Assessing the quality of OpenStreetMap building data and searching for a proxy variable to estimate OSM building data completeness

Claire Fram[1], Katerina Chistopoulou[2] and Claire Ellul[3]

[3] Dept. of Civil, Environmental and Geomatic Engineering, University College London

Gower Street, London, WC1E 6BT

Tel. +44 (0) 20 7679 4118 Fax +44 (0) 20 7380 0453

9 January 2015

1. Introduction

OpenStreetMap (OSM) is an open data, geospatial information (GI) project that relies on contributions from volunteers to create a digital, on-line, map of the world. The use of OSM is also free and available under the Open Database License (Hecht, et al., 2013). As the amount of information available in the OSM database continues to grow, interest has grown regarding the application of OSM in industry and according to scientific standards (Haklay, 2010; Hecht, et al., 2013; Koukoletsos, et al., 2012). However the quality of OSM building data is largely unknown and thus the practical applications for OSM building data remain limited.

This study was taken on with the support and supervision of Risk Management Solutions (RMS) to investigate the quality of OSM building data with the purpose of assessing OSM build data's potential application in RMS products, specifically natural catastrophe exposure models.

To incorporate OSM building data into exposure modelling, the quality of OSM data must be known or the uncertainty of OSM data completeness must be quantified. Unless OSM data quality can be quantified, applying OSM data to any commercial products would introduce an unacceptable, unknown quantity of uncertainty.

Two objectives were defined in this study. The first objective is to understand OSM building data quality. Multiple case studies were used to reveal the similarities and differences between OSM data quality in different places. The second objective of this study is to test whether OSM building data quality might be estimated with a proxy variable. Identifying an appropriate proxy variable might offer RMS, and use-cases like RMS's, to quantify the quality of OSM building data in areas without official reference data.

[1] claire.fram.13@ucl.ac.uk
[2] Katerina.Christopoulou@rms.com
[3] c.ellul@ucl.ac.uk

2. OSM building data assessment

OSM building data quality was tested in three study areas Leeds, London and Sheffield using a *unit-based* assessment defined by Hecht et al. (2014). To highlight the heterogeneity of OSM completeness and accuracy, the case study areas were divided into smaller sub-sections. Within each sub-section, OSM data was compared against data from OS Street View. Using this approach, OSM data completeness is quantified at the spatial resolution of the sub-unit. “Spatial units” (Hecht, et al., 2013, p. 1073) for this study were defined as1km^2 cells. This grid size was used in Koukoletos et al.’s study (2012).

Hecht et al.’s unit-based method (2013) for assessing OSM polygon completeness was applied. In this method the aggregate area of building footprints’ per-spatial unit was used to quantify data completeness. The 1km^2 grid was applied to each study area; buildings that were partially located in multiple spatial units were subdivided by the spatial unit boundaries (illustrated in Figure 1).

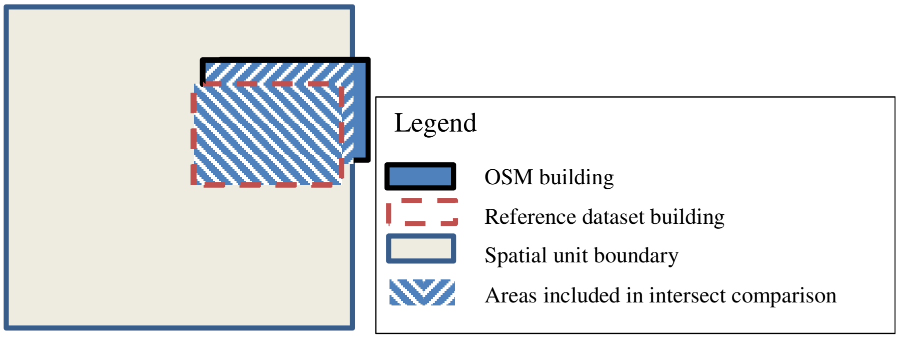

Legend
OSM building
Reference dataset building
Spatial unit boundary
Areas included in intersect comparison

Figure 1: Example of completeness calculation based on footprint intersections with spatial unit boundaries

