A Quality Analysis of OpenStreetMap Data

M.Eng. Dissertation

Aamer Ather
28 May 2009

GEOM 4021/4026
M.Eng. Geoinformatics
Department of Civil, Environmental & Geomatic Engineering
University College London

Abstract

The past few years have seen an influx of open source websites due to advances in broadband communication and the advent of Web 2.0. An example of one such project in the geo-domain is OpenStreetMap, a free editable map of the world created by volunteers. However, as with all open source projects questions are asked about the quality and validity of the data provided by OpenStreetMap. In response to this, initial research has been carried in this area and in particular, one study looked at the positional accuracy of OpenStreetMap data through comparison with the Ordinance Survey's Meridian 2 dataset. This was performed on a sample of motorway segments.

This dissertation explains further research carried out by the author on OpenStreetMap data quality, in order to provide a more detailed insight as to how good the data is and help determine its fitness for use. This project has built upon the original positional accuracy study by extending the same analysis to A-roads and B-roads whilst also using a higher level dataset for comparison: OS MasterMap. Further map quality tests were also conducted in terms of a completeness study of road name attribution, and an analysis of number of users per area.

The results of this analysis found the positional accuracy of OpenStreetMap data to be very good in comparison to OS MasterMap, with over 80% overlap between most the road objects tested between the two datasets. The results also found there to be a positive correlation between road name attribute completeness and number of users per area.

Acknowledgements

I would like to thank my project supervisor, Muki Haklay, for all the help and support he has given me throughout this project, and for providing numerous initial datasets to help me get started with the research.

I would also like to thank Naureen Zulfiqar, whose buffer analysis program was used carry out a large portion of this research, and Liz Jones, who provided me with one of the department's GPS receivers to facilitate my own attempts at OpenStreetMap mapping.

Finally, I would also like to acknowledge the various websites through which the datasets were obtained, namely the Ordinance Survey, Edina, CloudMade and of course OpenStreetMap.

Contents

List of Figures

List of Tables

1. Introduction

In the past, geographic technologies and geo-data had been a closed shop affair with near-monopolies dominating the geo domain, making it an expensive and exclusive business. However, recent years have seen this field open up with increased availability of geographic data through the likes of Google and Microsoft, including geo-data that is genuinely free and open through projects like OpenStreetMap (O'Reilly Where 2.0, 2009). Even the Ordnance Survey are now making its own data freely available through its OS OpenSpace API portal in response to this. The Ordnance Survey have long been perceived as having a monopoly on UK geographic data, although this has now been undermined by initiatives like OpenStreetMap (Personal Computer World, 2009).

The OpenStreetMap (OSM) project is an online open source editable map of the world, created by volunteers worldwide through the collection and contribution of geographic data. Anyone can contribute to the map by simply registering on the OSM website. Why is OSM data useful? Firstly, the data is completely free with an open content licence. Secondly, it is current as it constantly being updated by the subscribed users who can also add points of interest import to them. Finally, OSM has the potential to establish volunteers from all over world including less developed regions, where obtaining data in such places can be difficult for most commercial mapping companies.

However as with other open source projects like Wikipedia, questions are asked about the quality and validity of the information delivered by OSM. Hence there is a requirement for an understanding of how good OSM data actually is in order to identify areas for improvement and determine how fit for use the map currently is. In response to this requirement, this project aims to carry out a quality evaluation of OSM data by primarily focusing on the positional accuracy of the dataset.

Previous work on OSM map quality has been carried by out Dr Mordechai (Muki) Haklay, whose research involved an initial assessment of positional accuracy and completeness through numerous quality tests. The positional accuracy was assessed through the comparison of motorway segments between OSM and Ordnance Survey (OS) Meridian 2 datasets, although it was identified that further research was still required here. Hence the primary aim of this project is to build upon the initial positional accuracy assessment, by extending this analysis to A-road and B-roads whilst also comparing OSM data to a higher quality dataset than OS Meridian 2, namely OS MasterMap data. The secondary aim of this project is to carry out further forms of map quality assessment based on datasets chosen for the main analysis. This will involve an assessment of attribute accuracy and completeness, and an analysis of the numbers of users per area. It is hoped that the results of this project will shed further light on the quality of OSM and ultimately provide an insight into the accuracy of the data at a more local level, due to the use of a more accurate dataset for comparison and a greater number of roads sampled.

The first part of this report is a literature review where methods of assessing map quality data are outlined followed by an assessment of Volunteered Geographic Information (VGI); a relatively new area of research, which OSM can be categorised under. The literature review then concludes with a summary of the research carried out in this field. The next section of this report is a description of the OSM and OS data used, which is then followed by outline of the methodology employed for the analysis. The next section then details the results and conclusions of the initial pilot test carried out, after which the procedure of the main quality analysis is explained. The results are then shown followed by discussion and analysis of them. Finally, the report ends with the conclusions where suggestions for future research directions are given.

2. Literature review

2.1 Assessing map data quality

Data quality can be defined as fitness for purpose, or how suitable some data is in satisfying particular needs or fulfilling certain requirements to solve a problem (Coote & Rackham, 2008). Quality is a major concern as it determines the limits of use for any dataset, and it is key in putting GIS products into an understandable form. (Paradis & Beard, 1994). As identified by Van Oort (2006), spatial data quality has been an increasing concern due to two reasons, (1) the emergence of Geographical Information Systems (GIS) in the 1960s and (2) from the 1970s onwards, a strong increase of available spatial data from satellites. He also states that the number of users from non-spatial disciplines have grown due to the large-scale adoption of GIS. This is certainly the case for Volunteered Geographical Information (VGI) and neogeography applications, as explained in the next section of this literature review (section 2.2).
The quality of geographic data can be assessed against both subjective and quantitative quality elements. Based on the ISO standards for the quality principles of Geographic information[1], Coote and Rackham (2008) outline how both these quality elements can be assessed:

Subjective elements provide a valuable initial indication as to how useful a particular data is going to be for certain purposes. They usually fall under three headings:

- *Purpose* – the rational for creating the dataset
- *Usage* – the application to which the dataset has been put
- *Lineage* – the history of the dataset

Quantitative elements imply a quality evaluation involving measurement and an objective result. They are categorised as follows:

- <u>*Positional accuracy*</u> – the accuracy of the position of features or geographic objects in either two or three dimensions. Positional accuracy can be expressed either as the *absolute accuracy;* the closeness of coordinate values to values accepted as true, *relative accuracy;* closeness of the relative positions of objects in a dataset to those relative positions accepted as true, or *gridded data position accuracy;* the closeness of gridded data position values to those accepted as being true.

[1] ISO 19113:2003 Geographic information – Quality principles

- *Temporal accuracy* – This is the accuracy of temporal attributes, such as dates and time, and the temporal relationships of features, such as 'later' or 'earlier than' relationships. Temporal accuracy can be expressed as the *accuracy of time measurement*; i.e. if the stated recorded dates of objects are correct, *temporal consistency*; the correctness of ordered events, or *temporal validity;* the validity of data with respect to time.
- *Thematic accuracy* – This is the accuracy of *quantitative attributes;* such as population, *non-quantitative attributes;* such as geographic names, and *classifications*; how correct classes assigned to attributes are in relation to ground truth.
- *Completeness* – This is the presence and absence of objects in a dataset at a particular point in time. These can be errors of *omission*; data missing from the dataset which should have been included at the time of capture (such as missing streets or street names) or *commission;* Data that is present in the dataset but should have been omitted (such as buildings now demolished).
- *Logical consistency* – This is the level of adherence to logical rules of data structure, attribution and relationships. This can be characterised as *conceptual consistency*, *domain consistency*, *format consistency* and *topological consistency*.

As this project is concerned with the assessment of OSM map quality in particular, techniques that have been directly used in assessing some of these quality elements for OSM are explained section 2.3 of this literature review.

Once the quality of map data has been assessed it then needs to be communicated to the user. Spatial data quality is usually implicitly implied in mapping and traditionally the implicit measures of quality, transferred from surveyor to the cartographer, were understood by experts. However, the nature of digital data requires an explicit approach in communicating the overall quality of map data, hence the expertise and knowledge of the surveyor, cartographer or geographer needs to be passed on to the GIS user (Coote & Rackham, 2008). In the wake of VGI this extends to any user of such information, who are most likely untrained in such fields. Much research has taken place in development of methods to communicate map data quality and uncertainty in the domain of digital 0map data; Coote & Rackham (2008) give an example of this by use of a data quality filter, and Hunter and Goodchild (1996) in their paper "Communicating uncertainty in spatial databases" discuss various methods of doing so. Also, different applications place different levels of importance on data quality elements and different approaches to visualising the uncertainty of each of these may need to be adopted (Hunter & Goodchild, 1996). However, discussion of the various techniques used for the communication and visualisation of spatial data quality goes beyond the scope of this project.

Another factor to be considered once an assessment of data quality has been carried out is assessing fitness for use. As mentioned at the beginning of this chapter, quality can be defined as fitness for purpose. Van Oort (2006) outlines three steps in how this can be achieved:

1. *To search for a spatial dataset that contains the information needed for the intended application;*
2. *To explore whether there are legal or financial constraints to access or particular use of the spatial data;*
3. *Finding out if, given the spatial data quality, risks are acceptable.*

Whilst these steps may be more applicable to end users of the datasets, data providers may want to look at these steps 'inversely' in order to "sell" their data; i.e. to identify what applications the data can be used for, convey what the legal or financial constraints of using the dataset are, and convey what possible risks there could be for each of the applications identified (maybe by testing the data for each of these applications).

The spatial data quality factors dealt with in this section of the literature review are applicable to all forms of geographic information. The next section deals with a more specific and very recent form of geographic information: Volunteered Geographic Information.

2.2 Volunteered Geographic Information and OpenStreetMap

Volunteered Geographic Information (VGI) is a term used to describe the voluntary act of creating geographic information, often by users who are largely untrained with little or no formal qualifications in the creation of such data. This is sometimes also referred to as "Neogeography", which can be defined as combining the complex techniques of GIS and cartography and placing them within the reach of users and developers, and is essentially about people using and creating their own maps, by combining elements of an existing toolset (Turner 2006). This has also been described as the 'wikification' of GIS where the masses are quietly being transformed from being passive consumers to active producers of geospatial information (Sui, 2008).

The term "VGI" was labelled and defined by Goodchild (2007) as a special case of the more general Web 2.0 phenomena of user-generated content, and is a relatively new area of research. In his paper "Citizens as sensors: the world of volunteered geography" Goodchild indentifies the technologies that have made VGI possible, namely *Web 2.0;* the two-way interaction between users and webpages, *Georeferencing;* the ability of users to specify locations on the Earth's surface, *Geotags;* standardised code that can be inserted into information to specify geographic location, *GPS;* enabling users to identify their own position and movements, *Graphics;* the ability of current household computers to support the dynamic visualisation of 3D objects, and *Broadband communication;* wide spread, high speed connection to the internet.

Goodchild (2007) then states how VGI fits into the idea of using spatial data infrastructures (such as the National Spatial Data Infrastructure [NDSI] in the US) and patchworks, where (p.217) "A collection of individuals acting independently, and responding the needs of the local communities, can together create a patchwork coverage", and that this can be achieved using a central server with the appropriate tools. This leads on to the notion of using humans as sensors, where there is a global network of 6 billion humans each the ability to interoperate local information. So what motivates a user to voluntarily contribute geographic information, or other user-generated content, to open source websites? Goodchild states that this could be due to self-promotion but also the personal satisfaction one might gain from seeing their own contribution as part of a patchwork. Another reason may simply due to the fact that neogeography is fun, as stated by Turner (2006).
Over the past few years there has been rapid growth in the number of sites implementing VGI in a Web 2.0 framework. Examples (as shown in Figure 1) include WikiMapia: an online editable map that allows users to describe any place on earth (with over 9,997,000 places marked at the time of

writing), FixMyStreet: a site which allows users to locate and report local problems like graffiti and broken pavement slabs, and OpenStreetMap, the focus of this thesis.

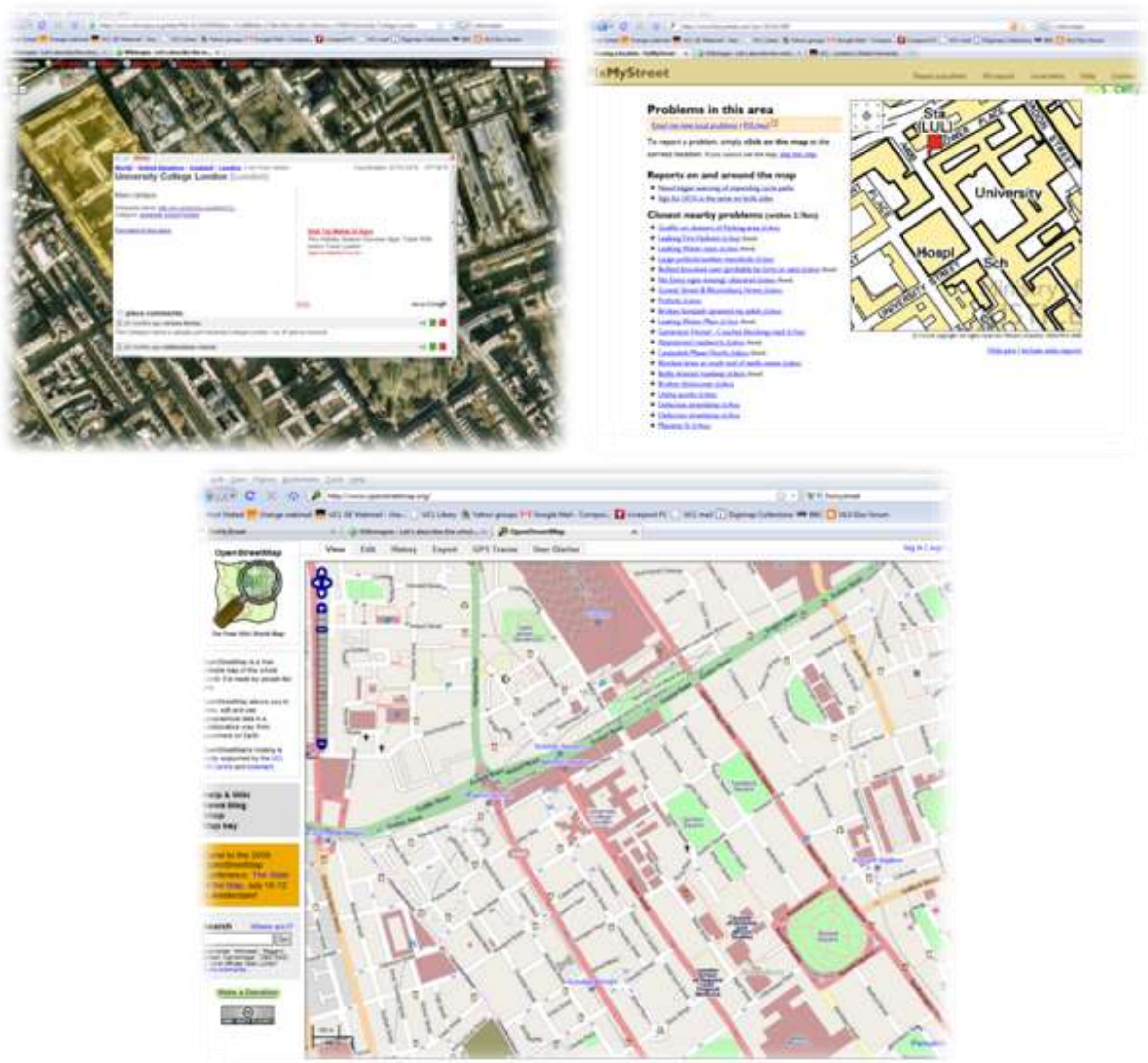


Figure 1 - WikiMapia (top-left), FixMyStreet (top-right), OpenStreetMap (bottom). These all show same region (UCL and the Bloomsbury area) but exemplify different uses of VGI.

OpenStreetMap (OSM) is an open source and open content licence mapping project where any user around the world can contribute to help create a global street map of the world either by collecting their own GPS data tracks, digitally tracing aerial images from Yahoo! Imagery, or by obtaining data from other free sources. One of the key motivations behind OSM is to provide free access to current digital geographic information, often considered to be expensive from other sources (Haklay & Weber, 2008). Users can actively add data to the map using either a light-weight online Flash-based editor, Potlatch, or using the Java OpenStreetMap Editor (JOSM), a more advanced offline editing suite which provides more functionalities than Potlatch. The main server infrastructure holding the data is hosted by UCL (Haklay & Weber, 2008). Hence OSM is clear example of the VGI concepts introduced by Goodchild (2007) in that the OSM project provides a central server (at UCL) with the appropriate tools (Potlatch and JOSM), and relies upon a network of humans as sensors and as interpreters of local information, to collect and contribute various pieces of the OSM patchwork. Using humans as sensors in this context can also be considered a 'crowdsourcing' activity, which defined by Haklay and Weber (2008, p.13) "is how large groups of users can perform functions which are either difficult to automate or expensive to implement".

As of March 2009, OSM surpassed 100,000 registered users with over 750 million Track points uploaded as shown in Figure 2. This illustrates the increasing popularity of VGI. OSM data is increasingly being used for a number of purposes and is available through a number of website in different formats. For example, the data is currently being used to highlight cycle routes (www.opencyclemap.org), routing and navigation facilities (openrouteservice.org), and to map out ski lifts and features of ski resorts (openpistemap.org) (Wikipedia, 2009a).

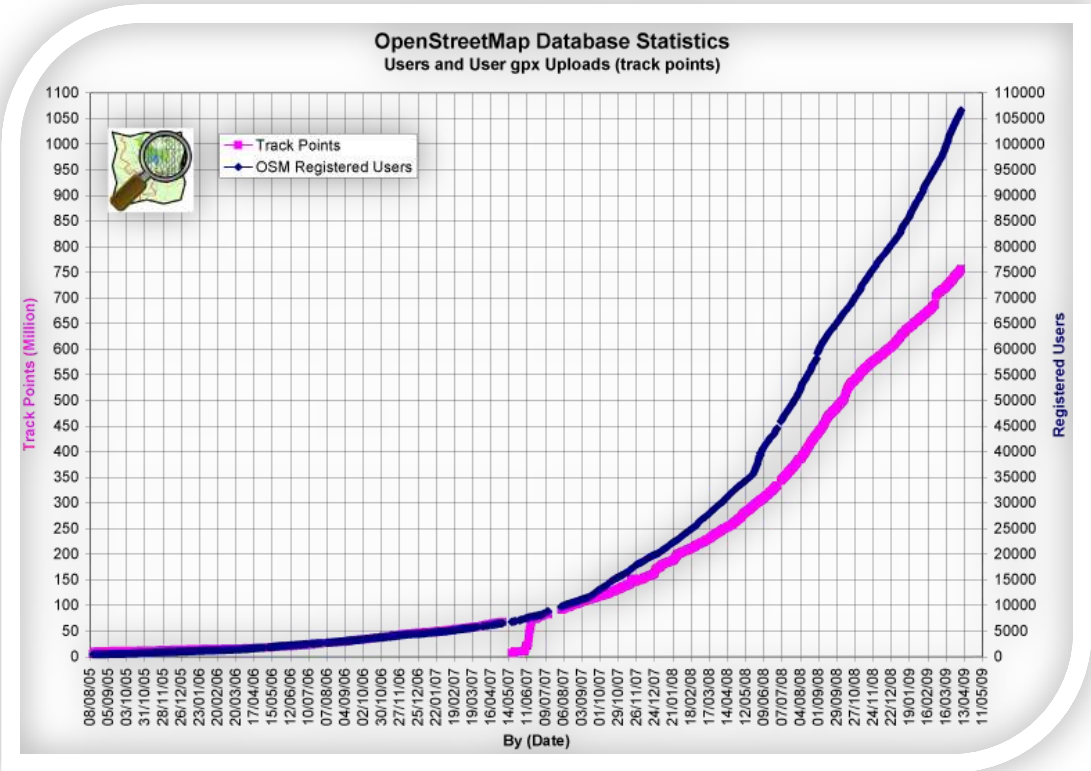


Figure 2 - Monthly growth of users and track points (source: http://wiki.openstreetmap.org/wiki/Stats)

The rise of VGI and the availability of user-generated geographic data has clear benefits. Data from websites like OSM can be obtained for free and is often the cheapest source of geographic information, and in areas where access to geographic information may be regarded as an issue of national security, data from OSM may sometimes be the only source (Goodchild, 2007). There is also greater availability of geographic and environmental knowledge in the form of place-based data, images and other geographically relevant information. Moreover, the idea of crowdsourcing has exploited the clear advantage of utilising individuals who are in many cases the in the best position to provide information that requires indigenous experience, local expertise, and current information about local conditions (Flanagin & Metzger, 2008).

On the other hand, there are concerns with VGI with regards to its quality, reliability and value as an information resource (Flanagin & Metzger, 2008). Professional mapping agencies, like the Ordenance Survey, have elaborate standards and specifications to govern the production of geographic information and employ highly qualified cartographers, and have developed a reputation of quality and authority (Goodchild, 2007). Hence the credibility of user-generated geographic information comes into question. The concept of credibility is complex and revolves around the believability of a piece of information based on its levels of trustworthiness or expertise. The definitions of credibility also vary between different fields. In the scientific community credibility is considered as the degree to which information can be considered to be accurate, whereas in the fields of communication and social psychology credibility is considered more as a perceptual variable, as the same piece of information may be judged differently by different people. Hence, credibility as perception highlights the notion of believability and is dependent upon trust and expertise (Flanagin & Metzger, 2008).

