cream |
| gasometer | studio | | nursing_home | crane | car_rental | lawyer |
| services | | | fuel | public_building | post_office | newspaper |
| wastewater_plant | | | fountain | cafe | library | park |
| beauty | | | hostel | pub | dentist | parking_entrance |
| bridge | | | | fast_food | doctors | prison |
| ford | | | | school | marketplace | register_office |
| reservoir | | | | restaurant | atm | road |
| shower | | | | tram_stop | college | shop |
| turntable | | | | bar | | social_centre |
| | | | | community_centre | | subway entrance |
| | | | | | | subway_entrance |

## 5 Conclusions and Future Research Directions

In this chapter we looked into what has been done, in terms of research, to use VGI, with particular emphasis on OSM data, for LULC database production and we extended our previous investigations in this area. The explored studies have shown that we are still at the beginning, and more research is needed to address the identified issues. Based on the literature, OSM data have been investigated in terms of their suitability to be used as a source of ground-truth or ground measurements for validation purposes, with some studies focusing on urban areas, one of the CLC level 1 classes. Other studies have tried to derive information from the set of unstructured attributes of OSM data.

In the practical part of this study, we explored the OSM PoI dataset and demonstrated its suitability for the purpose of helping in the LULC production process. From the total of 26,191 points classified as class 1, almost 78 % of them matched the classification of the CLC level 1 at the same location. For some point types, the classification accuracy was actually 100 %. In contrast, those points falling into the WB class did not correspond well to the Corine LC databases and therefore more research is needed to compare these results with others at different locations.

All the reviewed literature, along with the practical part, demonstrated the suitability of OSM data to be used in the process of LULC database production. The main contributions in this activity would be in validation or helping the classification process when needed (e.g. when the remote sensing images or aerial photographs are not clear in a particular location). This suitability is different for different classes and for different locations. Urban and touristic areas, for instance, are more likely to have more data available in comparison to other places. This phenomenon is related to many factors including the amount of people living in larger cities, the amount of people visiting touristic places, and the Digital Divide already reported by Goodchild (2008a), among others.

Future research needs to be oriented towards a higher level of detail regarding the nomenclature of CLC. More research is needed to find ways to use these data, not only from OSM but also from other VGI sources, within CLC levels 2 and 3. The integration of different sources would increase the quantity of available data, and in some cases the coverage, proving a way to have a certain level of uncertainty, at least among the available resources. It is also important to compare results among different countries and continents to understand if they are location-dependent. This would help understand whether the potential of using OSM data for LULC production can be generalized or not.

## References

Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic OpenStreetMap quality analysis. Trans GIS. doi:10.1111/tgis.12073

Baulies X, Szejwach G (1997) Survey of needs, gaps and priorities on data for land use/land cover change research. Report presented at LUCC data requirements workshop. Barcelona, Spain, 11–14 Nov 1997

Caetano M, Mata F, Freire S (2006) Accuracy assessment of the Portuguese CORINE Land Cover map. Glob Dev Environ Earth Obs Space:459–467

Cihlar J (2000) Land cover mapping of large areas from satellites: status and research priorities. Int J Remote Sens 21(6–7):1093–1114. doi:10.1080/014311600210092

Clark ML, Aide TM (2011) Virtual interpretation of earth web-interface tool (VIEW-IT) for collecting land-use/land-cover reference data. Remote Sens 3(3):601–620. doi:10.3390/rs3030601

Corine Land Cover Nomenclature (2011) Corine land cover nomenclature illustrated guide. In: Joint meeting Geoland2—EAGLE. Málaga, Spain, 23–24 June. http://sia.eionet.europa.eu/EAGLE/EAGLE_6thMeeting_g2_Malaga/04d_Nomenclature_CLC.pdf. Accessed 5 Jun 2014

Ellis E (2013) Land-use and land-cover change. In: The encyclopedia of earth. http://www.eoearth.org/view/article/51cbee4f7896bb431f696e92. Accessed 10 May 2014

Elwood S, Goodchild MF, Sui DZ (2012) Researching volunteered geographic information: spatial data, geographic research, and new social practice. Ann Assoc Am Geogr 102(3):571–590. doi:10.1080/00045608.2011.595657

Estima J, Painho M (2013a) Exploratory analysis of OpenStreetMap for land use classification. In: Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information—GEOCROWD '13. ACM Press, pp 39–46. doi:10.1145/2534732.2534734

Estima J, Painho M (2013b) Flickr geotagged and publicly available photos: preliminary study of its adequacy for helping quality control of corine land cover. In: Murgante B, Misra S, Carlini M, Torre CM, Nguyen HQ, Taniar D, Gervasi O (eds) ICCSA 2013: computational science and its applications. The 13th international conference on computational science and its Applications, Ho Chi Minh City, Vietnam, 24-27 June 2013 . Lecture notes in computer science, vol 7974. Springer, Heidelberg, pp 205–220. doi:10.1007/978-3-642-39649-6_15

Estima J, Painho M (2014) photo based volunteered geographic information initiatives: a comparative study of their suitability for helping quality control of corine land cover. Int J Agric Environ Inf Syst 5(3):75–92. doi:10.4018/ijaeis.2014070105

European Environment Agency (2014) http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2. Accessed 5 Jun 2014

