Exploratory analysis of OpenStreetMap for land use classification

Jacinto Estima
ISEGI – Universidade Nova de Lisboa
Lisboa
Portugal
Jacinto.estima@gmail.com

Marco Painho
ISEGI – Universidade Nova de Lisboa
Lisboa
Portugal
painho@isegi.unl.pt

ABSTRACT
In the last years, volunteers have been contributing massively to what we know nowadays as Volunteered Geographic Information. This huge amount of data might be hiding a vast geographical richness and therefore research needs to be conducted to explore their potential and use it in the solution of real world problems. In this study we conduct an exploratory analysis of data from the OpenStreetMap initiative. Using the Corine Land Cover database as reference and continental Portugal as the study area, we establish a possible correspondence between both classification nomenclatures, evaluate the quality of OpenStreetMap polygon features classification against Corine Land Cover classes from level 1 nomenclature, and analyze the spatial distribution of OpenStreetMap classes over continental Portugal. A global classification accuracy around 76% and interesting coverage areas' values are remarkable and promising results that encourages us for future research on this topic.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications---Spatial databases and GIS.

General Terms
Management, Experimentation, Design.

Keywords
Volunteered Geographic Information, GIS, OpenStreetMap, Land Use.

1. INTRODUCTION
Volunteered Geographic Information (VGI), term coined by Michael Goodchild in 2007 to describe geographic information produced by large numbers of engaged private citizens [6], has become exponentially available over the Web in the last years. The inventory made by Elwood et al. in 2009 identified ninety-nine VGI initiatives running [3]. In order to explore the potential application of this big quantity of spatial data in the solution of real world problems, research has already been conducted in some areas such as crisis and emergency response [7, 17], vernacular geography [9], navigation [10], land use/cover (LULC) validation [4, 5], among others. To our best knowledge, there is no study trying to explore data from OpenStreetMap (OSM), one of the best known and most studied VGI initiatives [3], for land use/cover production.

GEOCROWD '13:, November 05 - 08 2013, Orlando, FL, USA

The aim of this study is to conduct an exploratory analysis of the OSM database for land use/cover production, using the European Land Cover database titled Corine Land Cover (CLC) as reference data. Our main contributions to the research community are as follows:

- We establish a tentative to relate both nomenclatures, for the purpose of this paper;
- We evaluate the quality of OSM land use classification over continental Portugal taking CLC as reference data, to assess if it can be used as ground truth for LULC validation in the future.

This paper is structured as follows. After a brief introduction some related work is presented. We then describe the data and methods used followed by the results and discussion. The paper ends with some conclusions and possible future research directions.

2. RELATED WORK
VGI along with Neogeography [16] and Crowdsourcing geospatial data [11] are all terms related with a "spatial" type of User Generated Content (UGC) contributed by volunteers, a function that for centuries has been endorsed exclusively to official agencies [6]. Although the participation of citizens is not new, it has been exponentially growing in the last years, mainly due to the evolution of some important technologies like the Web 2.0, Google Maps, broadband communications, cheaper positioning devices integrated in cameras and smartphones, among others, and the disposition of private citizens to contribute for many reasons [3, 8]. 70 percent of the initiatives counted in the already mentioned inventory made by Elwood et al. in 2009 [3] started after 2005, year when Google Maps was launched. While the major issues of this type of data are related with their heterogeneity, absence of formal structures and quality control procedures, absence metadata, etc., the major advantages are related with their quantity, temporal coverage [14] and the local knowledge of its contributors that know their surroundings better than any outsider [8].

Initiatives can come from a variety of entities like government agencies, academia, individuals, commercial or for-profit companies, etc. HD TrafficTM initiative from TomTom[1] is a good example of a private company collecting data from users to provide information about traffic in real time more accurately [8]. Other important and well known initiatives are OSM[2] already

[1] http://www.tomtom.com/en_gb/services/live/hd-traffic/

[2] http://www.openstreetmap.org

mentioned before, Wikimapia[3], Flickr[4], Map Tube[5], "Did You Feel it?"[6], among many others.

The portal ''Did you feel it?'' is an initiative started around 1999 by the United States Geological Survey (USGS) Earthquake Hazards Program for earthquake mapping that aims to collect and provide information about peoples' feelings and experiences on earthquake activities based on their location [8].

Based on Google Maps, Wikimapia, adapted from the successful Wikipedia project, is one of the first initiatives using this platform, where people with an Internet connection can select any place in a world map by drawing their boundaries, and provide a description and other relevant information about that location [6].