The OSM building footprint area within each cell was aggregated and compared with the aggregate OS Street View building footprint located in the same cell (Equation 1). The product of this equation serves to quantify OSM building completeness and can be compared across sub-units.

$$C_{Area} = \frac{\sum Building\ footprint_{OSM}}{\sum Building\ footprint_{Ref}} \times 100$$

Equation 1: Unit based method: area (Hecht, et al., 2013, p. 1076)

Of the three case study areas, Leeds had the lowest level of OSM building completeness, as a measure of aggregate footprint coverage per square kilometre. When the completeness percentages of all sub-units was averaged within each study area, Leeds had an average completeness ratio of 30%, while Sheffield has the highest at 75% (Table 2). London also had a low average completeness ratio (33%, Table 2). However, London is a larger study area. 1456 km^2 were included in London’s assessment of OSM data quality, compared to 469 km^2 in Leeds and only 292 km^2 in Sheffield. The OSM completeness ratio for each city is represented at a 1km^2 resolution in Figure 3, Figure 4 and Figure 5. Results describing the distribution of OSM completeness estimates by 1km^2 are seen in Figure 2.

In calculating the average OSM completeness ratios, each spatial unit (km^2) was given equal weight. However, building density was not always equally distributed spatially. The number of buildings represented in each spatial unit (km^2) is presented in Table 3.

If areas with high OSM quality are considered, only 13% of Sheffield's buildings are located high quality OSM areas. In Leeds, 2% of buildings are in high quality OSM areas, and only 3% of London buildings are in areas of high quality OSM. However, London has the largest actual count of buildings (29352) located in these high quality areas (Figure 2 and Table 3).

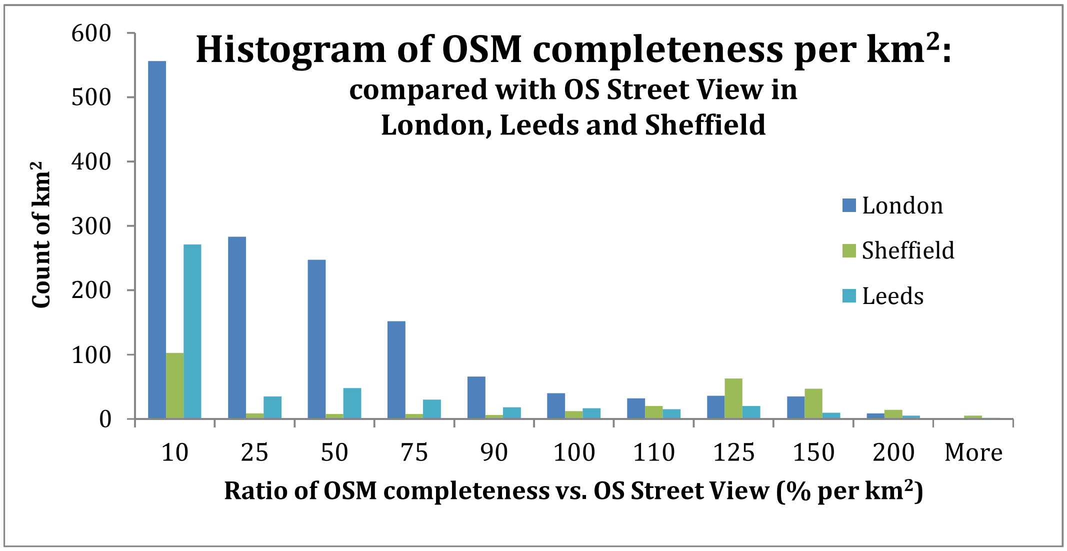


Figure 2: Histogram of OSM completeness ratio results for Leeds, London and Sheffield

