ISPRS Int. J. Geo-Inf. **2013**, *2*, 704-728; doi:10.3390/ijgi2030704

OPEN ACCESS
ISPRS International
Journal of
Geo-Information
ISSN 2220-9964
www.mdpi.com/journal/ijgi/

Article

Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network

Amir Pourabdollah *, Jeremy Morley, Steven Feldman and Mike Jackson

Nottingham Geospatial Institute, The University of Nottingham, Jubilee Campus, Nottingham NG7 2TU, UK; E-Mails: jeremy.morley@nottingham.ac.uk (J.M.); steven.feldman@nottingham.ac.uk (S.F.); mike.jackson@nottingham.ac.uk (M.J.)

* Author to whom correspondence should be addressed; E-Mail: amir.pourabdollah@nottingham.ac.uk; Tel.: +44-115-951-5449; Fax: +44-115-951-3881.

Received: 5 June 2013; in revised form: 25 July 2013 / Accepted: 30 July 2013 / Published: 5 August 2013

Abstract: The quality aspects of OpenStreetMap (OSM), as the global representation of crowd-sourced mapping, have always been of priomary concern to academics. While the methodologies for checking its quality against the national maps have been implemented by a number of studies, there are minimal works on how to practically improve the quality of OSM towards being an authoritative map source. This paper presents a method for conflating road attributes, namely the name and reference code, of OSM with the Open Data provided by Ordnance Survey (the British national mapping agency). The added values in the proposed methodology include the daily updates and serving of the conflated maps via open Web Services. More importantly, the OSM crowd correction is facilitated by frequently highlighting and web-serving the individual differences. There are currently over 5,800 differences in matching road names and references between the two datasets. In addition to describing the conflation methodology, the different geographic distribution patterns of the identified differences are discussed. A negative effect of the road density on the ratio of the mismatched features between the two datasets is observable, evidenced by their different geographical distribution over the map. It is shown that the best correspondence between attributes exists in the very dense areas, followed by the very low density areas, and lastly in the middle to large sized cities.

Keywords: OpenStreetMap; national maps; road network; conflation

---

1. Introduction

Trust in a geographic dataset is based upon a sense of authority and the belief that the data are accurate. As interest in and usage of OpenStreetMap (OSM) has expanded, discussion on OSM's accuracy and fitness for purpose has largely been focused on the geometric accuracy and completeness of road networks rather than the quality of attribution. The release of OS OpenData within Britain has provided an open reference dataset against which the attribution of OSM can be calibrated and subsequently enhanced to produce a potentially more complete and accurate dataset. Whilst retaining the unique thematic content of OSM, the data inherit some authority from the national mapping agency's OpenData products. This paper sets out the methodology used to conflate OSM and OS VectorMap District (VMD) and provides an analysis of the distribution of differences between the two datasets.

This enhanced OSM product (available as raw data or as OGC web services) has been one of the outputs of the OSM-GB project [1] within the Nottingham Geospatial Institute. One of the primary objectives of this project is to identify quality improvements for the OSM and to encourage their incorporation into the main OSM dataset. The methodology of this research forms an integral part of the development of the OSM-GB project, which aims at promoting the potential usability of the OSM in authoritative contexts by making the OSM more trustworthy for professionals. The OSM-GB architecture is designed to perform the following tasks:

(a) Making a local mirror of OSM that is updated daily
(b) Serving the OSM contents (raster and vector) through standard Web Services
(c) Checking data quality by identifying auto-detectable and reference-based errors in a "rule/action-based engine"
(d) Fixing the errors in the local OSM mirror wherever possible
(e) Visualizing and serving the individual quality check results through standard Web Services for the purpose of actual error corrections by the community.

In this paper our analysis has been based upon conflating OS OpenData road network into OSM to increase the authority of OSM. However the same methodology can be applied in the opposite direction to potentially enhance an authoritative dataset with some of the rich thematic content (footpaths, cycleways, points of interest, building footprints and detailed attribution) from OSM to create new datasets combining authoritative and crowd sourced data.

2. Background

In this section the characteristics of the two source datasets will be briefly reviewed. This review will be used to design the conflation scenario and the methodology in the next sections.

2.1. OpenStreetMap

OpenStreetMap is a collaborative world mapping project. The users can freely map any area of the world in a Web 2.0 manner, and the resultant maps become instantly available for free public access across the globe. Users map the world using GPS traces, aerial imagery or their local knowledge. Moreover, the unrestricted use of key-value pairs for tagging all the features provides an excellent

means of customized annotations which is an approach suitable for thematic applications. OSM which was started in 2004 as a project and in 2006 as a foundation, has attracted over a million users thus far [2]. The annual growth rate of OSM worldwide (based on the number of point, lines and polygons) was approximately 75% in 2011 [3].

