How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England

Dr Mordechai (Muki) Haklay, August 2008

Abstract

Within the framework of Web 2.0 applications, the most striking example of a geographical application is the OpenStreetMap project. OpenStreetMap aims to create a free digital map of the world and is implemented through the engagement of participants in a mode similar to software development in Open Source projects. The information is collected by many participants, collated on a central database and distributed in multiple digital formats through the World Wide Web (Web).

Within Geographical Information Science (GIScience) research, Mike Goodchild suggested a term for this type of information: 'Volunteered Geographical Information' (VGI). However, to date there has been no systematic analysis of the quality of VGI. This paper aims to fill this gap by analysing the quality of OpenStreetMap information for London and England by comparing it to Ordnance Survey datasets. As OpenStreetMap started in London in August 2004, the analysis of the information for the London area provides the best understanding of the achievements and difficulties of VGI.

The analysis shows that OpenStreetMap information can be fairly accurate: on average within about 6 metres of the position recorded by the OS, and with approximately 80% overlap of motorway objects between the two datasets. In the space of four years, OpenStreetMap has captured about 29% of the area of England, of which approximately 4% are digitised lines without a complete set of attributes. Importantly, most of the data capture (80%) was carried out by 90 participants and a very large group of users disengaged from the project after minimal contribution. The paper concludes with some suggestions for future developments and research directions.

1. Introduction

While the use of the Internet and the World Wide Web (Web) for mapping applications is well into its second decade, the landscape has changed dramatically since 2005 (Haklay *et al.*, forthcoming). The evidence for this change is the emerging neologism that follows the rapid technological developments. While terms such as Neogeography, Mapping Mash-ups, Geotagging and Geostack may seem alien to veterans in the area of Geographical information Systems (GIS), they are not from another planet. With

Contact details: Dr Muki Haklay m.haklay@ucl.ac.uk
Department of Civil, Environmental and Geomatic Engineering, UCL

over four decades of research and development, most of these seemingly ‘new’ concepts have been visited at one time or another within GIS research and practice, so Mash-up is interoperability, Geotagging is Georeferencing or Geocoding, the Geostack is a GIS and Neogeography is the sum of these terms in an attempt to divorce the past and conquer new (cyber)space.

Yet, it is hard not to notice the whole range of new websites and communities – from the commercial Google Maps to the grassroots OpenStreetMap, and to applications such as Platial – that have emerged. The sheer scale of new mapping applications is evidence of a step change in the Geographical Web (GeoWeb). For example, by mid 2007, less than two years from the introduction of the Google Maps application development toolkit (API), Tran (2007) reports 50,000 maps that are based on it. Mapping has gained prominence within the range of applications known as Web 2.0, as exemplified by the series of conferences ‘Where2.0’, which were started in 2006 by O’Reilly Media – one of the leading promoters of hi-tech knowhow: ‘GIS has been around for decades, but is no longer only the realm of specialists. The Web is now flush with geographic data, being harnessed in a usable and searchable format.’ (Where 2.0, 2008)

Haklay, Singleton and Parker (Forthcoming) provide an overview of this Web Mapping 2.0 landscape, but one of the best examples of what it is possible to achieve with the range of technologies that enabled it is the OpenStreetMap project.

OpenStreetMap (OSM) is a ‘Crowd Sourcing’ activity (Howe, 2006). Crowd Sourcing is one of the most significant and potentially controversial developments in Web 2.0. This term developed from the concept of outsourcing where business operations are transferred to remote cheaper locations (Friedman, 2006). Similarly, Crowd Sourcing is how large groups of users can perform functions that are either difficult to automate or expensive to implement. Tapscott and Williams (2006) note that ‘in many peer production communities, productive activities are voluntary and non-monetary’; content is created for free, for the benefit of the community.

OSM aims to create map data that are free to use, editable and licensed under new copyright schemes. A key motivation for this project is to enable free access to current digital geographical information across the world. For example, in European countries this information is considered to be expensive. Even in the US, where basic road information is available through the US Census Bureau TIGER/Line programme, the details that are provided are limited (streets and roads only) and do not include green space, landmarks and the like. Also, due to the cost of updates, the update cycle is slow and does not take into account rapid changes. Thus, even in the US, there is a need for detailed free geographical information.

OSM information can be edited online through a wiki-like interface where, once a user has created an account, the underlying map data can be viewed and edited. A number of sources have been used to create these maps including uploaded Global Positioning System (GPS) tracks, out of copyright maps and, more recently, Yahoo! aerial imagery which was made available through collaboration with this search engine. Unlike Wikipedia, where the majority of content is created at disparate locations, the OSM community also organises a series of local workshops (called ‘mapping parties’), which aim to create and annotate content for localised geographical areas (see Perkins and Dodge, 2008). These events are designed to introduce new contributors to the community with hands-on experience of collecting data, while positively contributing to the project overall by generating new information and street labelling as part of the exercise. The OSM data are stored on servers at University College London, and Bytemark, which contributes the bandwidth for this project. Whilst over 50,000 people have contributed to the map as of August 2008, it is a core group of about 40 volunteers who dedicate their time to create the technical infrastructure for a viable data collection and dissemination service. This includes the maintenance of the servers, writing the core software that handles the transactions with the server in adding and editing GI, and creating cartographical outputs. The project includes two editing tools that participants have developed with a lightweight editing software package that is working within the browser and another stand-alone version, more akin to a GIS editing package. For a detailed discussion of the technical side of the project, see Haklay and Weber (Forthcoming).

The potential of Crowd Sourced geographical information has captured the attention of researchers in GIS (including Goodchild, 2007a, 2007b; Sui, 2008). Goodchild has coined a term to describe this activity as 'Volunteered Geographic Information' (VGI). One of the significant core questions within the VGI framework is how good is the quality of the information? This is a crucial question about the efficacy of VGI activities and the value of the outputs for a range of applications, from basic navigation to more sophisticated applications such as site location planning.

With OSM, it is possible to answer this question by examining the dataset against Ordnance Survey (OS) datasets in the UK. As OSM started in London, and thus the city represents the place that received the longest ongoing attention from OSM participants, it stands to reason that an examination of the city and of England will provide an early indication about the quality of VGI.

This paper discusses an analysis of the quality of the OSM dataset, evaluating its positional and attribute accuracy, completeness and consistency. In light of this analysis, the paper suggests the fitness for purpose of OSM information and some possible directions for future developments. However, before turning to the analysis, a short discussion of evaluations of geographical information quality will help to set the scene.

2. *How to evaluate the quality of geographical information*

The problem of understanding the quality of geographical databases was identified many years ago, and received attention from surveyors, cartographers and geographers (van Oort, 2006). Van Oort identified work on the quality of geographical information dating back to the late 1960s and early 1970s.

With the emergence of Geographical Information Systems in the 1980s, this area of research experienced rapid growth, receiving attention from leading figures in the area of Geographical Information Science including Peter Borough, Mike Goodchild, Peter Fisher, Nick Chrisman and many others (see van Oort for a comprehensive review of the area). By 2002, quality aspects of geographical information had been enshrined in the International Organisation for Standards (ISO) codes 19113 (Quality principles) and 19114 (Quality evaluation procedures) under the aegis of Technical Committee 211. Based on these standards, Kresse and Fadaie (2004) identified the following aspects of quality:

completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy, purpose, usage and lineage.

In his synthesis of various quality standards and definitions, van Oort identifies the following aspects:

- Lineage – this aspect of quality is about the history of the dataset, how it was collected and evolved.
- Positional accuracy – this is probably the most obvious aspect of quality and evaluates how well the coordinate value of an object in the database relates to the reality on the ground.
- Attribute accuracy – as objects in a geographical database are represented not only by their geometrical shape but also by additional attributes, this measure evaluates how correct these values are.
- Logical consistency – this is an aspect of the internal consistency of the dataset, in terms of topological correctness and the relationships that are encoded in the database.
- Completeness – this is a measure of the lack of data, which does not record objects that are expected to be found in the database, or excess data that should not be included.
- Semantic accuracy – this measure links the way in which the object is captured and represented in the database to its meaning and the way in which it should be interpreted.
- Usage, purpose and constraints – this is a fitness-for-purpose declaration that should help potential users in deciding how the data should be used.
- Temporal quality – this is a measure of the validity of changes in the database in relation to real-world changes and also the rate of updates.

Naturally, the definitions above are shorthand and aim to explain the principles of geographical information quality. The burgeoning literature on geographical information quality provides more detailed definitions and discussion of these aspects.

To understand the amount of work that is required to achieve a high-quality geographical database, the OS provides a good example of monitoring completeness and temporal quality. To achieve this goal, the OS has an internal quality assurance process, known as 'The Agency Performance Monitor'. This is set by the UK government and requires that 'Some 99.6% significant real-world features are represented in the database within six months of completion'. Internally to Ordnance Survey, the operational instruction that is based on this criteria is the maintenance of the OS current large-scale database currency at an average of no more than 0.7 House Units of unsurveyed major change, over six months old, per Digital Map Unit (DMU). DMUs are inherently map tiles, while House Units are a measure of data capture, with the physical capture of one building as the basic unit. In practical terms this means that every six months the OS analyses the result of auditing over 4000 semi-randomly selected DMUs for missing detail by sending semi-trained surveyors with printed maps on the ground. This is a significant and costly undertaking but it is an unavoidable part of creating a reliable and authoritative geographical database. Noteworthy is that this work focuses on completeness and temporal quality, while positional accuracy is evaluated through a separate process.

As this type of evaluation is clearly not feasible for OSM, a desk-based approach was taken using two geographical datasets: the OS dataset and OSM dataset. The assumption is that, at this stage of OSM development, the OS dataset represents higher accuracy and overall quality (at least positional and attribute). Considering the lineage and investment in the OS dataset, this should not be a contested statement. This type of comparison is common in geographical information quality research (see Hunter, 1999; Goodchild *et al.,* 1992).