Table 1: Results uses to calculate histogram of OSM area-completeness by km^2

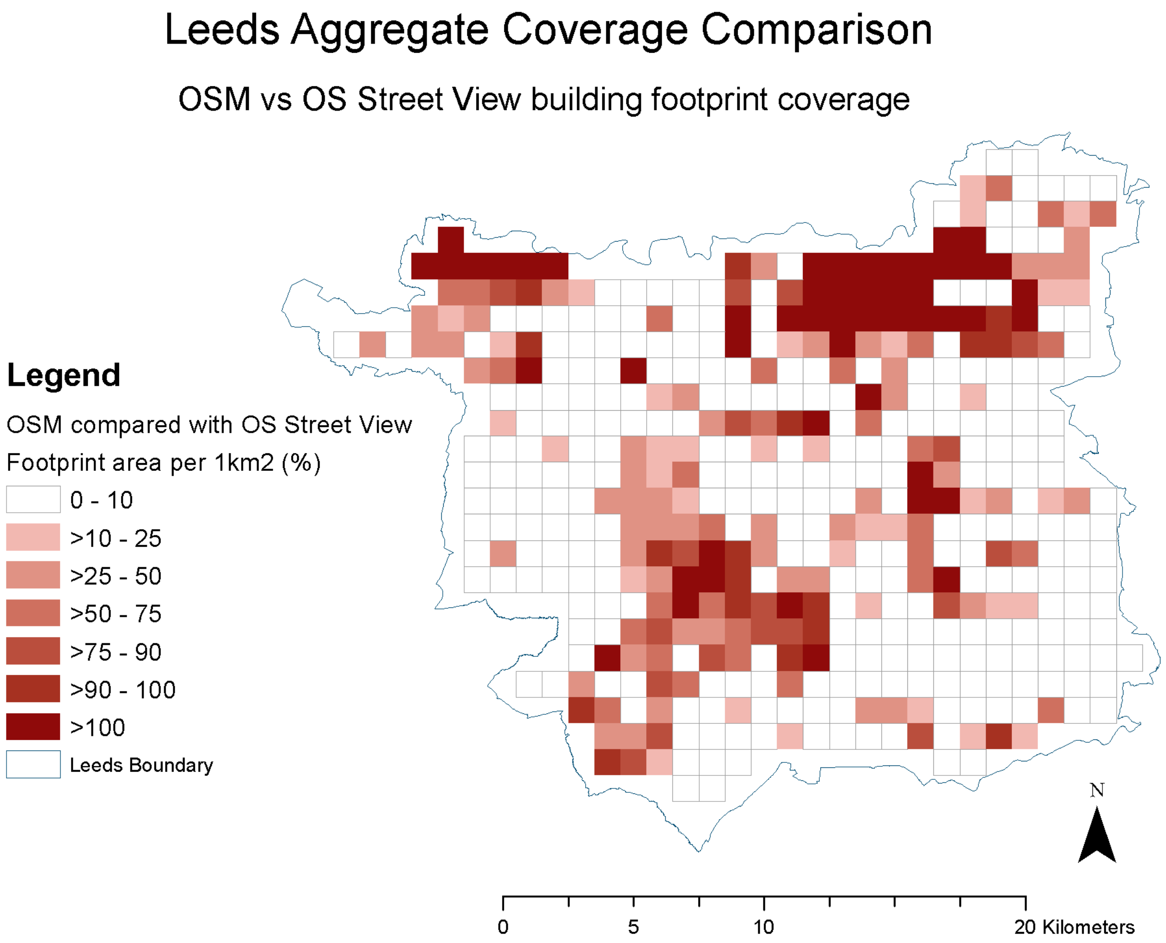
	London		***Sheffield***		***Leeds***	
Bin	*Count of Km^2*	*Percentage of total Km^2*	*Count of Km^2*	*Percentage of total Km^2*	*Count of Km^2*	*Percentage of total Km^2*
10%	556	38%	103	35%	271	7%
25%	283	19%	9	3%	35	10%
50%	247	17%	8	3%	48	6%
75%	152	10%	8	3%	30	4%
90%	66	5%	6	2%	18	4%
100%	40	3%	12	4%	17	3%
110%	32	2%	20	7%	15	4%
125%	36	2%	63	21%	20	2%
150%	35	2%	47	16%	10	1%
200%	9	1%	14	5%	5	0%
More	1	0%	5	2%	2	100%

Table 2: Average OSM completeness ratios (compared with OS Street View)

	London	Sheffield	Leeds
Average OSM completeness ratio	33%	75%	30%

Table 3: Assessment of OSM completeness ratios in the context of building density

