A METHODOLOGY FOR ASSESSING OPENSTREETMAP DEGREE OF COVERAGE FOR PURPOSES OF LAND COVER MAPPING

A. Ribeiro [a, b] *, C.C. Fonte [c]

[a] Civil Engineering Department, Coimbra Institute of Engineering, Coimbra, Portugal – alexr@isec.pt
[b] Centre for Informatics and Systems (CISUC), University of Coimbra, Coimbra, Portugal
[c] Department of Mathematics, University of Coimbra / INESC Coimbra, Coimbra, Portugal – cfonte@mat.uc.pt

Commission II, WG II/4

KEY WORDS: Volunteered Geographic Information (VGI), OpenStreetMap (OSM), Coverage, Land Cover Map Creation, Land Cover Map Validation

ABSTRACT:

The data available in the collaborative project OpenStreetMap (OSM) is in some locations so detailed and complete that it may provide useful data for Land Cover Map creation and validation. However, this degree of detail is not uniform along space. Therefore, one of the first requirements that needs to be assessed to determine if the creation and validation of Land Cover Maps using data available in OSM may be feasible, is the availability of data to provide a relatively complete coverage of the region of interest. To provide a fast and automatic quantitative assessment of this requirement a methodology is presented and tested in this article. Four study areas are considered, all located in Europe. The results show that the four regions presented very different coverages at the time of data download and its spatial distribution was not uniform. This approach enabled the identification of the most problematic regions for land cover mapping, where low levels of data coverage are available. Since the proposed methodology can be automated, it enables a fast identification of the regions that, in a preliminary analysis, may be considered fit for further analysis to assess fitness for use for Land Cover Map creation and/or validation.

1. INTRODUCTION

Land Cover Maps (LCM) are fundamental for many applications, such as environmental planning, climate change analysis or hydrologic modelling (Foody, 2002; Verburg et al. 2011; Nie et al. 2011). These maps are usually created through the classification of satellite imagery and are validated using reference databases that are created either through photo-interpretation of satellite or aerial images and/or field visits. However, alternative approaches both for LCM creation and validation have been tried, using alternative data sources (e.g. Fritz et al, 2012, See, et al., 2013, Foody and Boyd, 2013, Jokar Arsanjani et al., 2013, Estima et al., 2014).

The term Volunteer Geographical Information (VGI) is a term proposed by Goodchild (2007) that refers to geographical information provided voluntarily by individuals. This type of data can be also referred to as, for example, Collaboratively Contributed Geospatial Information (Birsh and Kuhn, 2007; Keßler et al., 2009), or Contributed Geographical Information (Harvey, 2013).

OpenStreetMap (OSM) is one of the most well-known VGI projects. It includes vector data about a large diversity of features, such as Buildings, Highways, Waterways, Landuse, Natural features and Points of Interest (OSM Wiki, 2014). The data created is open and can be copied, distributed and changed as long as credit is given to OSM. The data is created and edited continuously by the volunteers, and therefore the available information has a dynamic nature, which has the potential to enable a fast adaptation to the changing world. However, the data available at OSM presents very heterogeneous characteristics, regarding both the amount of data available and its quality (Mooney et al. 2010). This heterogeneity goes from regions with an impressive quantity and quality of information, which can even be more complete than authoritative data (Neis et al., 2011), to regions with no data at all.

The data available in OSM is so detailed in some regions that it enables the creation of LCM (Jokar Arsanjani et al., 2013). Its use for LCM validation was already made as auxiliary information (Bontemps et al., 2011; Fonte et al., 2015). Its potential use as the only source of data was also already analyzed and tested (Martinho and Fonte, 2015; Estima and Painho, 2013). However, the use of OSM for these applications requires that the data available has enough quality; which can be assessed in its several aspects, such as positional and thematic quality, completeness, currency and logical consistency. Since the assessment of data quality in all these dimensions is not an easy and fast process, before starting the assessment of the traditional data quality aspects, a preliminary analysis may be done to determine if enough data is available. Therefore, a first step to determine the fitness for use of OSM data for LCM purposes may be to assess its availability. To make this initial assessment, since LCM are spatially exhaustive, and therefore no empty space is supposed to exist, the degree of spatial coverage (used in this article to express the percentage of space with available data) is the aspect used to determine if the data may be considered or not as potentially usable for LCM creation and/or validation.

