Assessing OSM Road Positional Quality With Authoritative Data

Francisco ANTUNES[1], Cidália C. FONTE[1], Maria Antonia BROVELLI[2], Marco MINGHINI[2], Monia MOLINARI[2] and Peter MOONEY[3]

[1] Universidade de Coimbra (Portugal)
[2] Politecnico di Milano (Italy)
[3] Maynooth University (Ireland)

(fnibau@dei.uc.pt; cfonte@mat.uc.pt; maria.brovelli@polimi.it; marco.minghini@polimi.it; moniaelisa.molinari@polimi.it; Peter.Mooney@nuim.ie)

Keywords: Positional accuracy, OSM, Authoritative data, Road network, Transportation networks, Quality control.

Abstract: Online collaborative mapping projects, such as OpenStreetMap (OSM), have been developed not only to provide public and free information about many types of geospatial features, including communication and transportation networks such as roads, trails and railways, but also to give its users the chance to contribute with their local knowledge about the places. There must be, however, a special concern for the quality assurance within these community driven maps.

The aim of this paper is to assess the positional differences between the road network available in OSM for some regions of the Coimbra Municipality, Portugal, and the data provided by the Coimbra City Hall, considered as reference. The assessment is made by computing the distance between the features' lines, extracted from OSM and the reference data, using two approaches. One regards the application of a workflow, which uses tools already available in GIS software. The other approach applies directly the FOSS4G-based procedure developed by Brovelli *et al.* (2015).

1. Introduction

The Internet is currently the ultimate medium of communication and therefore it plays an important role in agglomerating a variety of information created and publicly shared by the users. Amongst these platforms, online collaborative mapping projects, such as OpenStreetMap (OSM), have been developed not only to provide public and free information about many types of geospatial features, including communication and transportation networks such as roads, trails and railways, but also to give its users the chance to contribute with their local knowledge about the places. This kind of projects resulted, and will continue, as a natural response to the prohibitively expensive or difficult access to high-quality and official geospatial data. There must be, however, a special concern for the quality assurance within these community driven maps. If, on the one hand, some of these maps may be more up-to-date and detailed than the authoritative ones, on the other hand, as they are mostly build by non-experts volunteers, they may also contain low quality information.

The study of geospatial data quality is not a recent concern. Both the emergence of Geographic Information Systems (GIS) and the increasing use of satellite-based data and equipment boosted this concern (Oort, 2005). Nowadays, even the most standard smartphone is equipped with Global Positioning System (GPS) sensor and includes a variety of geographical information tools. Moreover, the fact that these portable and rather cheap devices are network-enable, coupled with the strong use of online collaborative mapping platforms, considerably increases the availability, ease of exchange and use of spatial data. This data growth has, in fact, greatly enhanced the awareness of geospatial quality assurance and control, with a considerable number of publications and even conferences dedicated to this topic.

The current easy access, creation and manipulation of geospatial data have consequently culminated in the definition of the concept of Volunteer Geographical Information (or VGI) (Goodchild, 2007). As Coleman (2010) points out, the VGI provides opportunities, but also risks, of updating and enriching the authoritative geographic information systems maintained by both public and private sector providers. The geographic information voluntarism has the potential to relocate and redistribute the mapping services traditionally offered by official agencies to networks of non-state volunteer actors.

Currently, OSM is one of the most significant projects for voluntary mapping work, which is hourly updated by world-wide amateurs, enthusiasts and professionals. The vast, dynamic and multi-layer data of OSM, makes its general positional quality quite difficult to study with a high degree of confidence and in a short-time period. An alternative is to use geographically constrained data sets of different types of features, which are analyzed separately. In a very interesting work, Girres *et al.* (2010) studied the quality of the French OSM, using several elements to assess its quality, such as, geometric, attribute and temporal accuracy, completeness and logical consistency.

In another comparative study, Haklay (2010) assessed the quality of VGI using OSM and Ordnance Survey datasets in terms of transport networks (motorways, A-roads and B-roads) and areas, focusing his research in London and England. Within his particular case study, the author concluded that the OSM information can reach very interesting quality performances.

Cipeluch *et al.* (2010) used five Irish cities and towns, where they analyzed the spatial coverage, currency and ground-truth positional accuracy of OSM, Google Maps and Microsoft Bing Maps. Generally, the authors concluded that there was not a clear winner, as all mapping system had their strengths and weaknesses.