Flickr is another well know initiative, started in 2004, composed by an online application where people can upload photos and save them in a database along with some additional information in the form of tags. Special "geo" tags are also available to store latitude and longitude values, automatically retrieved from smartphones, making them a source of geographical data [9]. In 2010 Kisilevich et al. downloaded a total of 86,314,466 geotagged photos from Flickr initiative to study peoples' activities [13]. A set of tools allowing the non-professional user to integrate their data have been developed by the Centre for Advanced Spatial Analysis (CASA). MapTube is one of those tools, known as a "*place to put maps*" and based on the generic idea of YouTube, where users can share their own information as thematic maps [11].

OSM is one of the best known and most studied VGI initiatives [3]. It is a project developed by the OpenStreetMap Foundation that aims to provide geographic data, such as street maps, for free to anyone. Users can contribute in two ways: 1) by accessing OSM website and add, edit, delete or update data directly or using desktop applications or; 2) using GPS enabled equipment's to collect data directly from the field that is afterwards uploaded to the OSM servers directly from the application or via the OSM website. Both way, metadata like descriptions, names, among many other tags, can be saved along with the geographic features and everything can be edited, deleted or updated at any time. The access and use of the data can be made in various ways: 1) by accessing the OSM website that has its own rendering; 2) through other websites that have OSM maps embedded; 3) by embedding OSM maps in their own websites using their Applications Programming Interface (API); 4) by downloading the raw data or other derivative subsets, respecting always their license[7].

The research community already conducted several studies exploring VGI data for different purposes. In 2010, Leung and Newsam conducted some experiments to derive maps of what-is-where from large collections of georeferenced photos in an automated way achieving almost 75% classification accuracy with their approach [14]. In 2013, Estima and Painho explored the possibility of using Flickr photos as a source of truth data to help in the accuracy assessment phase of land use/cover production [4].

Geo-Wiki.Org is another remarkable project that uses a global network of volunteers to help improving the quality of global land cover maps [5]. It is a platform based on Google Earth (GE), where areas of disagreement between three global land cover databases ("GLC-2000", "MODIS", and "GlobCover") are identified. The registered volunteers help then in the process of validation, comparing the areas of disagreement with their local knowledge, high resolution imagery from GE and also georeferenced pictures from other VGI projects.

Regarding the investigation using OSM data, of more interest for this study, Over et al. studied in 2010, for the first time, the possibility of generating interactive 3D City Models based on free geo-data available from OSM, and public domain height information provided by the Shuttle Radar Topography Mission [15]. They investigated also the possibility of adding value to OSM using Location Based Services (LBS).

Al-Bakri and Fairbairn used OSM and Ordnance Survey (OS) to give one step towards the integration of geospatial datasets from varied sources. They focus on measuring semantic and structural similarities between categories of formal and VGI data [1].

3. MATERIALS AND METHODS

In this chapter we introduce the exploratory part of this study. We start by presenting and describing the study area as well as the data used for this study. We finish explaining the methodology used to accomplish our objective.

3.1 Study area and datasets used

The defined study site is Continental Portugal, located in the southwestern side of Europe, which is constituted with 18 districts and 278 municipalities covering a total area of 8908220.16 Ha. The land cover is mainly composed by 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[8]. Although we could have downloaded the database in the original format, for this exploratory analysis we decided to download the shapefile format that, according to the website, is constituted by a selection of layers where the most important features get exported (road and railway network, forests, water areas and some points of interest). This database is current as of July 23, 2013, and is divided in 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. For the purpose of this study, as one of the objectives was to quantify areas, only the polygon based datasets were considered, e.g. only the levels buildings, land use and natural areas. We are aware that, by leaving behind point and line features, we might be losing important information about some landuse classes, but to include them it would be required a specific study that is outside the scope of this work.

The nomenclature used to classify features in the OSM datasets is available in wiki Website[9], along with pictures and descriptions for each class. Table 1 shows the OSM nomenclature classes identified over continental Portugal for natural areas and landuse classes. Regarding the buildings dataset, as the majority of the features do not have a class defined, it was decided to assign a

[3] http://wikimapia.org

[4] http://www.flickr.com

[5] http://www.maptube.org/

[6] http://earthquake.usgs.gov/earthquakes/dyfi/

[7] The legal terms of the OSM data license can be viewed at http://www.openstreetmap.org/copyright

[8] http://www.geofabrik.de/data/download.html

[9] http://wiki.openstreetmap.org/wiki/Map_Features

generic class "urban" to all of them. It is important to refer that the generalization we are doing for this specific case can have a negative impact, mainly in rural areas and this should be further investigated in the future.