In this article an automated methodology is presented to determine the OSM data coverage for a grid of cells with user defined size. The proposed operator is applied to several case

* Corresponding author

studies, the obtained results are presented and conclusions are drawn.

2. METHODOLOGY

The diversity of features available in OSM enables the classification of the majority of phenomena occurring at the earth surface. Therefore, the aim of the operator developed is to assess the proportion of space occupied by the spatial elements available in OSM for a grid of cells chosen by the user, providing the here called degree of coverage of the earth surface for each cell.

The data available in OSM is in a vector format, consisting of points, lines and polygons. Since point and linear features do not occupy any area in the terrain, those that represent phenomena that in reality have area extents and are relevant for land cover, such as roads and waterways, need to be assigned to an area feature. Therefore, some pre-processing of this type of data is necessary to assess the degree of coverage. This is done considering buffer regions around these features, with an extent dependent on the type of feature and local characteristics (Jokar Arsanjani et al., 2013).

The proposed operator to assess OSM data coverage was developed for data downloaded using the methodology adopted by Geofabrik for freely available shapefiles, consisting of the following levels: Roads (which is a sub-feature of Highway), Railways, Waterways, Landuse, Natural and Buildings. The proposed methodology includes the following steps, indicated in Figure 1:

1. Extract data from the considered OSM levels;
2. Create buffers around the linear features that should have an area extent, namely roads, railways and water lines. The buffer widths applied to roads and water lines are based on their category. For the roads different widths are considered for primary roads, secondary and residential roads and for the water features different widths are associated to rivers, streams and canals;
3. Merge OSM features (polygons) of all levels;
4. Create a grid of rectangular cells and compute the area of each cell;
5. Overlay the grid and the OSM map layers, identifying the intersection of the grid cells with the OSM features;
6. Dissolve OSM features boundaries by cell ID;
7. Compute the area of the OSM generated feature included in each cell and join this information to the grid, using the cell ID;
8. For each grid cell compute the coverage index (CI) given by Equation (1).

$$CI = \frac{Area\ occupied\ by\ OSM\ features\ included\ in\ the\ cell}{cell\ area} \qquad (1)$$

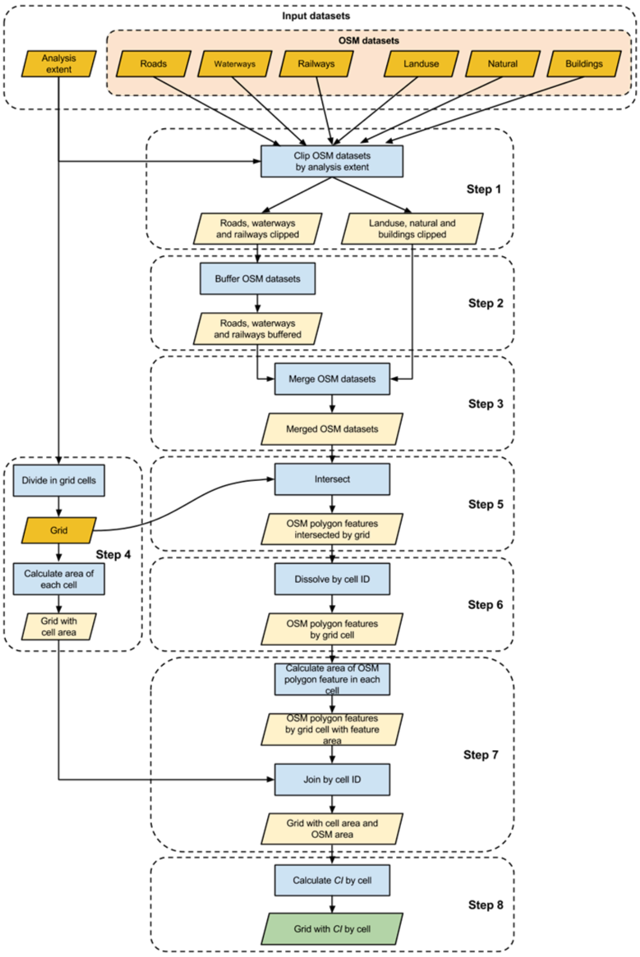

Input datasets
OSM datasets
Analysis extent
Roads
Waterways
Railways
Landuse
Natural
Buildings
Clip OSM datasets by analysis extent
Step 1
Roads, waterways and railways clipped
Landuse, natural and buildings clipped
Buffer OSM datasets
Step 2
Roads, waterways and railways buffered
Merge OSM datasets
Step 3
Merged OSM datasets
Divide in grid cells
Grid
Step 4
Calculate area of each cell
Grid with cell area
Intersect
Step 5
OSM polygon features intersected by grid
Dissolve by cell ID
Step 6
OSM polygon features by grid cell
Calculate area of OSM polygon feature in each cell
OSM polygon features by grid cell with feature area
Step 7
Join by cell ID
Grid with cell area and OSM area
Calculate CI by cell
Step 8
Grid with CI by cell