	London		Sheffield		Leeds	
	Count	%	Count	%	Count	%
Total OS Street View Building Count in study area	864916	100%	83310	100%	137632	100%
Total OS Street View buildings in grids with completeness ratios <=10%	364866	42%	5535	7%	85018	62%
Total OS Street View buildings in grids with completeness ratios >10%	500050	58%	77775	93%	52614	38%
Total OS Street View buildings in grids with completeness ratios >=90% and <=110%	29352	3%	10307	12%	3392	2%

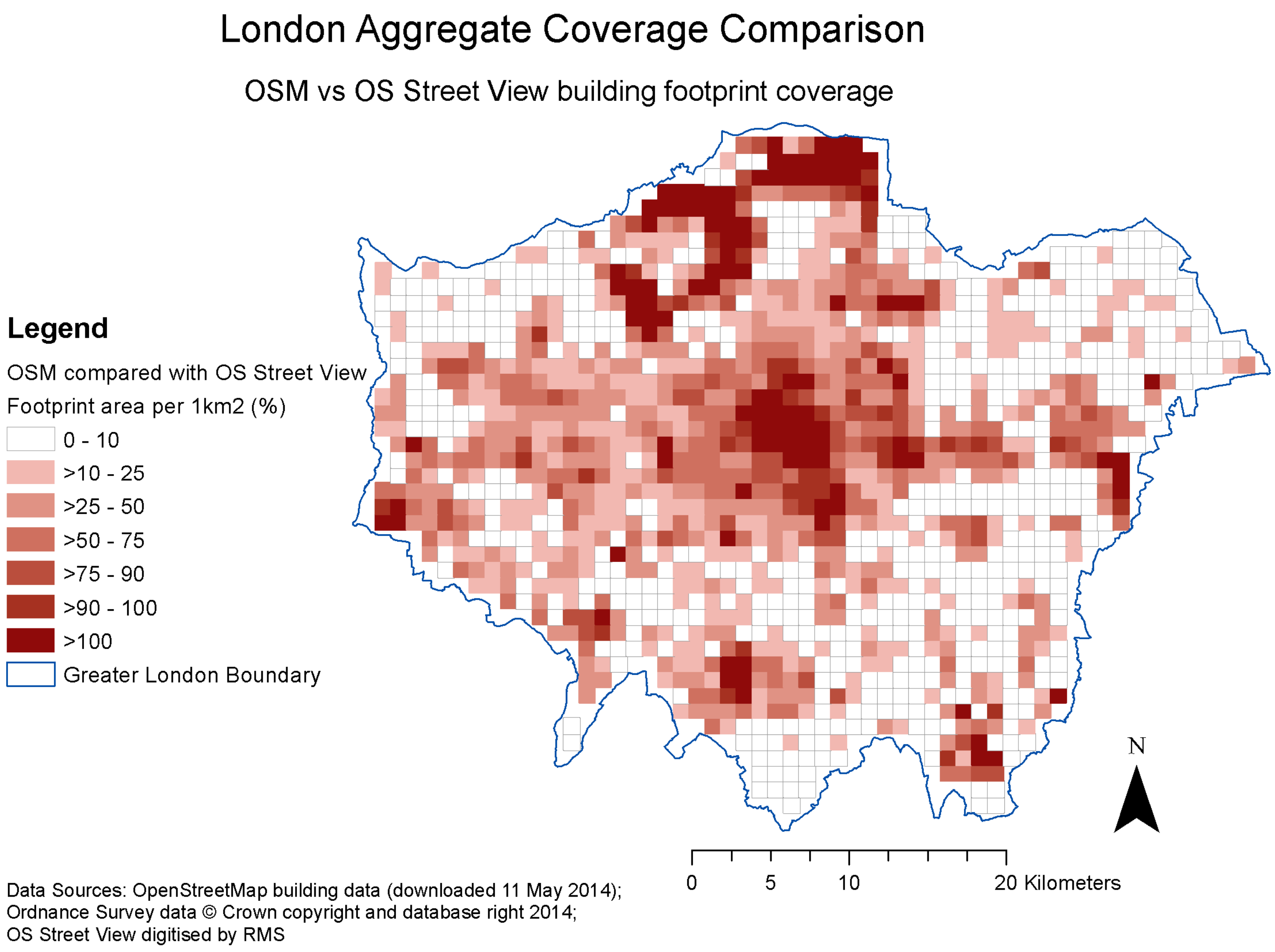


Leeds Aggregate Coverage Comparison

OSM vs OS Street View building footprint coverage

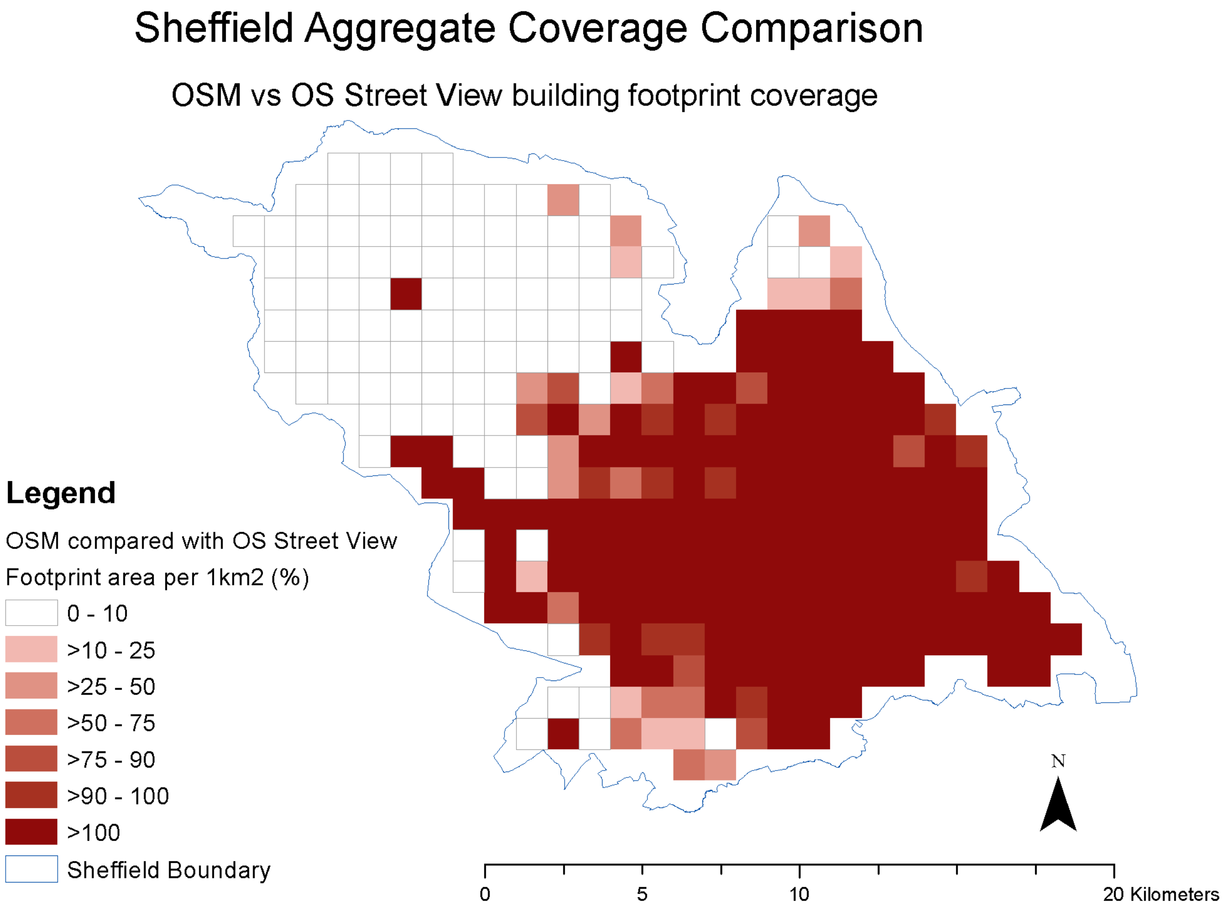
Data Sources: OpenStreetMap building data (downloaded 11 May 2014); Ordnance Survey data © Crown copyright and database right 2014

Figure 3: Leeds, map of aggregate footprint coverage comparison between OSM and OS Street View