Table 1 - OSM datasets' classes over continental Portugal

"Landuse" classes				
Abutters	Farm	Harbour	Park	Scrubs
Allotments	Farmland	Industrial	Public	University
Basin	Farmyard	Landfill	Quarry	Village_green
Beach	Field	Leisure	Railway	Vineyard
Brownfield	Garages	Meadow	Reservoir	Waste_water_plan
Cemetery	Garden	Military	Residential	Water
Commercial	Grass	Museum	Retail	Wood
Conservation	Greenfield	Not_known	Salt_pond	Greenhouse_horti
Construction	Greenhouse	Orchard	Scrub	Recreation_groun
"Natural areas" classes				
Forest	Park	Riverbank	Water	

The CLC database is composed by the version 16 (04/2012) of Corine Land Cover (CLC) database for the CLC2006 inventory, downloaded from the European Environment Agency (EEA)[10]. 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. The land cover is classified according to the CLC nomenclature, which is hierarchically divided in three levels of classes. Table 4 shows the categories for each level along with the respective covered area.

3.2 Assumptions

For the correct understanding of this study, it is important to refer that we assume that the time difference between CLC and OSM databases (2006 for CLC and 2013 for OSM) would not represent a major issue. Considering a yearly average change value of land cover in Europe of 0.23% [2], for the purpose of this exploratory analysis, we believe that the impact of such change rate between both periods does not depreciate this study. In a more in depth analysis, data from similar periods shall be used. We also assume the CLC database as the truth classification that is therefore our reference data.

3.3 Methods

The adopted methodology to conduct this exploratory analysis, summarized in Figure 1, was as follows:

1. Analysis of the defined OSM datasets. We have explored the three polygon based OSM datasets defined in the previous section in terms of nomenclature and area of coverage. We have also analyzed the areas of overlap to identify eventual existing inconsistencies;
2. Analysis and establishment of a relationship between the classification nomenclatures used by the different databases (CLC and OSM). In this step we tried to establish a correspondence between CLC and OSM classes defined by their respective nomenclatures, extremely important to develop the subsequent steps in this methodology;
3. Analysis of the coverage of each OSM class using CLC level 1 as reference. As shown in Figure 1, and according to the relationship between OSM and CLC established in the previous step, we first merged all the OSM datasets and gave each OSM class the corresponding CLC level 1 class. We have then dissolved all the polygons by each CLC class value to have a resultant map with only 5 classes plus the areas without corresponding CLC class. In the last step we have removed overlapping areas in conflict. Then a comparison between the resultant areas and the correspondent ones from the CLC database was made;
4. Analysis of the matching degree between related classes. In this step, the area covered by each class that matched the correspondent CLC level 1 class was determined by intersecting both datasets, and the accuracy of OSM classification calculated;
5. Analysis of the OSM spatial distribution. In this final step, we intersected the dataset resultant from the previous step with a dataset representing the Portuguese districts, an administrative division that splits the country in 18 areas.

Figure 1 shows a flowchart describing all the spatial analyses developed in this methodology. It is important to refer that in steps 3, 4 and 5 the developed analyses were restricted to the level 1 of the CLC. This was due to multiple correspondence issues detected in the step 2. Solutions to solve this multiple correspondences need further investigation that is outside the scope of this study.

4. RESULTS AND DISCUSSION

In this chapter we present and discuss the results of our study.

4.1 Analysis of OSM datasets

Landuse, natural areas and buildings were the defined OSM datasets to be used in this study. In this first step we will explore these datasets in terms of nomenclature, area of coverage and overlapping areas to identify eventual existing inconsistencies. Table 2 describes the areas of coverage of each dataset in Ha and the representative percentage relative to continental Portugal.

Table 2 - Areas of coverage of OSM datasets

Dataset	Area in Ha	Country coverage (%)
Natural areas	140006.95	1.57%
Landuse	144350.23	1.62%
Buildings	7057.61	0.08%

Table 3 - Existing classification differences whithin the three OSM datasets

Natural areas dataset	Landuse dataset	Buildings dataset	Area (Ha)
Forest	Military	None	5.24
	Residential	Reservoir_cover	0.02
	Recreation_ground	Hospital	0.25
Park	Commercial	None	0.01
	Residential	Museum	0.39
		Cafe	0.05
		Chapel	0.01
		Church	0.00
		House	0.03
		Library	0.03
		Museum	0.08
		Public	0.02
		Public_building	0.37
		Restaurant	0.03
		Roof	0.01
		Theatre	0.03
		Toilets	0.00
		Yes	0.01