In the context of VGI, Flanagin and Metzger (2008, pp.141-142) explain that many forms of VGI are more about which information, opinion or perspective people believe rather than scientific data accuracy; “while credibility-as-accuracy is an appropriate concept for those who have a “factual” relationship with geospatial information, credibility-as-perception is more useful for those who use VGI for social, communal, or political purposes”.

This concept of credibility-as-perception may be true for VGI applications like WikiMapia and FixMyStreet, where opinions and descriptions about physical locations can be purely subjective, and hence assessment of credibility in such situations may be assessed upon the reputation of users; i.e. how believable the information provided by a particular user is. However, OpenStreetMap is a fact-based project as it deals with information which is purely objective. Therefore unlike other VGI applications, credibility-as-accuracy is more appropriate for an assessment of how good the information is, and ultimately how credible OSM is perceived to be. Flanagin and Metzger (2008) identify that researchers have barely began to examine the credibility of VGI and that pressing questions include whether users and professionals will accept systems populated largely by volunteered input as credible, and if they did, what purposes would they use it for and with what effects? With regards to OSM, this question of acceptability could be simply be answered by providing some measure of accuracy of the data. After all, some idea of OSM data quality is important in order to determine the fitness of use for various applications (Haklay & Weber, 2008). Steps for assessing the fitness for use were outlined previously in section 2.1, but before this can even taken place an assessment of OSM data quality is required. The next part of this literature review looks at previous work carried out in this field.

2.3 Previous research of OpenStreetMap quality

Dr Mordechai (Muki) Haklay (2008) in his paper "How good is OpenStreetMap information?" identified that (prior to his paper) there has not been any systematic analysis of the quality of VGI. His research aimed to fill this gap by performing various forms of quality analysis of OpenStreetMap information by comparing it to Ordnance Survey vector datasets (OS Meridian 2). His research addressed many of the Map data quality elements mentioned in section 2.1 of this literature review; analysis was carried out on the positional accuracy, completeness, number of users per area and the equality in which the data was collected. The analysis techniques used and the main findings of his research shall now be summarised, including areas that were identified as requiring further research.

Research on the positional accuracy of OSM comprised of two parts, a statistical comparison and a visual comparison. The statistical comparison was based upon analysis carried by Naureen Zulfiqar (2008) as part of her M.Eng. project, where the positional accuracy of OSM was compared to OS Meridian 2 using a sample of matching motorways segments in and around London. OS Meridian 2 was used because the nodes of this dataset are derived from the Ordnance Surveys high resolution topographical dataset and are therefore highly accurate. However, a point by point analysis was not carried due to the fact OSM represented motorways as line objects in each direction, whereas Meridian represented them as one line, hence deeming point by point analysis inappropriate. Instead the analysis was based on a buffer technique developed by Goodchild and Hunter (1997) illustrated by Figure 3 below:

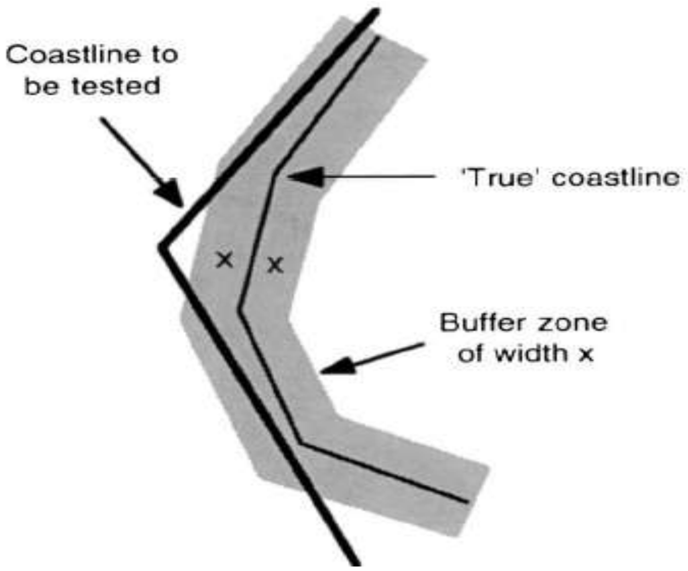


Figure 3 - Goodchild & Hunter buffer comparison technique (source: Goodchild & Hunter 1997)

This buffer technique requires the use of a reference source, illustrated by the ‘True’ coastline in Figure 3, and a tested source, represented by the thicker line. A buffer of width *x* is then created around the reference object and the proportion of the tested source that lies within the buffer is then calculated. Depending on the size of the buffer chosen, the level of accuracy of the tested source can then be determined.

For Naureen’s project, the Meridian dataset was used as the reference source and OSM data as the tested source. Analysis between the two datasets was then carried out using a program she created in MapBasic, which allowed a user to assign buffers to both datasets. The buffer around the Meridian data was set to 20 metres, due to the fact that this was size of the filter used by the OS in the creation of the meridian line object, and a 1 meter buffer was applied to OSM dataset in order to calculate the overlap. The results showed an average overlap of around 80% with a variability between 60%-89%, hence indicating that OSM dataset provided a good representation of motorways. However, she identified that there were limitations as to what information could be derived this analysis as it was only carried out for motorways. It was therefore suggested this same analysis needs to be carried out using OS MasterMap as the reference source with the buffer sizes set to represent the true width of the motorway, and that further analysis should be carried out on A-roads and B-roads.

The next part of Dr Haklay’s research was a visual accuracy assessment carried out across five OS 1:10,000 raster tiles, from which 100 samples were visually inspected. The total average distance from all five tiles was found to be 5.83 metres, with the analysis identifying differences in accuracy and attention to detail between each of the test areas. He concluded that this could be attributed to the digitisation and data collection skills and the patience of the person who carried out the work.

An assessment of completeness was also undertaken by comparison of the OSM and Meridian datasets for the whole of England. This was achieved by splitting the datasets to a square grid of 1km resolution and adding up the length of lines that were contained in the grid cells or intersected by them. The results showed that, at the time this analysis was carried out, only 29% of England had been captured by OSM users of which approximately 4% were digitised lines without a complete set of attributes.

Also, an assessment of the number of users per areas was carried out. Such type of analysis is a good indication of quality as it highlights the number of people who carried out the work; if more users are involved that there is a greater likelihood of errors being identified and corrected (Haklay, 2008). This relates to the principle of 'Linus's Law', a term coined by Raymond (1999, p.29), who states that "given enough eyeballs, all bugs are shallow". This principle is true for all Open source applications; a study of how users determine the credibility of user-volunteered information, by Flanagin and Metzger (2008, p.144), found that most people "agreed that more testimonials, or more authors or contributors to a Wikipedia article, would produce less biased and, thus more credible information". The results Dr Haklay's user analysis showed that 80% of data captured was carried out by 90 participants, with hundreds of users only contributing a few nodes. User analysis was also carried out for the five OS tiles used for the visual accuracy assessment, and the results showed that 51.3% of the OSM features within these tiles were captured by one person, and cumulatively 89.5% were captured by 3 users.

Finally, an assessment was made of the equality in which the data was collected, by comparison with the UK government's Index of Deprivation 2007 (ID 2007). The results found that there was a bias of data capture towards more affluent areas, shunning socially marginal places.

One of the main issues highlighted in his paper is the inconsistent nature of OSM data in terms of its quality, due to differences in digitisation. The research showed that OSM data could potentially be used for the same functions OS Meridian 2 is used for, although still a long way from being a viable replacement for a dataset like Meridian. Based on the results of his research some suggestions were given for future developments of OSM, including one suggestion of having a small core group of committed contributors with light support from a wider group of participants to identify and guide the core group to regions requiring coverage or where errors have been identified.

2.4 Summary

Section 2.1 of this literature review highlighted some of the factors that are usually considered in assessing map data quality, whilst introducing the issue of communicating map quality and uncertainty. Steps for assessing fitness of use were also identified. Section 2.2 provided an overview of VGI including some of the benefits and challenges it brings, with OpenStreetMap cited as a classic example of VGI with the potential of truly exploiting the concept of using humans as sensors. A gap in the research of VGI map quality was identified. Finally Section 2.3 then provided an overview of VGI research carried out thus far using OSM. It was identified that further research is required in assessing the positional accuracy of the data, by comparing it to a dataset that offers greater positional accuracy than OS Meridian 2; i.e. OS MasterMap. The analysis also needs to be extended to A-roads and B-roads.

In light of this Literature review there is a clear need for an understanding of OSM data quality in order to provide an assessment of the level of accuracy of the current OSM dataset, indentifying where improvements are required and to determine applications for which the data could potentially be used. Whilst an initial quality study has been carried out, this project aims to build on this by comparing OSM data to OS MasterMap data, and to also extend the analysis to A-roads to B-roads within chosen test areas. This project also aims to carry out further map quality tests through analysis of attribute completeness, and test the principle of Linus's law by carrying out a user analysis in the test areas chosen.

3. Data

3.1 OS MasterMap & the Integrated Transport Network (ITN) Layer

The Ordnance Survey's MasterMap is a database that records every fixed feature of Great Britain larger than a few metres in one continuous digital map (Wikipedia, 2009b). OS MasterMap is not a product but rather a framework on which future Ordnance Survey products will be based (EDINA, 2006). It comprises of four separate but complementary layers:

- The Topography Layer,
- Integrated Transport Network (ITN) Layer,
- Address Layer(s) and
- Imagery Layer

Between each layer over 450 million geographic features have been mapped, represented either as points, lines, text and polygons, each with their own unique reference number (a TOID) (Ordnance Survey, 2009a).

The OS MasterMap data used for this project is taken from the ITN layer, which according to the ITN user guide (p.16) "consists of a fully topologically structured link-and-node network representing the road network of Great Britain, from motorways to pedestrianised streets". As well as a Road Network theme, the ITN layer consists of a Road Routing Information (RRI) theme which includes features such as height and width restrictions, traffic calming, turn restrictions and one-way roads (Ordnance Survey, 2009b). The layer includes road classifications, road names, forms of roads, motorway junctions, information potentially relevant to routing and references to the intersecting polygons from OS MasterMap Topography Layer. Figure 4 illustrates an example of ITN layer data.

The ITN layer Road Network was originally sourced from the Ordnance survey ROAD dataset in 2002-2003. The ROAD dataset was created from Land-Line data (at a scale of 1:1250, 1:2500 and 1:10000) where it was available, and in areas where there was no digital coverage the data was captured by digitising centrelines from 1:10000 scale derived graphic sheets. The ROAD dataset was maintained until 2006 to support the OSCAR product range, which was then withdrawn by the OS. It was used to create the ITN road network, which was then subsequently improved by matching to the OS MasterMap Topography Layer. The RRI theme was collected by field surveyors during 2002, and was incorporated into the data (Ordnance Survey, 2009c).

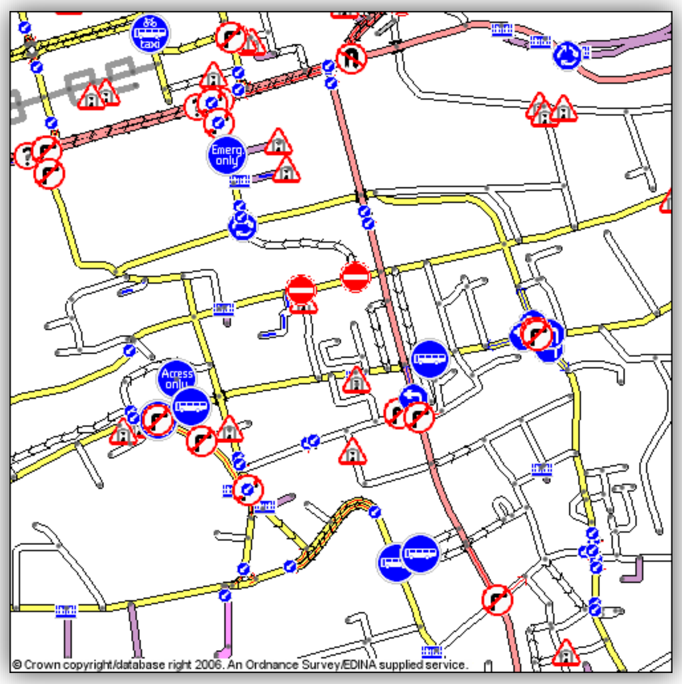


Figure 4 - ITN layer data: Edinburgh (source: EDINA DigiMap help page)

It is difficult to find any quoted figure on the overall accuracy of ITN; the attribution of the OS Topography layer contains an ‘accuracyOfPosition’ field, where for any feature the accuracy of horizontal position in metres is given at the 95% confidence level (EDINA, 2006). Therefore the accuracy of a particular feature in the ITN layer is specified by the attribution of the corresponding feature in the OS MasterMap Topography layer. To this extent, all major urban areas were surveyed to an accuracy of 1 metre, representing urban data capture standards (Ordnance Survey, 2009d). Either way, it is often quoted by the Ordnance Survey that the ITN layer is the most accurate and up-to-date geographic reference for Great Britain’s road structure (any real world changes are updated within 6 months (Ordnance Survey, 2009e)).

The ITN layer has been designed to support a range of different applications, such as transport management systems, road routing, emergency planning and the co-ordination of street works. Table 1 below lists some of the most common applications:

Accident Analysis	Asset recording and inventory management
Catchment-area analysis	Command and control for emergency services
GIS analysis, indexing and mapping	Highway design, planning and engineering
In-vehicle navigation and guidance	Derivation of street gazetteers
Logistics management	Real-time traffic control

Road and highway maintenance	Road-user charging schemes
Route planning and vehicle tracking	Scheduling and delivery
Site location	Traffic management

Table 1 - Common OS ITN applications (source: OS ITN user guide)

Ultimately, the OS MasterMap data is being used for this project because the ITN layer is a very accurate, up-to-date and complete dataset and is used for numerous applications. In contrast to Meridian 2, motorway objects are represented by two line features (one for each direction), the same way OSM represents motorways, and complex junctions are retained whereas they are collapsed to single nodes in Meridian 2. This should make the comparison between the OSM and ITN datasets easier, and should thus lead to more accurate results.

All the OS ITN data used for this project was obtained from the EDINA Digimap service, and was downloaded in GML format.

3.2 OpenStreetMap data

As outlined in the literature review, all the map data used to create OpenStreetMap has been collected and uploaded by users who have registered on the OSM website. The collection is carried out through the use of handheld GPS devices, or traced from Yahoo! Imagery or other free map sources, and digitised using either the online or offline editing tools Potlach and JOSM. To best understand the exact process involved in contributing data to OpenStreetMap, some data collection of my own was carried out. This would also help in gauging level of difficulty involved and also to identify where potential source of error could occur.

The area selected for this data collection was my local area, which was largely incomplete in OSM as illustrated by Figure 5. The GPS device used for the data collection was a THALES MobileMapper CE (shown in Figure 6), describe as having sub-metre accuracy with a minimum of 5 satellites and PDOP less than 4 (Thales, 2005). Kinematic use of the device was tested to have an accuracy of about 2 metres 95% of the time, assuming 100% view of the sky, and 3-5.5 metres with a 75% view of the sky (Starpal, 2005). On the whole, this was a very accurate GPS handset.

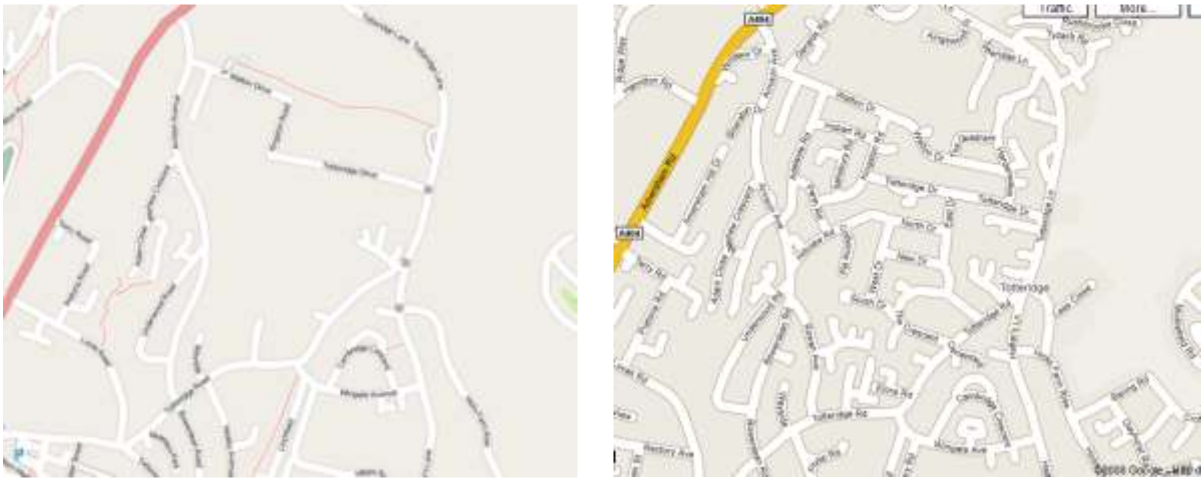


Figure 5 - Comparison of my local area in OSM (left) compared to Goolge Maps (right)

The OSM Wiki page provides detailed instructions for beginners on how to collect and contribute data to the website. The first step was to collect the GPS measurements. GPS trails were collected using Pocket GIS software installed on the handheld MobileMapper. Driving around my local area, the data was collected as GPS point trails. Once all the data had been collected it was then uploaded to a computer in CSV format, which was then converted to GPX format; the standard xml format for exchanging GPS data between applications (TopoGrafix, 2009) and the format required for uploading GPS traces to the OSM website. Once the traces had been uploaded, it was then possible use them as the underlying data for the OSM map. These traces were then digitised using the online editor

Potlach, which requires a user to draw vectors over the GPS tracks whilst forming nodes by clipping the vectors to adjacent and neighbouring roads where required. The corresponding road attribution such as road type (primary, secondary etc) and road name can then be added, however these fields could be left blank if unknown. Other features, like road-a-bouts, can also be added as points. This whole process is illustrated by Figure 6 below:

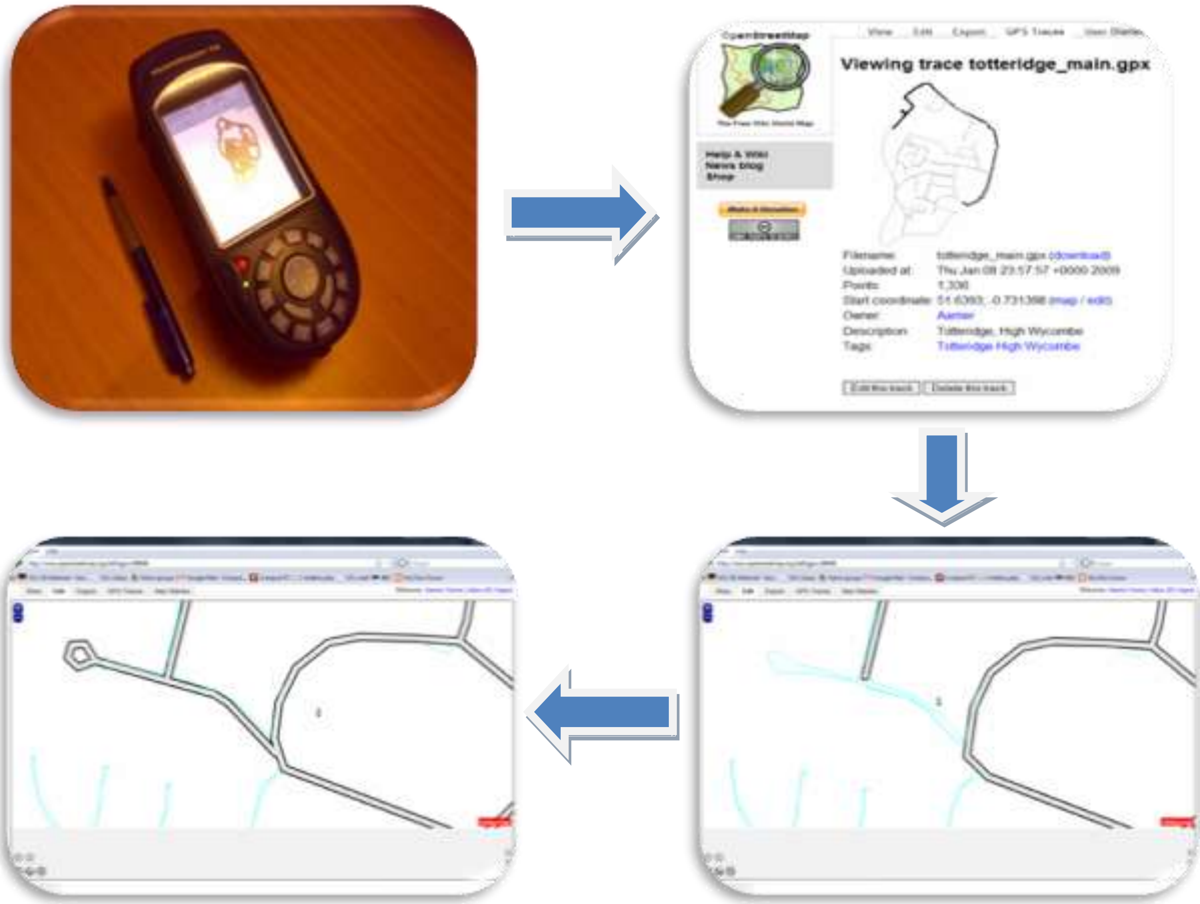


Figure 6 - OSM data collection process: Top left – data collected using the THALES GPS MobileMapper, Top right - GPS traces uploaded to OSM in GPX format, Bottom left & Right – digitising the GPS traces in Potlatch

Once complete, the information is then rendered for display on the main OSM map (the slippy map). The final results, as displayed on the OSM website, are shown by Figure 7 below. A brief visual comparison of this data with OS ITN data of the same region, as shown by Figure 8, shows that the results appear to be fairly accurate with many of the OSM lines (red) falling directly on top the ITN lines (black). However, a closer inspection reveals some discrepancies between some of the road junction positions, as shown by Figure 9, where a difference of 6.6 metres was recorded in this instance.