Figure 1. Workflow of the methodology used to assess OSM data degree of coverage.

3. CASE STUDIES

The proposed operator was applied to four regions. The chosen regions are all located in Europe, namely in the United Kingdom (UK), France, Ireland and Portugal. Regions were considered where different degrees of coverage were expected to exist, according to the data provided by Geofabrik (http://www.geofabrik.de/).

3.1 Study Areas

The studied areas were: Greater London in the UK; the city of Dublin in Ireland; the city of Paris in France; and the municipality of Coimbra in Portugal. Figure 2 shows the OSM data available for the four regions. For the Greater London and the municipality of Coimbra, the spatial extent given by the bounding box was cut by the administrative boundary. For Dublin, the bounding box was cut with the coastline.

3.2 OSM Data Processing

The datasets were obtained from the Geofabrik website (http://www.geofabrik.de/), and consist of the free version of shapefiles of the levels roads, railways, waterways, landuse, natural and buildings. These files were pre-processed by Geofabrik from XML OSM file (.osm). Table 1 summarizes the main characteristics of the datasets used.

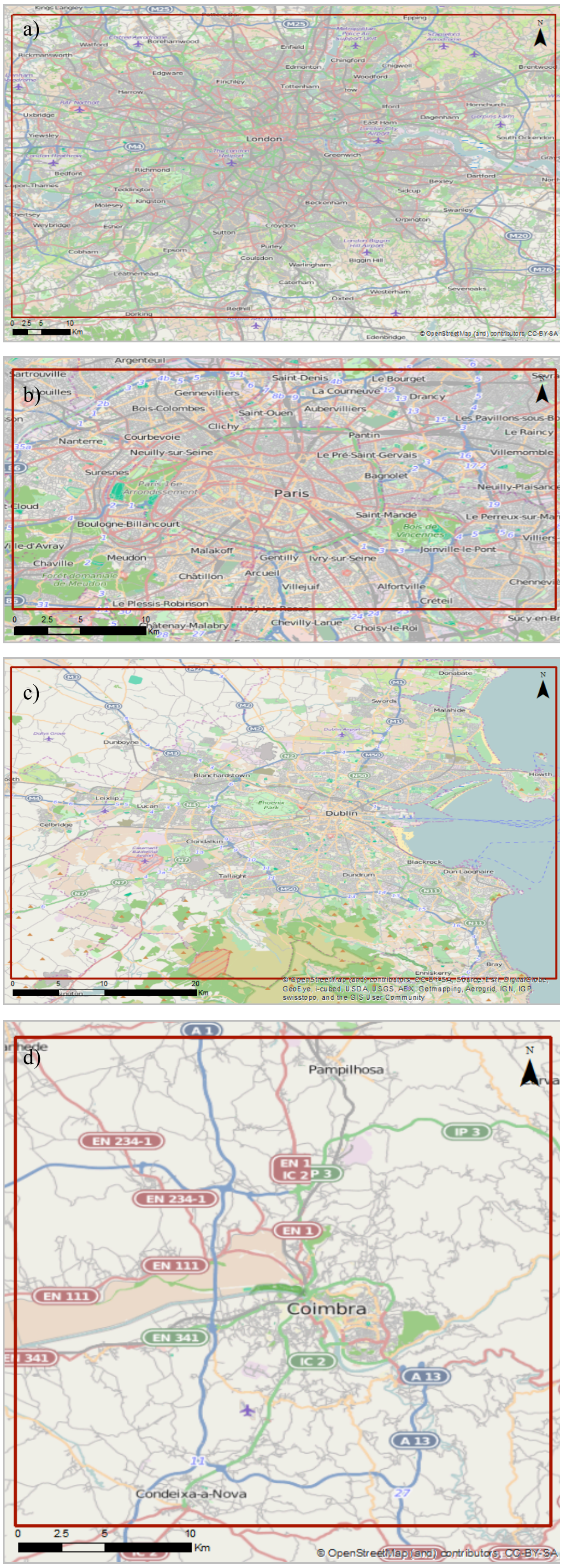


Figure 2. Study areas: a) London, United Kingdom; b) Paris, France; c) Dublin, Ireland; d) Coimbra, Portugal.