3. Datasets used

For the comparison of the Ordnance Survey (OS) vector dataset with the OpenStreetMap dataset, the OS Meridian 2 (for the sake of simplicity, 'Meridian') dataset was used. Meridian 2 is a vector dataset that provides coverage of Great Britain with complete details of the national road network: 'Motorways, major and minor roads are represented in the dataset. Complex junctions are collapsed to single nodes and multi-carriageways to single links. To avoid congestion, some minor roads and cul-de-sacs less than 200m are not represented ... Private roads and tracks are not included.' (OS, 2007, p. 24.) The source of the road network is high-resolution mapping (1:1250 in urban areas, 1:2500 in rural areas and 1:10,000 in moorland).

The dataset notional resolution is 1:50,000 and the official resolution statement is 1 metre. The way in which Meridian is constructed is that the node points are kept in their original position while, through a process of generalisation, the road centre line is filtered to within a 20m region of the original location. The generalisation process decreases the number of nodes to reduce clutter and complexity.

The OS describes Meridian as a dataset suitable for applications from environmental analysis to design and management of distribution networks for stores and warehouses to health planning.

All these characteristics make this dataset the most suitable for comparison with the OSM dataset. The main reason to justify this comparison is that, due to the dataset collection method, the OSM dataset cannot be more accurate than the quality of the GPS receiver (which usually captures a location within 6-10m) and the Yahoo! imagery, which outside London provides about 15m resolution. This means that we can expect the OSM dataset to be within a region of about 20m from the true location.

The OSM dataset that was used in this comparison was from the end of March 2008, and was based on roads information created by Frederik Ramm and available on his website Geofabrik. The dataset is provided as a set of thematic layers (building, natural, points, railways, roads and waterways), which are classified according to their OSM tags.

Two other sources were used to complete the comparison. First, the 1:10,000 raster files from the OS. These are based on detailed mapping, and went through a process of generalisation that leaves most of the features intact. It is a highly detailed map, which again is suitable for locating attribute information and details of streets and other features that are expected to be found in OSM.

The second source is the Lower Level Super Output Areas (SOA), which is provided by the OS and the Office of National Statistics and is based on the Census. SOAs are about the size of a neighbourhood and are created through a computational process by merging the basic Census units. This dataset was combined with the Index of Deprivation 2007, created by the Department of Communities and Local Government and which indicates the socio-economic status of each SOA.

4. Comparison framework

It is important to understand that a comparison between Meridian and OSM is not like with like. A more appropriate comparison would have been with Navteq or TeleAtlas information, where comprehensive street level information without generalisation is available. Yet, it is exactly these characteristics of Meridian that allow a general comparison of coverage between the two datasets.

The main hypothesis behind the comparison is:

Because Meridian is generalised, excludes some of the minor roads, and does not include foot and cycle paths, in every area where OSM has a good coverage, the total length of OSM roads must be longer than the total length of Meridian features.

Furthermore, the fact that Meridian's nodes are derived from high-resolution datasets means that it can be used for positional accuracy analysis.

This aspect can be compared across the whole area of Great Britain, but as OSM started in England (and more specifically in London) a comparison across England was more appropriate and manageable.

The process of comparison started from an evaluation of positional accuracy, first by analysing motorway objects in the London area, and then by closely inspecting five OS tiles at 1:10,000 resolution, covering 113 square kilometres. After this comparison, an analysis of completeness was carried out: first through a statistical analysis across England, followed by a detailed visual inspection of the 1:10,000 tiles. Finally, statistical analysis of SOAs and ID 2007 was carried out.

5. Detailed methodology and results

Of the various quality aspects that were mentioned above, positional accuracy and completeness stand out as the two aspects that are most commonly associated with geographical information quality in popular understanding. Thus, the evaluation of the quality starts with these aspects.

5.1 Positional accuracy: motorway comparison[1]

The evaluation of the positional accuracy of OSM can be carried out against Meridian, since the nodes of Meridian are derived from the high-resolution topographical dataset and thus are highly accurate. However, the fact that the number of nodes has been diluted by the application of a filter and the differing digitising methods means that the two datasets have a different number of nodes.

[1] This section is based on the M.Eng. report of Naureen Zulfiqar

Furthermore, OSM represents motorways as a line object for each direction, whereas Meridian represents them as a single line (Figure 1). This means that matching on a point-by-point basis would be inappropriate. Therefore, the analysis was carried out on the linear features.

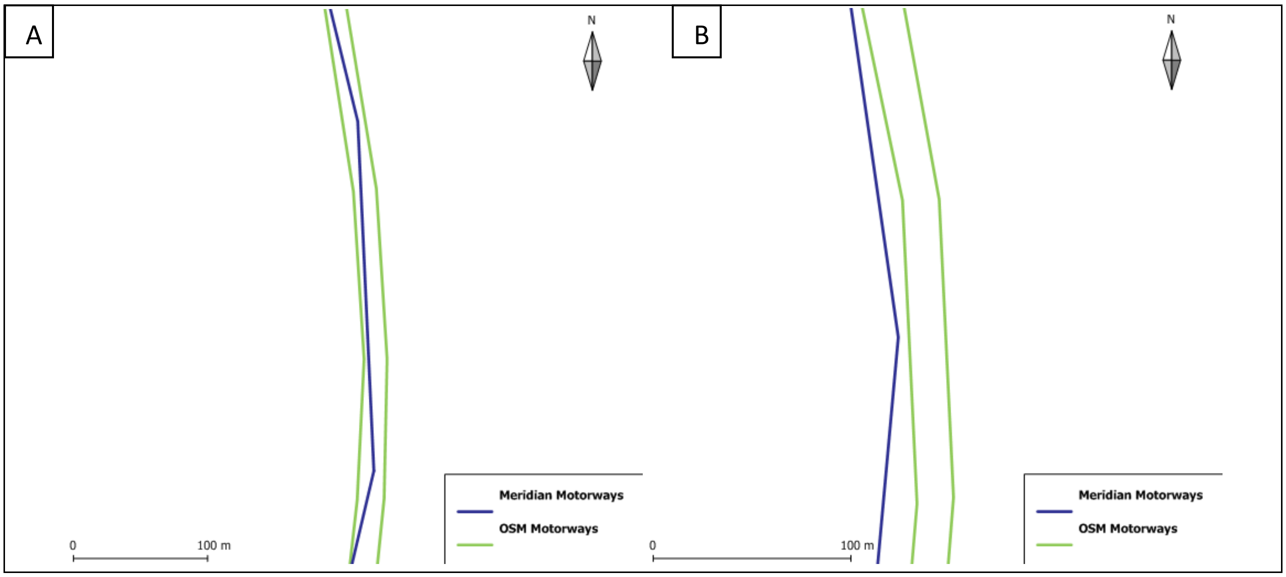

Figure 1 – Two stretches of the M11 Motorway: (A) Expected position of OS and OSM, (B) OSM dataset not matching OS

The methodology used to evaluate the positional accuracy of motorway objects across the two datasets was based on Goodchild and Hunter (1997) and Hunter (1999). The comparison is carried out by using buffers to determine the percentage of line from one dataset that is within a certain distance of the same feature in another dataset of higher accuracy (Figure 2).

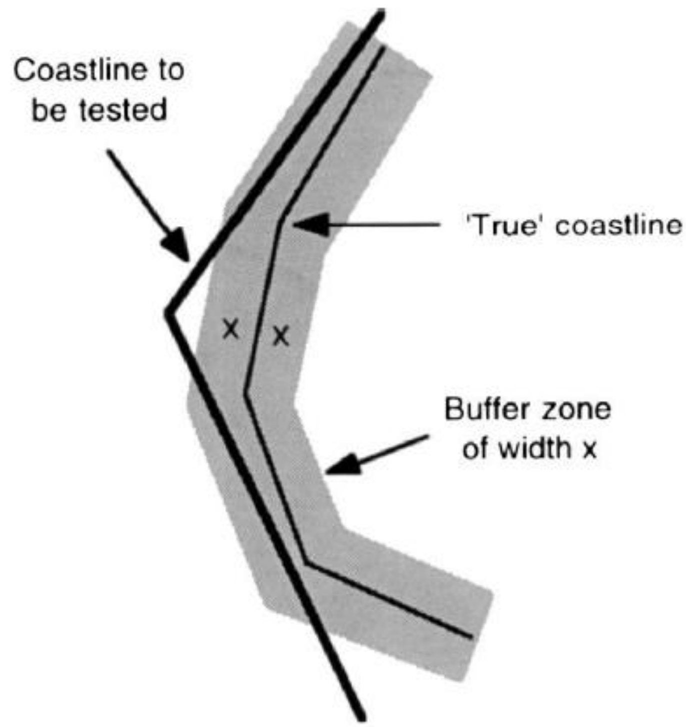

Figure 2 – Goodchild and Hunter buffer comparison method. The buffer of width x is created around the high-quality object, and the percentage of the tested object that falls within the buffer is evaluated (Source: Goodchild & Hunter, 1997)

The preparation of the datasets for comparison included some manipulation. First, the comparison was carried out for the motorways in the London area. Both datasets so they were representing roughly the same area and length. Complex slip road configurations were edited in the OSM dataset to ensure that the representation was similar. The rest of the analysis was carried out by creating a buffer around each dataset, and then evaluating the overlap. As the OS represents the two directions as a single line, it was decided the buffer around the OS data should be set at 20 metres (as this is the filter that the OS applies in the creation of the line) and, to follow Goodchild and Hunter's method, the OSM dataset was buffered with a 1-metre buffer to calculate the overlap.

The results are displayed in Table 1.