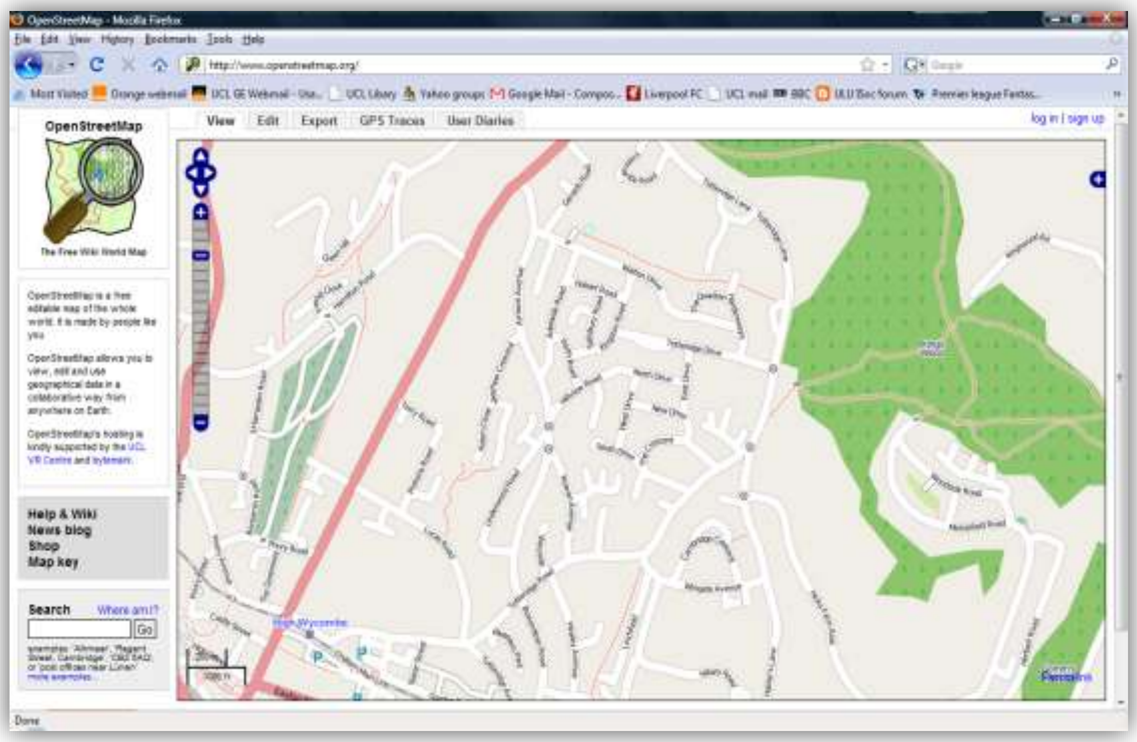


Figure 7 – Final results of OSM data collection on OSM website

Figure 8 - Comparison of OSM data created from GPS tracks (red line) with ITN data of same area (black line)

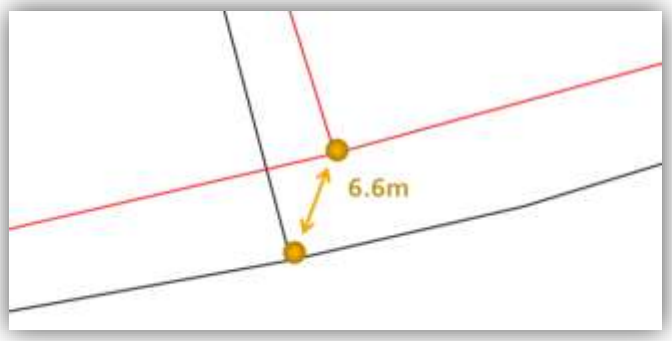


Figure 9 - Discrepancy between two road intersection positions (OSM = red line, ITN = black line)

After carrying out my own OSM data collection, it was clear where potential sources of error in the dataset could occur. Overall, the entire process of collecting and uploading data to OSM was fairly straight forward and should not be difficult for most users. The OSM Wiki site provides a lot of information to support the whole process, such as numerous reviews for different GPS receivers that can be used for data collection, and links for useful software such as GPSBabel - a program for converting data to GPX format. However, there are number of foreseeable problems with the OSM data collection process.

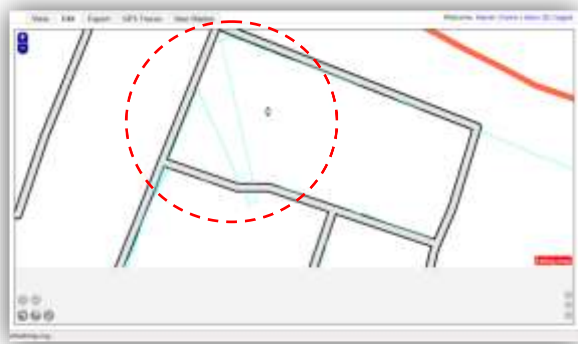


Figure 10 - GPS measurement error

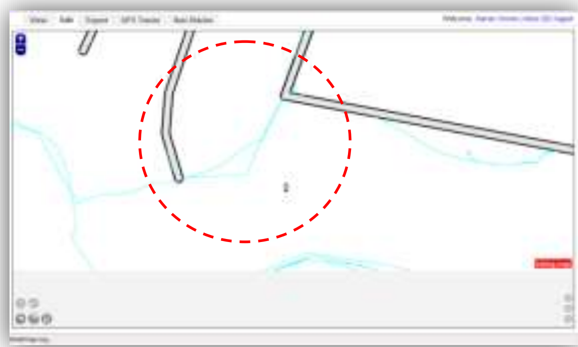


Figure 11 – GPS track interpolation problems

Errors in the data can result from inaccurate measurements made by the GPS device. Such GPS errors could be caused by atmospheric or multipath effects. Figure 10 illustrates one such error from my own data collection which shows a large kink in the GPS track (light blue line). In this instance however, the digitised road line could be easily interpolated from the GPS measurements either side of the kink. This may not be the case in all situations though, and would depend on a user interpolating the correct line whilst digitising. An example of this is shown in figure 11, where two different paths of the same stretch of road were measured; it would be up to the user to decide which line to digitise over, or to create a line somewhere in the middle of both. Hence another source of error can result from human interpolation of results, which also leads to the issue of inconsistency as different users will make different decisions over similar sources of error in their own measurements.

The issue of inconsistent results and differing levels of accuracy in the OSM dataset is also partially due to the fact that different GPS devices are used by each user, and therefore the measurements are of different levels of accuracy depending on the device used. The level of accuracy is also highly dependent upon the nature of data collection. For a user collecting data whilst driving, the number of points collected for any particular road would vary upon the speed of the car; driving slowly allows more GPS points to be measured and would therefore lead to greater accuracy. Therefore, users who collect data by cycling or by walking are more likely to obtain more accurate results as they are travelling slower.

Numerous human errors can also occur during the digitisation process. Users may accidently make changes to the map whilst digitising their own points, for example by moving road points created by other users. This may certainly be the case for first time users using Potlatch, who may be experimenting at first, although a practice option is provided where no edits to the map are saved. Another problem is the ease of sabotage, as once some registers to OSM they are free to make any changes they like to the OSM map. This may particular be the case for boundary disputes.

OSM data can be viewed and obtained from the slippy map on the main webpage. Some metadata is provided as the edit history of all the information can be viewed. Data from the slippy map can be selected and downloaded in XML format or in various image formats. Other websites also provide OSM information in other data formats. The majority of the OSM data used for this project was obtained from the CloudMade website (cloudmade.com) in shapefile (.shp) format.

4. Methodology

As already mentioned, the main aim of this project is to assess the positional accuracy of OpenStreetMap data by conducting Goodchild and Hunter's buffer analysis procedure on sample ITN datasets. Before the main analysis was carried out an initial poilot test was conducted, explained in section 5. This section outlines the main procedure carried out for both the poilot test and the main analysis, although the specific methodolgy pertaining to both tests shall be elaberted on further in sections 5 and 6 respectivaly. Figure 12 below outlines this procdure:

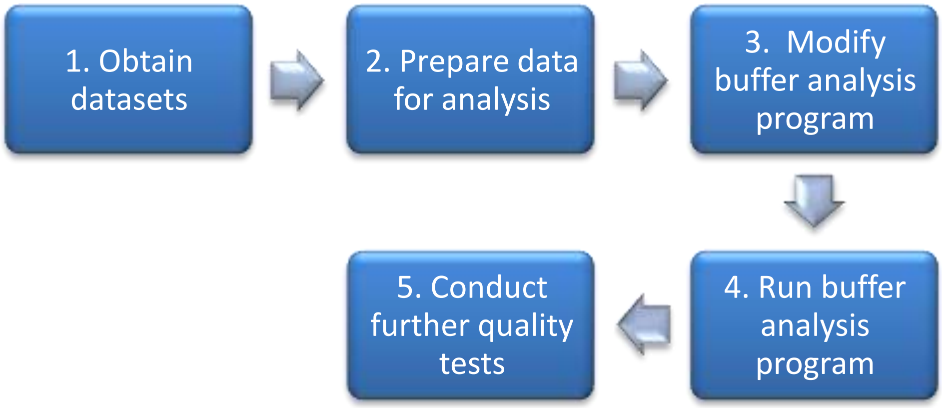


Figure 12 – Method outline

1. **Obtain datasets**

   The OSM and ITN datasets were obtained from CloudMade and EDINA DigiMap respectively. The area chosen for the test analysis was the M4 Heathrow Spur and neighbouring roads within this region, and areas chosen for the main analysis were four 25km^2 regions in London. The reasoning behind the choice of these areas shall be explained in sections 5 and 6.

2. **Prepare data for analysis**

   For both the pilot test and main analysis, the A-roads, B-roads and Motorways had to be extracted from the ITN and OSM datasets, as the tests were only to be conducted on these roads. Also, both datasets had to be edited so that they matched in coverage and in road name attribution, so that feature comparison could be carried out between the two datasets using the buffer analysis program. A separate text file also had to be created listing all the road names of the road features to be compared by the program.

3. Modify Buffer analysis program

The program used to carry out the buffer analysis was developed by Naureen Zulfiqar for her MEng project, and was written in MapBasic. This program was originally designed for OSM comparison with OS Meridian 2 datasets, so for this project the program therefore had to be adapted to work with OS ITN datasets instead. The final program can be found in the appendix. Also before each analysis, the directories of the relevant datasets had to be set in the coding, as well as the buffer values to be used for each dataset. Therefore, for each of the three road types to be analysed, the buffer widths had to be chosen. As mentioned in the literature review, it was concluded from Meridian motorway analysis that the buffers around the OS data should represent the true widths of the roads as this would provide more meaningful results. With this in mind, the following buffer widths were chosen:

Road Type	Buffer size
Motorways	8 metres
A-roads	5.6 metres
B-roads	3.75 metres

Table 2 - Buffer widths

Using the assumption that each motorway lane is about 3 metres wide, the width of the motorway was estimated to be 16m (with four lanes, a hard shoulder and 1 metre for the central reserve). Hence a buffer size of 8m would be appropriate. These dimensions are illustrated in Figure 13. For A-roads, an estimated road width of 11.2 metres was used hence resulting in a buffer width of 5.6 metres, and for B-roads, an estimated road width of 7.5metres resulting in a buffer width of 3.75 metres. The road width estimations were taken from the House of Commons Hansard written answers (Parliament, 2003). The exact size of the buffer used for the OSM data road objects varied between the pilot test and main analysis, and shall be discussed further in section 5.

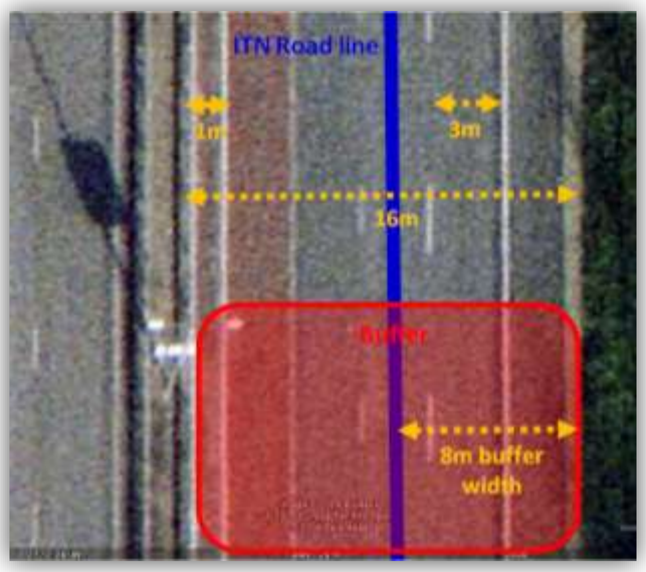


Figure 13 - Estimation of motorway width with buffer

4. **Run buffer analysis program**

   With the datasets prepared and the program adapted with the required buffer parameters set, the program could then be run. All the processing was carried out in MapInfo. The program would firstly read the text file containing the list of road objects to be compared. It would then find the corresponding road objects from both the OSM and ITN datasets and create buffers around them both at the specified sizes. It would then save these buffer objects in a separate table, the output parameters of which included an ID number, road type, road name and length of road.

   Once all the roads in the datasets were buffered, the program would them begin the feature comparison by selecting the corresponding road buffer objects created from the two datasets, and then calculate the percentage overlap between the two. These results would then be written to a separate table, the output parameters of which included an assigned ID number, road name and the percentage overlap value. The program had to be run each time for each of the test areas chosen, and for each of the three road types.

5. **Conduct further quality tests**

   For both the pilot tests and main analysis, further forms of map quality assessment were made on the datasets to help provide answers for the results of the buffer analysis. For the Pilot test this involved comparison between ITN and Meridian data, and for the main analysis this involved an investigation of road name attribute completeness and a user analysis. Further details of each shall be explained in sections 5 and 6 respectively.

5. Pilot Test

An initial pilot test was carried out in order to test the modified program and to also obtain some preliminary results, so that any improvements to the modified program and buffer analysis procedure could be identified.

5.1 Method

The area selected for this pilot test was a 3.7km by 2.3km area around the M4 Heathrow spur as shown in Figure 14. This area was chosen as the M4 Heathrow spur was included as part of the previous OS Meridian analysis, and therefore results could be compared. The analysis was upon all the motorway and A-road segments in this region (there were no B-roads here). To begin with, the data for this test area was downloaded for both OSM and ITN datasets, and were in different data formats and different projections; ITN as a GML file in OSGB36, and OSM as a SHP file in WGS84.

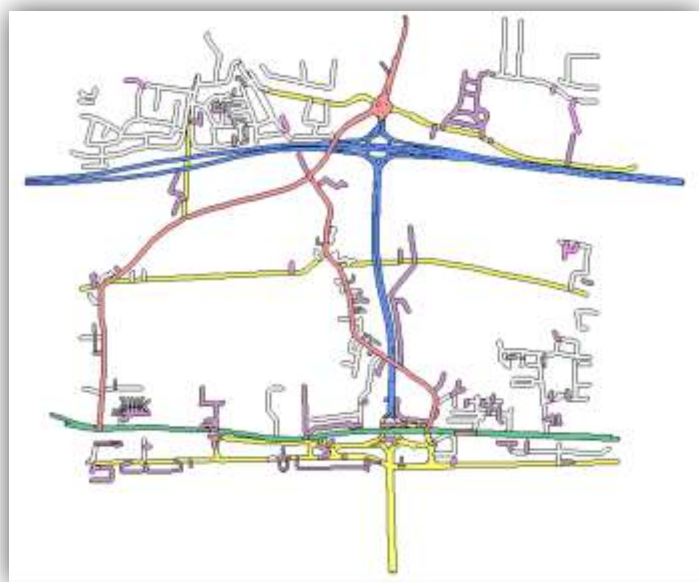


Figure 14 - Test area M4 Heathrow Spur (OS ITN)

The first task would be to edit both datasets so that all roads, except motorways and A-roads, were deleted and that both datasets were covering the exact same extent. Before this could be achieved the ITN data was converted to a shapefile so that editing in ArcGIS could take place, and the OSM data had to be converted to OSGB36; both datasets were then in the same format and projection and now ready to be edited. Using ArcGIS the A-roads and motorways were then extracted from both datasets, and any extra road segments were split and deleted so that both datasets were covering the exact same area. The attribute tables for both datasets also had to be modified so that identical road labels were used for both, as the buffer analysis program would calculate the percentage overlaps according to corresponding road names. Figure 15 below shows both datasets with all editing complete.

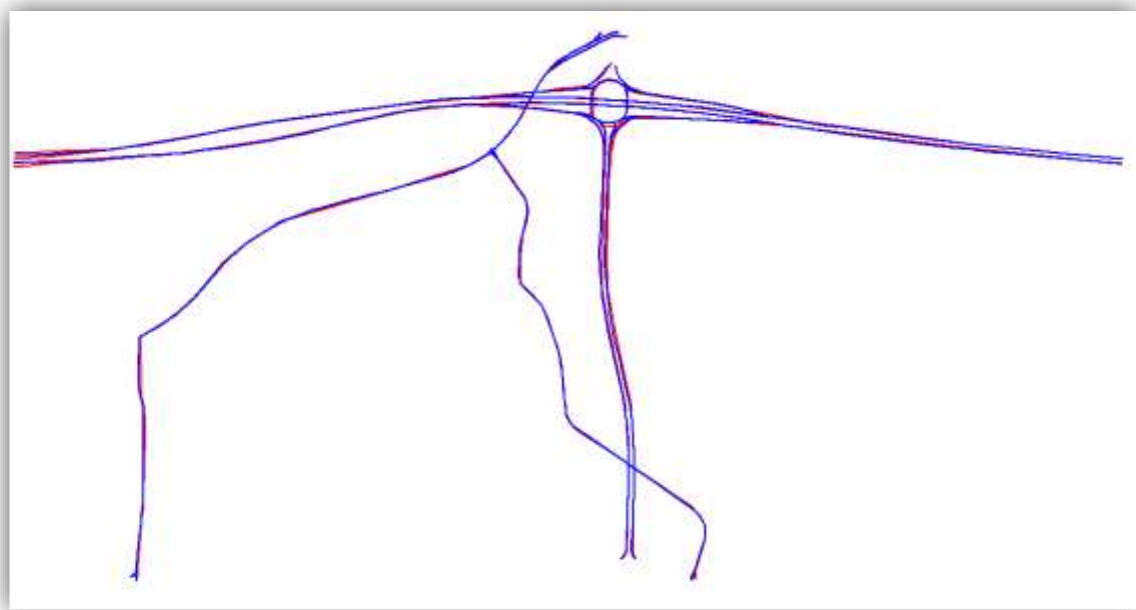


Figure 15 - Edited ITN and OSM datasets

The datasets were then ready for the buffer analysis, but before this could take place the size of the buffers to be had to be set. Initially a buffer of 10 metres was applied to all roads in the ITN dataset, as it was felt that this was a reasonable figure to test programme, and a 1m buffer was used for the OSM data (as this was buffer size used for the previous OSM/Meridian analysis). A second test was then carried out using buffer sizes shown in Table 2, to represent the true width of the roads.

5.2 Results

First Analysis (10m ITN buffer, 1m OSM buffer):

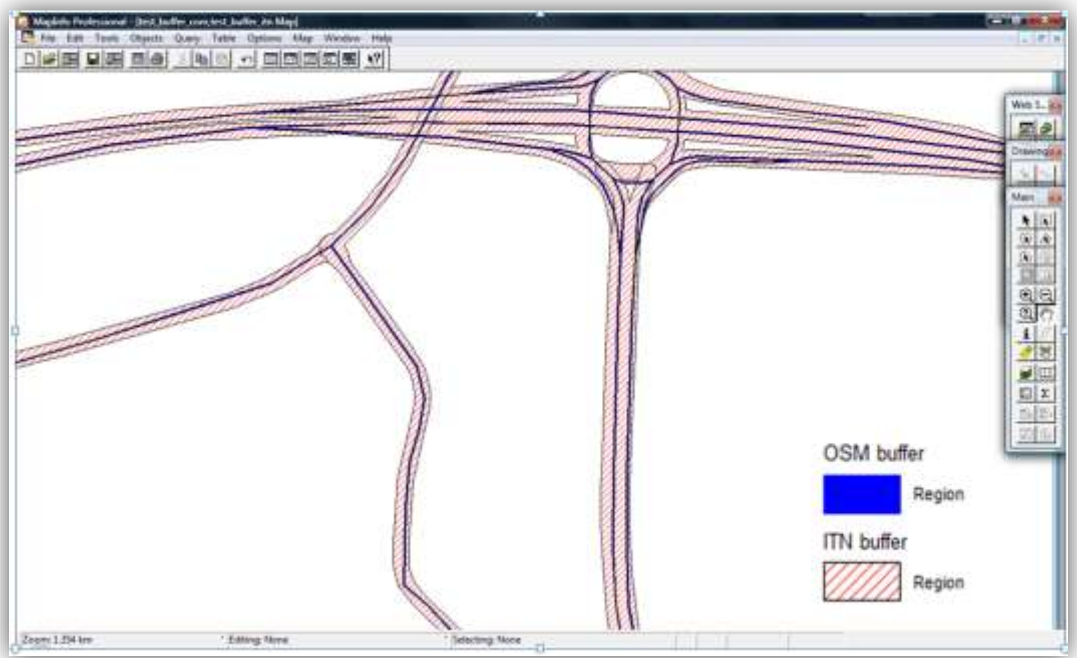


Figure 16 - ITN and OSM buffers

Road	Percentage Overlap
M4	98.44%
M4 HSpur	100%
A3044	100%
A408	99.32%

Table 3 - First analysis results: percentage overlap between OSM and ITN buffers

Second Analysis (8m motorways & 5.6m A-roads ITN buffer, 1m OSM buffer):

Road	Percentage overlap	Difference from 1st analysis
M4	96.49%	1.95%
M4 HSpur	99.59%	0.41%
A3044	93.30%	6.70%
A408	98.82%	0.50%

Table 4 - Second analysis results: percentage overlap between OSM and ITN buffers

5.3 Analysis

The results for the first analysis (Table 3) clearly show there is a very high percentage overlap between the OSM and ITN datasets, including 100% overlaps for the M4 Heathrow Spur and the A3044. This indicates that the OSM data collected was very accurately mapped as there is almost a direct overlap. The results of the second analysis further confirm this, where the buffer sizes were smaller to represent the true road widths. Table 4 shows that the difference between these results and that of first analysis is very little, except for the A3044 where the percentage overlap was 6.7% lower in the second analysis. Regardless of this, all the results are still above 90% overlap which shows a very high level of accuracy.