London Aggregate Coverage Comparison
OSM vs OS Street View building footprint coverage
Legend
OSM compared with OS Street View
Footprint area per 1km2 (%)
0 - 10
>10 - 25
>25 - 50
>50 - 75
>75 - 90
>90 - 100
>100
Greater London Boundary
N
0
5
10
20 Kilometers

Data Sources: OpenStreetMap building data (downloaded 11 May 2014);
Ordnance Survey data © Crown copyright and database right 2014;
OS Street View digitised by RMS

Figure 4: London, map of aggregate footprint coverage comparison between OSM and OS Street View


Sheffield Aggregate Coverage Comparison
OSM vs OS Street View building footprint coverage
Legend
OSM compared with OS Street View
Footprint area per 1km2 (%)
0 - 10
>10 - 25
>25 - 50
>50 - 75
>75 - 90
>90 - 100
>100
Sheffield Boundary
N
0
5
10
20 Kilometers

Data Sources: OpenStreetMap building data (downloaded 11 May 2014); Ordnance Survey data © Crown copyright and database right 2014
OS Street View digitised by RMS

Figure 5: Sheffield, map of aggregate footprint coverage comparison between OSM and OS Street Vie

3. Testing a proxy for OSM building data completeness

In the applied, unit-based method a reference dataset was required to benchmark OSM building data quality. However, OSM building data has the most potential to reduce uncertainty in RMS exposure models in areas where reliable reference datasets are not available. While OSM building data quality is likely heterogeneous in all cities (see results in section 2), this section investigates if there is a relationship between OSM building data quality and an independent, proxy variable.

This section explores the relationship between LandScan gridded population data and OSM building data completeness. Population data was chosen as a potential proxy for OSM data completeness for two reasons. First, across international case studies (Germany, The United Kingdom and America), a strong correlation between population density and OSM road network density has been proven (Haklay, 2010; Hecht, et al., 2013). Second, population information is available globally at a 1km^2 resolution via LandScan gridded population data (Oak Ridge National Labratory, 2014). The global coverage offered by this dataset allows for any relationship discovered between OSM completeness and population density to be tested and applied internationally.

To compare OSM data completeness with population density, a centroid was extracted from each 1km^2 grid of the unit-based assessment, and the completeness result metric was spatially joined with LandScan population density information. A simple correlation assessment was used to assess the feasibility of using population as a proxy variable or OSM area-completeness.

There was no significant correlation discovered between OSM area completeness and population density in any of the cities. In London correlation of population density to OSM completeness estimates had an R-squared value of 0.1753. The relationship R-squared value was 0.3044 and for Leeds it was 0.0010.