Study area	Date	Bounding box[1]	Area (km2)
Greater London (UK)	21/10/2014	(0.525°W/51.702°N, 0.331°E/51.227°N)	1867
Paris (France)	18/11/2014	(2.161°W/48.941°N, 2.533°E/8.777°N)	500
Dublin (Ireland)	28/10/2014	(6.584°W/53.493°N, 6.050°W/53.18°7N)	1057
Municipality of Coimbra (Portugal)	25/11/2014	(8.592°W/40.352°N, 8.313°W/40.099°N)	372

Table 1. Main characteristics of the datasets

The original OSM data are in the WGS84 geographic reference system. The grids created in the fourth step of the workflow were also created in this reference system. For each case study two grids of square cells were generated, with dimensions of 0.01 and 0.005 degrees, respectively, which were overlaid with the OSM data. The geodesic cell's area was computed (in km^2) and therefore there are some area variations for the regions located in different regions of the Earth and also among cells of the same region, when the region is relatively large, as for the London area. Table 2 shows the number of cells considered for all study areas for both grid sizes as well as their mean and total area.

Grids used per study area		Number of cells	Total grid area (km^2)	Cell's mean area (km^2)
London	Grid 1	2367	1829	0.773
	Grid 2	9275	1791	0.193
Paris	Grid 1	592	483	0.816
	Grid 2	2475	505	0.204
Dublin	Grid 1	1374	1019	0.741
	Grid 2	5521	1023	0.185
Coimbra	Grid 1	417	394	0.945
	Grid 2	1507	356	0.236

Table 2. Characteristics of the grids used for each study area.

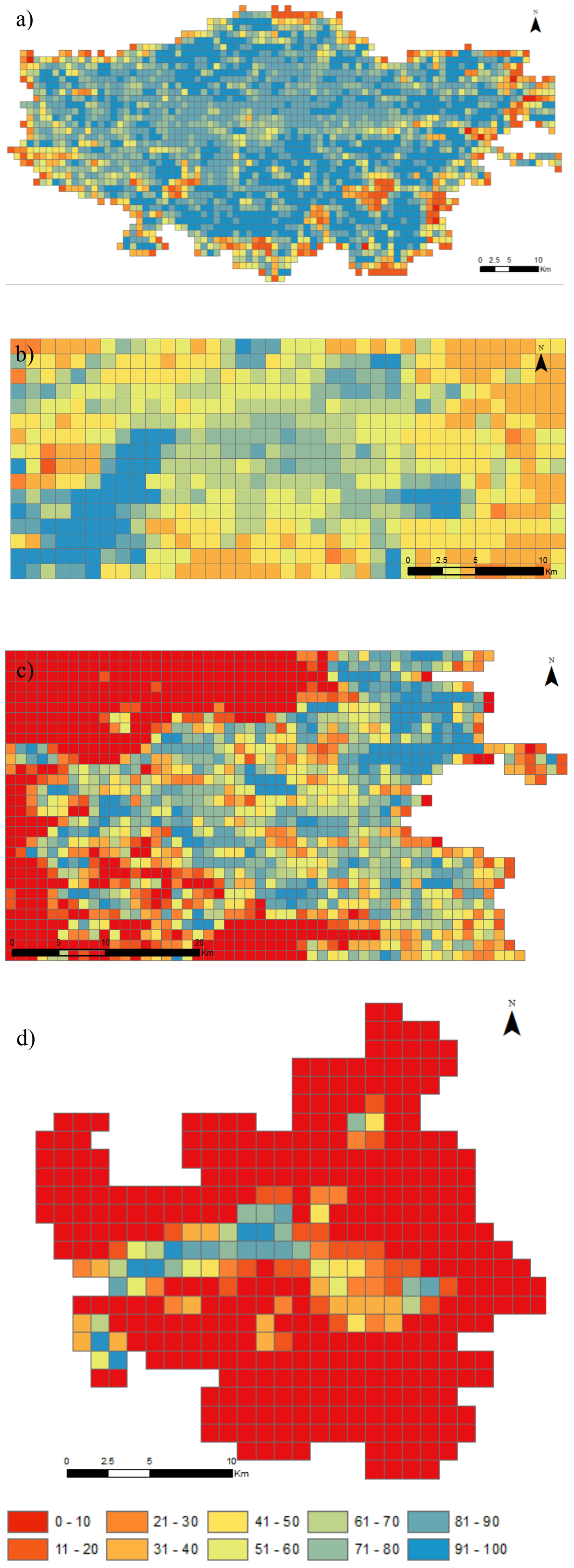
The buffers width, used is step 2 of the workflow to generate areas from the linear features, were estimated by visual inspection over an image layer. Equal values were used for features with the same characteristics in the four study areas.