Motorway	Percentage Overlap
M1	87.36%
M2	59.81%
M3	71.40%
M4	84.09%
M4 Spur	88.77%
M10	64.05%

M11	84.38%
M20	87.18%
M23	88.78%
M25	88.80%
M26	83.37%
M40	72.78%
A1(M)	85.70%
A308(M)	78.27%
A329(M)	72.11%
A404	76.65%

Table 1 – Percentage overlap between OS and OSM buffers

With an average overlap of nearly 80%, and variability from 60% up to 89%, the OSM dataset provides a good representation of motorways. A further analysis is required to compare segments of the motorway as it is represented in OSM in comparison to OS MasterMap, as this can be done with the use of buffers around the OS dataset only to represent the true width of the motorway. Also, the same analysis should be carried out for A-roads and B-roads, where the ability to compare the two datasets is more straightforward than the case of motorways.

5.2 Positional accuracy: urban areas in London

In addition to the statistical comparison, a more detailed, visual comparison was carried out across 113 square kilometres in London using five OS 1:10,000 raster tiles (TQ37ne – New Cross, TQ28ne – Highgate, TQ29nw – Barnet, TQ26se – Sutton, and TQ36nw – South Norwood). In each one of them, the tiles were inspected visually and 100 samples were taken to evaluate the difference between the OS centreline and the location that is recorded in OSM.

The average differences between the OS location and OSM are provided in Table 2.

Area	Average difference (m)
Barnet	6.77
Highgate	8.33
New Cross	6.04
South Norwood	3.17
Sutton	4.83
Total	*5.83*

Table 2 – Positional accuracy across five areas of London

Notice the difference in averages between the areas. In terms of the underlying measurements, in the best areas many of the locations are within a metre or two of the location, whereas in Barnet and Highgate distances of up to 20 metres from the OS centreline were recorded. Figure 3 provides examples from New Cross (A), Barnet (B) and Highgate(C), which show the overlap and mismatch between the two datasets.

Figure 3 – Examples of overlapping OSM and OS maps for New Cross (A), Barnet (B) and Highgate (C). The green lines which overlay the map are OSM features.

The visual examination of the various tiles shows that the accuracy and attention to detail differs between areas. This can be attributed to digitisation and data collection skills and the patience of the person who carried out the work.

5.3 Completeness: length comparison

After gauging the level of positional accuracy of the OSM dataset, the next issue is the level of completeness. While Steve Coast, the founder of OSM, stated 'it's important to let go of the concept of completeness' (GISPro, 2007), it is important to know which areas are well covered and which require further work by the OSM community.

To prepare the dataset for comparison, a grid at a resolution of 1km was created across England. Next, as the comparison is trying to find a ratio between OSM and Meridian objects, and to avoid the inclusion of coastline objects and small slivers of grid cells, all incomplete cells with an area less than a square kilometre were eliminated. This meant that out of the total area of England of 132,929 sq km, the comparison was carried out on 123,714 sq km (about 93% of the total area).

The first step was to project the OSM dataset onto the British National Grid, to bring it to the same projection as Meridian. The grid was then used to clip all the road objects from OSM and from Meridian in such a way that they were segmented along the grid lines. This step enabled the comparison of the two sets in each cell grid across England.

The rest of the analysis was carried out through SQL queries, which added up the length of lines that were contained or intersected the grid cells. The clipping process was carried out in MapInfo, whereas the analysis was in Manifold GIS.

The results of the analysis show the current state of OSM completeness. At the macro level, the total length of Meridian roads is 302,349,778 metres, while OSM is 209,755,703 metres. Thus, even at the highest level, the OSM dataset total length is 69% of Meridian. It is important to remember that in this and in the following comparisons that Meridian is an incomplete and generalised coverage, and thus this is an under-estimation of the real value for England. Yet, considering the fact that OSM has been around for a mere four years, this is a significant and impressive rate of data collection.

There are 16,300 sq km in which neither OSM nor Meridian has any feature. Out of the remainder, in 70.7% of the area, Meridian provides a better, more comprehensive coverage than OSM. In other words OSM volunteers have provided an adequate coverage for 29.3% of the area of England in which we should expect to find features.

Table 3 – Length comparison: OSM and Meridian (sq km)

Empty cells	16,300 (13.2%)
Meridian more detailed than OSM	75,977 (61.4%)
OSM more detailed than Meridian 2	31,437 (25.4%)
Total	**123,714**

Naturally, the real interest lies in the geography of these differences. The centres of the big cities of England (such as London, Manchester, Birmingham, Newcastle, and Liverpool) are well mapped using this measure. However, in the suburban areas, and especially in the boundary between the city and the rural area that surrounds it, the quality of coverage drops very fast and there are many areas that are not covered very well.

The following series of images provide examples for these differences across the major areas of England.

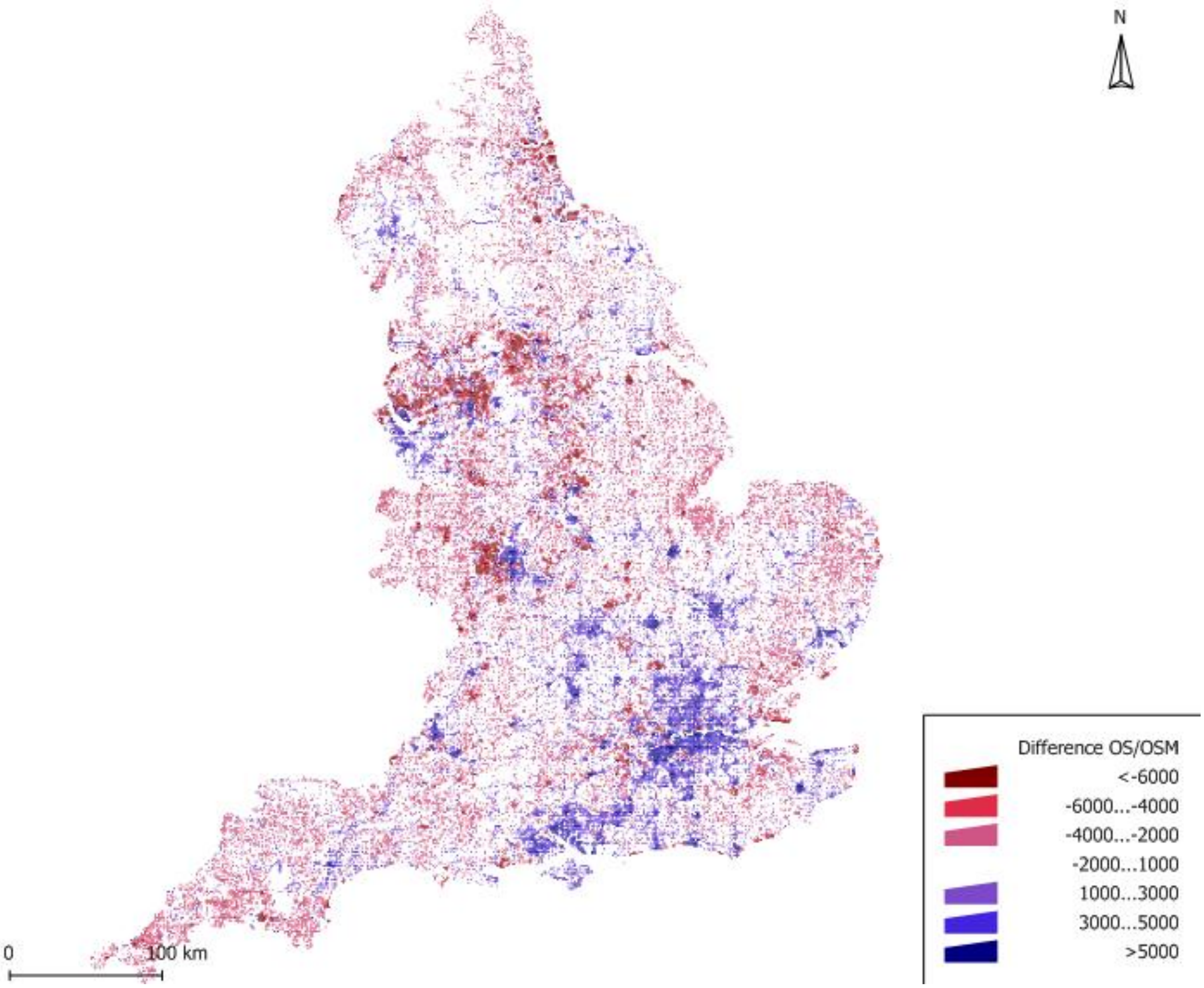

Figure 4 – Length difference between OS Meridian dataset and OSM dataset. The bigger the difference, the more incomplete the OSM dataset is. The blue tinge shows where OSM is likely to be complete, while the red tinge indicates incompleteness. The white areas are the locations where it is difficult to make a judgement using this indicator.