Another interesting way of approaching the problem of positional accuracy is to apply the traditional photogrammetric methods. Zuzelski *et al.* (2013), proposed a rigorous photogrammetric approach to determine the positional accuracy of OSM roads using aerial images and vector adjustment model. Eventually, the authors concluded that this vector adjustment model was able to capture 95% of the OSM actual positional displacement.

In this paper we focus our study in comparing the position of OSM road linear features within some regions of the Municipality of Coimbra – Portugal, to the road dataset kindly provided by the Coimbra City Hall, which was considered as reference data. The assessment is made by computing the distance between the features' lines extracted from OSM and the reference data, using two approaches. One regards the application of a workflow, which uses tools already available in GIS software. The other approach applies directly the automated FOSS4G-based procedure developed by Brovelli *et al.* (2015).

2. Methodology

The aim of the presented methodologies is to compare the position of OSM road linear features with the corresponding lines present in the reference data. For the sake of simplicity, we forwardly refer to the OSM data and reference data only by OSM and REF, respectively.

2.1 Workflow A: Based on available GIS tools

This workflow consists of a sequential use of several tools available in ArcGIS. It can, however, in principle, be applied using the majority of all other GIS options. The steps are the following:

I. Load both OSM and REF data;

II. Create a square grid extended to the limits of OSM/REF;

III. Convert the OSM and REF lines to points. The used tool should convert each line vertex to a point. Then compute the distances in two ways. Firstly, calculate the distances between the OSM points and the nearest line feature in REF, and, secondly, between the REF points and the nearest linear feature in OSM. The tool should compute the distance from a point the nearest line. A search buffer radius is set. Usually, when no feature is found within the specified buffer, the value -1 or NULL is returned, thus the points with no corresponding line feature is removed when computing the statistics;

IV. Independently intersect both layers obtained in III with the grid from II;

V. Calculate distance-based statistics grouped by each grid cell. In the end of this step, two tables should have been created;

VI. Join the tables obtained in V with the grid;

VII. Display the results by classifying each cell according to the statistics calculated in V.

Finally, in the end of the presented workflow, we should have a set of square grid cells associated with a set of statistics, which can be represented graphically. The obtained results from this procedure are presented in Section 3.2.

2.2 Workflow B: FOSS4G-based procedure

The methodology based on a novel approach developed by Brovelli *et al.* (2015) is built with Free and Open Source Software for Geospatial (FOSS4G), which makes it completely reusable and extensible. This fully automated procedure is expressly focused on spatial accuracy and completeness. It is currently implemented in Python as three Geographic Resources Analysis Support System (GRASS) GIS modules. Each of these modules accomplish different and independent sequential steps, which are intrinsically related with the proposed methodology:

I. Preliminary comparison between OSM and REF and computation of global statistics;
II. Geometric preprocessing of OSM in order to extract a subset directly comparable with REF;
III. Assessment of the OSM spatial accuracy using a grid-based approach (same as in Section 2.1).

In the first step, a preparation of both OSM and REF is conducted and several measures of spatial similarity are computed. Also included in this step is the length and percentage length computation of OSM and REF that are included within a user-specified buffer. Next, in step 2, a subset of the original OSM is extracted so that its line features have a direct correspondent in REF. The features that do not have any homologous REF are removed. This is accomplished not only by, again, applying a buffer around REF but also through a comparison between the angular coefficients of REF and OSM line features. The OSM features placed outside the buffer and the ones whose angular coefficient exceeds a user-specified threshold are then discarded from analysis. In the end of this step we should obtain two fully comparable datasets. Finally, the last step regards the evaluation of the OSM spatial accuracy through two possible outputs: by returning, per cell, a) the length and length percentage of OSM having a deviation smaller than a user-specified threshold value or b) the maximum deviation between OSM and REF datasets. Please refer to Brovelli *et al.* (2015) for more and deeper details on this procedure. The results from this methodology are presented in Section 3.2.

3. Case Study

3.1 Data

The REF data set comes from an ongoing survey, conducted by the City Hall of Coimbra, for documenting the main road axis of the city. Seven administrative regions (parishes) of the Municipality were analysed. The corresponding OSM data was downloaded from the Geofabrik servers. Only the road layer is used. Figure 1 presents both data sets and the considered regions and Figure 2 shows a detail of the data. The PT-TM06/ETRS89 projected coordinate system was used.