3.3 Results and Discussion

The application of the presented methodology identifies the percentage of each grid cell that is covered with OSM data. Figure 3 and Figure 4 show the results obtained per cell for the four considered regions for grids 1 and 2, respectively.

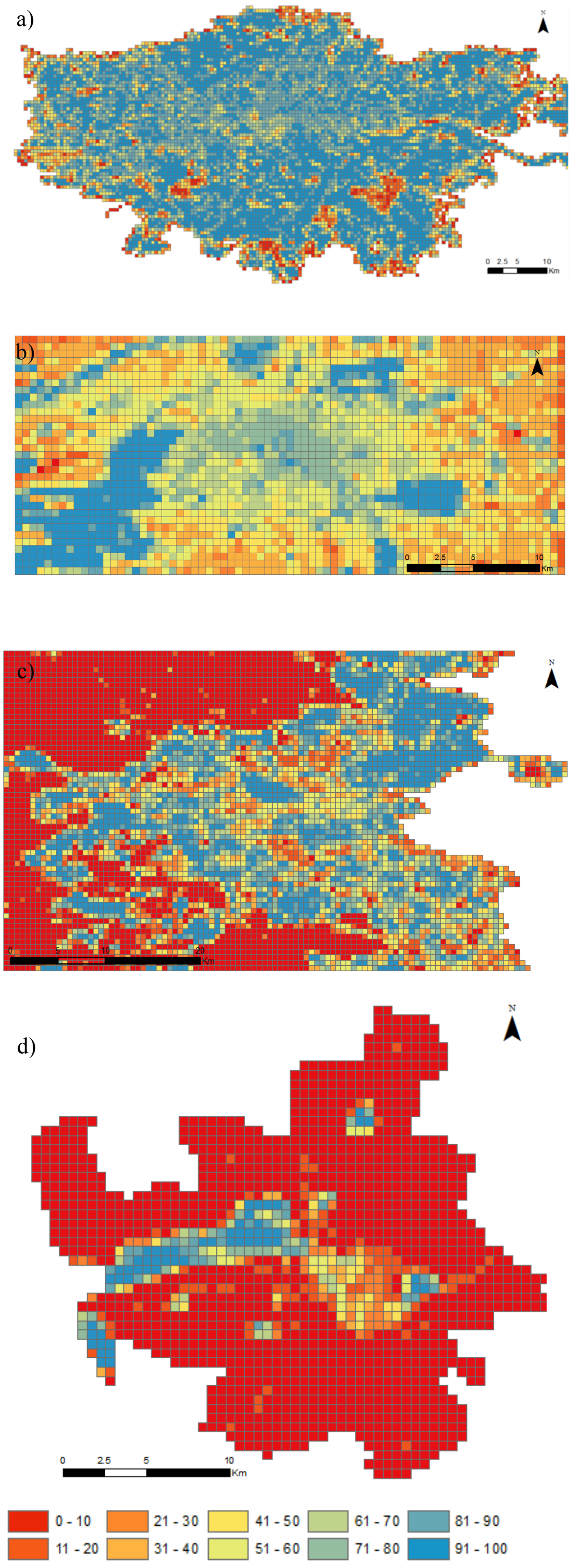
Figure 5 shows the number of cells with the percentage of coverage in each of the bins, with a range of 10%, for the study areas for both grids. Table 3 and Table 4 show some statistical indicators obtained for the regions under analysis, namely the mean coverage, standard deviation, maximum and minimum values, as well as the 25%, 50% and 75% quartiles for each study area, respectively for grids 1 and 2.