Fan H, Zipf A, Fu Q, Neis P (2014) Quality assessment for building footprints data on OpenStreetMap. Int J Geogr Inf Sci 28(4):700–719. doi:10.1080/13658816.2013.867495

Flickr (2014) https://www.flickr.com/. Accessed 5 Jun 2014

Foody GM, Boyd DS (2013) Using volunteered data in land cover map validation: mapping West African forests. IEEE J Sel Top Appl Earth Obs Remote Sens 6(3):1305–1312. doi:10.1109/JSTARS.2013.2250257

Fritz S, McCallum I, Schill C, Perger C, Grillmayer R, Achard F, Obersteiner M (2009) Geo-Wiki. Org: The use of crowdsourcing to improve global land cover. Remote Sens 1(3):345–354. doi:10.3390/rs1030345

Fritz S, McCallum I, Schill C, Perger C, See L, Schepaschenko D, Obersteiner M (2012) Geo-Wiki: an online platform for improving global land cover. Environ Model Softw 31:110–123. doi:10.1016/j.envsoft.2011.11.015

Geofabrik (2014) http://www.geofabrik.de/. Accessed 5 Jun 2014

GMapCreator (2014) http://www.bartlett.ucl.ac.uk/casa/latest/software/gmap_creator. Accessed 5 Jun 2014

Goodchild M (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4):211–221. doi:10.1007/s10708-007-9111-y

Goodchild M (2008a) Assertion and authority: the science of user-generated geographic content. In: Proceedings of the Colloquium for Andrew U. Frank's 60th birthday. Department of Geoinformation and Cartography, Vienna, Austria

Goodchild M (2008b) Commentary: whither VGI? GeoJournal 72(3–4):239–244. doi:10.1007/s10708-008-9190-4

Goodchild M, Glennon JA (2010) Crowdsourcing geographic information for disaster response: a research frontier. Int J Digit Earth 3(3):231–241. doi:10.1080/17538941003759255

Google MyMaps (2014) https://www.google.com/maps/d/. Accessed 5 Jun 2014

Hagenauer J, Helbich M (2012) Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. Int J Geogr Inf Sci 26(6):963–982. doi:10.1080/13658816.2011.619501

Hecht R, Kunze C, Hahmann S (2013) Measuring completeness of building footprints in OpenStreetMap over space and time. ISPRS Int J Geo-Inf 2(4):1066–1091. doi:10.3390/ijgi2041066

Hollenstein L, Purves R (2010) Exploring place through user-generated content: using Flickr to describe city cores. J Spat Inf Sci 1(1):21–48. doi:10.5311/JOSIS.2010.1.3

Holone H, Misund G, Holmstedt H (2007) Users are doing it for themselves: pedestrian navigation with user generated content. In: International conference on next generation mobile applications, services and technologies. IEEE, pp 91–99. doi: 10.1109/NGMAST.2007.4343406

Hudson-Smith A, Batty M, Crooks A, Milton R (2009) Mapping for the masses: accessing web 2.0 through crowdsourcing. Soc Sci Comput Rev 27(4):524–538. doi:10.1177/0894439309332299

Jokar Arsanjani J, Helbich M, Bakillah M (2013a) Exploiting volunteered geographic information to ease land use mapping of an urban landscape. In: International archives of the photogrammetry, remote sensing and spatial information sciences. 29th Urban data management symposium, vol XL-4/W1. London, United Kingdom, 29–31 May 2013

Jokar Arsanjani JJ, Helbich M, Bakillah M, Hagenauer J, Zipf A (2013b) Toward mapping land-use patterns from volunteered geographic information. Int J Geogr Inf Sci 27(12):2264–2278. doi:10.1080/13658816.2013.800871

Jokar Arsanjani J, Vaz E (2015: in-press) An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. Int J Appl Earth Obs Geoinf

Leung D, Newsam S (2010) Proximate sensing: Inferring what-is-where from georeferenced photo collections. In: Conference on computer vision and pattern recognition CVPR. IEEE, San Francisco, CA, pp 2955–2962, 13–18 June 2010. doi:10.1109/CVPR.2010.5540040

London Profiler (2014) http://128.40.111.250/casa/websites/profiler.asp. Accessed 5 Jun 2014

MapTube (2014) http://www.maptube.org/. Accessed 5 Jun 2014

Mooney P, Corcoran P, Winstanley A (2010) A study of data representation of natural features in OpenStreetMap. In Proceedings of the 6th GIScience international conference on geographic information science, vol 150. Zurich, Switzerland, 14–17 Sept 2010

OpenStreetMap Map Features (2014) http://wiki.openstreetmap.org/wiki/Map_Features. Accessed 5 Jun 2014

Perger C, Fritz S, See L, Schill C, Van Der Velde M, Mccallum I, Obersteiner M (2012) A campaign to collect volunteered geographic information on land cover and human impact. In GI Forum 2012: Geovizualisation, Society and Learning. pp 83–91

Turner AJ (2006) Introduction to neogeography. Sebastopol, CA

Wikimapia (2014) http://wikimapia.org/. Accessed 5 Jun 2014

Zook M, Graham M, Shelton T, Gorman S (2010) Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian Earthquake. World Med Health Policy 2(2):6–32. doi:10.2202/1948-4682.1069