It is worthy to note that while the REF dataset is only constituted by road axis regardless of the road direction or number of traffic lane, the OSM roads are generally represented by one line per direction track. This means that wider roads like highways or highway-like roads will tendentiously have two axis, one for each direction. This characteristic can be clearly seen in Figure 2. On the other hand, for smaller roads the REF axis and OSM lines will generally coincide.

For the methodology described in Section 2.1, the Workflow A, several parameters were set for our experiments. The dimensions of the square grid cells are 250 by 250 meters and the search buffer used in the distance tool was of 10 meters. Finally, we use the minimum (MIN), maximum (MAX), average (MEAN) and standard deviation (STD) statistics to classify each cell in terms of the obtained distances. For the Workflow B, the same cell-grid and buffer were used. However, being a more complete and sophisticated approach it requires additional input parameters that must be specified such as the Douglas-Peucker and the angular coefficient thresholds for the comparison between each REF segment and all OSM segments inside the buffer.

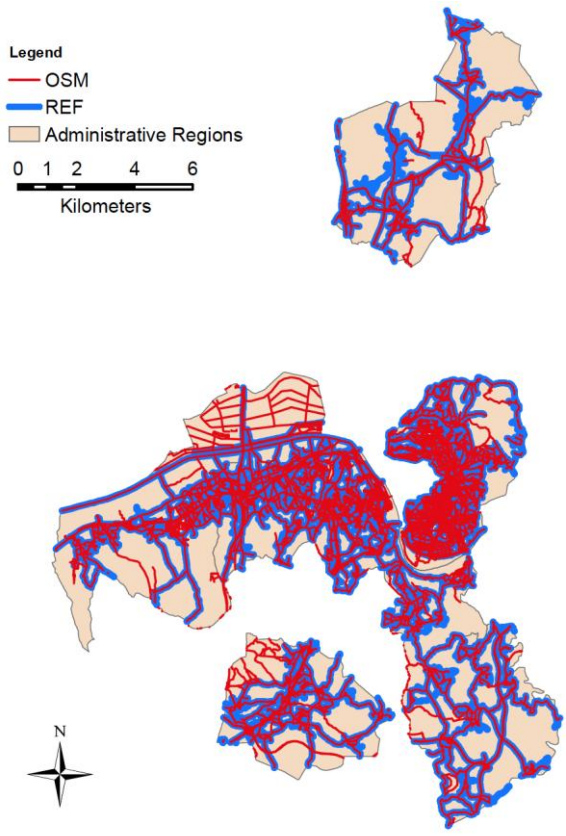

Figure 1 – OSM (red) and reference lines (blue) with the Administrative Regions (beige)

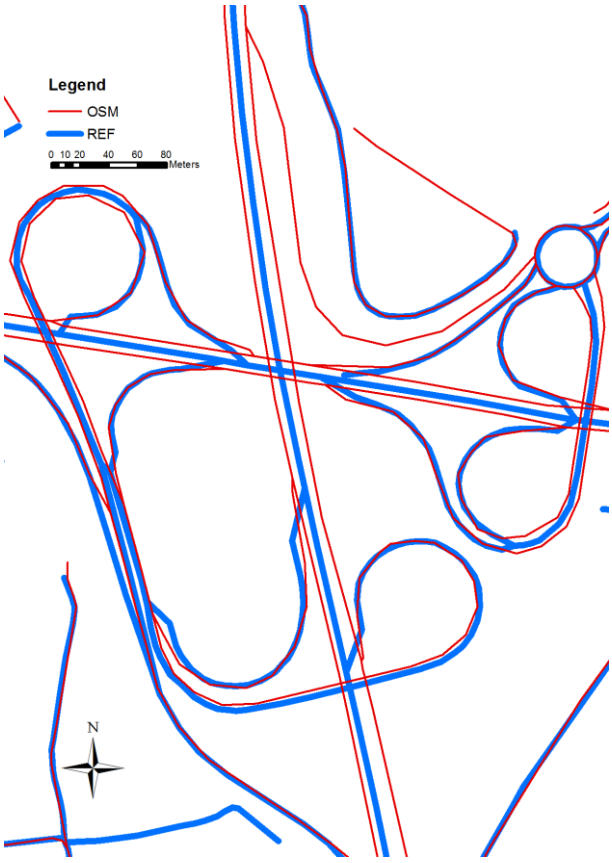

Figure 2 – Detail of the intersection between A1 and N341

3.2 Results

In this section we present the results obtained from the application of the mentioned methodologies to road data of Coimbra. Figures 3 to 6, and Table 1 relate to the results from Workflow A, while Figures 7 and 8 correspond to the results obtained from Workflow B.