[1] (upper left latitude/longitude, lower right latitude/longitude)



Units: Percentage

Figure 3. Maps representing the percentage of OSM data coverage for grid 1 for: a) London; United Kingdom, b) Paris; France, c) Dublin, Ireland; d) Coimbra, Portugal.



Units: Percentage

Figure 4. Maps representing the percentage of OSM data coverage for grid 2 for: a) London; United Kingdom, b) Paris; France, c) Dublin, Ireland; d) Coimbra, Portugal.

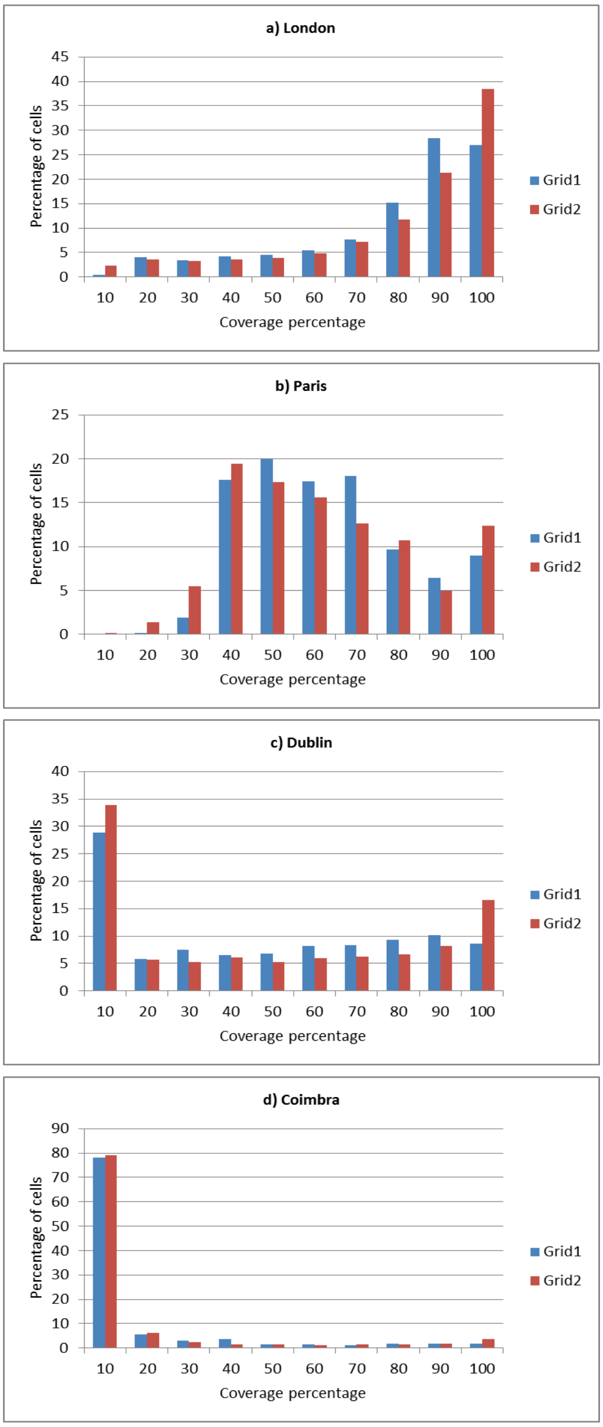


Figure 5. Histograms representing the number of cells by percentage of coverage for: a) London, United Kingdom, b) Paris, France, c) Dublin, Ireland, d) Coimbra, Portugal.

The results show that the region with larger coverage of OSM data is the region of Greater London, where the 25% quartile corresponds to a coverage of 65% and 66% for grids 1 and 2 respectively and the 75% quartile to a coverage of 91% an 95% (Tables 3 and 4). The 90% and 100% bins contain more than 50% of the cells for both grid sizes and only a few regions within the study area show very low levels of coverage (Figure 5a)). An analysis over a satellite image of what exists in those regions indicates that they correspond to low density residential areas, whose buildings are not represented in OSM and therefore most of the coverage comes from the streets that were already inserted in OSM. However, a closer analysis of the OSM data shows that there are "landuse" features of the type "residential" available that were not in the data downloaded using Geofabrik, since they are represented as a relation, and the free shapefiles generated by Geofabrick do not support this type of data (Geofabrick, 2015). The results obtained for the two grids considered present in general low differences, except larger differences for the 80%, 90% and 100% bins. For the 100% bin the cell's coverage increased from 27% to 38 % with the decrease of the grid size.