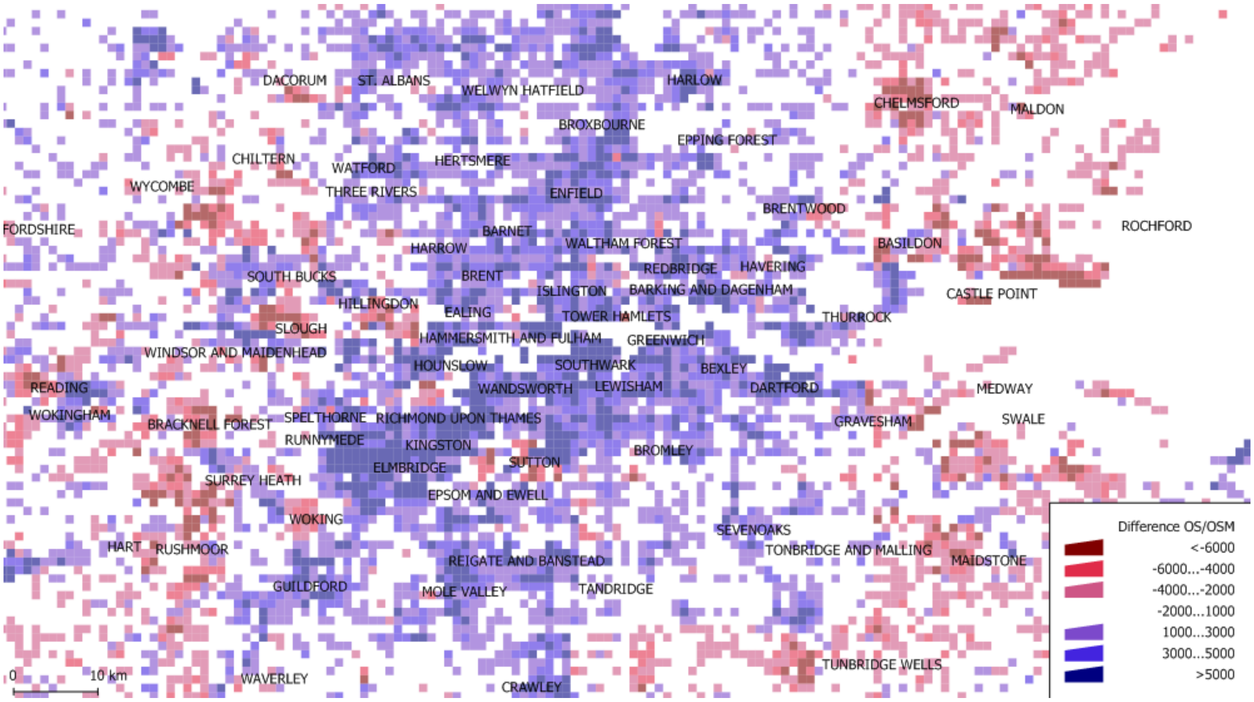

Figure 5 – London area

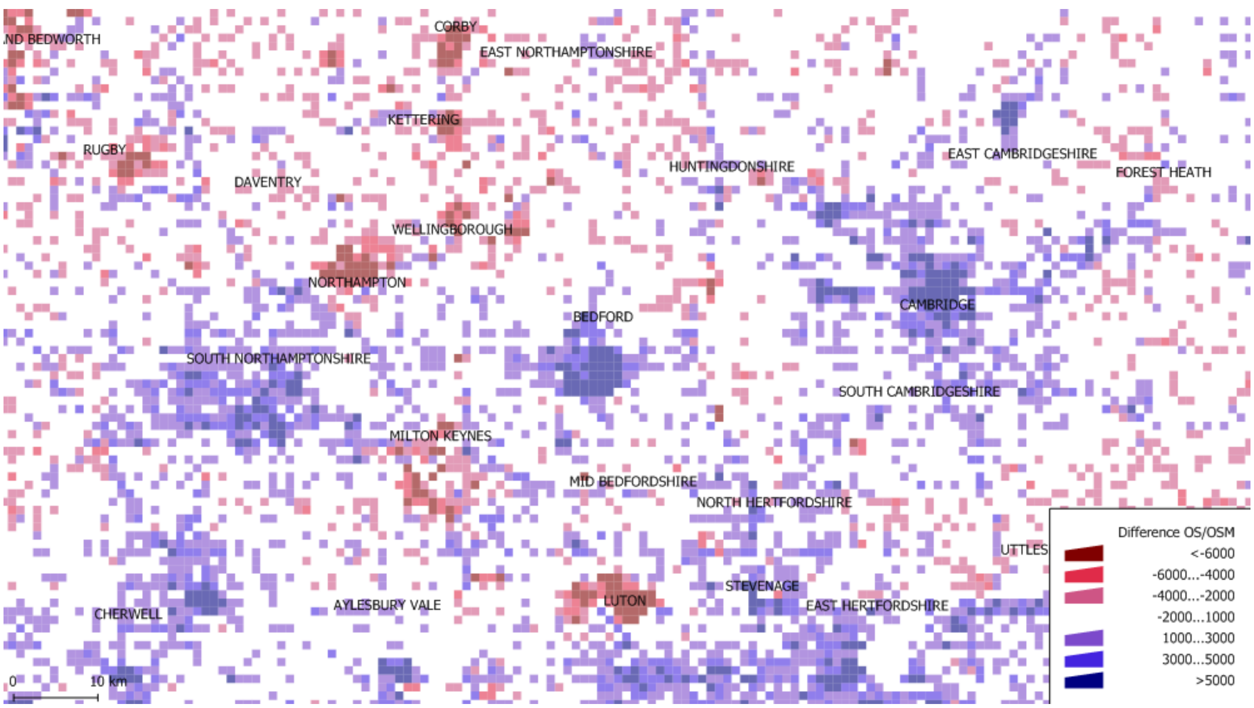

Figure 6 – Milton Keynes and Cambridge area

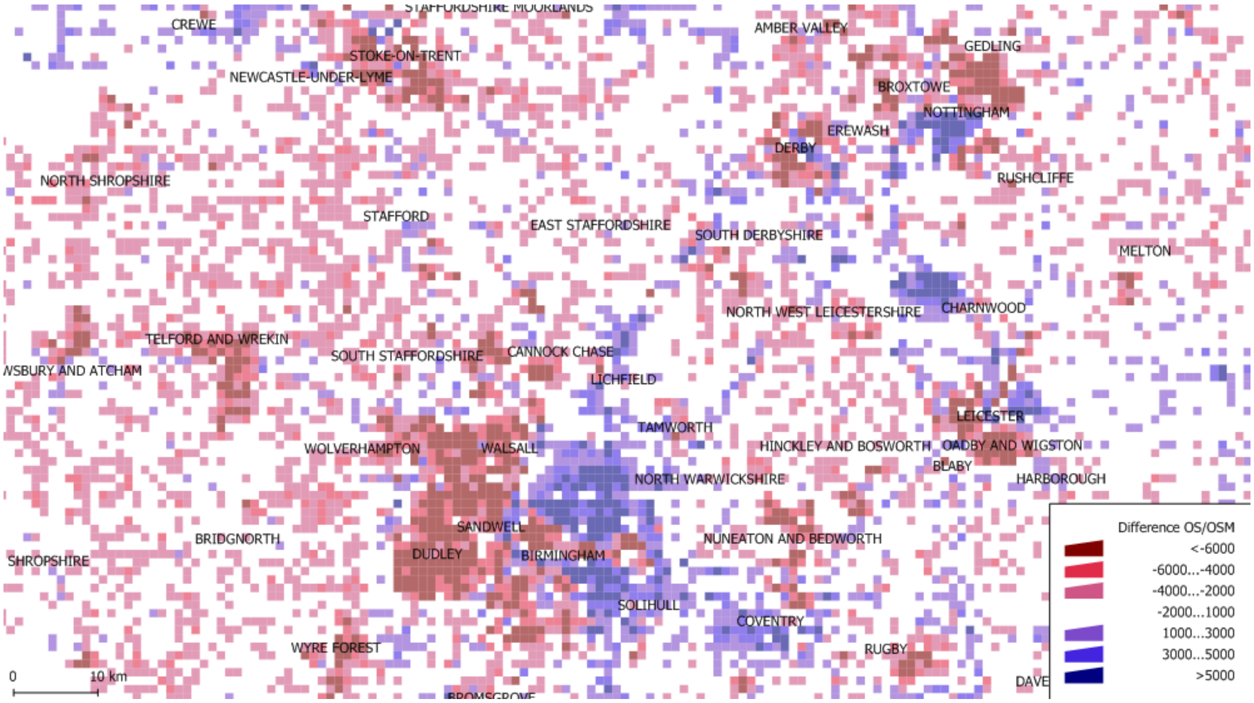

Figure 7 – Birmingham and Nottingham area

Figure 8 – Manchester and Liverpool area

Figure 9 – Newcastle Upon Tyne area

Following this comparison, which inherently compares all the line features that are captured in OSM, including footpaths and other minor roads, a more detailed comparison was carried out. This time, only OSM features that were comparable to the Meridian dataset were included (e.g. motorway, major road, residential).

Noteworthy is that this comparison moves into the area of attribute quality, as a road that is included in the database but without any tag will be excluded. Furthermore, the hypothesis that was noted above still stands – in any location in which the OSM dataset has been captured completely, the length of OSM objects must be longer than Meridian objects.

Table 4 – Length comparison: OSM and Meridian (sq km)

Empty cells[2]	17,632 (14.3%)
Meridian 2 more detailed than OSM	80,041 (64.7%)
OSM more detailed than Meridian 2	26,041 (21.0%)
Total	**123,714**

[2] The rise in the number of empty cell is due to the removal of cells that contain OSM information on paths and similar features.

Notice that under this comparison, the OSM dataset is providing coverage for 24.5% out of the total area that is covered by Meridian. Figure 10 provides an overview of the difference in the London area.