[10] http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2

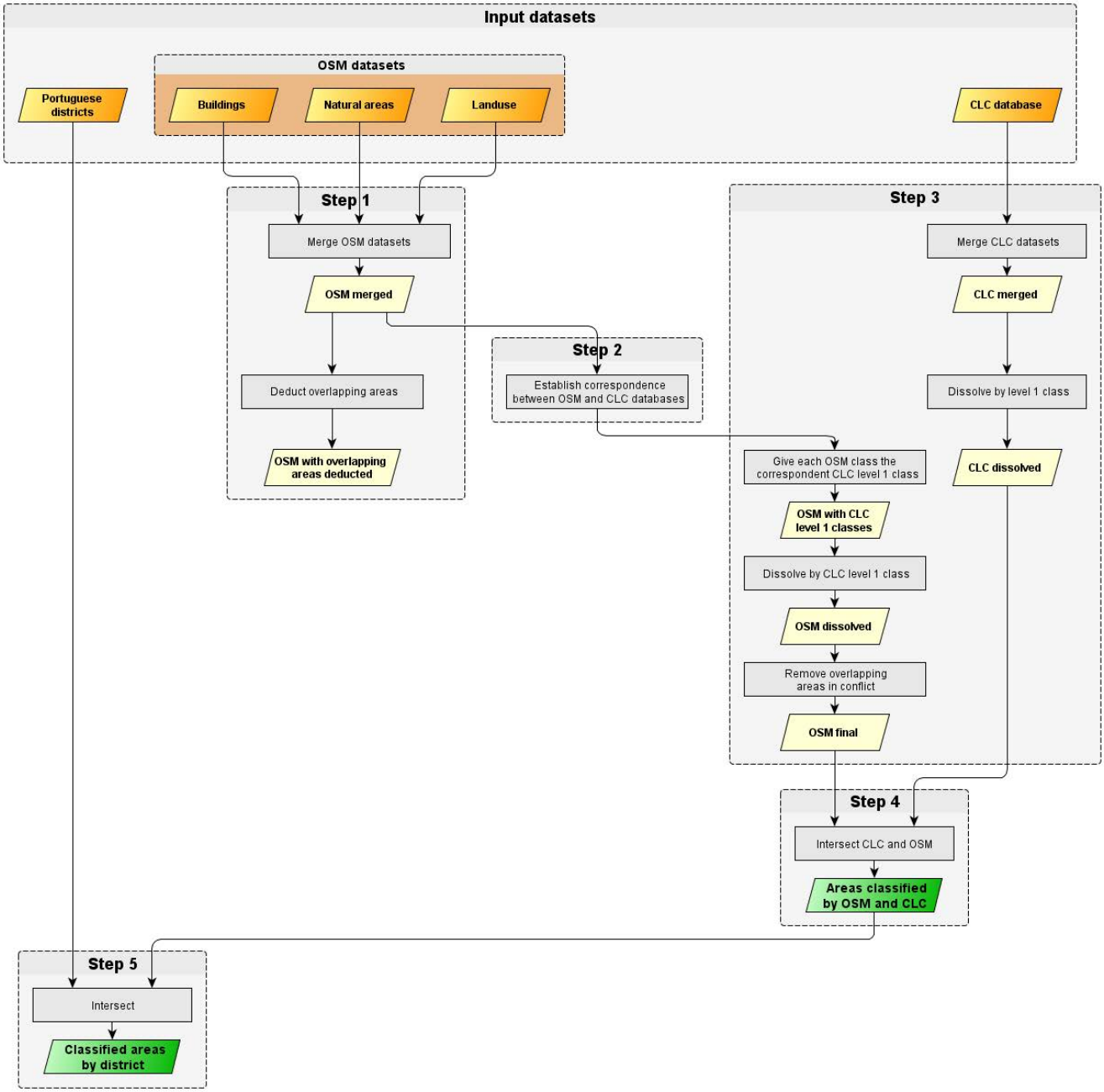

Figure 1 - Flowchart of the spatial analyses developed in this methodology

Landuse is the dataset with bigger coverage, covering 1.62%, followed by the natural areas dataset covering 1.57% and the buildings dataset covering 0.08% of the country. These three datasets together cover a total of 3.27% of the study area. In order to have a more realistic value, once some of the features represented in these datasets are totally or partially superimposed, the overlapping areas were deducted. The determined overlapping area was approximately 3017.18 Ha representing 0.03%, making the real coverage area to decrease by 3.24%.

Before deducting the overlapping areas, the three OSM datasets were also intersected to identify existing classification inconsistencies in those areas. Table 3 summarizes the different classifications recognized in those common areas. These different classifications do not represent a real conflict but rather the combination of different features/classes in the same location, seen probably by their contributors at different scales. A good example of that, extracted from Table 3, would be a place classified as park in natural areas, residential in landuse and café, church or museum, etc. in buildings. This example represents actually something that happens in reality with these datasets.

The total value of overlapping areas with different classification shown in Table 3, 9.47 Ha, is significantly lower that the total area of overlapping areas show above, 3017.18 Ha, which gives us a good indicator that the classification has some consistency.