Statistical indicators	London	Paris	Dublin	Coimbra
Mean coverage (%)	74	58	42	12
Standard deviation (%)	23	19	34	22
Maximum (%)	100	100	100	100
Minimum (%)	3	17	0	0
25% Quartile (%)	65	42	4	1
50% Quartile (%)	83	56	42	3
75% Quartile (%)	91	70	73	8

Table 3. Statistical analysis of the OSM coverage in the study areas obtained for grid 1.

Statistical indicators	London	Paris	Dublin	Coimbra
Mean coverage (%)	76	57	43	13
Standard deviation (%)	25	22	37	25
Maximum (%)	100	100	100	100
Minimum (%)	0	7	0	0
25% Quartile (%)	66	39	2	1
50% Quartile (%)	85	54	38	3
75% Quartile (%)	95	73	80	8

Table 4. Statistical analysis of the OSM coverage in the study areas obtained for grid 2.

For the region of Paris, from Figure 5b) and Tables 3 and 4, it can be seen that lower levels of coverage exist, when compared to the results obtained for London. The percentage of cells with coverage between 90% and 100% is much lower (9% for grid 1 and 12% for grid 2 – Figure 5b)) and the 25% quartile corresponds to a coverage of 42% for grid 1 and 39% for grid 2, while the 50% and 75% quartile are respectively 56% and 70% for grid 1 and 54% and 73% for grid 2. It can also be seen that most cells have percentage of cover around 50%, which is very different from what was observed in the London area, where most cells have high levels of coverage. From Figures 2b), 3b) and 4b), it can be seen that the regions with better coverage in Paris correspond to vegetated areas. An analysis of the regions with less coverage also showed that some features available in OSM for the type "leisure" are missing from the data extracted from Geofabrick, such as golf-courses, even though they are available in the original OSM data. Since these features usually correspond to relatively large regions their omission generates some cells with low coverage.

For the Dublin region there is a great number of cells with no information, corresponding mainly to cells that are further away from the city of Dublin (Figure 2c), 3c) and 4c)). The 25% quartile corresponds only to 4% and 2% respectively for grids 1 and 2. Cells with larger percentage of coverage do not correspond to the city centre but to some regions where there is more detailed land-use and natural information. The percentage of cells in all bins, except the first one, is relatively evenly distributed; with a small increase for higher levels of coverage percentage and a considerable increase for the 90% to 100% bin for grid 2 (see Figure 5c)), which was also observed for the bin 0% to 10%. For this reason, the 75% quartile corresponds to a coverage value of 73% and 80% for grids 1 and 2 respectively, which is slightly larger than the one obtained for Paris.

For the Coimbra region, the results show (Tables 3 and 4 and Figure 5d)) that the degree of coverage is very low. The 25%, 50% and 75% quartiles only take values of respectively 1%, 3% and 8% for both grids. The cells with larger coverage values are natural regions, corresponding most of them to rice fields along the Mondego river. These results show clearly that the data in most of Coimbra municipality is not enough to use for LCM purposes.

The results obtained for both grid cell's size provide similar information, showing that the analysis may be started with larger cell sizes. If more refined information is needed the cell size may be decreased, providing more resolution on the spatial variability of the coverage.

4. CONCLUSIONS

The proposed approach enables to assess the degree of coverage of OSM data for a grid of cells with user defined dimensions. The approach is easily automated, requiring only the choice of the cell size and the prior identification of the typical width of the linear features that need to be converted to area features. In this study these values were chosen using a previous visual inspection for each region, but pre-defined values can be used for a full automation of the process. It should be stressed that this procedure raises problems of accuracy, since the real width of the elements represented by linear features may change considerably within the same area and from region to region. However, for this initial analysis of the percentage of coverage with OSM data, the errors generated by this lack of accuracy have in general low influence over the degree of coverage when the cells size is larger than the order of magnitude of the typical error committed with this approach, and therefore are not relevant.