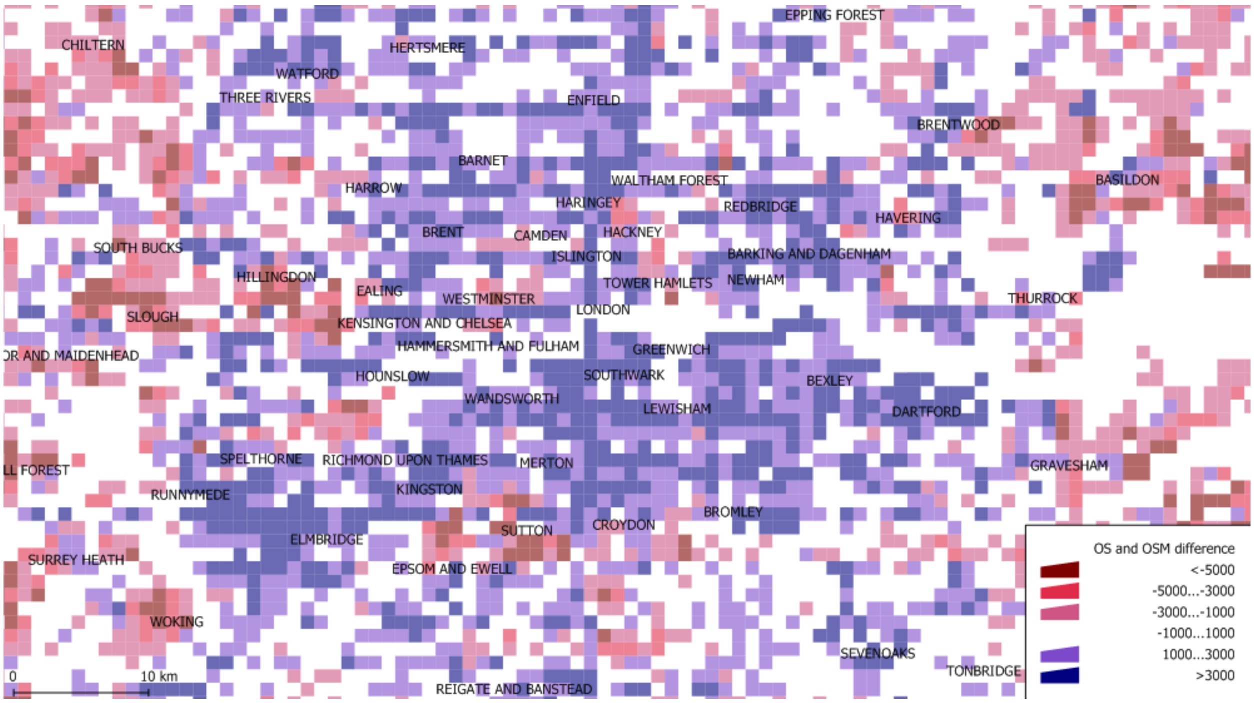

Figure 10 – difference between OS and OSM, including attributes that indicate that the feature is in a comparable category to Meridian.

5.4 Completeness: urban areas in London

Another way to evaluate the completeness of the dataset is by visual inspection of the dataset against another dataset. Dair Grant, an OSM contributor, has developed a method to do such a comparison with Google Maps (see http://www.refnum.com/osm/gmaps/). However, the details that are available on Google are partial as there are no landmarks and other features. Similar to the method that was described above for the detailed analysis of urban areas, 113 square kilometres in London were examined visually to understand the nature of the incompleteness in OSM. The five 1:10,000 raster tiles are shown in Figure 11, and provide a good cross-section of London from the centre to the edge. Each red circle on the image indicates an omission of a detail or a major mistake in digitising (such as a road that passes through the centre of a built up area).

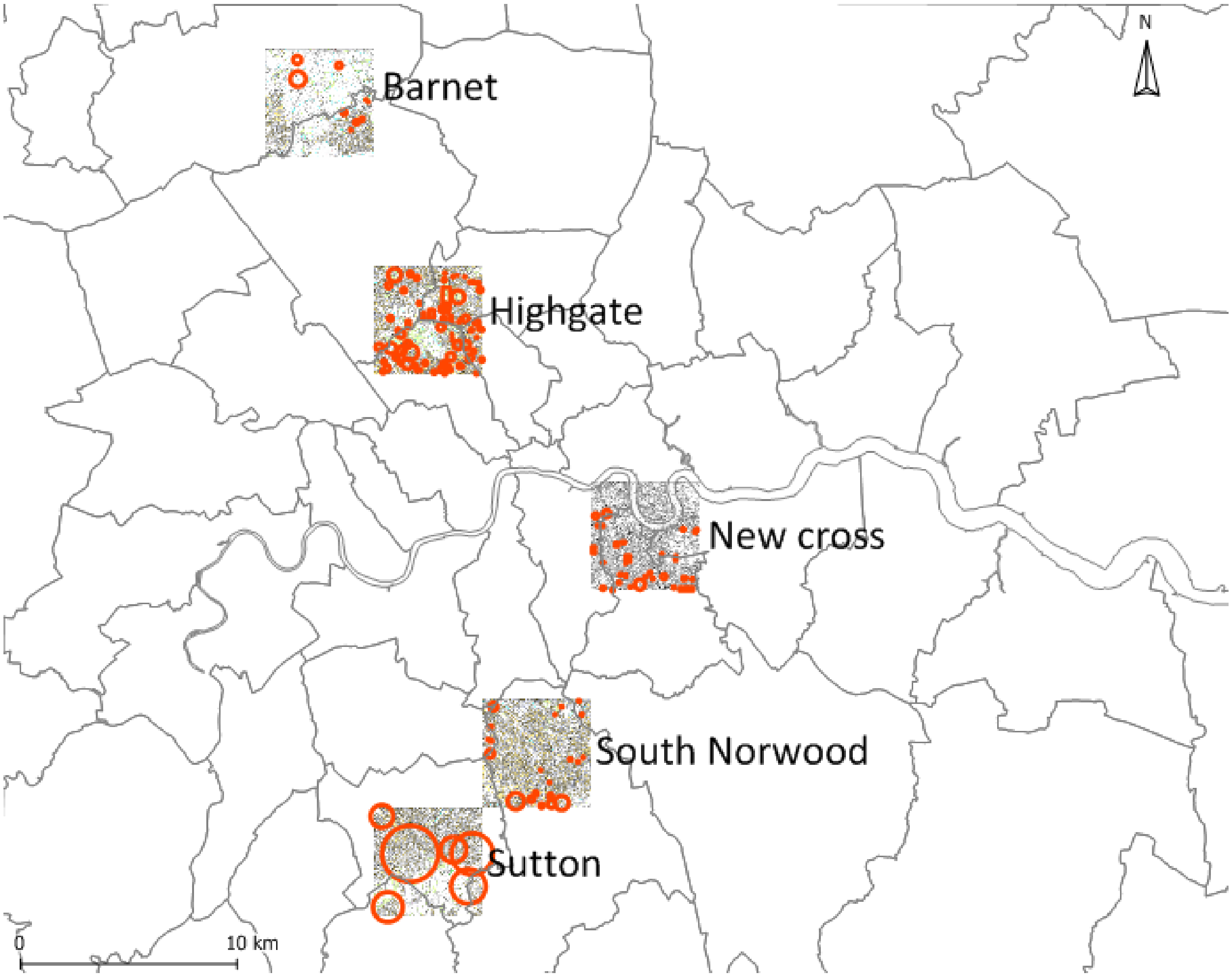

Figure 11 – Overview of completeness across five areas in London

Out of the five tiles, two stand out dramatically. The Highgate tile includes many omissions, and, as noted in the previous section, also examples of sloppy digitisation, which impact the positional accuracy of the dataset. As Figure 12 shows, open spaces are missing, as well as minor roads. Notice that some of OSM lines are at the edge of the roads and some errors in digitising can be identified clearly. The Sutton tile contains large areas that are completely missing – notice the size of the circles.

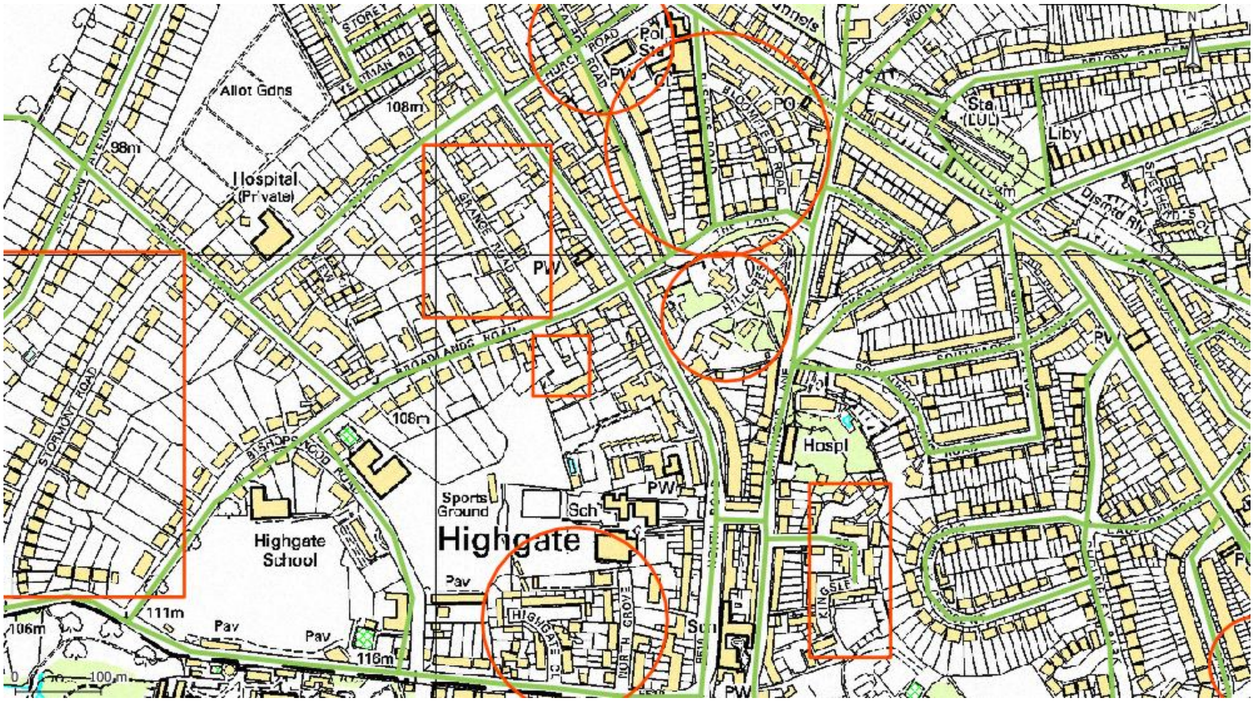

Figure 12 – Highgate area. The green lines are OSM features, and the red circles indicate omissions.

The South Norwood tile, on the other hand, shows small areas that are missing completely, while the quality of the digitising is high (Figure 13).

