An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap

Thomas Koukoletsos[1], Mordechai (Muki) Haklay[2], Claire Ellul[3]

[1,2,3] University College London,
Gower Street, London, WC1E 6BT, UK
+44 20 7679 2745
[1] thomas.koukoletsos.09@ucl.ac.uk , [2] m.haklay@ucl.ac.uk , [3] c.ellul@ucl.ac.uk

1. Introduction

OpenStreetMap (OSM) is an open source mapping application that is based on volunteered effort to create a free and worldwide spatial database. The increasing density, importance and acceptance of OSM increase the importance of understanding data quality, so that potential users can evaluate fitness-for-purpose. When spatial quality analysis is performed through comparison with a reference dataset, a data matching procedure is necessary for the comparison to be meaningful. This matching is usually performed manually at data preparation stage. After this, methods need to be applied to measure quality elements of completeness, positional and attribute accuracy, which should be capable of dealing with OSM's heterogeneity in accuracy, density and attribute information.

So far, research in the UK for OSM (Haklay 2010, Basiouka 2009, Ather 2009), provided valuable information on OSM for selected areas. However, all these studies include manual procedures and methods that hinder repetition of the evaluation in a different and larger area or in the future when OSM data is updated. Furthermore, they measure positional accuracy using a simplified version of the Increasing Buffer Method (Goodchild and Hunter, 1997).

We slightly modify and integrate the Increasing Buffer Method in an automated method that performs data matching and evaluates data completeness and positional accuracy of OSM data, taking into consideration heterogeneity of Volunteered Geographic Information (VGI). We apply the proposed method to the area of greater Liverpool.

2. Method

2.1. Data selection

As reference dataset, the ITN dataset of Ordnance Survey's (OS) MasterMap was used, as the most accurate official dataset covering the whole country. The method is applied in the greater area of Liverpool (1780 km^2) (fig. 1).

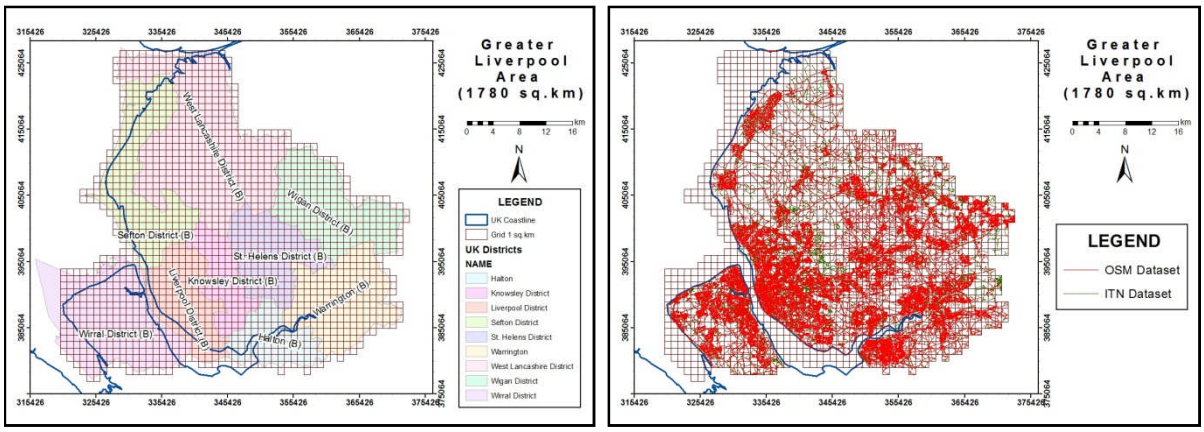

Figure 1. Area studied

2.2. Dealing with VGI heterogeneity

Data was split along the OS 1 km^2 National Grid and examined individually. In this way, possible variations in data density and accuracy will produce different results for each area, providing a more representative quality evaluation for VGI.

2.3. Data matching

As a first step, it is essential to remove any data that is not present in both datasets, so that any further evaluation will refer to corresponding data. The proposed data matching method combines geometric and attribute restrictions in a multi-stage approach (table 1).