Table 4 - CLC nomenclature and respective areas for continental Portugal (Source: http://www.igeo.pt/gdr/pdf/CLC2006_nomenclature_addendum.pdf)

Level 1	Area (Ha)	Level 2	Area (Ha)	Level 3	Area (Ha)
1 Artificial surfaces	309716.89	11 Urban fabric	227482.56	111 Continuous urban fabric	12234.34
				112 Discontinuous urban fabric	215248.23
		12 Industrial, commercial and transport units	47821.49	121 Industrial or commercial units	33895.51
				122 Road and rail networks and associated land	7678.06
				123 Port areas	1945.27
				124 Airports	4302.65
		13 Mine, dump and construction sites	21149.09	131 Mineral extraction sites	13659.71
				132 Dump sites	971.58
				133 Construction sites	6517.80
		14 Artificial, non-agricultural vegetated areas	13263.75	141 Green urban areas	1763.71
				142 Sport and leisure facilities	11500.04
2 Agricultural areas	4199177.27	21 Arable land	1245009.51	211 Non-irrigated arable land	981677.22
				212 Permanently irrigated land	210509.59
				213 Rice fields	52822.70
		22 Permanent crops	592974.48	221 Vineyards	228965.31
				222 Fruit trees and berry plantations	100983.22
				223 Olive groves	263025.95
		23 Pastures	41871.11	231 Pastures	41871.11
		24 Heterogeneous agricultural areas	2319322.18	241 Annual crops associated with permanent crops	404000.98
				242 Complex cultivation patterns	607041.55
				243 Land principally occupied by agriculture	686819.25
				244 Agro-forestry areas	621460.40
3 Forest and semi natural areas	4259642.22	31 Forests	2016515.84	311 Broad-leaved forest	1007003.84
				312 Coniferous forest	533981.79
				313 Mixed forest	475530.21
		32 Scrub and/or herbaceous vegetation associations	2074423.48	321 Natural grasslands	171861.61
				322 Moors and heathland	284552.04
				323 Sclerophyllous vegetation	206613.41
				324 Transitional woodland-shrub	1411396.42
		33 Open spaces with little or no vegetation	168702.90	331 Beaches, dunes, sands	11148.98
				332 Bare rocks	23862.88
				333 Sparsely vegetated areas	100830.47
				334 Burnt areas	32860.57
				335 Glaciers and perpetual snow	0.00
4 Wetlands	28777.11	41 Inland wetlands	1138.71	411 Inland marshes	1138.71
				412 Peat bogs	0.00
		42 Maritime wetlands	27638.40	421 Salt marshes	18457.26
				422 Salines	7228.50
				423 Intertidal flats	1952.64
5 Water bodies	110906.66	51 Inland waters	72859.65	511 Water courses	19874.09
				512 Water bodies	52985.56
		52 Marine waters	38047.01	521 Coastal lagoons	8521.46
				522 Estuaries	26680.68
				523 Sea and ocean	2844.87

4.2 Correspondence between OSM and CLC nomenclatures

Each database (CLC and OSM) uses different nomenclatures for classification. It is therefore necessary to find some correspondence between both systems before proceeding to the next steps. Although the OSM wiki page already has a possible correspondence[11], some of the tags present in the study area are not mentioned there. Thus, in Table 5 we propose a tentative to relate both CLC and OSM nomenclatures, developed based on the description of each CLC and OSM class available at the OSM wiki Website mentioned before and the CLC illustrator guide[12], respectively.

Difficulties arouse trying to establish a direct relation between some classes from the two nomenclatures. In this sense, three types of issues occurred: 1) two OSM classes were not identified at all due to absence of any description (case of OSM classes "field" and "not_known") in the OSM wiki; 2) one OSM class didn't match with the description of any CLC (the "military" class); and 3) some OSM classes did not fit in the description of only one CLC class resulting in multiple correspondences. In the first and second cases, a unique correspondence was not possible to provide.

It is noticeable that the difficulty in finding correspondence rises when the level of detail increases, e.g. more multiple correspondences can be verified in the level 3 than in the level 1 of CLC. Actually, for the level 1 only one case of multiple correspondence was identified: the "grass" class. In the description of this class it is stated that it should be used to represent "*areas covered with grass*" and, as a complement, it is

[11]http://wiki.openstreetmap.org/wiki/Corine_Land_Cover

[12]http://sia.eionet.europa.eu/EAGLE/EAGLE_6thMeeting_g2_Malaga/04d_Nomenclature_CLC.pdf

also specified that the user should "*consider landuse=meadow for meadow and landuse=pasture for pasture*". According to the description of CLC level 1 classes, two CLC classes can match this OSM class: agricultural and forest and semi natural areas making it a multiple correspondence case.