From Figures 3 and 4 we can see that, on average, the results are generally good. The differences between both approaches within the Workflow A are quite small. However, the results presented in Figures 5 and 6, where the maximum distance (or error) per cell is displayed, show that for a large percentage of cells values the maximum error is close to the considered buffer width. Table 1, which shows the overall statistics grouped by parish, validates what is shown in Figures 5 and 6 (the OSM columns refer to the relation between the OSM points the REF lines, whereas the REF columns refer to relation between the REF points to the OSM lines). The results show that the mean values for both OSM and REF for all parishes are between 1 m and 2 m, with standard deviations of around 1.3 m and 1.6 m. For all parishes maximum values closer to the buffer value were obtained, which shows that for all parishes there were cases in which lines relatively far from the axis were considered in the workflow. The differences amongst the parishes are also not significant. However, we notice that, generally, the "REF points to OSM lines" approach achieved lower MEAN and STD values. This may be due to the following fact. In terms of line features numbers, the REF has 2757 and the OSM has 5153, while the number of generated points is, respectively, 50709 and 39731. Thus, when computing the distance between a certain REF point and an OSM line, the chances of finding one near feature are greater than in the "OSM points to REF lines" approach. Nevertheless, a deeper analysis is required on this matter, which is skipped on this text due to the space restrictions.

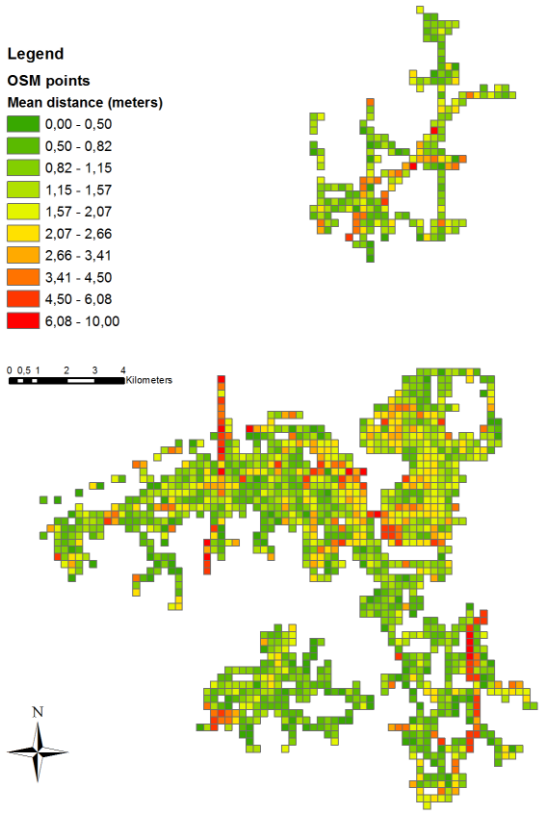

Figure 3 – Mean distances, per cell, between the OSM points and the nearest line feature within a 10 meter buffer (Workflow A)