Stage	Basic Unit	Constraints (in order of importance)
1	ITN Segment	Geometric (Distance,Orientation,Length)
2	ITN Segment	Attribute and geometric (name,type,Distance,Orientation)
3	ITN Segment	Attribute and geometric (name,type,Distance,Orientation)
4	ITN Segment	Geometric (Distance,Orientation)
5	OSM & ITN Feature	Geometric (Length)
6	OSM Feature	Attribute and geometric (name,type,Distance)
7	OSM Feature	Geometric and attribute (Distance,Length,type)
8	OSM & ITN Feature	Geometric (Length)

Table 1. The proposed multi-stage approach

We start by splitting features into segments. Stage 1 deals with corresponding segments based on distance, orientation and length when there is only one possible candidate. Stages 2 and 3 look for an exact and similar name matching accordingly. Stage 4 deals with segments with no name attribute. Stage 5 recomposes features and classifies them as matched or not, based on the information gathered so far. Stages 6, 7 address non matched features to solve cases not covered in previous stages. Stage 8 moves away from the tile-by-tile examination and deals with datasets as a whole, to cover matching errors in cases of corresponding features that because of their proximity to the tile border, they lie in different tiles.

A manual evaluation of data matching is performed in a randomly selected area of 80 km^2 (fig. 2). The lengths of the misjudged features are calculated and compared with the dataset's length for each tile and dataset. Results prove the efficiency of the data matching method (table 2).

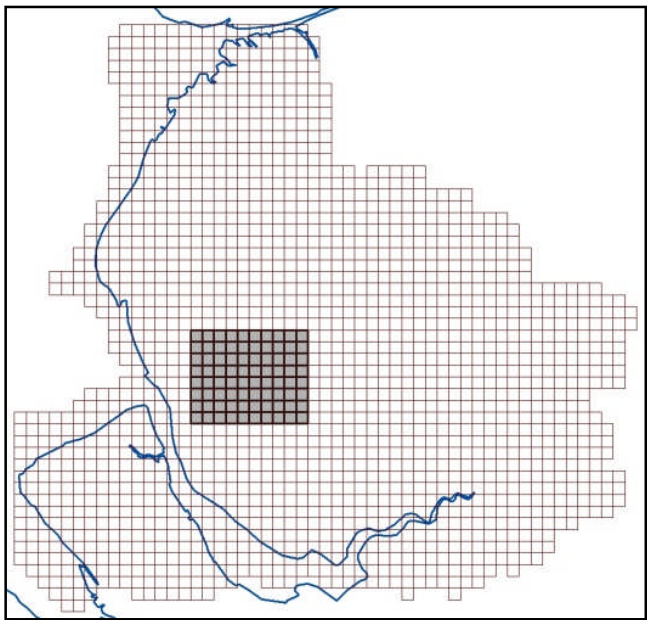

Figure 2: Data matching evaluation area

Dataset	Total length(km)	Length evaluated (km)	Missing data length (km)	Surplus data length (km)	Total matching error (km)
OSM	9042.138	694.469(7.68%)	1.575(0.23%)	2.298(0.33%)	**3.873(0.56%)**
ITN	10863.845	898.855(8.27%)	0.105(0.01%)	30.911(3.44%)	**31.016(3.45%)**

Table 2. Evaluation results: Total matching errors

2.4. Data completeness

The length of matched features is calculated and compared with the total dataset length for each tile and for each dataset, producing a data matching percentage for OSM and ITN. Table 3 provides a rough classification of the possible matching scenarios. Classification however depends on the percentages' distribution and the crisp boundaries of table 3 cannot always be appropriate for visualisation. Fuzziness due to spatial correlation may demand more classes with variable size to represent the matching percentage distribution; in the studied area for example, 90 % of the examined tiles achieved percentages above 50 % for both datasets.

Case	OSM matching percentage	ITN matching percentage	Mixed percentage	Meaning
1	High	High	High	Datasets agree with each other
2	High	Low	Low	ITN is denser
3	Low	High	Low	OSM is denser
4	Low	Low	Very Low	Datasets contain different data

Table 3. General cases of matching score for each tile

Since OSM dataset contains footpaths, steps, bridleways etc, the data matching results show the agreement rather than the completeness between the two datasets. For the results to be more representative of completeness, certain OSM road types are removed before the matching process (e.g. steps, bridleways, footpaths, tracks).

2.5. Positional accuracy

After removing data not present in both datasets, we address positional accuracy. According to Goodchild and Hunter (1997), if an increasing buffer is applied on a reference line, it will accordingly cover increasing percentages of the tested line (fig. 3). The buffer could then be considered as the accuracy of the reference dataset for the specific overlap percentage. We can either provide a buffer value to calculate the percentage, or provide a desired percentage to calculate the buffer (accuracy) using and iterative method. For the second option, which is not applied in any study so far, we use the binary search algorithm rather than the suggested formula by the authors.