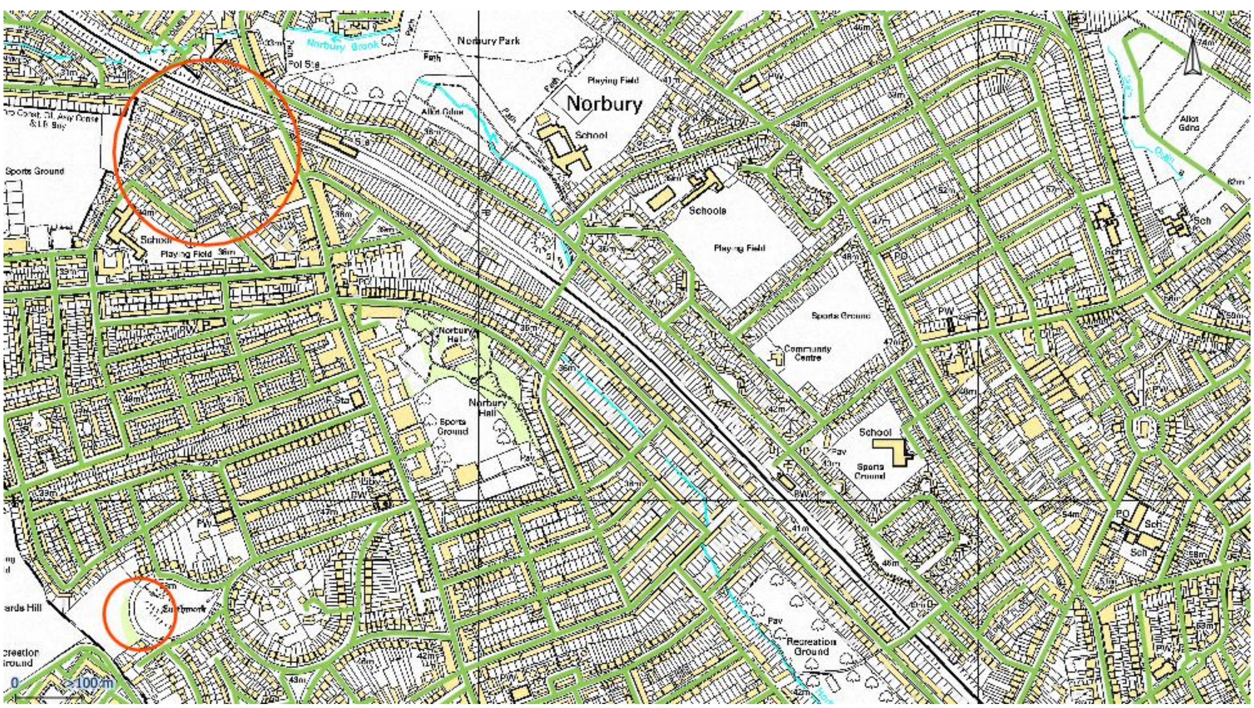

Figure 13 – South Norwood area: high-quality data capture and good match with OS features.

6. *Number of users per area*

Another aspect of quality is the number of users that digitised each area. The number of users for each grid cell is a good indication of quality as it indicates whether all the work was carried out by one individual or there were more people involved so that the likelihood of identifying errors and correcting them increased. This follows the principle of Open Source software development, which highlights the importance of 'Given enough eyeballs, all bugs are shallow' (Raymond, 2001, p.19). For mapping, this should be translated as the number of contributors that worked on an area and therefore removed 'bugs' from it.

To analyse this aspect, the nodes dataset from OSM, which contains the identification of user for 92.25% of the nodes in the area of Great Britain, was used. To gain some understanding about the nature of OSM data collection, it is worth noticing the percentage of nodes which each user collected (Figure 14).

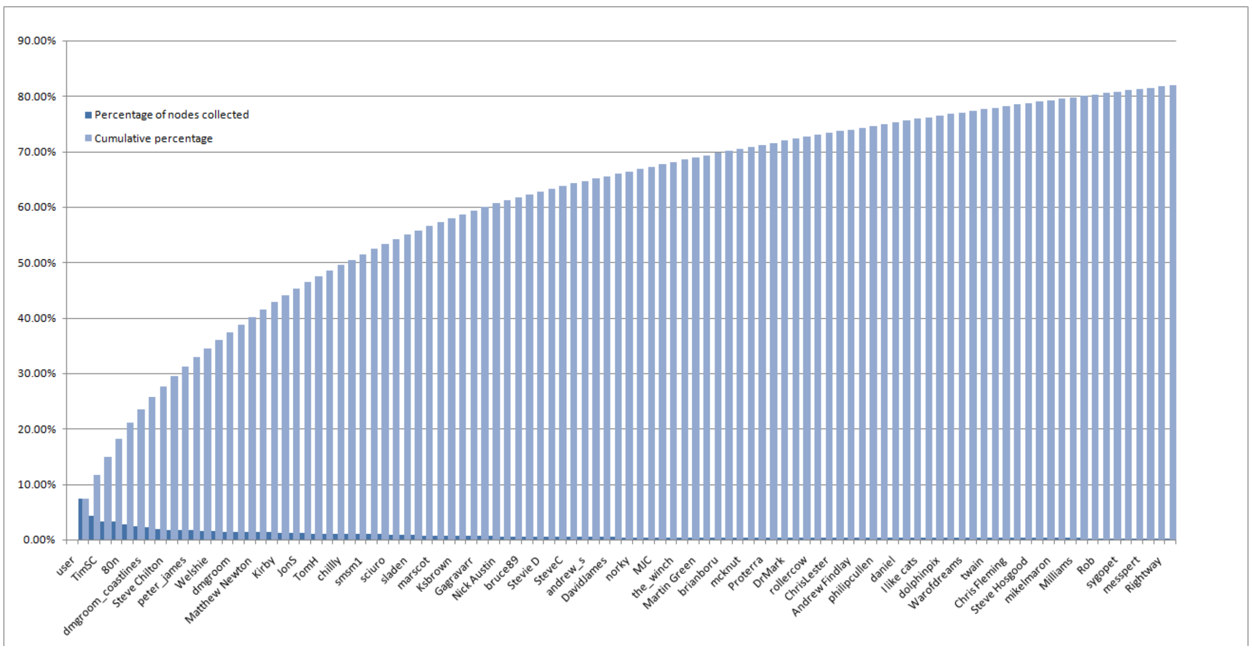

Figure 14 – percentage of nodes collected by a user and cumulative data collected

The OSM dataset records about 1100 users who contributed to the dataset, and the number might be higher if we assume that some of the anonymous nodes were also recorded by specific users. However, the contribution to data collection is not spread evenly across the users. Of the users, 26 contributed over 50% of the data, and 92 contributed 80% of it. There are many hundreds of users who only contributed a few nodes.

In the next stage of analysis, the nodes dataset was overlaid with the grid described in previous stages, and the number of users for each grid square was calculated. Table 5 provides a summary of the number of users for each cell.

Table 5 – Number of users for grid cells

Number of Users	Area covered (sq km)
1	40021
2	20720
3	9136
4	4184
5	1986
6	936
7	448
8	269
9	139
10 and above	246

Table 5 shows that 51.3% of the total 78,085 sq km that contain OSM features has been mapped by a single person. Cumulatively, the areas that were covered by very few users (up to 3) are 89.5% of the total area. In many cases, the contribution by two users is caused by a motorway feature that passes through the cell: no clear collaboration has taken place.

The contribution also has a spatial distribution and, as expected, central areas where mapping parties occur show a higher number of users. An example of this pattern is provided in Figure 15 for the area of London.