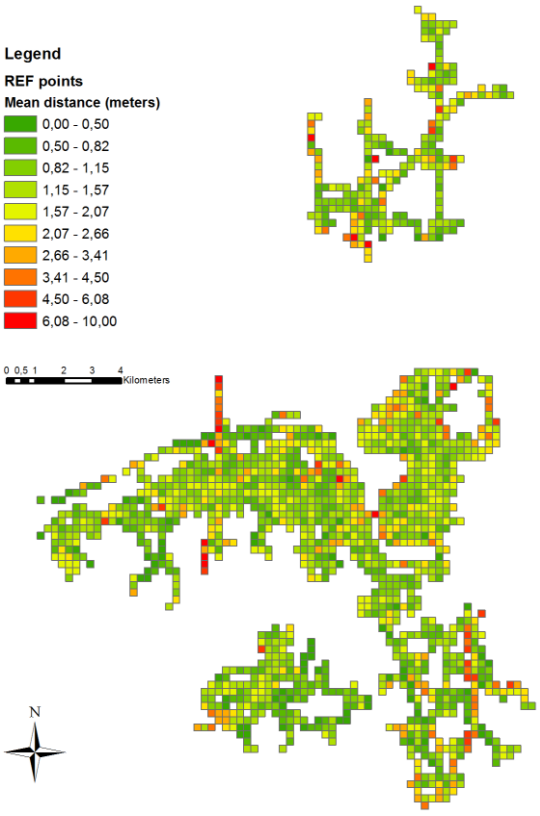

Figure 4 – Mean distances, per cell, between the REF points and the nearest line feature within a 10 meter buffer (Workflow A)

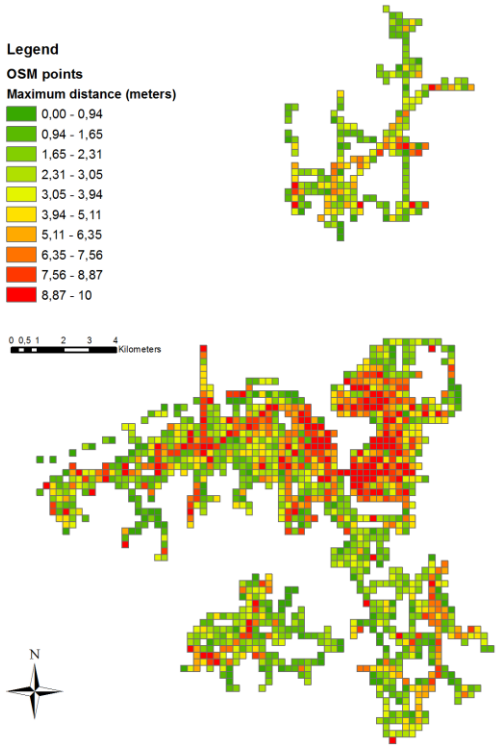

Figure 5 – Maximum distance, per cell, between the OSM points and the nearest REF line feature within a 10 meter buffer (Workflow A)

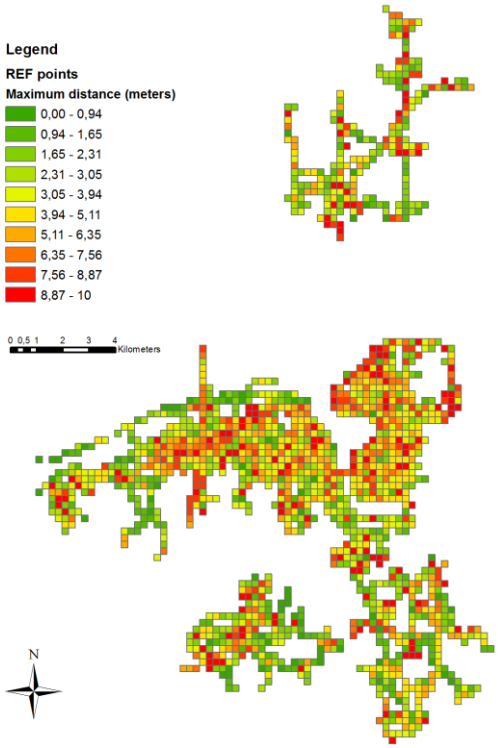

Figure 6 – Maximum distance, per cell, between the REF points and the nearest OSM line feature within a 10 meters buffer (Workflow A)

Table 1 – Resume of the distance-based statistics (meters) per administrative region (Workflow A)

Parish	MIN		MAX		MEAN		STD	
	OSM	REF	OSM	REF	OSM	REF	OSM	REF
Almalaguês	0.00	0.00	9.80	9.92	1.46	1.41	1.66	**1.57**
Cernache	0.00	0.00	9.81	9.96	1.19	1.18	1.38	1.33
Stº. António dos Olivais	0.00	0.00	**9.99**	9.95	**1.89**	**1.29**	**2.13**	1.37
Stª. Clara e Castelo Viegas	0.00	0.00	9.97	**9.98**	1.74	1.27	1.95	1.40
S. Martinho do Bispo e Ribeira de Frades	0.00	0.00	9.91	9.97	1.51	1.28	1.87	1.49
Souselas e Boto	0.00	0.00	9.91	9.97	1.40	1.26	1.49	1.40
Taveiro, Ameal e Arzila	0.00	0.00	9.96	9.86	1.49	1.26	1.76	1.35