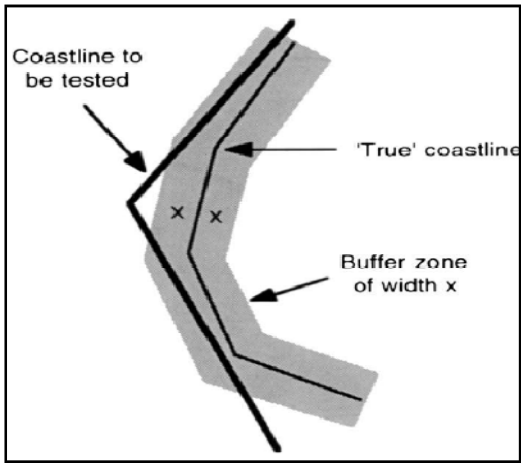

Figure 3: Increasing buffer method (from Goodchild and Hunter ,1997, p.301)

The user defines a desired overlap percentage. A first buffer of 8 m is applied on the ITN dataset and the OSM percentage falling into the buffer is calculated. If and as long as it is less than the user-defined desired overlap percentage, the buffer is doubled and calculations are repeated. When it percentage exceeds the desired one, the next buffer to be applied is half the distance between the two buffers previously used that achieved a lower and bigger percentage than the desired one correspondingly (table 4). The iteration process finishes when percentage is within 0.1 of the desired one, or when successive buffers differ less than 0.1 m.

Tile	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5	Iter.6	Iter.7
SD3612	8m-90.9%	16m-95.7%	12m-93.1%	14m-94.1%	15m-94.8%	15.5m-95.3%	15.25m-95.1%

Table 4. Example of the binary search algorithm, target percentage: 95%

To decide on a suitable 'desired percentage', tests were carried out in an area of 25 km^2 in central London (where OSM is proved to be accurate by previous research). The method was applied for various percentages and the corresponding buffer values were examined. A value of 95% was chosen to be used. Above this, differences in features' length between datasets

(due to varying data capture) as well as possible matching errors lead to unusually high buffer values.

3. Results

Fig. 4, 5 show data matching percentages for each dataset, as well as their combination. Generally ITN proves to be much more complete, as most of its data is not found in the OSM dataset (table 5).