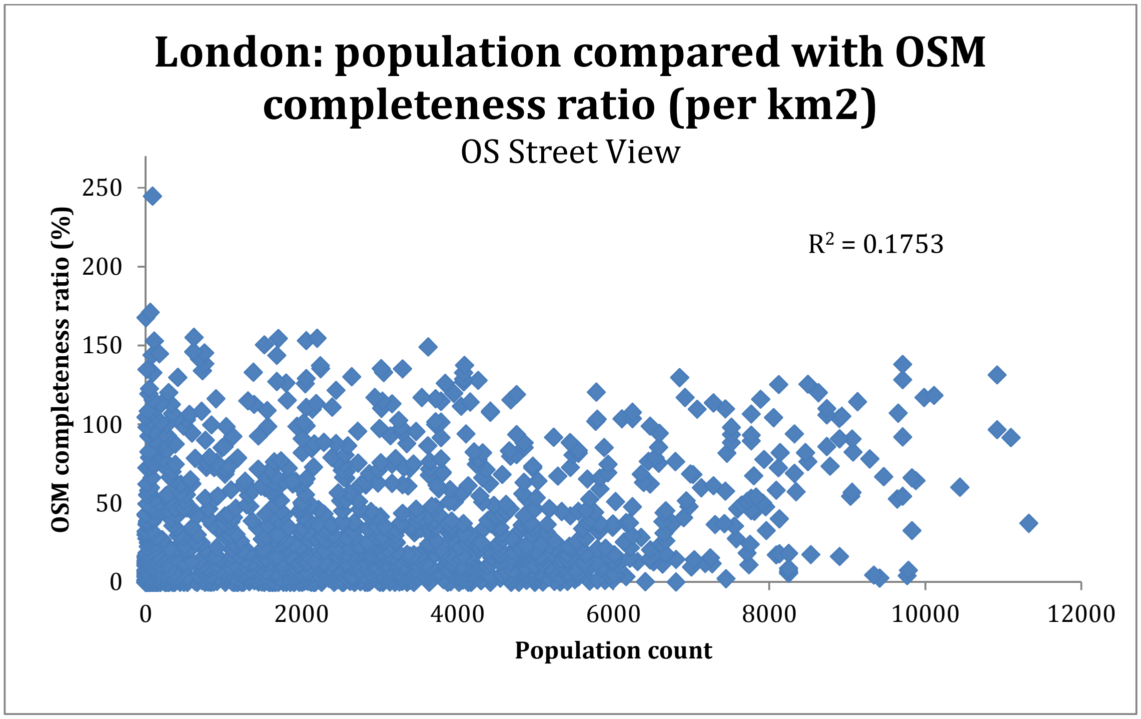


Figure 6:London, correlation between population density and OSM completeness (OS Street View as reference dataset)

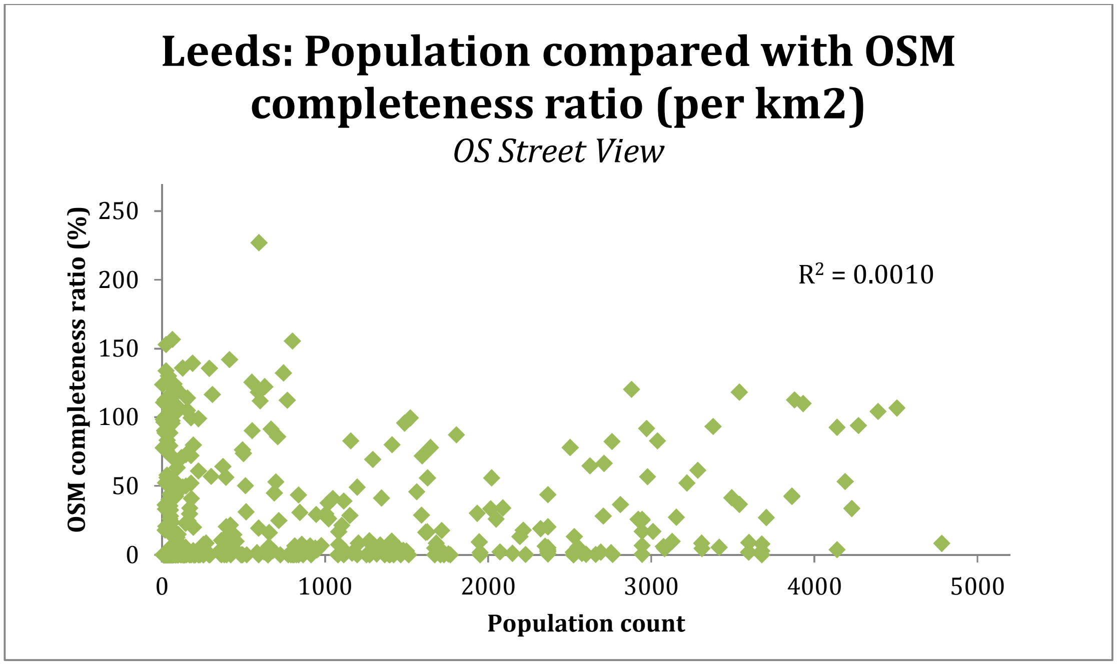


Figure 7: Leeds, correlation between population density and OSM completeness (OS Street View as reference dataset)