In comparison to the results of the previous OSM/Meridian study, an interesting observation can be made. Table 5 below shows the results of the top four percentage overlaps from the previous study, where the OSM data was compared with OS Meridian 2.

Motorway	% Overlap
M25	88.8
M4 HSpur	88.77
M23	88.73
M1	87.36

Table 5 - Percentage overlap between OSM and OS Meridian 2 datasets (Source: Haklay, 2008)

Here the M4 Heathrow Spur had the second highest percentage overlap with 88.77%; this is more than 10% lower than the results obtained from the second analysis of this pilot test. This could be due to one of two reasons:

1. That OS ITN is more accurate and precise than OS Meridian 2, which is a fare assumption considering that latter is a generalised dataset.
2. Or that the M4 Heathrow Spur has been edited or updated in the OSM dataset since the previous study.

Having checked the edit history of the OSM M4 Heathrow Spur from the OSM website, it appears that no changes have been made to this stretch of road since the last study. Therefore the first reason stated above is most likely the cause of the difference in results. To test if this first assumption is correct the Meridian dataset would need to be compared to the ITN dataset. The road this test was carried out on was the M4 Heathrow Spur. Using the Goodchild and Hunter buffer comparison method logic would suggest that Meridian dataset should be considered as the tested object, just like the OSM dataset, and therefore have a buffer of only 1m, and the ITN dataset consider as the high-level dataset with an 8m buffer (as carried out in the second analysis). However the problem with treating the ITN dataset as the high-level object in this scenario is that even with the 8m buffer there will be a gap between the two carriage ways, which will lower the percentage overlap value where the Meridian line crosses over this gap, leading to inaccurate results. This problem is illustrated in Figure 17, which shows the Meridian line running partially through the middle of the ITN buffer and partially through one of the carriageways.

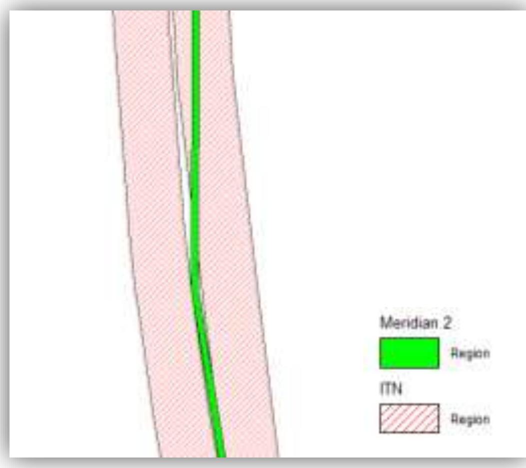


Figure 17 - ITN and Meridian buffers

It was therefore decided that best way to carry out this comparison would be to make ITN buffer the tested object (with a buffer of 1m) and treat the Meridian road line as the higher level object. The buffer size used for Meridian would be set to 20 meters, as this is the filter the OS used in the creation of the line. Although the ITN object is treated as the tested object here, it is still considered the higher level dataset and hence the benchmark to which Meridian data would be compared. The percentage overlap between the two datasets in this section of road would indicate how much of an effect the generalisation process used to create the Meridian reduced its accuracy; a result of 100% would indicate that no accuracy is lost from the generalisation.

The result of this percentage overlap is shown below:

Road	Percentage overlap
M4 HSpur	91.43%

Table 6 - Percentage overlap of ITN and Meridian buffers

With a percentage overlap just over 90% between the two datasets, the accuracy of Meridian is quite high as expected. However, as the overlap is not 100% or thereabouts, it still indicates that a certain level of accuracy is lost from the generalisation process used to create the Meridian 2 dataset. Hence, this explains why the OSM comparison with ITN of the M4 Heathrow spur produced better results than the OSM comparison with Meridian.

5.4 Conclusions

The results of this initial test show that the percentage overlap between OSM and ITN is very high, indicating that the accuracy of OSM is quite good. The accuracy of the M4 Heathrow Spur also shows that the comparison carried out in this analysis provides more accurate results than that of the previous analysis carried out with Meridian, and the test between ITN and Meridian proved that some accuracy was lost in the generation of the Meridian 2 dataset. The next step would be to apply this method to a greater coverage of roads to test if the results are true across a broader dataset.

Overall, the adapted buffer analysis program worked well and this pilot test acted as a good opportunity to iron out some bugs in the coding. Improvements that can be made for the main analysis would be to use a smaller buffer for the OSM dataset, as this is the test object and a buffer of 1 meter could reduce the validity of the results. Therefore an OSM buffer size of 0.1 metres would be used for the main analysis.

6. Main Quality Analysis

This section explains the process carried out for the main quality analysis based on the method outlined in section 4. The previous quality analysis carried by Dr Haklay (2008) primarily examined data from London, due to the fact the OSM started in this city and has received the longest ongoing attention from OSM contributors, and hence would provide a good initial indication of VGI quality. Following suit, this quality analysis also examined data sampled from London. The GIS software used to carry out this analysis were ArcGIS, Manifold GIS and MapInfo.

6.1 Obtaining the datasets

The areas of London chosen for this project were four 25 km^2 regions, based on Ordnance Survey 1:10000 raster grid tiles. The tiles used were chosen to represent the north, south, east and west of the city, with two of the tiles closer to the centre and two further out. Considering that the results of Dr Haklay's completeness analysis found that quality of coverage deteriorates with increasing distance from city centres, it was hoped the four tiles chosen would have represented a good sample of London with varying levels of VGI quality. Figure 18 illustrates the locations of the four tiles. All the tiles contained a mixture of A-roads and B-roads, and the only tile with a motorway segment (the M4) was TQ17ne. The total length of A-roads, B-roads and Motorways varied between each tile, but in total the roads analysed covered 329 km.

ITN datasets for the four tiles shown in Figure 18 were downloaded from the EDINA DigiMap service in GML format. Only the Road Network information was selected for download and not the Road Routing Information. These datasets were then converted the shapefiles so that they could be easily imported and edited in ArcGIS. Various output layers were created, such as "Road_Node" consisting of all the road node points, although only the "Road_Link" layer was used, which consisted of polylines connecting the road node points.

The OSM data was obtained in shapefile format from CloudMade as one dataset for the whole UK, although only the road and highway data was downloaded. The required test areas were extracted from this UK dataset, using the ITN datasets as a reference. All these datasets can be found on the CD-ROM attached with this thesis.

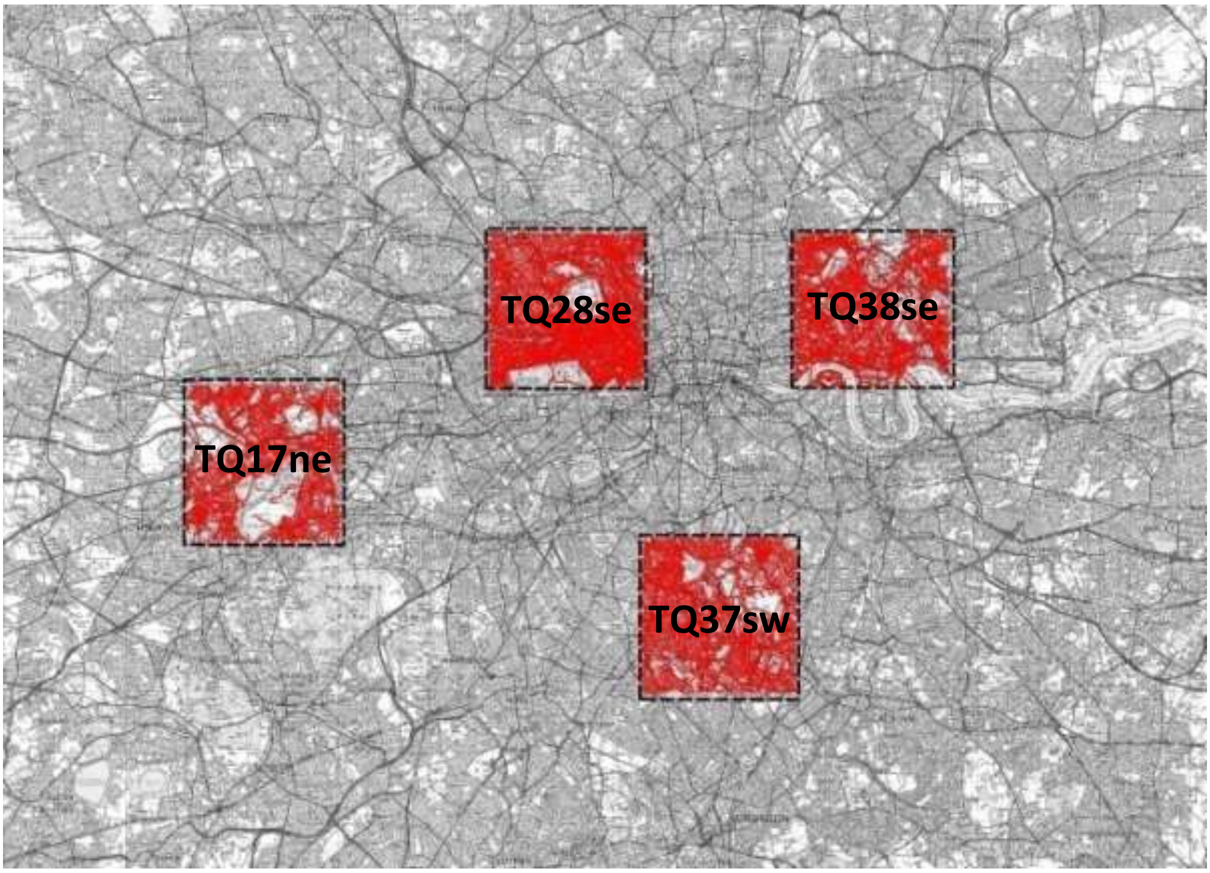


Figure 18 - Location of OS tile regions selected for the quality analysis (TQ28se - North/Central London, TQ38se - East London, TQ37sw - South London, & TQ17ne - West London)

6.2 Preparing the datasets

The first step in preparing the datasets was to ensure that the roads selected from both the OSM and ITN datasets matched exactly for the four test areas. As already mentioned, the OSM datasets were extracted from the main UK shapefile using the ITN datasets as a reference; this was achieved by overlaying both the OSM and ITN datasets in ArcGIS, and then selecting and extracting the required regions from the OSM dataset. However, this was done very roughly at first and the OSM datasets had to be further edited by deleting excess road segments, as shown in Figure 19 below.

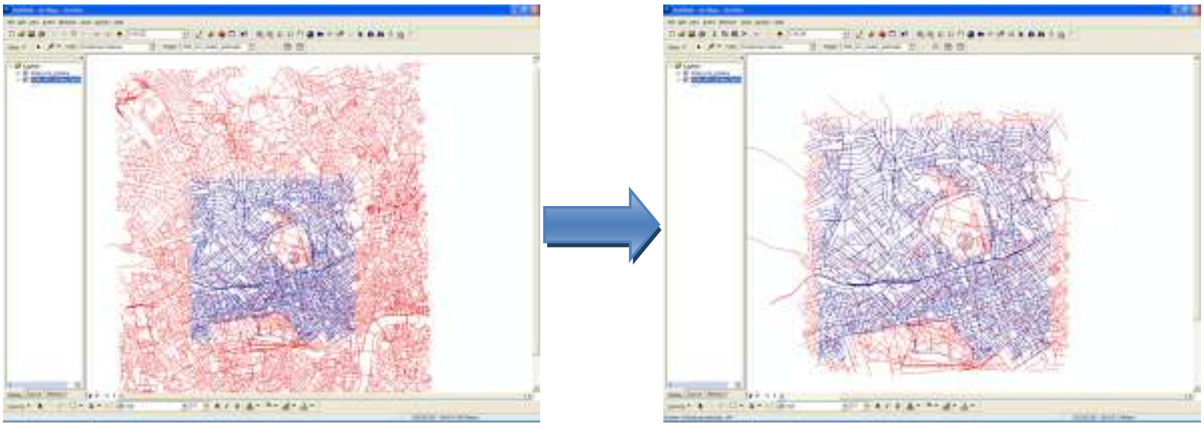


Figure 19 - Deleting excess OSM road segments, red lines = OSM roads, blue lines = ITN roads (example of tile TQ28se)

Figure 19 shows that at this stage the coverage was still not exact between the two datasets and due to the fact that an exact coverage was only required for the roads to be tested. Hence the next step was to extract the A-roads, B-roads and Motorways from the two datasets. This was done in ArcGIS by selecting the relevant roads from the ITN and OSM attribute tables and then exporting these as new shapefiles. Figure 20 shows an example of this, illustrating tile TQ28se with all the A-roads and B-roads selected.

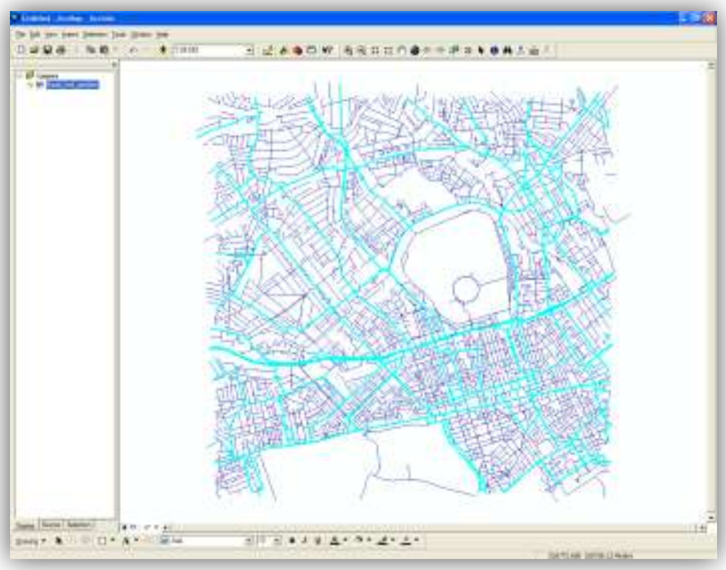


Figure 20 - Tile TQ28se with A-roads and B-roads selected for extraction

In the attribute tables of the ITN datasets, the required road types were labelled as either “Motorway”, “A-road” and “B-road”, whereas in the OSM datasets the corresponding road types were labelled as “Motorway”, “Primary” or “Trunk”, and “Secondary”. Whilst extracting the relevant roads from the ITN dataset was straightforward, there was an issue with extracting from the OSM dataset due mismatches in classification. For example, some roads classified as B-roads in the ITN dataset were classified as tertiary roads in OSM, and vice-versa. Considering that the ITN data is the higher level dataset and hence the benchmark, any major roads that did not appear in the OSM selection, but were in the ITN selection, were then manually selected and added to the selection of OSM roads to be exported. The road typed attribution for these OSM features were then re-edited to the correct attribution label.

With all the required roads exported for both datasets, the OSM road features were then edited to match the corresponding ITN road features. This was achieved using the split tool to delete extra road segments in the OSM dataset. With all the road features matched, the next step was to assign identical road names to both datasets. However, neither dataset had the required road names already attributed, with the ITN datasets having no road names at all, and OSM having some road names in places. Therefore, using Google Maps as a reference, the roads for both datasets were labelled using the road numbers rather than actual road names, for example the “Great West Road” in tile TQ17ne was labelled the “A4”.

One problem was encountered here whilst labelling the roads in the OSM datasets; in reality some roads have two different road names along what may appear to be one section of road. In OSM, these roads were sometimes digitised as single road segment when in reality they should have been digitised as two, in order to represent the change in road name. This would often occur on long road sections and at complex junctions, as illustrated by Figures 21 and 22. Figure 21 shows one single road segment leading to a junction, highlighted in red. This whole road section was labelled as the A205 in OSM, when reality it splits to the A315 as illustrated by the yellow road segment on the Google Maps image in Figure 22. This error occurred numerous times in the OSM dataset and could distort the results of the analysis if not labelled correctly with the ITN dataset. To solve this, rather than manually splitting and changing the OSM road segments, a simple solution was to change the road labels in the ITN dataset (as roads objects were segmented correctly here) so to the match the OSM data; although factually incorrect, it was fine for the purpose of this analysis which is concerned with the positional accuracy of the roads rather than attribute accuracy.

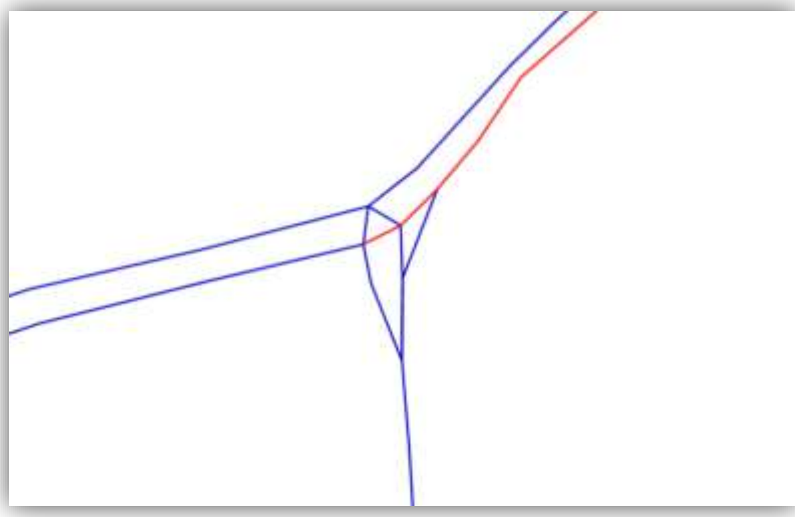


Figure 21 - OSM road segment labelling error – two road sections digitised as a single road segment

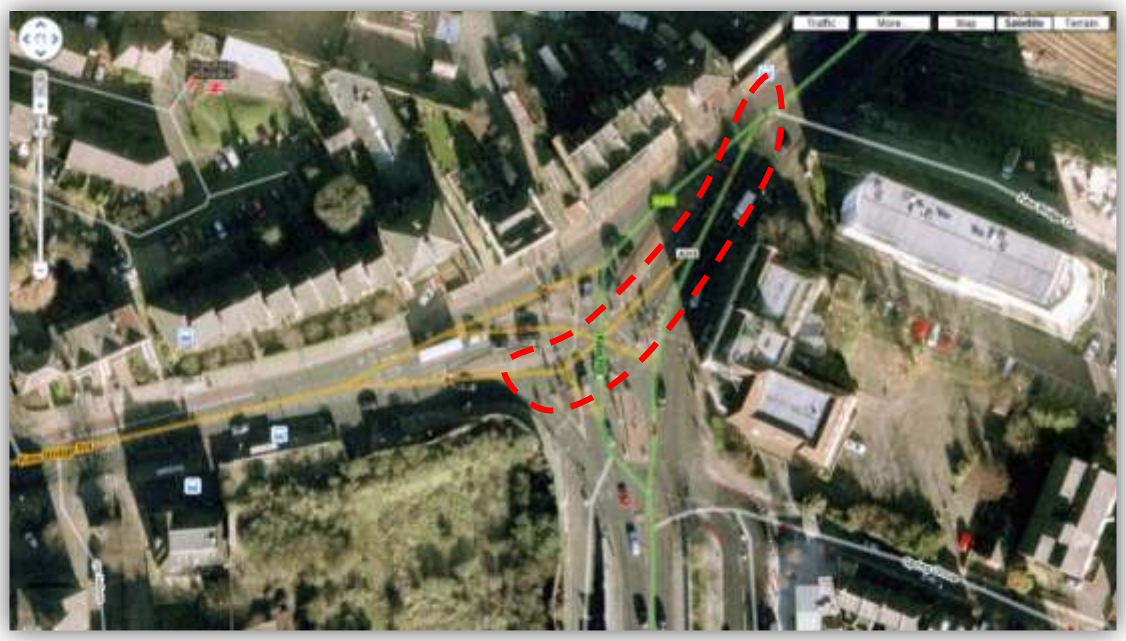


Figure 22 - Correct road segmentation (source: Google Maps)

With all the road features correctly labelled the road names were then listed in a separate text file for the buffer analysis program. Validation of datasets was also carried out by comparing the list of road names for both OSM and ITN datasets to each other. The corresponding pairs of road objects were also visually compared in ArcGIS. This validation helped identify and correct numerous human errors that occurred as part of this data preparation process. The final datasets can be found on the attached CD-ROM, and maps of the final datasets are shown in the appendix.