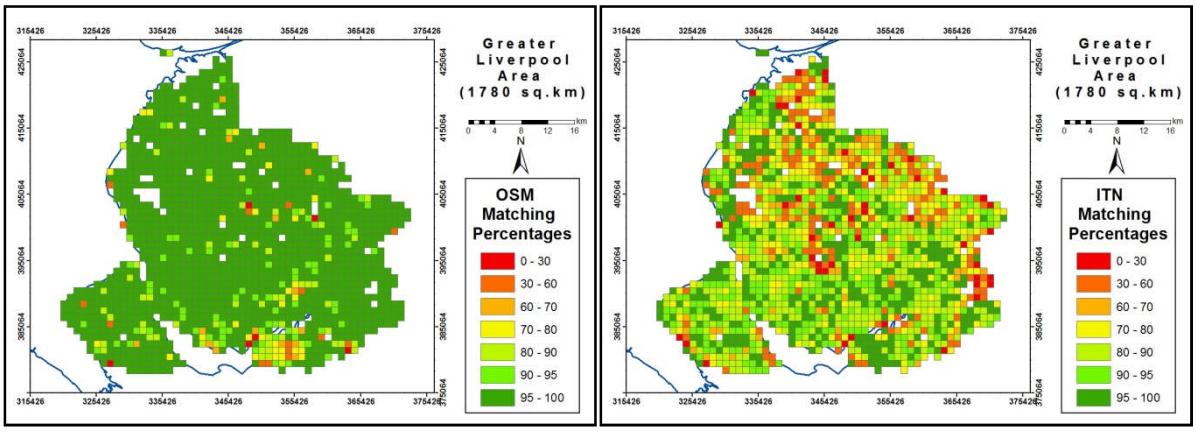

Figure 4: Data matching percentages for each dataset

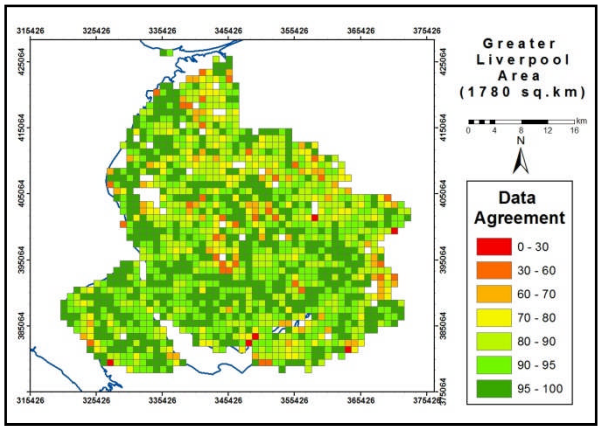

Figure 5: Data agreement between ITN and OSM

	OSM	ITN
Total length compared (km)	9175.903	10863.845
Total length matched (%)	96.62%	84.91%
Average pct matched (per tile)	96.77%	80.77%

Table 5. Data Completeness results

Fig. 6 shows the positional accuracy for 95% of OSM dataset per tile in the studied area (average accuracy 6.94 m, standard deviation 3.46 m). However, 19 tiles with buffer sizes up to 487 m had to be removed, as outliers. Due to different data capture methods, these tiles contain corresponding objects with the OSM feature extending much further than the ITN one, resulting in an increased buffer in order to reach the desired overlapping percentage, as shown in fig. 7.

The proposed method could also be used to compare other road network VGI sources and official datasets, provided that data structures include road name and road type attributes.

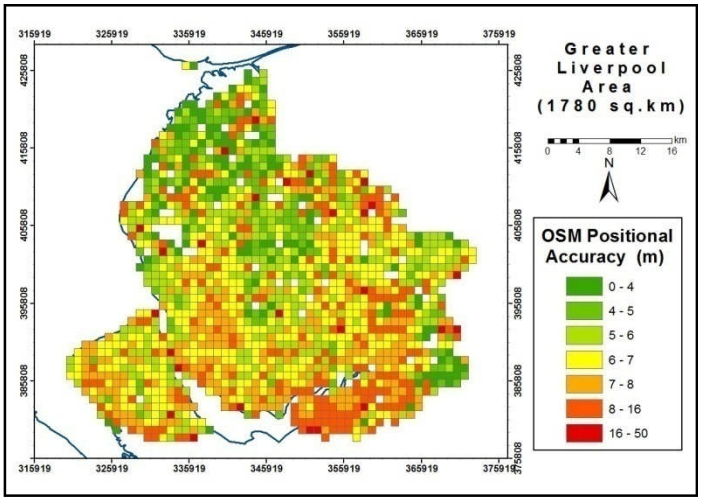

Figure 6: Positional accuracy of OSM

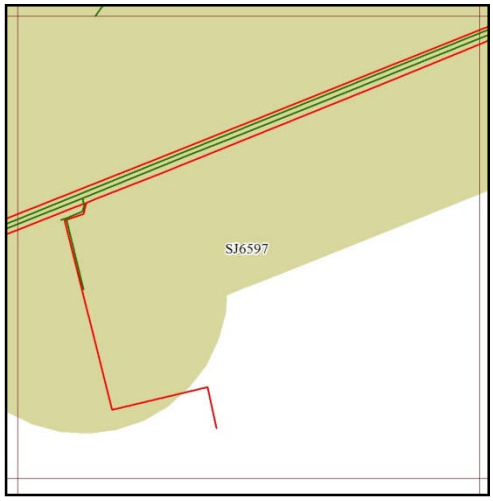

Figure 7: Buffering problems

4. Future Work

More areas need to be examined and a deeper statistical analysis of the results is necessary. Positional accuracy and data completeness results also need to be combined in search for a possible correlation. Finally, evaluation of other data quality elements needs to be integrated in the automated procedure as well.

5. Acknowledgments

We thank the Ordnance Survey and OSM for the data used in this work. All figures and tables using OS data are ©Crown Copyright/database right 2011, an Ordnance Survey/EDINA supplied service.

6. References

Ather, A., 2009. *A Quality Analysis Of Openstreetmap Data*. MSc, University College London

Basiouka, S., 2009. *Evaluation of the Openstreetmap Quality.* MSc, University College London

Goodchild, M.F. and Hunter, G.J., 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3):299-306

Haklay M (2010). How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, In *Environment and Planning*, 37(4):682-703