The results obtained with the proposed operator do not provide information about the accuracy of the OSM data but are useful to identify the regions where the data available may provide enough land cover information, and especially those that are not appropriate for further analysis, since too much data is missing. Therefore, it may be used as a rough indicator of data completeness to assess if OSM may be fit for use for LCM purposes. However, the fitness for use will then need to be analysed with additional methods to assess, for example, completeness in its traditional meaning and positional accuracy. Thus, the level of coverage acceptable to proceed to the next step of analysis needs to be identified for each application.

The data used in the study areas presented in this analysis were freely downloaded using Geofabrick. A more detailed analysis of the regions showing less coverage than their surrounding showed that in some cases there was data missing from the downloaded shapefiles that is available in the original OSM data. This aspect should be taken in consideration when using OSM data for LCM, and the herein proposed approach showed to be useful to identify missing features from the data.

Some studies showed that the use of OSM data for LCM creation or validation may be possible (Jokar Arsanjani et al., 2013, Martinho and Fonte, 2015), but this type of application of OSM data still presents many chalenges and several developments are necessary. One of the main problems is the conversion of OSM features to the LCM classes, particularly because of the diversity of attributes and attribute's values the volunteers can use in OSM. An additional problem is the positional accuracy of the data and the existence of overlapping features, which need to be transformed into only one feature class to produce traditional Land Cover information. Therefore, the approach proposed in this article represents only a contribution to the several developments that are still needed to enable the use of OSM data for LCM purposes.

ACKNOWLEDGEMENTS

This work was partially supported by “Fundação para a Ciência e a Tecnologia” (FCT) under the project grant UID/MULTI/00308/2013.

REFERENCES

Bontemps, S., et al., 2011. GLOBCOVER 2009: products description and validation report. European Space Agency. http://ionia1.esrin.esa.int/docs/GLOBCOVER2009_Validation_Report_2.2.pdf

Estima, J. and Painho, M., 2013. Exploratory analysis of OpenStreetMap for land use classification. In: Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. New York, NY, USA: ACM, 39–46.

Estima, Jacinto, C.C. Fonte, and Marco Painho. 2014. “Comparative Study of Land Use/Cover Classification Using Flickr Photos, Satellite Imagery and Corine Land Cover Database.” In 17th AGILE International Conference on Geographic Information Science. Castellon, Spain.

Fonte, C.C., Bastin, L., See, L., Foody, G., Lupia, F., 2015. Usability of VGI for validation of land cover maps. International Journal of Geographical Information Science. Accepted. doi:10.1080/13658816.2015.1018266

Foody, G. M. 2002. "Status of land cover classification accuracy assessment." Remote Sensing of Environment 80:185-201.

Foody, G.M., and D.S. Boyd. 2013. “Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6 (3): 1305–12

Fritz, Steffen, Ian McCallum, Christian Schill, Christoph Perger, Linda See, Dmitry Schepaschenko, Marijn van der Velde, Florian Kraxner, and Michael Obersteiner. 2012. “Geo-Wiki: An Online Platform for Improving Global Land Cover.” Environmental Modelling & Software 31 (May): 110–23.

Geofabrick, 2015. Geofabrick Data Extracts – Technical Details. http://download.geofabrik.de/technical.html. Accessed 14th April 2015.

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

Martinho, N., Fonte, C.C., 2015. Assessing the applicability of OpenStreetMap data to validate Land Use/Land Cover Maps. INESC Coimbra research report.

Mooney, Peter, Padraig Corcoran, and Adam C. Winstanley. 2010. "Towards Quality Metrics for OpenStreetMap." In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 514–17. New York, NY, USA: ACM.

Neis, P., Zielstra, D., and Zipf, A., 2011. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet, 4 (4), 1–21.

Nie, W., et al., 2011. Assessing impacts of landuse and landcover changes on hydrology for the upper San Pedro watershed. Journal of Hydrology, 407 (1–4), 105–114.

OSM Wiki Map Features, 2014. https://wiki.openstreetmap.org/wiki/Map_Features accessed in 23rd December 2014.

See, Linda, Ian McCallum, Steffen Fritz, Christoph Perger, Florian Kraxner, Michael Obersteiner, Ujjal Deka Baruah, Nitrashee Mili, and Napendra Ram Kalita. 2013. "Mapping Cropland in Ethiopia Using Crowdsourcing." International Journal of Geosciences 4 (6A1): 6–13.

Verburg, P. H., Neumann, K. and Nol, L., 2011. "Challenges in Using Land Use and Land Cover Data for Global Change Studies." Global Change Biology 17 (2): 974–89.