6.3 Running the buffer analysis program

The modified program used for the pilot test was again used for the main analysis. The program was copied four times and modified for each of the datasets representing the four test areas. Hence the directories for each of the datasets were already set in each copy of the program, which made it easier and quicker to carry out numerous runs of the program for each of the four test areas to fix any bugs. All that needed to be changed for each analysis was the ITN road buffer size, and to ensure the corresponding road lists contained the required road names. The OSM road buffer was permanently set to 0.1 metres as concluded from the pilot test.

Using MapBasic and MapInfo, the program was then run twice for tiles TQ28se (North/Central London), TQ38se (East London), and TQ37sw (South London), first time for the A-roads analysis and the second time for the B-roads analysis. The program was run one extra time for tile TQ17ne (West London) for the motorway segment this tile contained. The results tables for each of tiles were then exported from MapInfo as text files where they were imported and compiled in excel. The corresponding road lengths, compiled as part of the buffer object creation, were also exported to excel. All the copies of the program, output and input datasets used for this analysis can be found on the CD-ROM.

Only a couple of problems occurred with the program during this process. The first problem was that the program would stop halfway and not complete the analysis for some of the tiles. This was then found be caused by human error in the labelling of some of the roads, i.e. the program could not match some the ITN and OSM roads. The problem was fixed easily by re-labelling these roads. The second problem was due to some issues with parameters of the output table structures, where the widths of some of the characters were not long enough for some of the input field information. This prevented some of the output values being written and would stop the program. The output table structures therefore had to be modified in MapInfo to solve this.

6.4 Further quality tests: road name attribution completeness and user analysis

The purpose of this further analysis was firstly to add another dimension of OSM map quality other than positional accuracy, and secondly to provide possible answers to levels of accuracy found from the buffer analysis. Using the OSM datasets for the four test areas, further quality analysis was carried out in based upon the length of roads without road names and number of users who have contributed data. Both of these tests were carried out in Manifold.

A length analysis of the percentage of roads with road name attribution, in each test area, would help identify how much of the data was collected on the ground. This is due to the fact the road names can only be established by users who have conducted ground surveys, as identifying road names from other map sources would break copyright laws. It can therefore be assumed that any data without road names were likely to be obtained from aerial imagery. Hence, conclusions could be drawn as to where the accuracy of each sample was affected by the nature of the data collection. The hypothesis here is that areas with higher levels of incomplete road name attribution will have lower levels of positional accuracy. Secondly, this road length analysis would provide a measure of the level of attribute completeness.

This analysis was carried out using the OSM datasets that were saved before the road name editing took place, hence still containing the original attribute information. The road features without road names in the attribute tables were then identified and the total length of these roads were calculated. This was then divided by the total length of road features in the dataset and expressed as a percentage. The reason why the length of the roads were measured, as opposed to merely the number of unnamed road segments, was because it provides a more definitive measure attribute completeness whereas a road segment could be of any length.

With regards to the user analysis, this involved looking at the number of users who contributed in each of the test areas selected for the OSM datasets, in order to identify any trends to support the theory of Linus's Law. The hypothesis here is that areas with a greater number of users should have greater levels of completeness and higher positional accuracy.

The user name information did not come as part of the shapefile data obtained from CloudMade, and could only be obtained from the main OSM slippy map as XML files, which were then later converted to shapefiles. The required data was therefore exported directly from the OSM slippy map, with the areas chosen to cover the exact same regions as the OS tiles (although the extracted data was 'hairy' in that some road segments that lay outside the selected area were included and not cropped). A script provided by Dr Haklay was then used to read these xml files and extract all the point, line and polygon data, including all the attribute information (from which the user names were required) in Manifold GIS and was then saved as three sets of shapefiles for each dataset, one for the node data, one for the line data and the last for the polygon data. However, only the line data shapefiles were used for this analysis.

The user analysis consisted of two parts. The first looked at overall number users who contributed to each of the four test areas, in order to compare the results of overall positional accuracy against the number of users who have contributed to them. However, each test area contained different numbers of road segments which were not reflective of the road lengths within each area. Therefore to make this comparison meaningful, the number of users were expressed as the ratio of the total length of roads in each area to the total number of users. The datasets were exported to excel where the number of users were counted and the sum of the length of roads were calculated.

The second part of the user analysis was a more detailed investigation, where the datasets for the four test regions were broken down to 1km gird squares. Within each grid square, the number of users were counted along with the number of road features that were named and unnamed. This analysis was carried out in Manifold, where 1km grid squares were created for each of the four OS tile regions, and the required road objects were clipped to each grid square. The line dataset shapefile, that this analysis was carried out on, did not exclusively contain road features but also other features represented as lines, such as pathways, steps and cycle routes. Such features may not necessary have names and were therefore excluded from the analysis so that only road features were remaining; these should definitely have a road name and would therefore provide a more meaningful measure of completeness. The data for each of grid square was exported to excel where a filter was created to count the number of named and unnamed road attributes whilst excluding attributes which were not road features. The number of users were also counted here.

The average positional accuracy of road features in each grid square was also measured for comparison, and was calculated as an average of the results obtained from the A-road and B-road buffer analysis. The datasets used here were the OSM road buffers generated from the analysis program (with the percentage overlap results joint to it as a table relation), as all road segments were grouped such that each road was represented by one single line object segment, rather than by numerous segments (as with the input datasets). This ensured all the roads in any one grid cell were represented by one road segment for each. The percentage overlap results were joined to the OSM road buffers table as a table relation, and these datasets were again clipped to the 1km grid squares created for each test region, where the average position accuracy of each was then calculated using an SQL query.

7. Results

Tables 7 – 10 and Figures 23 – 28 show the results obtained from this analysis:

- Tables 7 & 8 show the results of the buffer analysis for each of the A-roads, B-roads and motorways in the four chosen test areas (tiles TQ38se, TQ28se, TQ17ne and TQ37sw). For each tile the total length and results of the percentage overlap are given for each road, along with the average percentage of overlap of each road type and the combined average percentage overlap for the whole tile.
- Figures 23 – 26 are the results maps created from the percentage over lap results for each of the four tiles. These results, represented by OSM road objects, are overlaid on OS 1:10000 raster maps (obtained from EDINA) for visual comparison. These maps can also be found on the attached CD-ROM
- Table 9 shows the results of the road name attribution and road length analysis. The total length of OSM roads are given along with the percentage of these roads without labelled road names. A small screenshot of OSM roads are also shown with the unlabelled roads highlighted in red.
- Table 10 shows the results of the 1st user analysis, where the total number of users for each test region are given along with the total length of roads, the calculated length of road per user, and the calculated ratio between number of users and total length of road (exaggerated by a factor of 10).
- Figures 27 and 18 are scatter diagrams illustrating the results of the 2nd user analysis, where the number of users per square kilometre are compared against road attribution completeness and positional accuracy. The full table of results for this analysis can be found in the appendix.

East London Tile (TQ38se)						North/Central London Tile (TQ28se)					
A-roads (5.6m buffer)	Total length of road (km)	Percentage overlap (%)	B-roads (3.75m buffer)	Total length of road (km)	Percentage overlap (%)	A-roads (5.6m buffer)	Total length of road (km)	Percentage overlap (%)	B-roads (3.75m buffer)	Total length of road (km)	Percentage overlap (%)
A118	1.22	99.70	B142	1.06	95.98	A4207	0.87	100.00	B412	0.15	100.00
A107	0.60	99.50	B126	1.45	92.86	A503	2.80	99.08	B451	0.51	100.00
A1020	2.01	99.34	B165	0.93	89.30	A4206	2.18	98.90	B415	0.29	98.29
A1011	4.94	97.43	B125	1.02	81.40	A5204	2.39	98.64	B512	0.54	96.59
A11	9.79	96.46	B119	2.38	80.20	A4200	1.51	98.22	B414	1.66	93.17
A124	1.78	95.43	B113	1.95	79.30	A502	2.44	94.81	B410	0.78	91.85
A112	2.60	94.13	B140	3.25	77.12	A5200	1.04	93.48	B511	1.11	90.91
A1205	4.03	93.61	B205	1.40	75.41	A400	8.26	92.21	B506	2.10	90.40
A106	4.66	92.49	B121	1.42	75.09	A4204	0.73	91.19	B405	0.52	89.70
A115	2.30	92.18	B164	1.64	67.95	A404	5.08	90.55	B507	3.13	88.71
A12	15.10	90.54	B120	0.76	65.55	A401	0.84	89.56	B411	1.52	85.64
A1261	6.03	88.94	B118	2.18	54.90	A501	7.37	89.12	B509	3.22	76.46
A13	9.34	86.81	B127	0.82	5.28	A4202	3.01	89.08	Oxford_St	2.88	75.42
A1206	4.01	85.65				A41	8.95	89.05	B517	0.86	74.77
A1203	7.98	75.62				A4209	1.57	87.85	B525	1.11	74.61
A101	1.69	74.82				A40	8.53	86.24	B406	1.07	74.39
A102	2.39	21.00				A4	4.54	86.13	B524	1.07	73.23
						A4205	1.09	81.45	B413	2.21	70.36
						A5205	3.03	79.65	B510	1.34	68.81
						A5	5.40	77.31	B520	0.84	19.45
						A5202	3.00	76.47			
						A4201	4.85	56.11			
Average % overlap:		**87.27**	**Average % overlap:**		**72.34**	**Average % overlap:**		**88.42**	**Average % overlap:**		**81.64**
			Combined average % overlap:		**80.80**				**Combined average % overlap:**		**85.19**

Table 7 – Results of buffer analysis (part 1)

South London Tile (TQ37sw)						West London Tile (TQ17ne)								
A-roads (5.6m Buffer)	**Total length of road (km)**	**Percent-age overlap (%)**	**B-roads** (3.75m Buffer)	**Total length of road (km)**	**Percent-age overlap (%)**	**A-roads** (5.6m Buffer)	**Total length of road (km)**	**Percent-age overlap (%)**	**B-roads** (3.75m Buffer)	**Total length of road (km)**	**Percent-age overlap (%)**	**Motor -way** (8m buffer)	**Total length of road (km)**	**Percent-age overlap (%)**
A204	2.25	98.46	B219	1.28	90.60	A310	2.32	99.74	B452	2.99	90.80	M4	13.21	98.85
A212	3.12	97.77	B222	0.80	88.13	A3000	0.94	98.92	B353	2.15	73.94			
A234	1.37	97.31	B237	0.83	77.94	A305	1.63	97.75	B455	2.69	72.48			
A2216	1.94	96.42	B223	1.15	76.24	A3004	1.15	97.19	B454	3.40	72.02			
A2214	2.82	95.49	B221	0.60	73.97	A3002	4.11	94.71	B363	1.23	61.33			
A215	6.89	93.92	B232	0.92	71.56	A315	5.91	92.44	B4491	2.10	58.57			
A2199	4.35	93.23	B238	1.06	70.65	A307	2.85	87.49						
A205	6.80	88.48	B231	0.32	69.14	A4	12.00	74.77						
A23	7.96	84.76				A205	4.53	72.75						
A214	5.05	80.33				A316	5.48	70.94						
						A406	2.50	67.16						
						A4000	0.83	57.66						
	Average % overlap:	**92.62**		**Average % overlap:**	**77.28**		**Average % overlap:**	**84.30**		**Average % overlap:**	**71.52**		**Average % overlap:**	**98.85**
				Combined average % overlap:	**85.80**								**Combined average % overlap:**	**81.03**

Table 8 – Results of buffer analysis (part 2)

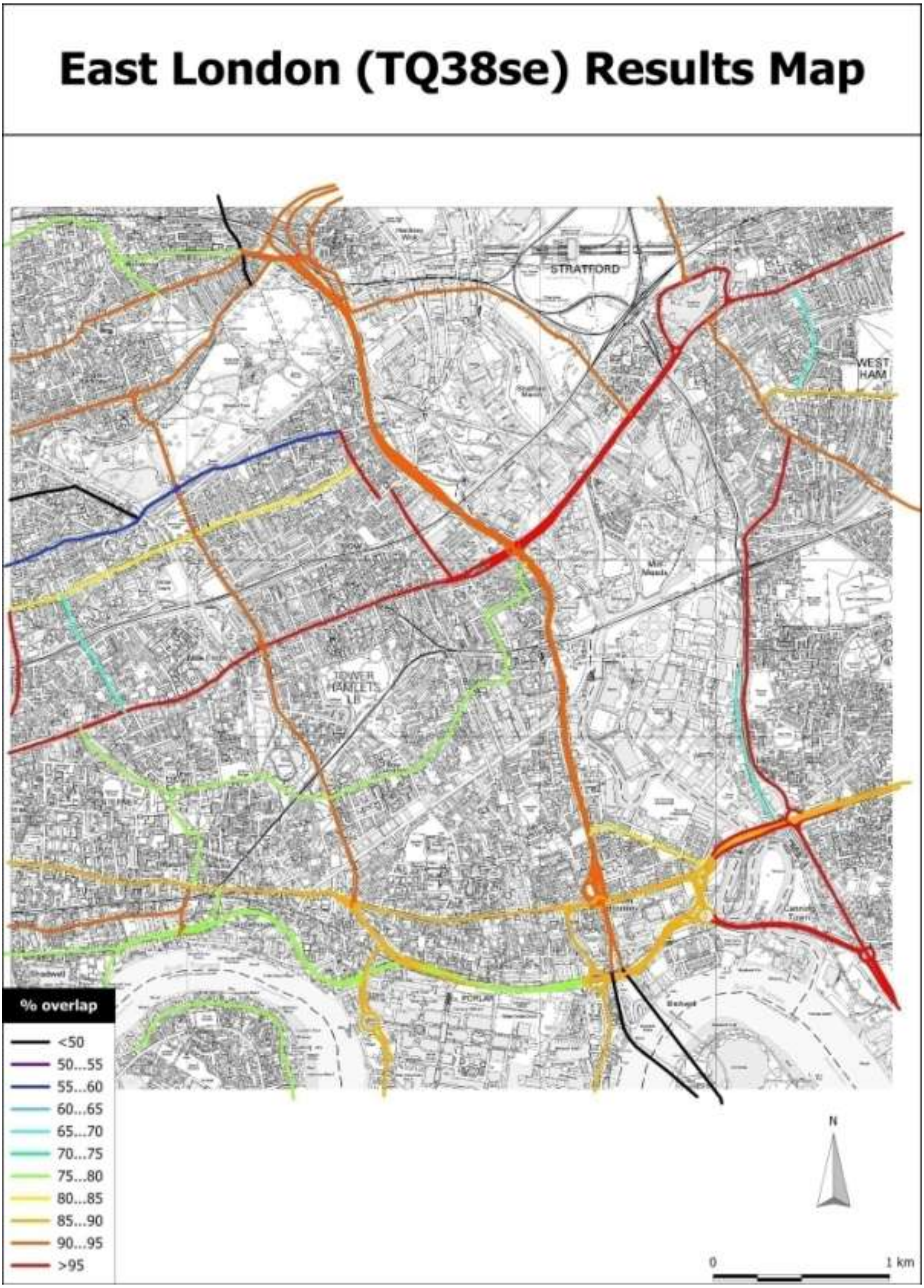


Figure 23 – Percentage overlap results for all A-roads and B-roads in tile TQ38se

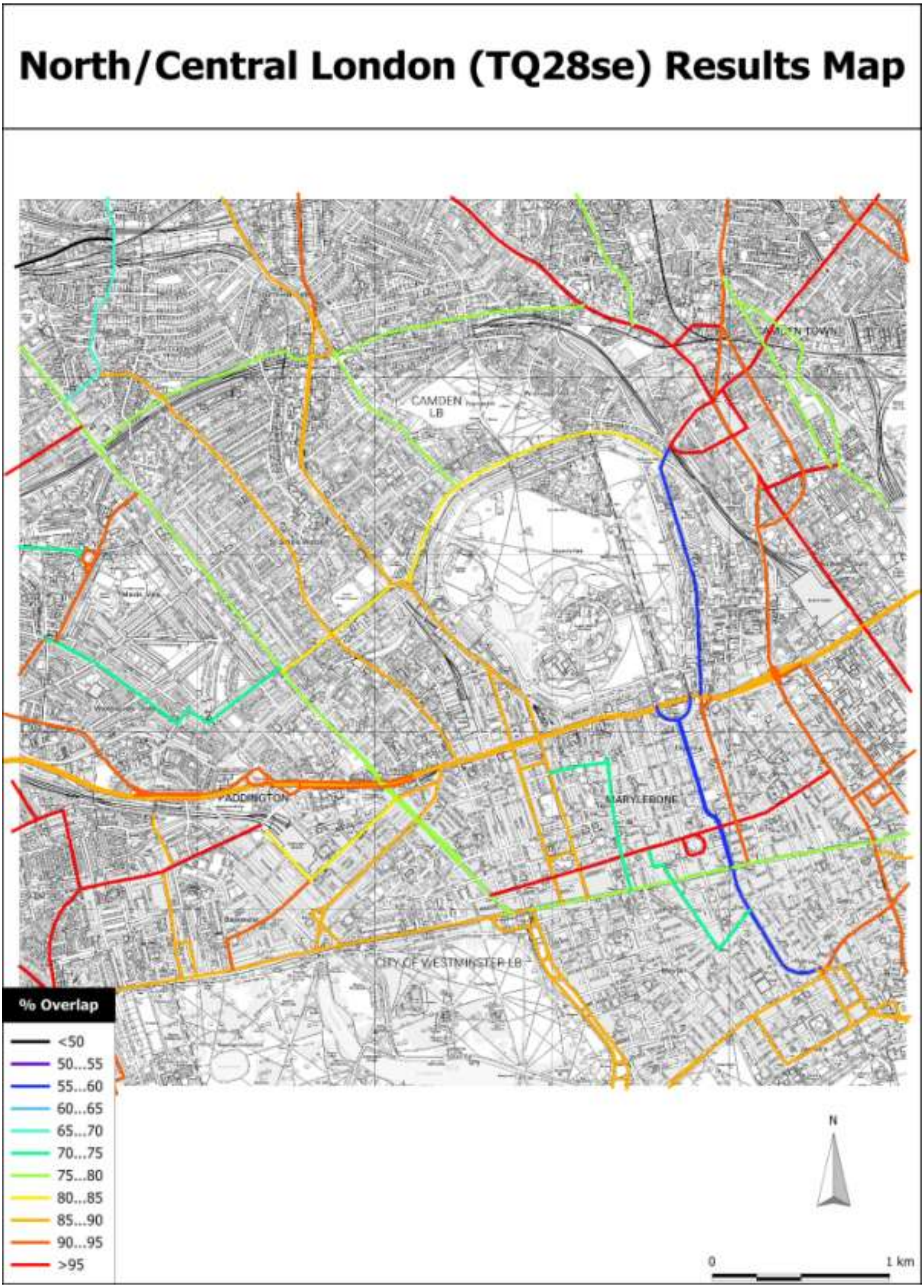


Figure 24 – Percentage overlap results for all A-roads and B-roads in tile TQ28se

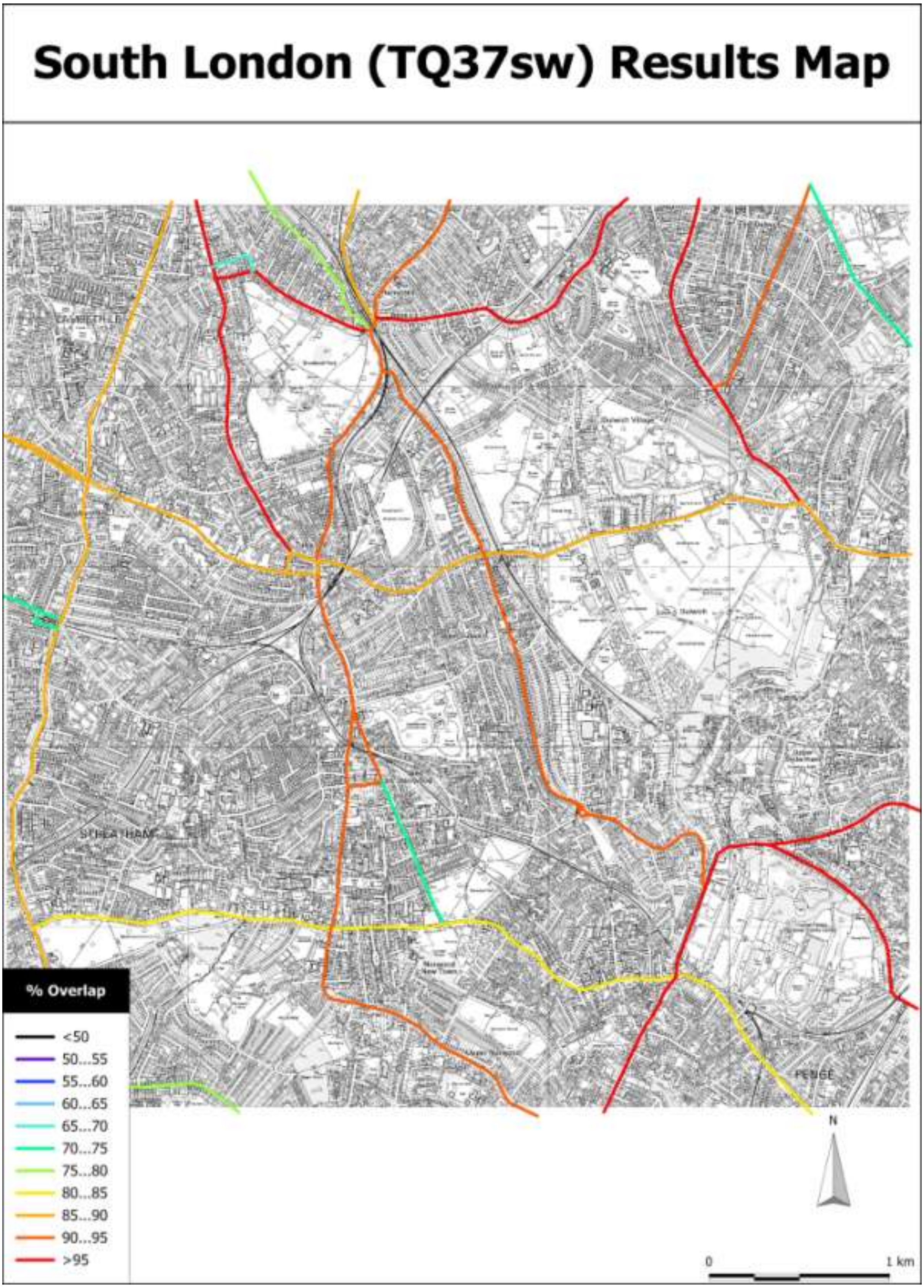


Figure 25 - Percentage overlap results for all A-roads and B-roads in tile TQ37sw