Figure 7 shows the maximum error for the Workflow B, which is similar to the results displayed in Figures 5 and 6 for the Workflow A, but with fewer cells with larger distances. Figure 8 shows, the length percentage of OSM roads included in a buffer of 6 meters around the REF dataset. The results are very good despite several locations being in the yellow and red classes. A deeper analysis shows that they correspond to highway-like roads similar to the ones highlighted in Figure 2.

Corresponding maps representing the OSM lengths for buffers of 6, 8 and 10 meters were created (see Figures 8, 9 and 10), which show that as the buffer increases, the length percentage also increases. This was expectable, as we are getting closer to the threshold of 10 meters, from which the total OSM length, per cell, was computed.

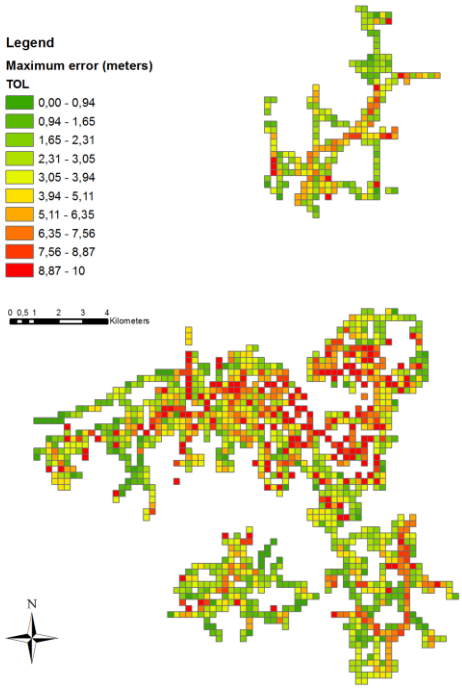

Figure 7 – Maximum error, per cell, of OSM within a 10 meter buffer (Workflow B)

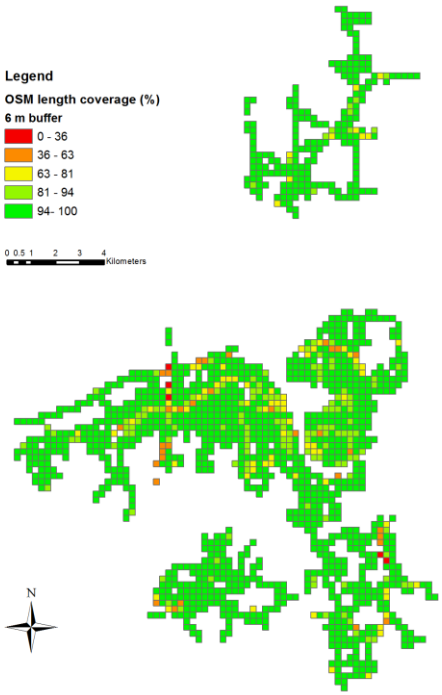

Figure 8 – OSM length coverage, per cell, within a 6 meter buffer (Workflow B)

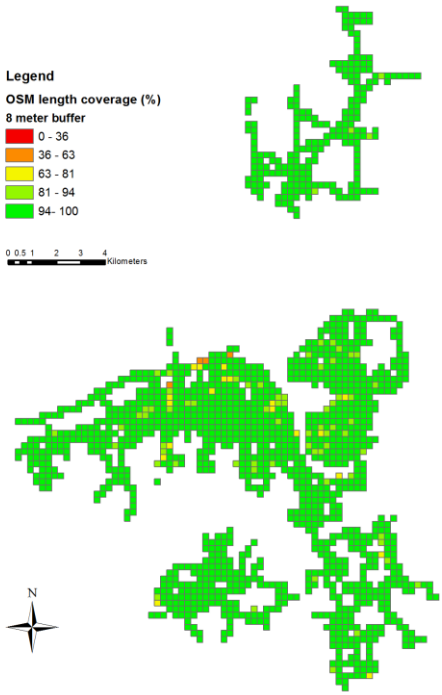

Figure 9 – OSM length coverage, per cell, within a 8 meter buffer (Workflow B)