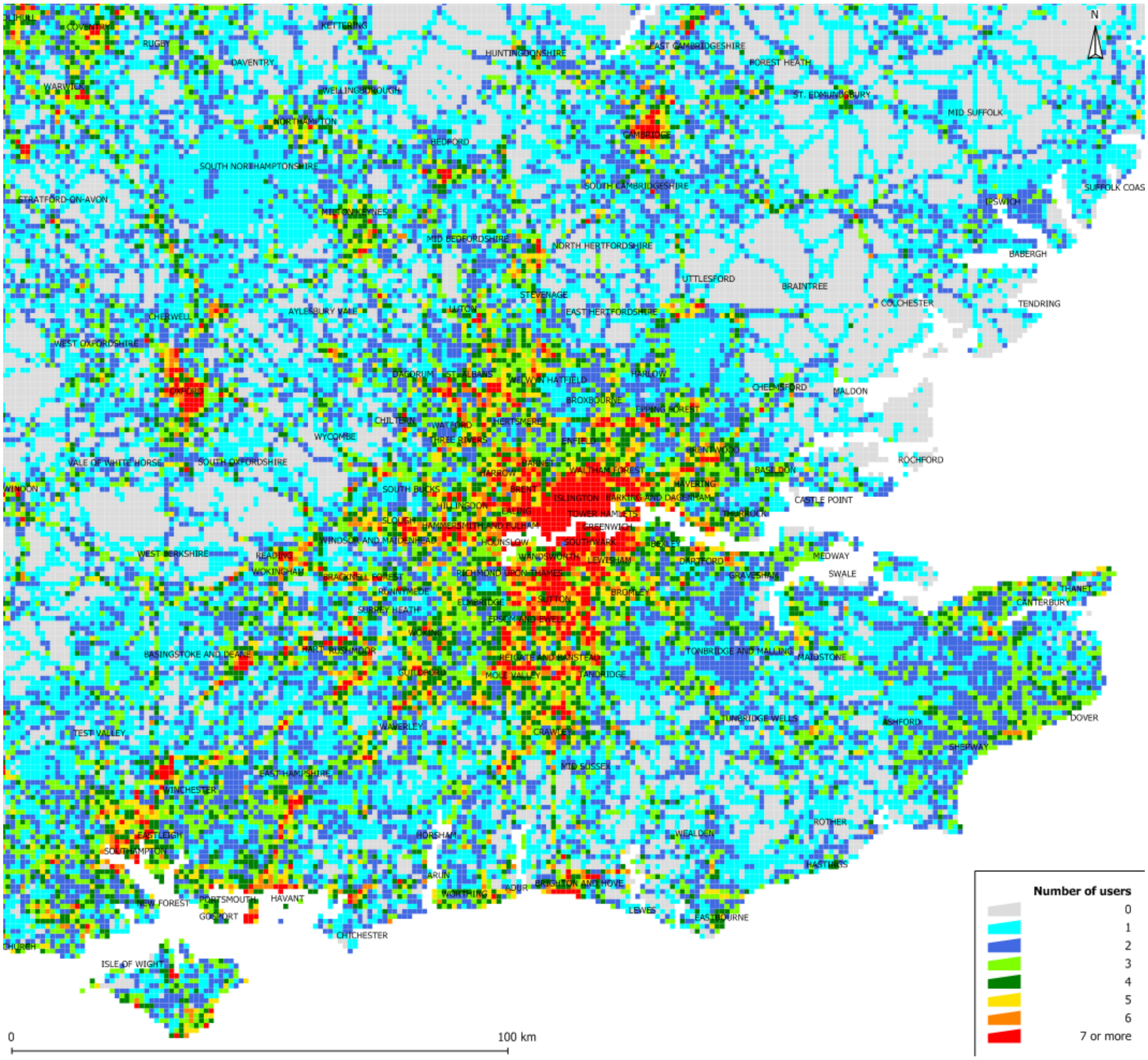

Figure 15 – Number of users in the London area

7. Social justice and OSM dataset

Another measure of data collection is the equality in which it is collected. Following the principle of universal service, governmental bodies and organisations like the Royal Mail or the Ordnance Survey are committed to providing full coverage of the country, regardless of the remoteness of the location. As OSM relies on the decisions of contributors about the areas that they would like to collect, it is interesting to evaluate the level in which deprivation influences data collection.

For this purpose, the UK government's Index of Deprivation 2007 (ID 2007) was used. The Index is provides a score for each Super Output Area (SOA) in England, and it is possible to calculate the percentile position of each SOA. Each percentile point includes about 325 SOAs. ID 2007 is calculated from a combination of governmental datasets and provides an index score for each area. Areas that are in the bottom percentiles are the most deprived, while those at the 99th percentile are the most affluent places in the UK.

Following the same methodology that was used with the 1-km grid, the road datasets from OSM and from Meridian were clipped to each of the areas for the purpose of comparison. In addition, OSM nodes were examined against the SOA layer.

As Figure 16 shows, we can see a clear difference between SOA at the bottom of the scale and at the top. While they are not neglected, the level of coverage is far lower, even when taking into account the variability in SOA size.

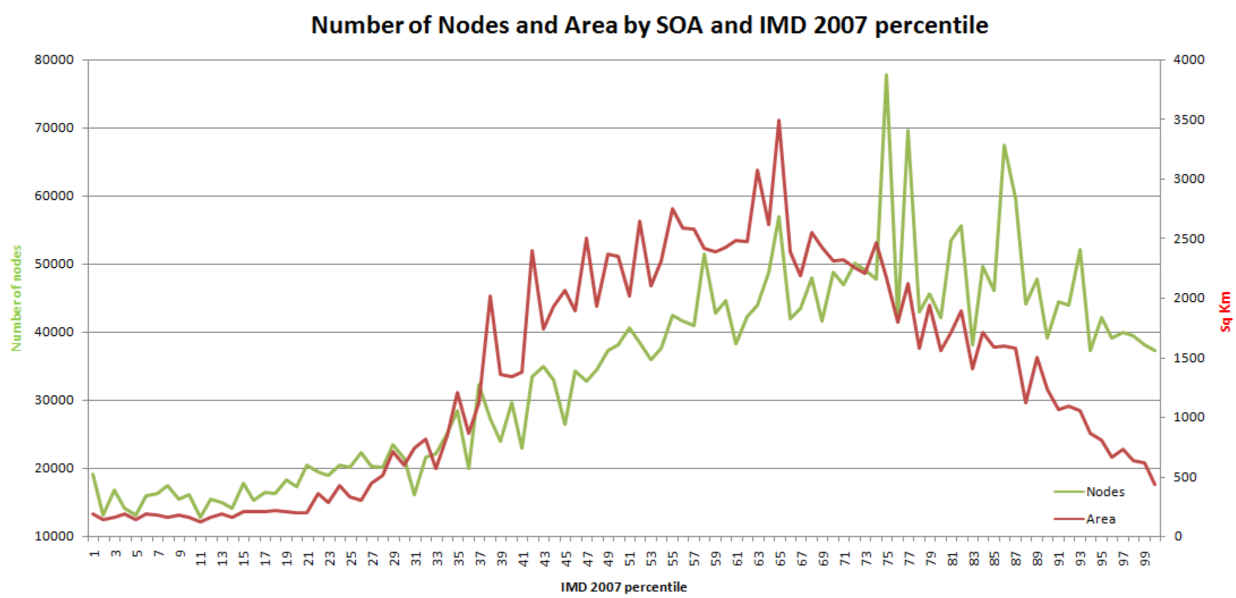

Figure 16 – Number of Nodes and Area by SOA and ID 2007 percentile. Notice that the area of each percentile, about 325 SOAs, is summed up in sq km on the right, while the number of nodes is on the left.

However, nodes provide very little indication of what is actually captured. A more accurate analysis of mapping activities is to measure the places where OSM features were collected. This can be carried out in two ways. Firstly, all the roads in the OSM dataset can be compared to all the roads in the Meridian dataset. Secondly, a more detailed scrutiny would include only lines with attributes that make them similar to Meridian features, and would check that the name field is also completed – confirming that a contributor physically visited the area as otherwise they would not be able to provide the street name[3]. This reduces the number of road features included in the comparison to about 40% of the objects in the OSM dataset. The results of this comparison are provided in Figure 17.

[3] Only out-of-copyright maps can be used as an alternative source of street names, but they are not widely used as the source of street name by most contributors.

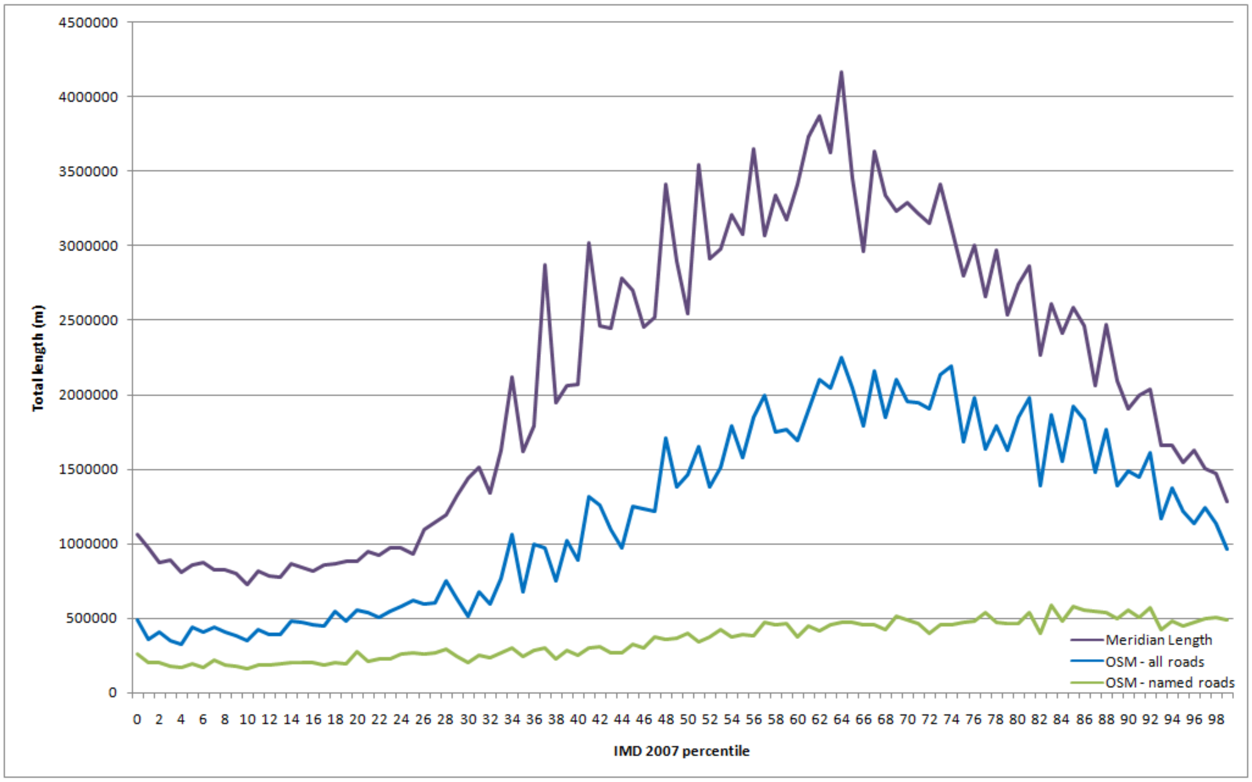

Figure 17 – Length of roads in Meridian (top line), OSM (all roads), and OSM (named roads only) by ID 2007 percentile.

Notice that while the datasets exhibit the same general pattern of distribution as in the case of area and nodes, the bias towards more affluent areas is clearer – especially between places at the top of the scale. As Table 6 demonstrates, at the bottom of the ID 2007 score the coverage in all roads is below the average for all SOAs – even though many of them are large and rural – and in named roads there is a difference of 8% between wealthy areas and poor areas.

Table 6 – Average percentage coverage by length in comparison to Meridian and ID 2007

ID 2007 percentile	All Roads	Named roads
1-10	46.09%	22.52%
91-100	76.59%	30.21%
Overall	57.00%	16.87%

This bias is a cause of concern as it shows that OSM is not an inclusive project, shunning socially marginal places (and thus people). While OSM contributors are assisting in disaster relief and humanitarian aid (Maron, 2007), the evidence from the dataset is that the concept of 'Charity begins at home' has not been adopted yet. This indeed verifies Steve Coast's declaration that 'Nobody wants to do council estates. But apart from those socio-economic barriers – for places people aren't that interested in visiting anyway – nowhere else gets missed' (GISPro, 2007).