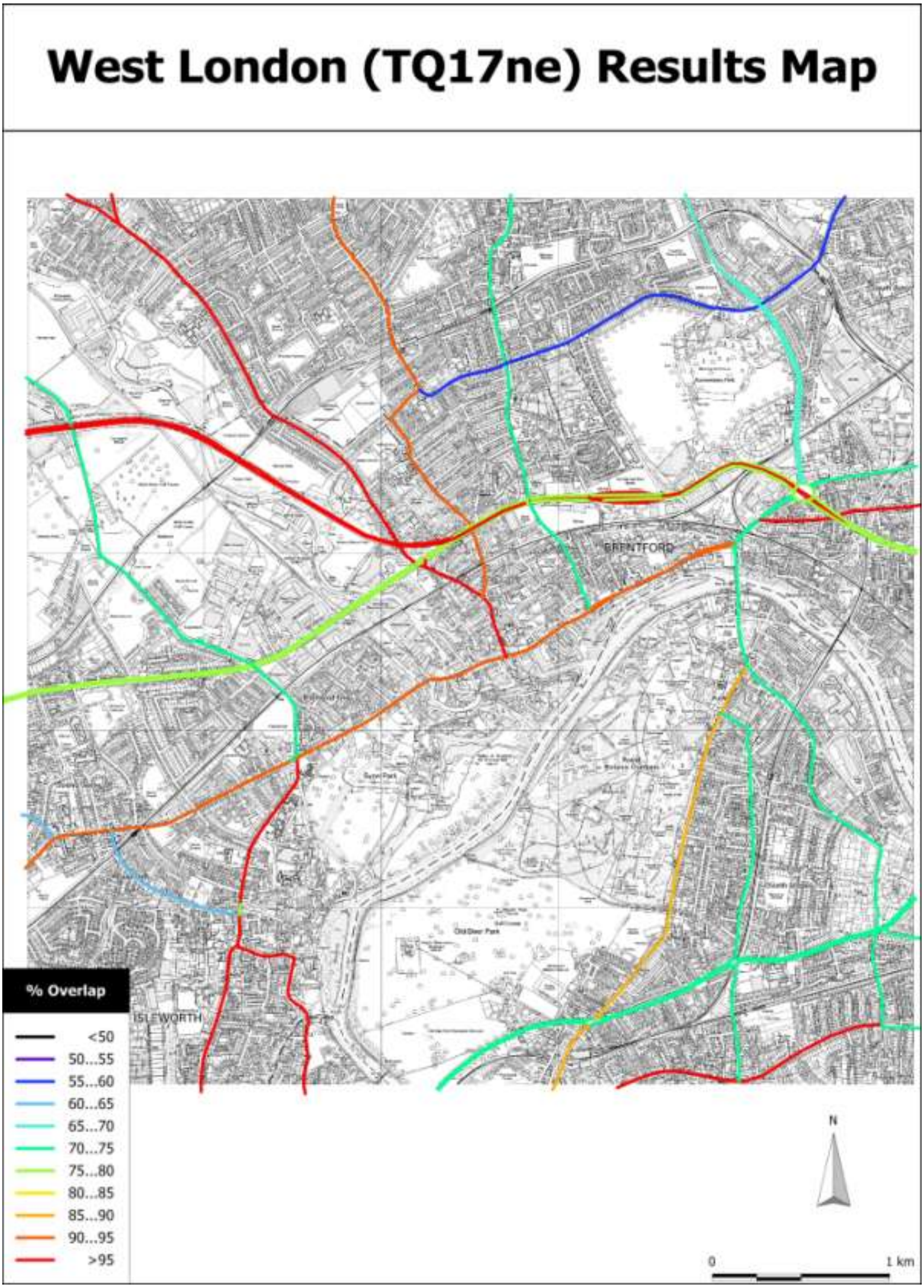


Figure 26 - Percentage overlap results for all A-roads and B-roads in tile TQ17ne

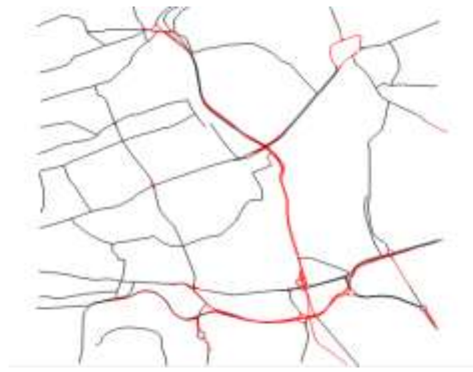
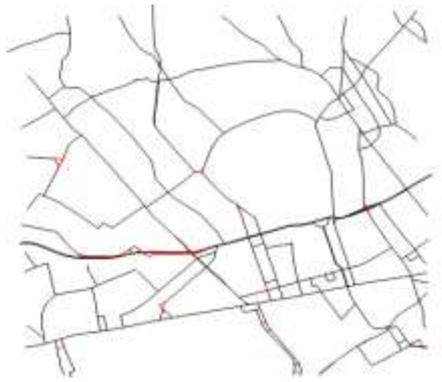
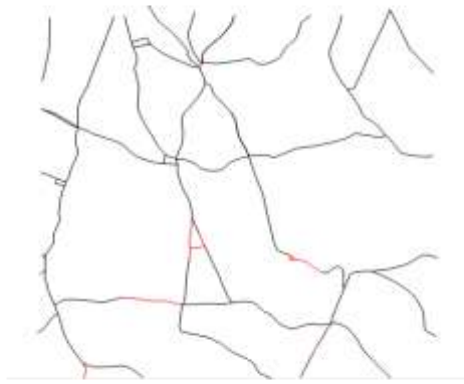
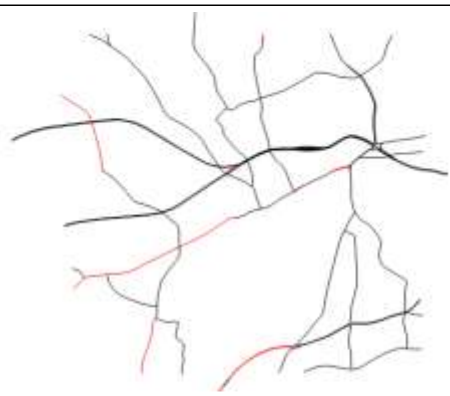
Tile	Total Road length (km)	Total length of roads without name attribute (km)	Percentage of roads without name attribute (%)	Screenshot (roads without names highlighted in red)
East London (TQ38se)	94.30	29.24	31	
North/ Central London (TQ28se)	98.53	4.99	5	
South London (TQ37sw)	48.75	2.99	6	
West London (TQ17ne)	73.96	8.76	12	

Table 9 – Length of OSM roads without road name attribution

Tile	Number of users	Total length of roads (km)	Length of road per user (km)	Ratio (*10)
East (TQ38se)	91	756.19	8.31	1.203
North (TQ28se)	145	819.97	5.65	1.768
South (TQ37sw)	56	465.01	8.30	1.204
West (TQ17ne)	81	796.34	9.83	1.017

Table 10 – 1st User Analysis: Overall results of user analysis for each tile

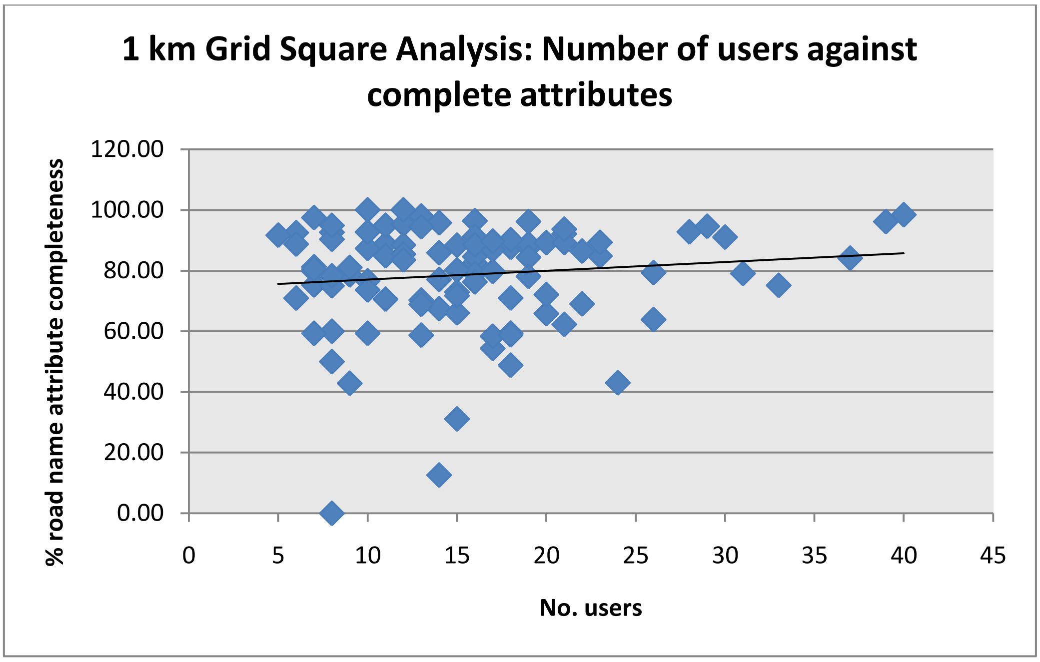


Figure 27 - 2nd User Analysis: Number of users per 1km grid square against road name attribution completeness

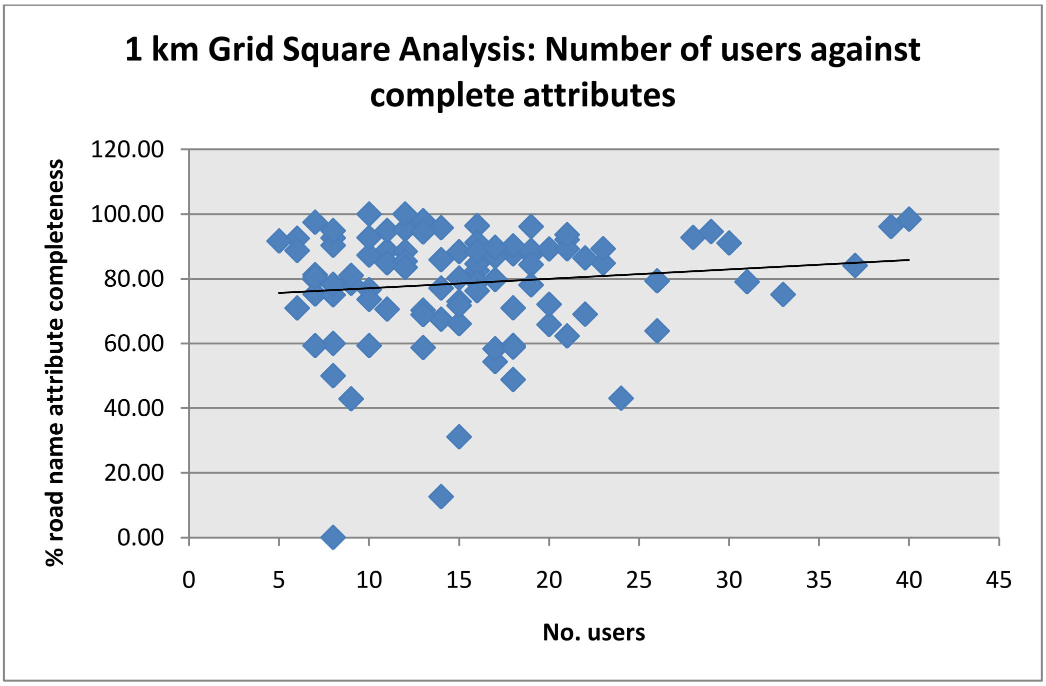


Figure 28 - 2nd User Analysis: Number of users per 1km grid square against positional accuracy

8. Discussion of results

8.1 Analysis of buffer overlap results

Overall, 109 different roads were examined covering over 328 km. The results of the buffer analysis (Tables 7 & 8) show that on the whole the percentage overlap for all the roads were very high. Each tile had a combined average percentage overlap above 80%, with the tile TQ37sw (South London) having largest percentage overlap of 85.8%. Three of the roads examined had an exact 100% overlap, the A4207, B412 and the B412.

Between each tile, the average percentage overlap for the A-roads ranged between 92.6% to 84.3% and for B-roads it ranged between 81.64% to 71.52%, and the only motorway to measured, the M4, had a very high percentage overlap of 98.9%. A pattern here is that the larger road type had better results, i.e. the percentage overlap of the motorway segment was higher than the average A-road results, which in turn was higher than the average B-road results. This is most probably due to the nature of the buffer analysis where larger buffer sizes were used for the wider road types, and hence there is a greater margin for error for any data collected on wider roads. Regardless of this, the average percentage overlap for each road type was still very high.

Looking closely at the distribution of the results, a large majority of the roads examined had percentage overlaps above 90%, as illustrated by the histogram below (Figure 29). The histogram clearly shows the results are skewed towards 100% overlap, with the cumulative percentage overlap results increasing dramatically at two stages, once after 70% overlap and a second time after 85% overlap. The last bin (95-100% overlap) contained the highest frequency of results, with 64 if the roads examined (57%) falling in the last three bins (85%-100% overlap).

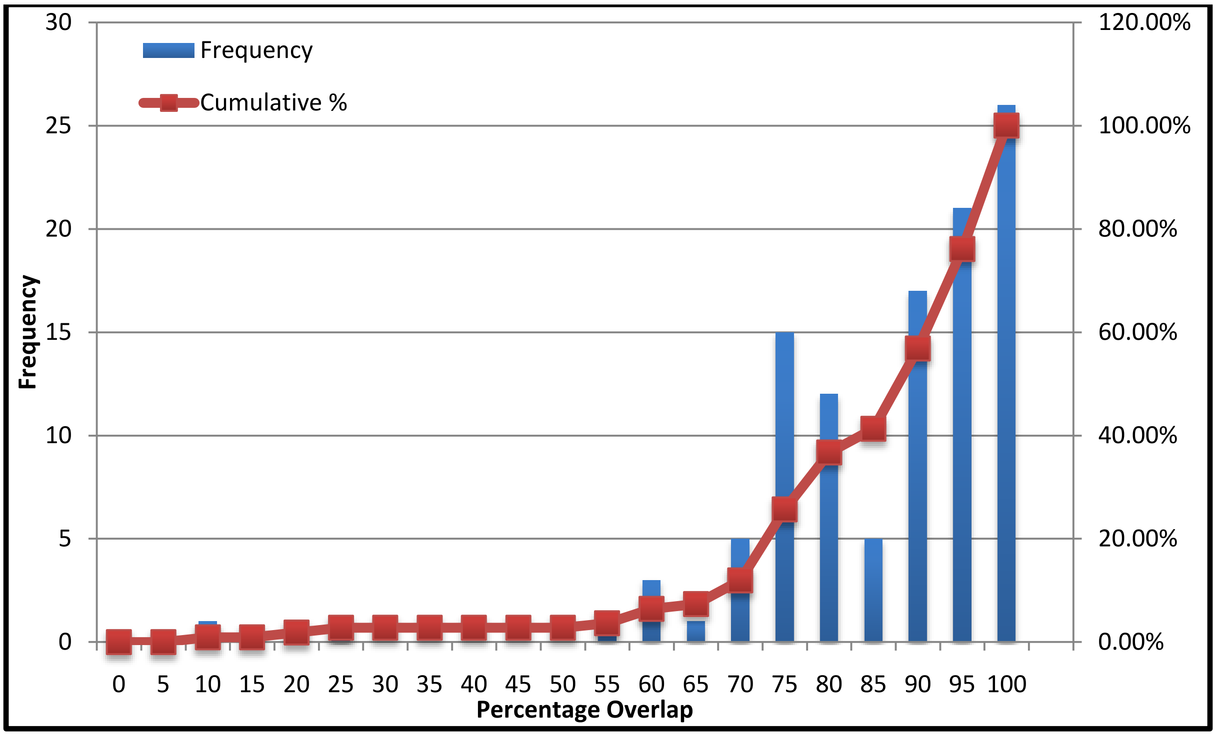


Figure 29 - Histogram illustrating the distribution of percentage overlap results

However, the length of each of the roads examined also needs to be taken into consideration in order to give a sense of proportion to percentage overlap results. The four results maps (Figures 23-26) clear show that the majority of the longest roads in each tile had over 85% overlap. Table 11 below is further proof of this observation, which shows the results for the 10 longest roads examined (consisting 40% of the total road length); other than the A4 and A1203, all the roads have a percentage overlap around or above 85%.

Road	Length of road (km)	Percentage overlap
A12	15.10	90.54%
M4	13.21	98.85%
A4	12.00	74.77%
A11	9.79	96.46%
A13	9.34	86.81%
A41	8.95	89.05%
A40	8.53	86.24%
A400	8.26	92.21%
A1203	7.98	75.62%
A23	7.96	84.75%
A501	7.37	89.12%
A215	6.89	93.92%
A205	6.80	88.48%
A1261	6.03	88.94%
A315	5.91	92.44%

Table 11 - Buffer overlap results for top 15 longest roads

A brief visual comparison of Figures 23 to 26 show that on the whole, the OSM datasets overlap the corresponding roads on the OS 1:10000 raster quite well; as expected, roads with high percentage overlap values tend fit through the centreline of the corresponding roads on the OS raster map, whereas roads with low percentage overlap values tend be near the edges and would sometimes cross the boundaries of the corresponding OS raster map roads. An example of this is shown in figure 30 below.

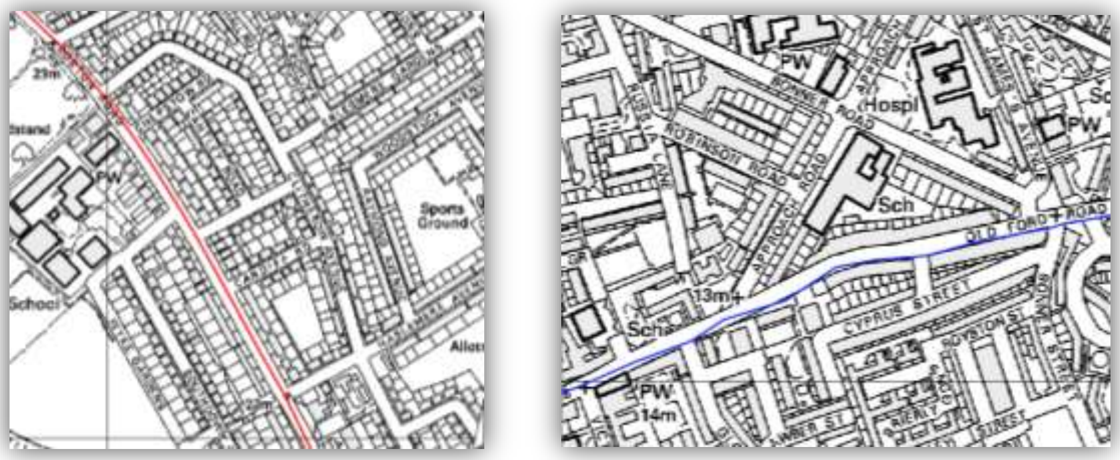


Figure 30 – Comparison of OSM buffer results and OS 1:10000 raster map (left: road with >95% overlap [A3002], right: road with <60% overlap [B118]

Whilst the results were good for the majority of the roads analysed, some had very poor percentage overlaps with a few clear anomalies; namely the A102 (21% overlap), the B520 (19% overlap) and the B127 (5.28% overlap). Comparing these OSM roads to the OS 1:10000 raster maps reveal reasons for the poor results. The B520 and the B127 are both near the edges of the corresponding roads on the OS raster map, and a comparison of the generated buffer maps for both roads show that OSM road line fractionally misses the ITN buffer, as shown in Figure 31. Therefore, if the ITN buffer widths had been slightly larger these roads would have registered significantly higher percentage overlaps. Secondly both these road segments are less than 1km in length and are therefore very short in comparison to the other roads in the dataset, and they both lie near edges of their respective datasets and have been cropped. If larger sections of these roads had been analysed the overall percentage overlap may have been higher.