Table 5 – Correspondence between CLC and OSM classes

OSM classes	CLC classes		
	Level 3	Level 2	Level 1
Landuse dataset			
Abutters	111-112-121	11-12	1
Allotments	242	24	2
Basin	512	51	5
Beach	331	33	3
Brownfield	133	13	1
Cemetery	111-112	11	1
Commercial	121	12	1
Conservation	313-312-311	31	3
Construction	133	13	1
Farm	222-231-241-242	22-23-24	2
Farmland	222-231-241-242	22-23-24	2
Farmyard	222-231-241-242	22-23-24	2
Field	**?**	**?**	**?**
Garages	122	12	1
Garden	142	14	1
Grass	231-321	23-32	2-3
Greenfield	321-322-323-324	32	3
Greenhouse	211	21	2
Greenhouse_horti	211	21	2
Harbour	123	12	1
Industrial	121	12	1
Landfill	132	13	1
Leisure	142	14	1
Meadow	231	23	2
Military	**?**	**?**	**?**
Museum	121	12	1
Not_known	**?**	**?**	**?**
Orchard	222-241	22-24	2
Park	142	14	1
Public	121	12	1
Quarry	131	13	1
Railway	122	12	1
Recreation_groun	142	14	1
Reservoir	512	51	5
Residential	111-112	11	1
Retail	121	12	1
Salt_pond	422	42	4
Scrub	324-323-322-321	32	3
Scrubs	324-323-322-321	32	3
University	121	12	1
Village_green	141	14	1
Vineyard	221	22	2
Waste_water_plan	121	12	1
Water	511-512	51	5
Wood	313-312-311	31	3
Natural areas dataset			
forest	313/312/311	31	3
park	313/312/311	31	3
riverbank	512/511	51	5
water	523/522/511/512/511	52/51	5

For the next steps we will use the level 1 classes of CLC database and we will assume that the OSM "grass" class only has one correspondent CLC level 1 class that is class 3, forest and semi natural areas.

4.3 Coverage analysis of OSM datasets

In this analysis we used the OSM merged dataset from the previous step and gave to each feature the corresponding CLC level 1 class. Then we dissolved the resultant dataset by CLC level 1 class and removed overlapping areas in conflict, e.g. all the overlapping areas with a different CLC level 1 class were removed. These areas perform a total of 4004.05 Ha representing 1.39% of the OSM area. It is important to refer that these areas were not deducted but totally removed from the analysis. We then calculate the coverage area of each new class group and compare them with those from CLC database.

Table 6 shows the results of this analysis. For each class we have the corresponding area from the CLC database in the second column and the area from OSM database in the third column. The fourth and fifth columns shows, the percentage covered by each OSM class over each respective CLC class and over continental Portugal, respectively.

Some interesting indicators can be seen in Table 6. Comparing the coverage area, by class, between OSM and CLC, class 5 has a very interesting value of 74.5% followed by class 1 covering 20.15%. Class 2, 3 and 4 have poor coverage with values under 10%. The "unclassified" areas, OSM classes without correspondent CLC level 1 class, represent a total of 7036.75 Ha that, comparing with the other values displayed in Table 2, covers 0.08 % of the country.

Table 6 - Coverage areas from CLC level 1 and OSM

CLC classes	Area from CLC (Ha)	Area from OSM (Ha)	Class coverage (%)
unclassified	---	7036.75	---
1	309716.89	62407.48	20.15
2	4199177.27	34309.93	0.82
3	4259642.22	98536.62	2.31
4	28777.11	64.59	0.22
5	110906.66	82621.61	74.50

4.4 Analysis of OSM classification accuracy

In this step the verification of classifications in overlapping areas was made. We based this analysis using a confusion matrix shown in Table 7. Values in shaded cells represent areas with the same classification in both databases.

Table 7 - Confusion matrix of CLC vs. OSM classifications

		OSM classes					
		1	2	3	4	5	Total
CLC classes	1	**44160.56**	1059.00	4086.69	0.00	663.20	52369.87
	2	12934.72	**31884.28**	10716.09	4.94	12088.20	68459.87
	3	5182.27	1214.07	**83362.66**	0.07	6322.15	99843.05
	4	42.27	114.77	238.65	**59.57**	4402.91	4870.53
	5	87.66	37.81	132.53	0.00	**59145.14**	59433.67
	Total	62407.48	34309.93	98536.62	64.59	82621.61	284976.99

Some calculations can be derived from Table 7 to have an idea about the classification provided by OSM comparing with the one obtained using CLC.

The accuracy index for each CLC class is an important indicator that shows which are the classes where the areas wrongly classified are higher. It is calculated dividing the area correctly classified in each OSM class (diagonal cell in the table) by the total area of each CLC class (sum of each line).

$$Class\ Accuracy = \frac{e_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}}$$

*(where **e** represents the value, **i** the line index and **j** the column index)*

The Global Accuracy (GA) represents the proportion of area where the classification matches in both databases over the total overlapping area, given by the formula:

$$Global\ Accuracy = \frac{\sum_{i=1}^{n} e_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}}$$

*(where **e** represents the value, **i** the line index and **j** the column index)*

Table 8 shows the resultant values for the accuracy of each class and the global accuracy. Class 4 obtained the worse result, around 1.2% followed by class 2 with an interesting value of 46.6%. All the other classes had very encouraging results with class 5 getting an impressive accuracy value of 99.5%. The GA value is also very interesting and promising around 76.7%.