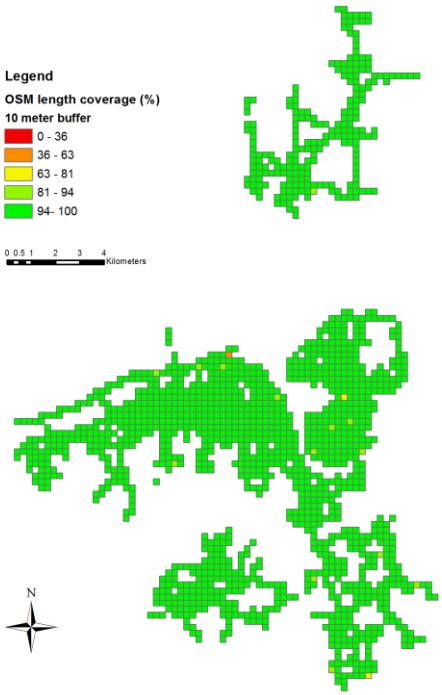

Figure 10 – OSM length coverage, per cell, within a 10 meter buffer (Workflow B)

4. Conclusions

In this paper we followed two different methodologies for OSM spatial accuracy assessment in some parishes of the administrative region of Coimbra City. One consists of a workflow applied through the sequential use of several tools already available in the majority of the GIS software. The other approach is a dedicated tool which was designed to assess the spatial accuracy of OSM linear features considering reference data.

Workflow A is based on distances between OSM and REF, within a pre-defined radius. Workflow B is based on OSM road lengths that are included in a pre-defined buffer. Thus, their outputs are not directly comparable, except for the maximum distance/error analysis. In Workflow A, for each point and each dataset, a straightforward search is conducted to find the nearest line on the other dataset. This search does not guarantee us, however, that the nearest feature (if it exists within the specified buffer) is, in fact, the natural corresponding feature. On the other hand, the Workflow B does exactly this during its second step: it extracts from the OSM dataset the corresponding, and therefore, comparable roads that build the REF dataset.

The strength of Workflow A is that it delivers a simple and very fast way to compare the OSM with an authoritative dataset. Its main drawback is the main strength of Workflow B: it does guarantee that corresponding features are actually being compared to each other. However, since Workflow B is much more complete, robust and customizable, it will require increased computational time to get the output results. On the other hand, the results obtained for Workflow A depend directly on the dataset segments, due to the conversion of these into points, and this is one of the reasons why the results obtained when computing the distances from OSM data to the REF are not equal to the ones obtained when computing the distance from REF data to the OSM.

Both procedures showed, however, to provide useful results. Generally, the Workflow A can be applied as a fast draft OSM spatial accuracy assessment. When a deeper and more careful analysis is required, it can be achieved with Workflow B.

Acknowledgements

The authors would like to thank the City Hall of Coimbra for providing the authoritative maps of the city used in this work and the support of COST Action TD1202 ‘Mapping and the Citizen Sensor’. http://www.citizensensor-cost.eu.

References

Brovelli, M.A.; Minghini, M.; Molinari, M; Mooney, P. (2015). A FOSS4G-based procedure to compare OpenStreetMap and authoritative road network datasets. Geomatics Workbooks 12, 235-238, ISSN 1591-092X.

Oort, P.V. (2005). Spatial data quality: from description to application. PhD Thesis, Nederlandse Commissie voor Geodesie (NCG), Netherlands Geodetic Commission, Delft, The Netherlands.

Coleman, D.J. (2010) Volunteered geographic information in spatial data infrastructure: an early look at opportunities and constraints, GSDI 12 World Conference.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4): 211-21. doi:10.1007/s10708-007-9111-y.

Girres, J.F.; Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. Transactions on GIS, 14.4.

Cipeluch, B.; Jacob, R.; Winstanley, A.; Mooney, P. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. Accuracy 2010 Symposium, July 20-23, Leicester, United Kingdom.

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

Zuzelski, R.C.; Agouris, P.; Doucette, P. (2013). A photogrammetric approach for assessing positional accuracy of OpenStreetMap roads. ISPRS International Journal of Geo-Information, 2, 276-301.