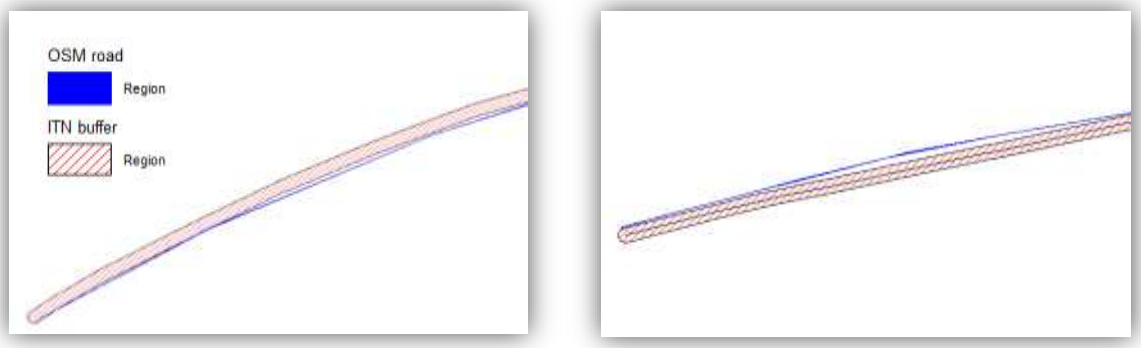


Figure 31 – Buffer map segments of the B520 (left) and the B127 (right)

With regards to the A102, the reason for the poor percentage overlap is very clear. This road comprises of both directions of the Blackwall Tunnel, which runs under the river Thames. Hence, no GPS measurements could have been collected in these tunnels nor could the road be visible from Yahoo! Imagery, and was therefore interpolated by the user who digitised these roads. This is illustrated by Figure 32, which shows the two roads leading up to the tunnels as having a very high percentage overlap, but where the tunnels begin the OSM roads completely miss the corresponding roads on the OS raster map.

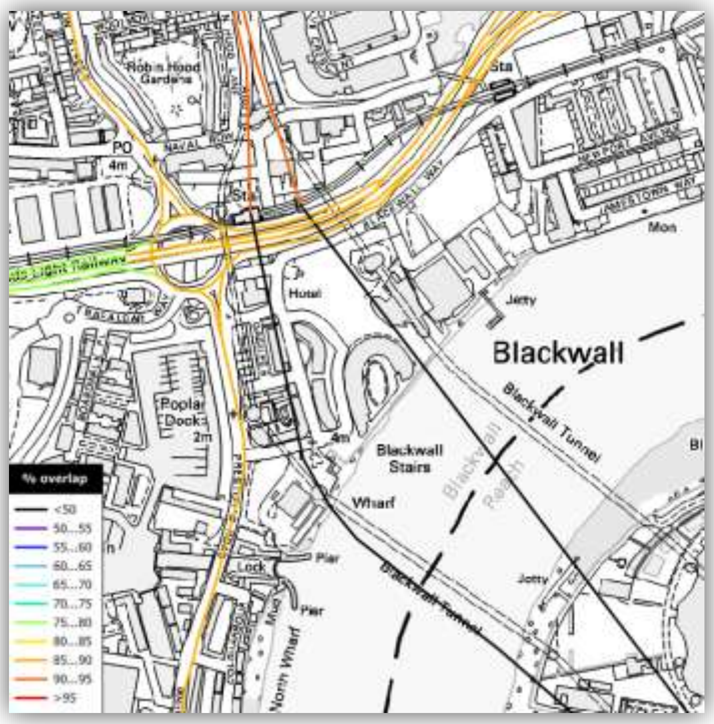


Figure 32- OSM buffer analysis anomaly: Blackwall tunnel

Overall, the results from the buffer analysis show that the positional accuracy of OSM data is very good in comparison to the OS MasterMap ITN layer. As the datasets used are sampled from across London, it is fair to assume that this level of positional accuracy could be found across the vast majority of the city, although further datasets would have to be analysed with the buffer analysis program to validate this. However, whilst the buffer analysis method is good for identifying how much of one datasets falls within the boundaries of another, it is not good for providing an exact measure of accuracy, i.e. if an OSM road segment falls outside the ITN buffer, this analysis does not provide a quantifiable measure of how far outside the buffer this road segment actually is.

8.2 Analysis of road name attribute completeness and user analysis results

The results of the road name attribution analysis (Table 9) showed that out of the four tiles, TQ28se (North/Central London) was the largest dataset with 98.53km of road and had the greatest level of road name attribute completeness with only 5% of roads unlabeled. Tq38se (East London), the second largest dataset with 94.3 km of road, was the most incomplete with 31% of road name attributes missing.

It was earlier hypothesised that regions with lower levels of road name attribute completeness should have lower positional accuracy results, as this is indicative of fewer users actually collecting results from the ground. The road name attribution results compared to the results of the buffer analysis, shown in Table 12 below, show that the tile with the lowest positional accuracy was TQ38se (80.80%), which was the tile with the lowest level of road name completeness. The opposite was also true in that the tiles with the highest average positional accuracy, tiles TQ28se (85.19%) and TQ37sw (85.8%), also had the highest percentages of road name attribute completeness. Therefore the hypothesis is proved correct here.

Tile	Average % overlap	% roads without name	Number of users	Road length to user ratio (*10)
TQ28se (North/Central)	85.19%	5%	145	1.768
TQ38se (East)	80.80%	31%	91	1.203
TQ37sw (South)	85.80%	6%	56	1.204
TQ17ne (West)	81.03%	12%	81	1.017

Table 12 - Results comparison table (ratio's increased by a factor of 10 for easier comparison)

The results of the first user analysis (Table 10) are also show in the above comparison table. It was hypothesised that regions with greater numbers of users should have greater levels of completeness and positional accuracy. Table 12 shows that tile TQ28se, as well as being the most complete dataset, also had by far the largest number of users (145) and also the greatest ratio of total length of road to number users (1.768). In other words, there were more users to cover all the roads in this tile than in any other tile, which explains why this tile had the greatest level of road name attribute completeness. This tile also had one of the highest levels of average percentage overlap. Also, the two tiles with the lowest ratio values of 1.203 and 1.017 are tiles TQ38se and TQ17ne respectively, with both tiles having the lowest average percentage overlap results of 80.8% and 81.03% respectively. Tile TQ37sw is the only exception to this trend as it has a low ratio value but a comparatively high average percentage overlap and high road name attribution completeness. On the whole it can be concluded that these results prove the stated hypothesis and support the theory of Linus's Law.

However these are very generalised results. The second user analysis sought to provide a greater insight into the possible trends between numbers of users and road name attribute completion and positional accuracy at a 1 km grid square resolution. In total, 100 gird squares were analysed across the four test regions. Figure 27 is a scatter diagram showing the results between the number of users and the road name attribute completeness. It shows that not a single grid square had less than 5 users who had contributed data within that region, with the vast majority of grid squares containing between 5 – 20 users. One grid square had as much as 40 users; not surprisingly this result was from the North/Central London tile (TQ28se).

The trend line in the scatter diagram shows a slight positive trend indicating the that as the number of users increase so does the road name attribute completeness, hence again supporting results of the 1st user analysis and Linus's Law. This is certainly the case for any grid square constituting more than 25 users, where the road name attribute completeness was no less than 60% with most results between 80-100%. However, the spread of results are very varied for all the grid squares with less than 20 users; in fact removing all results above 20 users would produce no correlation. Having said this, the vast majority of results appear to lie between 60-100% road name attribute completeness, and are therefore all good results. As mentioned in the literature review, the user analysis research carried by Dr Haklay found 89.5% of the whole of England covered by 3 or less users. Considering my results showed no grid squares as having less than 5 users, it is fair to say that these results are not a true reflection of the majority of England, with all my results being covered by more users than normal. This probably explains why the attribute completeness was generally quite high. Therefore it would have been useful to sample data from regions that are likely to contain 3 or less users, and see what effect this has on attribute completeness.

The results of the users analysis against average percentage overlap is shown by the scatter chart, Figure 28. There is no positive or negative trend here, and all that the results really shows is that positional accuracy is very high regardless of the number of users. This is probably due to a number of reasons. Firstly, it only takes a single user to achieve very high positional accuracy depending on the GPS equipment used and the nature of data collection. This was proved by my own GPS measurements as explained in section 3.2. Secondly, it was assumed that more users will increase accuracy as they will potentially add further measurements to what exists, and interpolate somewhere between their own results and the existing results. However, it is unlikely that users in same regions set out to measure features already mapped in OSM, and are far more likely collect new features than validate the positions of existing features, at least during the early stages of OSM where there is still a lot to map. Finally, the positional results were based on the averages of the results obtained from the buffer analysis which only consisted of major roads. Hence the issue of ecological fallacy is present here, where the positional accuracy of an entire 1km grid square is based only upon averages of a few of the roads that happen to run through it. A more accurate measurement would have been an average of the positional accuracies of all the road segments in any one grid square.

8.3 Evaluation of the project

Overall, this project has succeeded in providing an assessment of the positional accuracy of the OSM data sampled, and the results of the quality tests on datasets used for numbers of users and road name attribute completeness has provided a further assessment of OSM map data quality. However there were a number of things that could have been improved.

With regards to the datasets chosen, although they were spread across London and covered a large portion of road network the results of the user analysis showed that all these datasets were covered by high numbers of users. Ideally more tiles, particularly to the outskirts of London where there are likely to be fewer users, should have been tested with the buffer analysis. It would have been even better if OSM datasets in other areas of England were sampled as well, due to the fact that all positional accuracy analysis from this project and the previous Meridian project have been carried out in or around London. Having said this, matching both the OSM and ITN datasets was the most time consuming element of this project and it was therefore not possible to sample more datasets due to time restrictions. Also the results of the buffer analysis assessment were quite conclusive, and although further results would help to verify these findings more dataset samples would probably have been more useful for the user comparison analysis.

When carrying out the buffer analysis, a few issues with the OSM dataset became clear whilst preparing the datasets for the buffer overlap program particularly in terms of attribution. Firstly, some of the OSM roads were incorrectly attributed in terms of road type and this particularly occurred with B-roads, which were often incorrectly tagged as 'residential' roads rather than the correct label of 'secondary'. Secondly, as previously mentioned in the method (section 6.2) some of the roads were digitised incorrectly where one road segment was used to represent what should be two different roads with different road names. Secondly, it was clear that the OSM datasets lacked in detail in comparison to the ITN datasets in places, which for example had addition road segments to represent where a road widens or splits into further lanes at certain road junctions. An example of this at a round-a-bout junction is shown by the screenshots in Figure 33 below. However, this does not affect the positional accuracy of OSM but highlights lack of completeness.

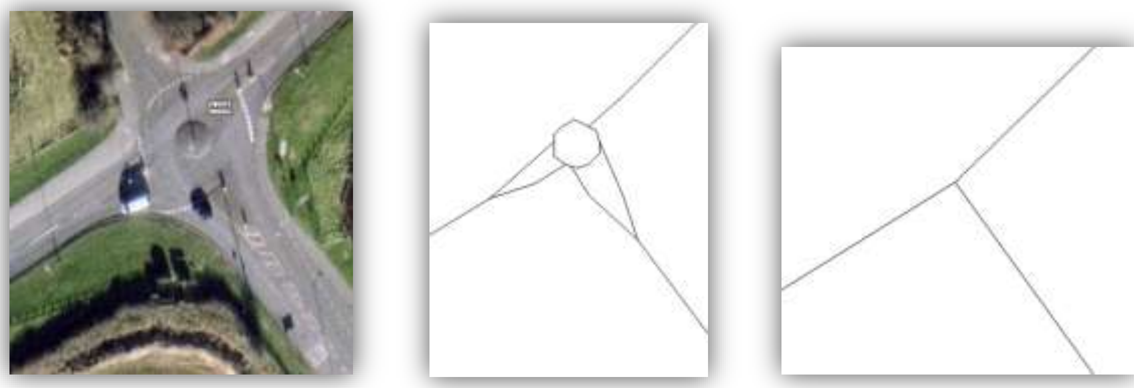


Figure 33 - Left: Round-a-bout junction (source: Google Earth), Middle: ITN representation, Right: OSM representation

The results of the buffer analysis could also have been affected to some extent due by human error in editing the datasets. For example, when the OSM data was edited some required road sections were accidently deleted. However, these usually were very short road sections and would have very little impact on the final results. It could be argued that some larger road sections may have been accidently deleted and not been noticed, but this unlikely as such missing sections would have been clearly visible from the validation that was carried out after both datasets were edited. Also, some errors could have occurred whilst labelling and matching road names between the two datasets as this was purely down to my own judgment. For example, at some complex road junctions where there were numerous slip roads leading off from major roads, it became difficult to match the corresponding roads to each other between the OSM and ITN datasets. On the whole however, most the labelling was found to be correct whilst carrying out the validation, so any errors here were likely to be minimal.

In terms of the road attribute completeness and user analysis, the OSM xml data obtained from the OSM slippy map was 'hairy' i.e. the roads were not clipped at the edges. Therefore, the attribute information for the extra road segments outside the test regions, defined by the OS tiles, may have been included as part of the results of the first user analysis, which looked at the datasets as a whole. Strictly speaking, only the data within the boundaries defined by the OS tiles should have been included for this analysis, in order to match the extents of the datasets used in the buffer analysis, as comparisons were drawn between the two results. Hence ideally, the 'hairy' OSM datasets should have been manually edited to size, but effects on the final results were judged to be marginal and the data was therefore left unedited.

With regards to the second user analysis, one problem already highlighted was that results of the positional accuracy were averaged from the buffer analysis results, which only included major roads. Ideally, the positional accuracy for each grid should have been based upon the average of all the roads in the gird square, although conducting a buffer analysis on all the roads in each grid square would have been far too time consuming. Finally, the comparison against road name attribute completeness could have also been carried out on the node and polygon shapefiles and on other attribute information, such as 'road type' which was blank in many cases. This would have provided a more complete picture of attribute completeness.

9. Conclusions

This project has succeeded in its aims of providing an understanding of OSM data quality whilst building upon the original OSM quality research previously conducted. The primary aim of this project was to build upon the positional accuracy analysis previously carried out on motorways, by extending the analysis to A-roads and B-roads whilst also using a higher level dataset for comparison; the OS MasterMap ITN layer. The secondary aim of this project was to carry out further map quality analysis on the datasets used for the positional accuracy assessment, by investigating the number of users per area and road name attribute completeness.

The initial GPS measurements collected by the author revealed that the OSM data collection and uploading process is fairly straight forward and accurate results are easily achievable depending on the GPS equipment used. The pilot test then found percentage overlaps above 90% for all the roads tested, and identified improved results from the comparison with OS Meridian 2. Four test regions around London were then chosen for the main quality analysis based on OS 25 km^2 grid tiles. The OSM and ITN datasets used for the four test regions were then downloaded and edited so that they matched exactly in terms of spatial extent and road name attribution. The buffer overlap analysis carried out on these datasets found the average percentage overlap to be greater than 80% for each of the four test regions with 57% of the roads examined registering results between 85-100% overlap. An evaluation of the road name attribution then found that test regions with the higher percentage overlap results also had greater levels of road name completeness. The results of the user analysis found that all the test regions chosen were covered by a large number of users, and a slight positive trend between number of users and road name attribute completeness was identified. However, the number of users did not appear to have any effect on the positional accuracy of the dataset.

On the whole these results proved that in terms of positional accuracy, OSM has the potential to deliver highly accurate results which can be achieved purely from the use of handheld GPS receivers. This is a good reflection of what can be accomplished by VGI applications; if such results are achievable from volunteers who may not be experts in the fields of surveying and cartography, then there is definitely great potential in utilising the concept of humans as sensors. OSM has most certainly capitalised on this global network of human sensors, as the project continuous to grow worldwide.

With regards to the map quality factors outlined in section 2.1 of the literature review, the findings of this project have proved that high levels of positional accuracy are achievable from OSM, in terms of comparability to OS MasterMap ITN layer, but there are still issues with regards to thematic accuracy and attribute completeness. This is due to the inconsistencies found in the attribution of the datasets tested. On this basis, there are still limitations as to what purposes OSM is fit for. Looking back at the list of common OS ITN applications (Table1, section 3.1) it is difficult to identify any one of these applications as being completely suitable for OSM, although this will vary from region to region. For well studied regions, like London, OSM could possibly be used for in-vehicle navigation due to the high levels of positional accuracy found from this research, but greater completeness in terms of road name and road type attribute completeness would still be required. OSM could also possibly be used for catchment-area analysis and site location, where the location of various roads is the most important factor.

Either way, it is important to remember that the main 'selling' factor of OSM is that the data is free and without licence constrictions, so until quality assurance can be made for the majority of the dataset it may only be perceived as acceptable by other open source developers, who do not take much liability for passing on any errors within the OSM data. But for organisations considering OSM for commercial purposes, such as selling GPS navigation products, a very strong set of quality assurance measures is will be required for them, as the consequences of passing on inaccurate information to customers may be severe.

There are a number of future research directions that could lead on from this project. Although further work is required in assessing OSM map quality over a greater sample of data, it is clear from both this project and previous OSM quality assessments that OSM is lacking in terms of attribute completeness and road network completeness. Whilst measures are already in place to for users to identify problems currently within the map (using tools like OpenStreetBugs), further research should focus on developing techniques that automatically highlight, or if possible, even fix problems when the data is originally uploaded and digitised. For example, any roads digitised without road names should be flagged before the data is rendered to the main slippy map. Such roads can then be viewed as a separate layer or even flagged in the main slippy map as incomplete, allowing browsers of the map to make corrections where required, hence utilising users with local knowledge.

In terms of positional accuracy, further research is required into the development techniques that can provide a more quantifiable measurement of map accuracy than the buffer analysis technique used for this project. For example, a technique could be developed which compares the distance between nodes at road junction, again using a higher level dataset like ITN for comparison.

As OSM grows, different regions will eventually fulfil the levels of map quality required for various applications at different times. Therefore, it would be useful if long term measures were in place to provide continued assessments of OSM map quality and then communicate these results back to users as they browse through the map. This could be a laborious task, but future research could involve investigation of techniques that would perform such quality assessments automatically. This would require a higher level dataset to sit in the background for comparison, but considering the Ordnance Survey are now starting to open up some their of data, through the likes their new OpenSpace API, this could be a possibility in the future.

10. References

COOTE, A. AND RACKHAM, L., 2008. Neogeographic data quality – is it an issue?. *Paper delivered at the AGI Geocommunity conference 2008*, ConsultingWhere Ltd.

FLANAGIN, A., J. AND METZGER, M., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72(3), 137-148.

GOODCHILD, M., F. AND HUNTER, G., J., 1996. Communicating uncertainty in spatial databases. *Transactions in GIS*, 1(1), 13.

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., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, (2007)(69), 211–221.

HAKLAY, M. AND WEBER, P., 2008. OpenStreetMap: User-Generated Street Map. *IEEE Pervasive Computing*, 7(4), 12-18.

HAKLAY, M., 2008. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. *Under review in Environment & Planning B.*

OORT, P.A.J. VAN, 2006. Spatial data quality: from description to application, PhD Thesis, Wageningen: Wageningen Universiteit, p. 125.

Ordnance Survey (2007) MasterMap Integrated Transport Network Layer User Guide and Technical Specification, Version 1.0 April 2007, Ordnance Survey, Southampton.

PARADIS, J. AND BEARD, K., 1994. Visualisation of Spatial Data Quality for the Decision- Maker: A Data-Quality Filter. *Journal of the Urban and Regional Inforamtion Systems Association*, 6(2), 25-34.

RAYMOND, E., 1999. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3), 23-49.

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(1), 1-5.

TURNER, J. A., 2006. Introduction to Neogeography. *Short Cuts, O'Reilly Media, 15th December 2006*, ISBN: 978-0-596-52995

ZULFIQAR, N., 2008. A Study of the Quality of OpenStreetMap.org maps: A comparison of OSM data and Ordnance Survey Data, MEng Thesis, London, University College London, p.54.

Website references:

EDINA. 2006. *Didimap OS MasterMap Userguide.* [online]
Available at http://edina.ac.uk/mastermap/support/MMUserGuide.shtml#Integrated [Accessed 11 May 2009].

ORDNANCE SURVEY. 2009a. *What is OS MasterMap.* [online]
Available at
http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/whatisosmm.html [Accessed 11 May 2009].

ORDNANCE SURVEY. 2009b. *OS MasterMap Integrated Transport Network (ITN) Layer.* [online]
Available at
http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/layers/itn/index.html
[Accessed 11 May 2009].

ORDNANCE SURVEY. 2009c. *The OS MasterMap Integrated Transport Network (ITN) Layer community forum.* [online] (Updated 15 April 2008)
Available at http://www.ordnancesurvey.co.uk/itnforum/thread.jspa?threadID=651&tstart=0
[Accessed 11 May 2009].

ORDNANCE SURVEY. 2009d. *Positional accuracy improvement programme.* [online]
Available at http://www.ordnancesurvey.co.uk/oswebsite/pai/faqmm.html [Accessed 11 May 2009].

ORDNANCE SURVEY. 2009e. *How often is the ITN Layer Roads Network updated?.* [online]
Available at http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/faqs/itn009.html [Accessed 11 May 2009].

O'REILLY.2009. *Innovation Through Open Location (Where 2.0 Conference).* [online] (Updated 20 May 2009)
Available at http://en.oreilly.com/where2009/public/schedule/detail/7245 [Accessed 26 May 2009].

PARLIEMENT, HOUSE OF COMMONS HANSARD WRITTEN ANSWERS. 2003. *Road Widths.* [online] (Updated 30 October 2003)
Available at
http://www.publications.parliament.uk/pa/cm200203/cmhansrd/vo031030/text/31030w08.htm#31030w08.html_wqn0 [Accessed on 17 May 2009].