Table 8 - Classification accuracy

Class	Classification accuracy (%)
1	84.3%
2	46.6%
3	83.5%
4	1.2%
5	99.5%
Global	76.7%

4.5 Analysis of the OSM spatial distribution

In this final step the spatial distribution of OSM areas were analyzed, using the dataset resultant from the previous step. Figure 2 and Figure 3 shows the spatial distribution of all OSM classified areas and the distribution of classes' coverage areas by continental Portuguese districts, respectively.

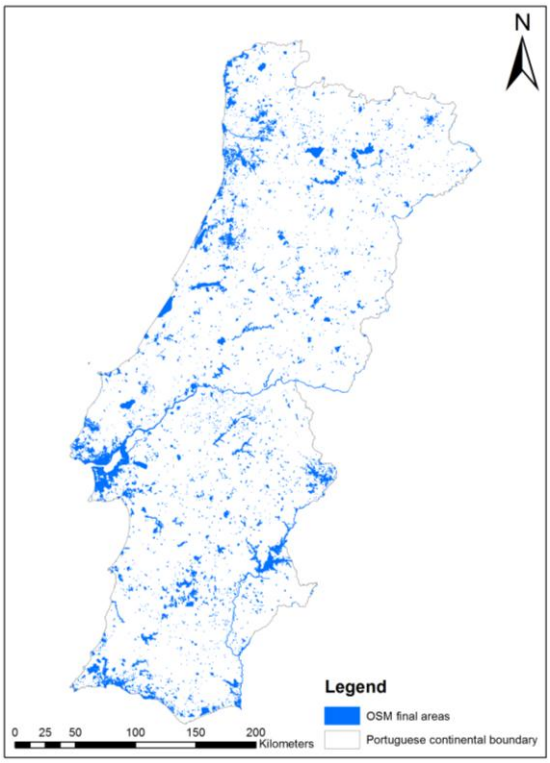

Figure 2 - Spatial distribution of OSM classified areas over continental Portugal

Figure 2 gives a visual insight about the overall distribution and Figure 3 displays, using bar charts, the distribution by class and district. The districts on the south and west coast have more coverage unlike the interior center and north that has significantly lower areas classified. This can be explained by: a) the asymmetric distribution of the Portuguese population over the country b) the seasonal concentration of tourists in the seaside regions during beach epochs. According to Statistics Portugal, the Portuguese official statistics institution, there is a higher concentration of young people on the west coast and bigger cities [12]. More population means more people able to act as volunteers in this type of initiatives while, at the same time, younger population is more likely to be opened and used to new technologies that are in the basis for these projects.

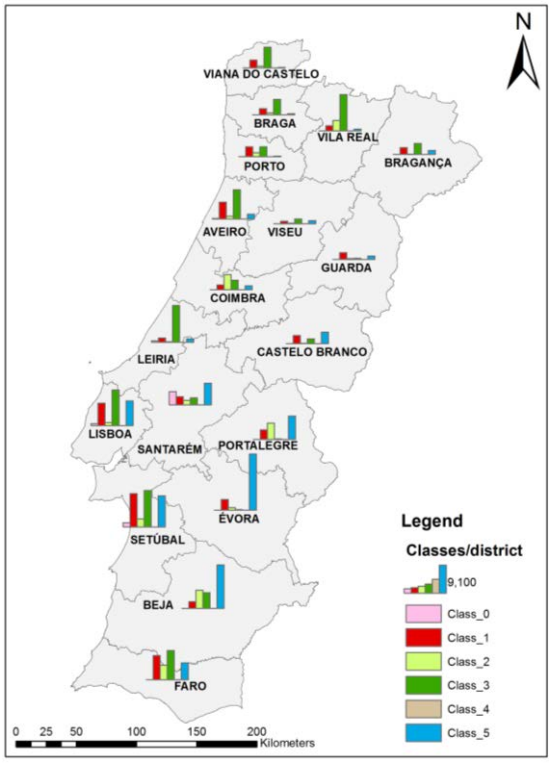

Figure 3 - Distribution of classes' coverage areas by continental Portuguese districts

5. CONCLUSIONS

In this paper we provided a first tentative to relate OSM and CLC nomenclatures, determined the accuracy of OSM polygon features classification based on CLC level 1 classes and continental Portugal area, and analyzed their spatial distribution.