Significantly, it is civic society bodies such as charities and voluntary organisations that are currently excluded from the use of the OS dataset due to costs. The areas at the bottom of the Index of Deprivation are those that are most in need in terms of assistance from these bodies, and thus OSM is failing to provide a free alternative to the OS where it is needed.

8. *For what purpose is the OSM dataset fit for?*

The analysis that was carried out here exposed many aspects of the OSM dataset. The most impressive aspect is the speed in which the dataset was collected – within a short period of time, about a third of the area of England was covered by a team of about 150 participants (with minor help from over 1000 others). With the help of Yahoo! imagery, the dataset provides reasonable positional accuracy of about 6 metres, and a good overlap of up to 90% of OS features for motorways. In places where the participant was diligent and committed, the information quality can be very good.

In terms of coverage, the centres of major cities in the UK are well mapped and covered, though, as you move from the centres to the edges, the quality of coverage deteriorates rapidly.

In the places where OSM information is complete and with full attribute information – an area of about 20% of England – it is estimated that OSM quality is such that it can be a replacement for Meridian; so the functions that are mentioned regarding Meridian uses are all relevant for OSM.

However, currently **OSM is a long way from being a viable replacement for a dataset like Meridian**, even though the latter is generalised and diluted.

The main issue with OSM is the inconsistency of the information in terms of its quality. Differences in digitisation – from a fairly sloppy approach in the area of Highgate to a consistent and careful approach in South Norwood – seem to be part of the price that is paid for having a loosely organised group of participants. The real problem that prevents OSM from being fit for use is that of completeness, not only as you move out of city centres (for example, the poorly covered Sutton just 12 km from very well-mapped Central London) but also within areas that look properly covered at first glance. While Meridian has some consistency in the criteria of removal of minor roads, with OSM this is largely an unknown element. As the case of Highgate demonstrates, these omissions can include a whole set of streets in a locality.

Furthermore, the fact that most of the area was covered by a single contributor means that very little quality assurance was carried out, if at all. The impact of this is variable positional accuracy, with up to 20-metre positional errors. The information is not covering socially marginal areas at the same quality that it covers affluent areas.

The result of the analysis means that, today, OSM is suitable for cartographic products that display central areas of cities (as has already been done in Cambridge and Oxford), but not yet for more sophisticated GIS analysis where detailed and more complete information is needed.

The comparison with OSM also helps to clarify why a price tag of about £1300 for a licence for Meridian is reasonable in terms of quality. Therefore, for practical GIS application, and although this will create a complex situation in terms of licence conditions, a current best solution seems to be to get hold of the two datasets and augment the fine details that are missing in Meridian with the OSM dataset.

This information should not come as a surprise to OSM participants. As OSM declares very clearly '**The map isn't finished yet**' (OSM, 2008). The project started in 2004 and, if it continues with its current growth, it is possible to see that England will be completed within a few years. When it reaches higher coverage of urban areas, it will be similar to Meridian and will provide suitable details at this level.

The comparison with the 1:10,000 raster helps to clarify the position that OSM fills in the marketplace of geographical information products. The comparison shows the advantages of deriving a generalised product from a highly accurate geographical dataset, and the richness of details that are provided is clear. OSM is more comparable to Meridian or the datasets that are provided by TeleAtlas and Navteq in terms of its offering – although all the other products are currently offering much better positional accuracy.

Significantly, there is clear attention within OSM community to the issue of quality of the information (e.g. Cherdlu, 2007). The introduction in June 2008 of OpenStreetBugs – a simple tool that allows rapid annotation of errors and problems in the OSM dataset – is also very welcomed and provides an easy way for non-participants to flag errors.

Yet, the analysis of the data collection patterns by users and across the country is a cause for concern. While clearly the number of users is rising steadily, out of the 1100 users that participated in the data too many have contributed little. As the project continues to evolve, there is a certain amount of learning that new users are required to put in; even at the most basic level, there are now dozens of tags that need to be read through and understood. This creates a growing obstacle for participation. Moreover, the hundreds of users that contributed very little are a missed opportunity for the project, as more contributors would have helped in completing the coverage as well as improving quality.

As noted in OSM's first State of the Map conference (Haklay, 2007), there are usability issues with OSM software and these problems partially explain the high number of users that found the process too difficult. This is a chasm that OSM must cross to become popular and accessible to a wider range of users.

An alternative interpretation of the development pattern of OSM suggests that due the nature of geographical information, the project should rely on a small group of committed contributors, with a light support from a wider group of participants that will use tools like OpenStreetBugs to guide the core group to places that require coverage or where errors have been identified. Using this model, it can be envisage that with a group of 200 or 300 committed contributors, the UK can be covered within two or three years, with quality improvement increasing steadily over this period through light contributions of many other occasional contributors.

It is also possible to suggest technical and organisational solutions that will help to improve information quality. Technically, it is possible to develop software that will use the ability of mobile phones to find their locations as a way to check the quality of the OSM dataset automatically and flag major errors (say over 10 metres). Organisationally, it might be necessary to set regional editors or find committed participants that are willing to help in improvement of information quality. The OSM community has proven to be resourceful so far, and it will be interesting to see how it turns quality assurance into an enjoyable and engaging process.

Finally, the analysis that is presented here was carried out with a generalised dataset, and the next step is to compare the quality of OSM with Navteq or TeleAtlas datasets, which are similar in function, and with OS MasterMap to evaluate the positional accuracy of the datasets. There is also a need to research methods that can assist VGI projects to ensure quality and to integrate it with data collection tools in a way that doesn't diminish the enjoyment from participating in the project.

Acknowledgements

Many thanks to Patrick Weber, Claire Ellul, and especially Naureen Zulfiqar who carried out part of the analysis of motorways as part of her M.Eng. project. Some details on geographical information quality are based on a report on the Agency Performance Monitor that was prepared for the OS in 2005. Some early support for OSM research was received from the RGS Small Research Grants programme in 2005. Apologies to the OSM system team for overloading the API server during this research.

Thanks to the Ordnance Survey External Research and University Liaison who provided the Meridian 2 dataset for this study in January 2008. The 1:10,000 raster and the SOA boundaries were provided under EDINA/Digimap and EDINA/UKBorders agreements. ID 2007 was downloaded from DCLG and is Crown copyright.

References

Cherdlu, E. (2007) *OSM and the art of bicycle maintenance*, Presented at State of the Map 2007, 14-15 July, Manchester, UK, http://www.slideshare.net/nickb/openstreetmap-and-the-art-of-bicycle-maintenance-state-of-the-map-2007/

Friedman, Thomas L. (2006) *The world is flat: A brief history of the twenty-first century, updated and expanded edition*. New York: Farrar, Straus and Giroux.

GISPro (2007) *The GISPro Interview with OSM founder Steve Coast*, GIS professional, Issue 18, October 2007, pp. 20-23.

Goodchild M.F. (2007a) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0, *International Journal of Spatial Data Infrastructures Research*, 2007, Vol. 2, 24-32.

Goodchild M.F. (2007b) Citizens as sensors: the world of volunteered geography. *GeoJournal*, Vol. 69, No. 4. (2007), pp. 211-221.

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.

Goodchild, M.F., Haining, R. and Wise, S. (1992) Integrating GIS and spatial data analysis: problems and possibilities, *International Journal of Geographical Information Science*, 6(5): 407-423.

Haklay, (2007) OSM and the Public – what barriers need to be crossed? Presented at State of the Map 2007, 14-15 July, Manchester, UK, http://www.slideshare.net/mukih/usability-engineering-for-osm-sotm-2007/

Haklay, M. and Weber, P. (2008) OpenStreetMap – User Generated Street Map, *IEEE Pervasive Computing.*

Haklay, M., Singleton, A. and Parker, C. (2008) Web mapping 2.0: the Neogeography of the Geoweb, *Geography Compass.*

Howe, J. (2006) The Rise of Crowdsourcing, *Wired Magazine,* June 2006.

Hunter, G.J. (1999) New tools for handling spatial data quality: moving from academic concepts to practical reality, *URISA Journal*, 11(2): 25-34.

Kresse, W. And Fadaie, K. (2004) *ISO Standards for Geographic Information*, Springer.

Maron, M. (2007) *OpenStreetMap: A Disaster Waiting to Happen*, Presented at State of the Art 2007, 14-15 July, Manchester, UK, http://www.slideshare.net/mikel_maron/openstreetmap-a-disaster-waiting-to-happen

Oort, P.A.J. van (2006) Spatial data quality: from description to application, PhD Thesis, Wageningen : Wageningen Universiteit, p. 125.

Ordnance Survey (2007) Meridian 2 User Guide and Technical Specification, Version 5.2 February 2007, Ordnance Survey, Southampton.

OSM (2008) *Places*, Available http://wiki.openstreetmap.org/index.php/Places [Accessed 28th July 2008]

Perkins, C. and Dodge, M. (2008) The potential of user-generated cartography: a case study of the OpenStreetMap project and Mapchester mapping party, *North West Geography*, 8(1), 19-32.

Raymond, E.S. (2001) The Cathedral and the Baazar, O'Reilly.

Sui, D., Z. (2008) The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems* 32 (2008) pp. 1–5.

Tapscott, D. and Williams, A.D. (2006) Wikinomics – How mass collaboration changes everything. Atlantic Books, London, UK.

Tran, T. (2007) Google Maps Mashups 2.0, Google Lat-Long Blog, posted 11/7/2007 http://google-latlong.blogspot.com/ [Accessed 2 November 2007]

Where 2.0 (2008) Where 2.0 Homepage http://en.oreilly.com/where2008/public/content/home [Accessed 23rd July 2008]