PERSONAL COMPUTER WORLD. 2009. *Ordinance Survey offers free use of map data (News article).* [online] (Updated 15 May 2009)
Available at http://www.pcw.co.uk/personal-computer-world/news/2242355/ordnance-survey-offers-free-map [Accessed 26 May 2009].

STARPAL. 2005. *Starpal dealer bulletin (MobileMapper CE).* [online](Updated 18 December 2005)
Available at http://www.starpal.com/Reports/MobileMapper_CE_Bulletin.html [Accessed 14 May 2009].

THALES. 2005. *MobileMapperCE (Details and Technical specifications).* [online]
Available at http://www.dgps4u.com/Details&Specs/MobileMapper_CE-Details&Spec.pdf [Accessed on 13 May 2009].

TOPOGRAFIX.2009. *GPX for Developers*. [online]
Available at http://www.topografix.com/gpx_for_developers.asp [Accessed on 14 May 2009]

WIKIPEDIA. 2009a. *OpenStreetMap*. [online] (Updated 16 May 2009)
Available at http://en.wikipedia.org/wiki/OpenStreetMap [Accessed 05 May 2009].

WIKIPEDIA. 2009b. OS MasterMap. [online] (Updated 27 February 2009)
Available at http://en.wikipedia.org/wiki/MasterMap [Accessed 11 May 2009].

11.Appendix

Buffer analysis program:

```
Include "MAPBASIC.DEF"

'/**********declare procedures**********/

Declare Sub Main
Declare Sub OpenTable
Declare Sub ReadTextFile
Declare Sub SelectRoadBuffer
Declare Sub SelectRoadBuffer2
Declare Sub PercentageOverlap
Declare Sub ClearBuffers

'/***************variables*************/

Dim str As String
Dim str2 As String
Dim road_buffer1 As Object
Dim road_buffer2 As Object
Dim overlap As String

'/************Main Procedure***********/

Sub Main
'close all interactive
Call OpenTable
Call ClearBuffers
Call ReadTextFile
Call PercentageOverlap

End Sub

'/**********OpenTable Procedure*********/

Sub OpenTable

'/***buffer file for OSM

Open Table "E:\OSM_analysis\East_london_tile\TAB_files\osm_buffer"
Interactive
Note "osm_buffer opened"

'/***buffer file for ITN

Open Table "E:\OSM_analysis\East_london_tile\TAB_files\itn_buffer"
Interactive
Note "itn_buffer opened"

'/***open OSM tab file

Open Table
"E:\OSM_analysis\East_london_tile\TAB_files\OSM_roads_final_E.tab"
Interactive
```

```
Map From OSM_roads_final_E
Set Map Zoom Entire

Note "OSM Table Opened"

'/***open ITN tab file

Open Table
"E:\OSM_analysis\East_london_tile\TAB_files\ITN_roads_final_E.tab"
Interactive
Map From ITN_roads_final_E
Set Map Zoom Entire

Note "ITN Table Opened"

'/***open PercentageOverlap table

Open Table "E:\OSM_analysis\East_london_tile\TAB_files\PercentageOverlap"

End Sub

Sub ClearBuffers

'/***ensure that the buffer tables are empty before new data is added to
them

Delete from osm_buffer
Commit table osm_buffer
Pack table osm_buffer graphic data

Delete from itn_buffer
Commit table itn_buffer
Pack table itn_buffer graphic data

End sub

'/*********ReadTextFile Procedure*******/

Sub ReadTextFile

'/***select a motorway based on the text file and create a buffer for that
object

Open File "E:\OSM_analysis\East_london_tile\OSM_roadlist_E.txt" For Input
As #1
Note "OSM List File Opened"
    Do While Not EOF (1)
        Line Input #1, str
        If Not EOF (1) Then
            Call SelectRoadBuffer
        End If
    Loop
Note "OSM buffer tables created"

Close file #1
Open File "E:\OSM_analysis\East_london_tile\ITN_roadlist_E.txt" For Input
As #1
Note "ITN List File Opened"
    Do While Not EOF (1)
        Line Input #1, str2
        If Not EOF (1) Then
```

```
            Call SelectRoadBuffer2
        End If
    Loop

Note "ITN buffer tables created"


End Sub

'/*******SelectRoadBuffer Procedure******/

Sub SelectRoadBuffer
Dim buffertext_OSM as string

Select * From OSM_roads_final_E
Where Name = str into Selection

'/***the following code was obtained using the MapBasic window in Mapinfo
and edited to function in the code

buffertext_OSM = "Create Object As Buffer From Selection Width 0.1 Units" &
chr$(34) & "m" & chr$(34) &  "Type Spherical Resolution 12" &
" Into Table osm_buffer Data
ID=sum(ID),NAME=NAME,TYPE=TYPE,LENGTHI=sum(LENGTHI)"
Run command buffertext_OSM
Commit Table osm_buffer Interactive

End Sub

'/*******SelectRoadBuffer2 Procedure******/

Sub SelectRoadBuffer2
Dim buffertext_ITN as string
Select * From ITN_roads_final_E
Where DFT_NUMBER = str2 into Selection


buffertext_ITN = "Create Object As Buffer From Selection Width 3.75 Units"
& chr$(34) & "m" & chr$(34) &  "Type Spherical Resolution 12" &
" Into Table itn_buffer Data
ID=sum(ID),DESC_TERM=DESC_TERM,DFT_Number=DFT_Number,LENGTHI=sum(LENGTH)"
Run command buffertext_ITN
Commit Table itn_buffer Interactive

End Sub

'/******PercentageOverlap Procedure******/

Sub PercentageOverlap

'/***calculate the percentage overlap between the same motorway in both
datasets

dim i as integer
dim buffer2obj as object
dim buffer3obj as object
dim themotorwayname as string

'Note(tableinfo("osm_buffer", TAB_INFO_NROWS))

For i = 1 to tableinfo("osm_buffer", TAB_INFO_NROWS)
```

```
    select * from osm_buffer where rowid = i into onetest_buffer2
    fetch first from onetest_buffer2
    buffer2obj = onetest_buffer2.obj
    select * from itn_buffer where rowid = i into onetest_buffer3
    fetch first from onetest_buffer3
    buffer3obj = onetest_buffer3.obj
    themotorwayname = onetest_buffer2.name
    overlap = ProportionOverlap (buffer2obj, buffer3obj)

    insert into PercentageOverlap (ID,MotorwayName,PercentageOverlap)
values (i,themotorwayname,overlap)
    'print overlap
Next

Browse * From PercentageOverlap

End Sub
```

Final dataset maps (OSM & OS ITN):

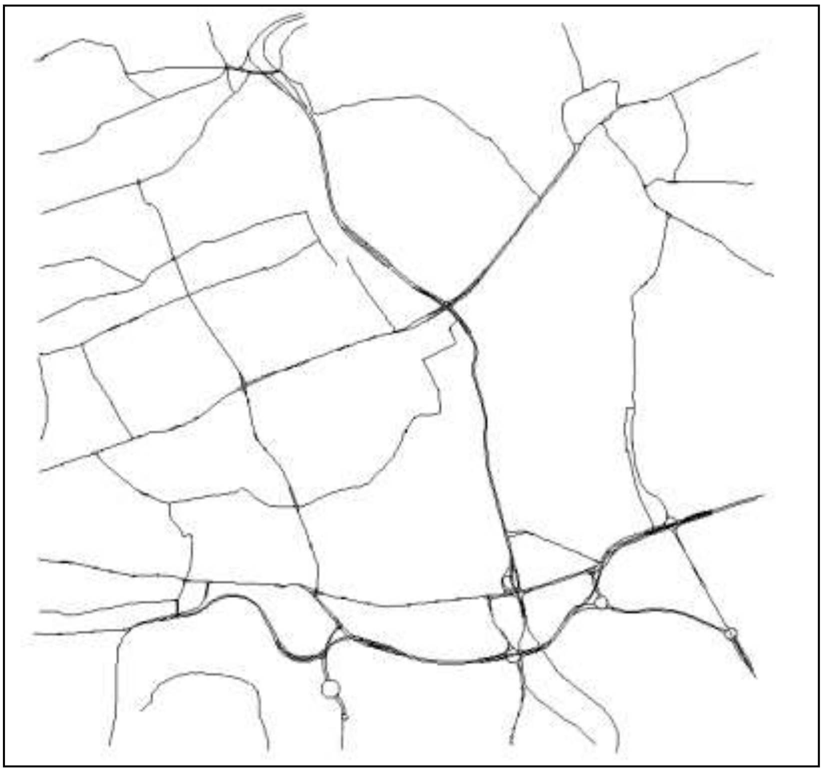


Figure 34 – East London (OS tile TQ38se)

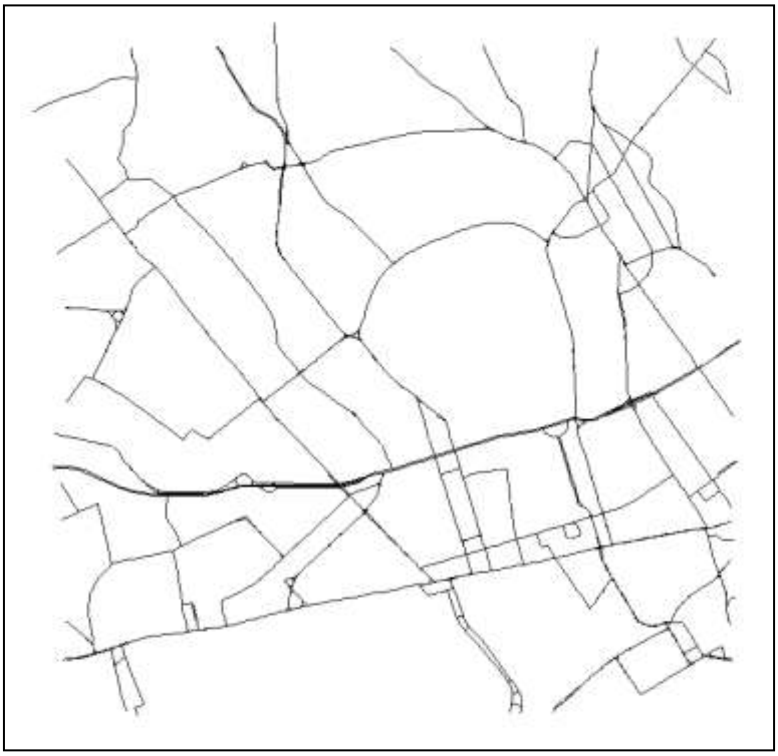


Figure 35 – North/Central London (OS tile TQ28se)

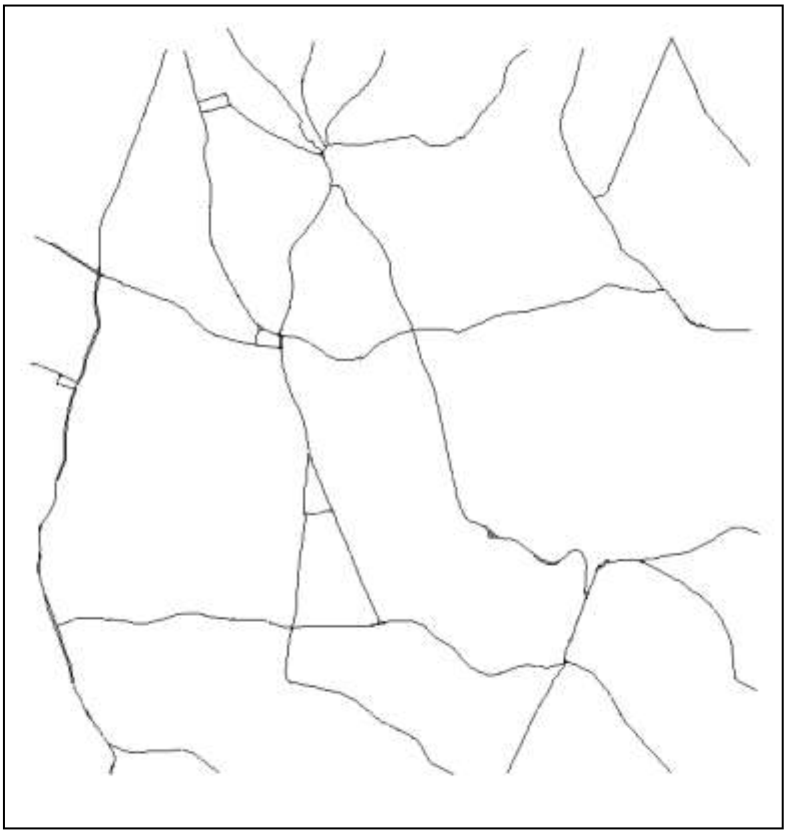


Figure 36 – South London (OS tile TQ37sw)

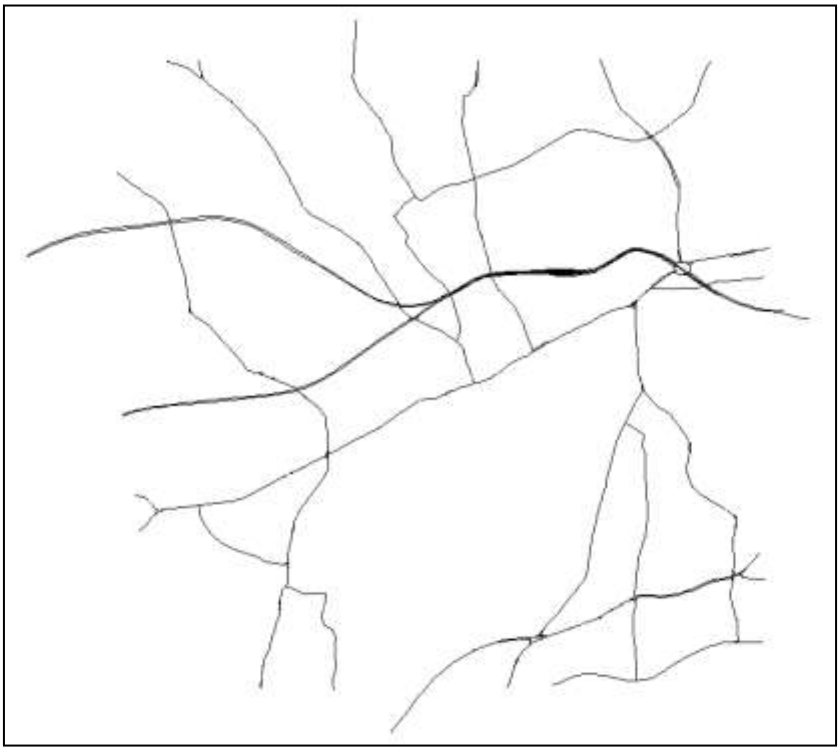


Figure 37 West London (OS tile TQ17ne)

Full table of results for 1km grid square analysis:

- The grid positions are given as Cartesian coordinates, where (1,1) represents the bottom left of the grid square.

Tile	Grid position	No. Users	No. road name features: named	unnamed	Total	% Attribute completeness	Average % overlap
TQ38se	1,1	14	56	27	83	67.47	79.75
TQ38se	1,2	18	57	39	96	59.38	80.2
TQ38se	1,3	26	76	43	119	63.87	84.5
TQ38se	1,4	24	37	49	86	43.02	77
TQ38se	1,5	17	25	21	46	54.35	98
TQ38se	2,1	14	14	97	111	12.61	85.6
TQ38se	2,2	15	97	36	133	72.93	83.25
TQ38se	2,3	10	82	25	107	76.64	82
TQ38se	2,4	18	41	43	84	48.81	88.17
TQ38se	2,5	17	63	45	108	58.33	87.2
TQ38se	3,1	13	64	45	109	58.72	78.5
TQ38se	3,2	18	64	9	73	87.67	95
TQ38se	3,3	13	52	22	74	70.27	90
TQ38se	3,4	15	14	31	45	31.11	91
TQ38se	3,5	8	39	26	65	60.00	82.5
TQ38se	4,1	16	71	15	86	82.56	65.2
TQ38se	4,2	15	76	10	86	88.37	83.2
TQ38se	4,3	20	52	27	79	65.82	94.33

TQ38se	4,4	18	30	21	51	58.82	94
TQ38se	4,5	16	88	16	104	84.62	87
TQ38se	5,1	16	79	21	100	79.00	85.5
TQ38se	5,2	19	91	13	104	87.50	75
TQ38se	5,3	16	27	1	28	96.43	91.5
TQ38se	5,4	21	38	23	61	62.30	94
TQ38se	5,5	19	103	13	116	88.79	89.5
TQ28se	1,1	17	102	12	114	89.47368421	91.67
TQ28se	1,2	21	70	6	76	92.10526316	88.67
TQ28se	1,3	20	31	12	43	72.09302326	81.75
TQ28se	1,4	28	128	10	138	92.75362319	83
TQ28se	1,5	40	186	3	189	98.41269841	79.6
TQ28se	2,1	23	95	17	112	84.82142857	93.67
TQ28se	2,2	31	117	31	148	79.05405405	87.71
TQ28se	2,3	37	164	31	195	84.1025641	85.18
TQ28se	2,4	30	203	20	223	91.03139013	80.63
TQ28se	2,5	39	224	9	233	96.13733906	84
TQ28se	3,1	13	69	2	71	97.18309859	82.75
TQ28se	3,2	17	86	13	99	86.86868687	81
TQ28se	3,3	22	78	35	113	69.02654867	86
TQ28se	3,4	26	96	25	121	79.33884298	78.33
TQ28se	3,5	33	142	47	189	75.13227513	93
TQ28se	4,1	19	50	2	52	96.15384615	82
TQ28se	4,2	11	71	9	80	88.75	82.25
TQ28se	4,3	19	54	10	64	84.375	77.5
TQ28se	4,4	22	70	11	81	86.41975309	84.4
TQ28se	4,5	29	105	6	111	94.59459459	92.4
TQ28se	5,1	20	33	4	37	89.18918919	63.5
TQ28se	5,2	21	58	7	65	89.23076923	82.75
TQ28se	5,3	16	53	2	55	96.36363636	85.5
TQ28se	5,4	21	103	7	110	93.63636364	82.8
TQ28se	5,5	23	100	12	112	89.28571429	90
TQ37sw	1,1	11	56	3	59	94.91525424	81
TQ37sw	1,2	10	35	24	59	59.3220339	84
TQ37sw	1,3	6	61	25	86	70.93023256	87
TQ37sw	1,4	8	113	9	122	92.62295082	89
TQ37sw	1,5	12	61	3	64	95.3125	88.5
TQ37sw	2,1	10	55	8	63	87.3015873	82.5
TQ37sw	2,2	15	74	38	112	66.07142857	87
TQ37sw	2,3	7	75	19	94	79.78723404	84.75
TQ37sw	2,4	7	73	24	97	75.25773196	95.5
TQ37sw	2,5	8	29	29	58	50	97.5
TQ37sw	3,1	18	64	7	71	90.14084507	79.5
TQ37sw	3,2	15	81	20	101	80.1980198	91
TQ37sw	3,3	10	53	19	72	73.61111111	90.5
TQ37sw	3,4	9	22	6	28	78.57142857	n/a
TQ37sw	3,5	7	35	24	59	59.3220339	n/a
TQ37sw	4,1	11	73	13	86	84.88372093	86.5
TQ37sw	4,2	15	66	26	92	71.73913043	93.33

TQ37sw	4,3	12	46	6	52	88.46153846	91.67
TQ37sw	4,4	8	28	3	31	90.32258065	91.67
TQ37sw	4,5	7	26	6	32	81.25	92
TQ37sw	5,1	14	73	12	85	85.88235294	85
TQ37sw	5,2	17	59	15	74	79.72972973	85.2
TQ37sw	5,3	10	51	4	55	92.72727273	89.2
TQ37sw	5,4	10	72	0	72	100	94
TQ37sw	5,5	6	37	3	40	92.5	81
TQ17ne	1,1	5	44	4	48	91.66666667	n/a
TQ17ne	1,2	16	52	5	57	91.22807018	83.25
TQ17ne	1,3	8	6	2	8	75	79
TQ17ne	1,4	16	46	6	52	88.46153846	82.5
TQ17ne	1,5	13	48	1	49	97.95918367	79
TQ17ne	2,1	6	63	8	71	88.73239437	76.5
TQ17ne	2,2	9	47	11	58	81.03448276	84.75
TQ17ne	2,3	8	0	3	3	0	n/a
TQ17ne	2,4	12	25	0	25	100	87
TQ17ne	2,5	12	41	7	48	85.41666667	72.67
TQ17ne	3,1	7	33	8	41	80.48780488	73.5
TQ17ne	3,2	8	36	10	46	78.26086957	79.67
TQ17ne	3,3	12	71	14	85	83.52941176	88.25
TQ17ne	3,4	14	37	11	48	77.08333333	79.6
TQ17ne	3,5	16	45	14	59	76.27118644	78
TQ17ne	4,1	9	3	4	7	42.85714286	85.5
TQ17ne	4,2	11	12	5	17	70.58823529	97
TQ17ne	4,3	17	61	7	68	89.70588235	81.83
TQ17ne	4,4	18	44	18	62	70.96774194	82.2
TQ17ne	4,5	19	57	16	73	78.08219178	82.6
TQ17ne	5,1	8	37	2	39	94.87179487	95
TQ17ne	5,2	7	77	2	79	97.46835443	93
TQ17ne	5,3	14	68	3	71	95.77464789	74
TQ17ne	5,4	13	31	14	45	68.88888889	63
TQ17ne	5,5	13	51	3	54	94.44444444	61.33

Table 13 - 1km gird square analysis results