As stated before, only the most important OSM features were used in this study. We are aware that by leaving behind features considered "less important", we might be influencing the results.

While the CLC level 1 classification is course, this was an exploratory study where the final results are showing that it might be worth to study the more detailed CLC level 2 and level 3.

The OSM and CLC correspondence showed some issues that need further research. Actions need to be taken to harmonize multiple correspondences, from OSM to all the three levels of CLC nomenclature, and ways to avoid classes like "not_known" or without any description need to be investigated. Also further investigation should be done to understand what is causing discrepancies between the two classification systems, mainly to see if there are errors in classification or if the issues are more related with differences between the diverse systems.

Although the conflicting overlapping areas, representing 1.39% of the OSM area, should not represent a significant impact, attention also needs to be drawn to these cases to understand their real effect. Measures providing a certain level of trust on the contributors of those specific classifications might help to decide if one class can be more reliable than the others, solving the conflict.

The coverage analysis of OSM datasets showed an impressive result for class 5 and a reasonable value for class 1. The other classes do not have significant representation. Nevertheless, further research is needed to verify, for instance, if these interesting results are more concentrated in some particular locations or if they are well distributed over the entire study area.

OSM datasets revealed remarkable results in terms of classification accuracy with a global value of 76.7%. Although there is still a value of 23.3% of error, we believe it can be used, for instance, as another source of ground truth data for the validation process of LULC databases. In this sense it is very important to identify and understand the causes of this error. The results have also shown that not all the classes have similar accuracy values and therefore some might be more suitable and reliable than others.

Regarding the spatial distribution, as expected also from previous studies [4], the asymmetry between west and east sides of continental Portugal and the concentration of people near the biggest cities and seaside regions was confirmed.

All these results suggest that this source of VGI information might be very useful for LULC classification, at least for classes with a good coverage and simultaneously interesting levels of accuracy, such as classes 1 and 5. Applications such as LULC validation, monitoring or even change detection might have some advantage in using this source.

In the future, our plan is to conduct research that allows us to find solutions for the identified issues. In this sense, we plan to conduct a more in depth study of the overlapping areas with classification conflict and the multiple correspondences in diverse nomenclatures. The study of CLC level 2 and level 3 is also strategic for a more in depth analysis.

6. REFERENCES

[1] Al-Bakri, M. and Fairbairn, D. 2012. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *International Journal of Geographical Information Science*. 26, 8 (Aug. 2012), 1437–1456.

[2] Büttner, G., Kosztra, B., Maucha, G. and Pataki, R. 2012. *Implementation and achievements of CLC2006*.

[3] Elwood, S., Goodchild, M.F. and Sui, D.Z. 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*. 102, 3 (May. 2012), 571–590.

[4] Estima, J. and Painho, M. 2013. Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. *Computational Science and Its Applications – ICCSA 2013*. B. Murgante, S. Misra, M. Carlini, C.M. Torre, H.-Q. Nguyen, D. Taniar, B.O. Apduhan, and O. Gervasi, eds. Springer Berlin Heidelberg. 205–220.

[5] Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F. and Obersteiner, M. 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*. 1, 3 (Aug. 2009), 345–354.

[6] Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*. 69, 4 (Nov. 2007), 211–221.

[7] Goodchild, M. and Glennon, J.A. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*. 3, 3 (Sep. 2010), 231–241.

[8] Heipke, C. 2010. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*. 65, 6 (Nov. 2010), 550–557.

[9] Hollenstein, L. and Purves, R. 2010. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*. 1, 1 (Jul. 2010), 21–48.

[10] Holone, H., Misund, G. and Holmstedt, H. 2007. Users Are Doing It For Themselves: Pedestrian Navigation With User Generated Content. *International Conference on Next Generation Mobile Applications, Services and Technologies* (2007).

[11] Hudson-Smith, A., Batty, M., Crooks, A. and Milton, R. 2009. Mapping for the Masses: Accessing Web 2.0 Through Crowdsourcing. *Social Science Computer Review*. 27, 4 (Apr. 2009), 524–538.

[12] INE 2011. *Censos 2011 – Resultados Provisórios*.

[13] Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N. and Andrienko, G. 2010. Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. *2010 14th International Conference Information Visualisation* (Jul. 2010), 289–296.

[14] Leung, D. and Newsam, S. 2010. Proximate sensing: Inferring what-is-where from georeferenced photo collections. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Jun. 2010), 2955–2962.

[15] Over, M., Schilling, A., Neubauer, S. and Zipf, A. 2010. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems*. 34, 6 (Nov. 2010), 496–507.

[16] Turner, A.J. 2006. *Introduction to Neogeography*.

[17] Zook, M., Graham, M., Shelton, T. and Gorman, S. 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*. 2, 2 (Jan. 2010), 6–32.