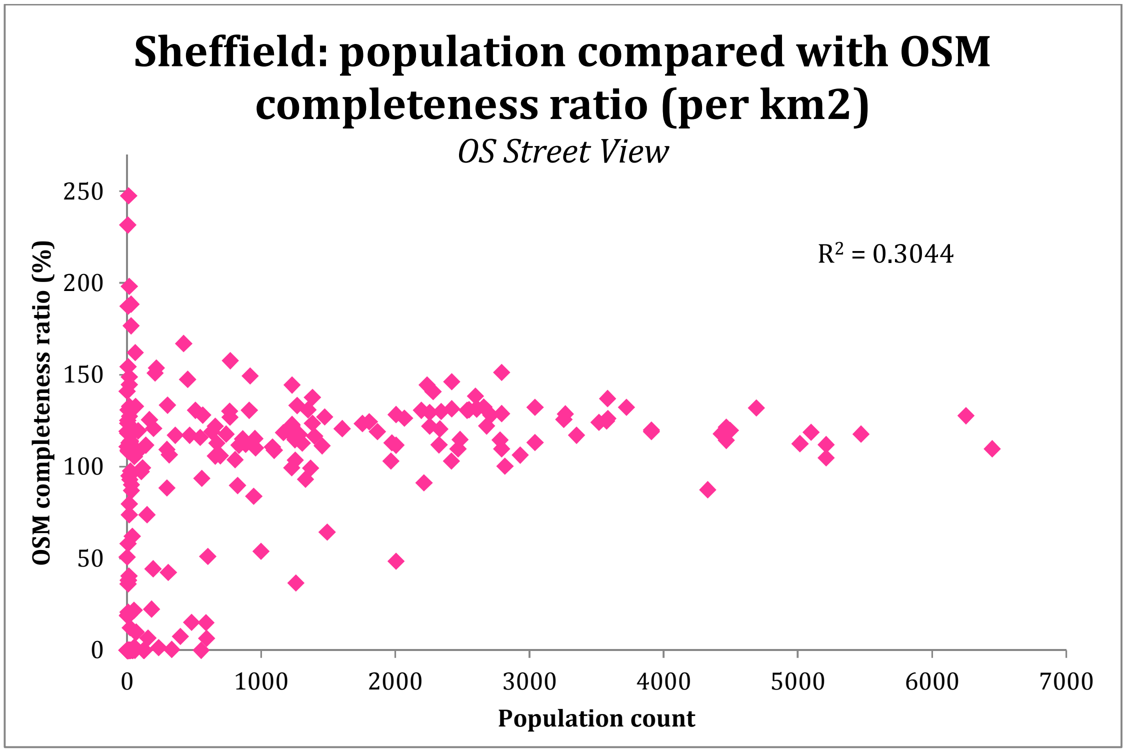


Figure 8: Sheffield, correlation between population density and OSM completeness (OS Street View as reference dataset)

4. Conclusion and further research

In order for OSM building data to add value to products like RMS's natural catastrophe exposure model, or other commercial products, OSM building completeness must be understood. This study proved that OSM building data completeness within UK cities and between UK cities is variable. However, without a reference dataset, estimating OSM building completeness is not be possible using population as a proxy variable. Alternatives proxy variables should be sought. Therefore this study did not discover a viable method for estimating OSM completeness for regions with limited official data.

5. Acknowledgements

This research was completed with the support of Risk Management Solutions. Katerina Christopoulou, my principal supervisor at Risk Management Solutions provided invaluable guidance to this study. I am very grateful for Katerina's early commitment to the project and persistent mentorship. Claire Elull, my academic advisor also provide a great amount of support to the completion of this study.

6. Biography

Claire Fram received her MSc in Geographical Information Science from UCL in 2014. She is now a Graduate Specialist in GIS at Arup in London. Her research interests include: open data, data visualisation and data analytics.

7. References

Haklay, M., 2010. How good is OpenStreetMap information? A comparative study of OpenStreet- Map and Ordnance Survey datasets for London and the rest of England. *Environment and Planning B: Planning and Design,* Volume 37, pp. 682-703.

Hecht, R., Kunze, C. & Hahmann, S., 2013. Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information,* 11 November, Volume 2, pp. 1066-1091.

Jackson, S. P. et al., 2013. Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information,* Issue 2, pp. 507-530.

Koukoletsos, T., Haklay, M. & Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS,* 14(4), pp. 477-498.

Mooney, P., Corcoran, P. & Winstanley, A. C., 2010. *Towards Quality Metrics for OpenStreetMap.* s.l., ACM, pp. 514-517.

Oak Ridge National Labratory, 2014. *LandScan.* [Online]
Available at: web.ornl.gov [Accessed 1 September 2014].