The simplicity of the OSM data structure originates from its integrated approach to modeling geographical features as well as being resolution-independent. Features are divided into nodes, ways and relations. A node is any one-dimensional feature such as a public amenity or a road vertex. Ways are any two-dimensional feature such as line boundaries (roads, coastlines, rivers, *etc.*) or polygons (lakes, farms, buildings, *etc.*). Finally a relation is a group of features of any type associated with each other in a defined relationship. Each feature has a unique identification number within its feature group. Nodes are the only features that have independent geometries (Latitude and Longitude) while the geometries of ways or relations are built from the nodes' geometries. For attribution, any feature can consistently have an unlimited number of key-value pairs which are called tags. Some of the key-value pairs that have been agreed by the OSM community (documented in the OSM Wiki [2]) have certain meanings for common OSM rendering tools (like Mapnik [4]). These community agreed tags are the most widely used, althouth the use of non-standard tagging is also permitted.

Focusing on the road network, OSM roads and their attributes are a subset of the OSM "ways". The "highway" key and its value are the most important road attribution, though the name "highway" is not always relevant. According to the OSM Wiki (at the time of writing) more than 40 standard values can be set for the "highway" key to define the road's type, varying from pathways to expressways. Some example values include "motorway", "primary", "residential", "footway" and "cycleway". Another useful standard key (without a standard value set) is "ref" which stores the reference number or code used nationally to identify a road (e.g., "M1"). The name of the road is also stored in the "name" key.

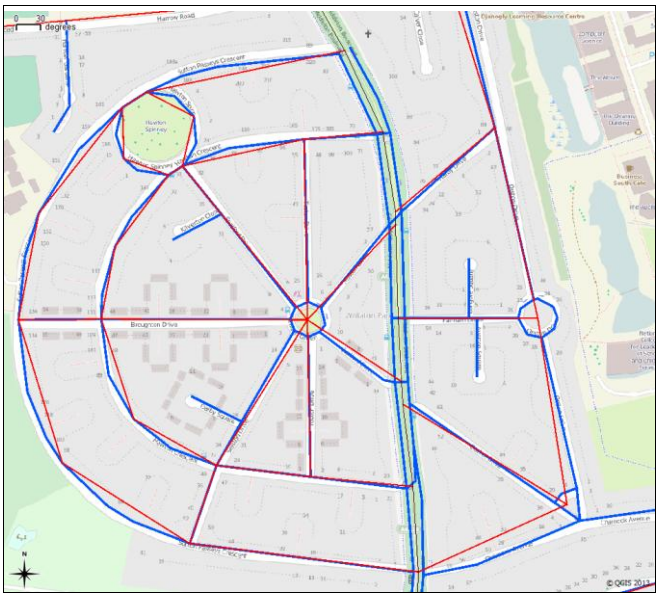
In this research, the road network is defined as the combination of all OSM "ways" having a non-null "highway" key. The "highway", "ref" and "name" are the three main attributes used in this research, alongside the geometries of the roads themselves.

2.2. OS Open Data

Since April 2010, Ordnance Survey has freely released to the public a set of raster and vector maps called OS OpenData [5]. This product set currently consists of three raster map products (MiniScale®, 1:250,000 Scale Raster and OS StreetView®) and eight vector map products (OS Locator™, 1:50,000 Scale Gazetteer, Boundary-Line™, Land-Form PANORAMA®, Code-Point® Open, Strategi®, Meridian™-2 and OS VectorMap® District). For conflating purposes, vector maps are required as these retain the feature details.

Meridian™-2 and OS VectorMap® District (called VMD hereafter) are the only two sets of thematic shape files that are relatively comparable in their detail with OSM. In general, OSM is more comparable to VMD than to the Meridian with reference to conflating purposes. Thus in this research, VMD is the main data set used among the OS OpenData products. An overlay of the Meridian and VMD roads datasets on top of an OSM background can be seen in Figure 1.

Figure 1. Comparing the roads in VectorMap District (VMD) (blue) and Meridian (red) on an OpenStreetMap (OSM) background. The two overlay maps are different in terms of completeness and accuracy.



In VMD, each road is defined by its geometry (in British National Grid projection), an ID number, classification (e.g., Primary Road or Local Street), DFT-Number (an alphanumerical code determined by the Department for Transport where applicable, e.g., M1 or A6514) and finally the road name.

2.3. Characteristics of OSM vs. *VMD*

Table 1 is a summary of characteristic differences and matching challenges between OSM and VMD, particularly when the road networks from the two datasets are going to be conflated.

Table 1. Comparing the characteristics of OSM *vs.* VMD.

		OSM	**VMD**
	Completeness	Generally higher details in urban areas. OSM has many more features than VMD (e.g., Footpaths, *etc.*)	Lower details in urban areas but uniformly distributed (as a reference map)
Quality Metrics	**Geometric/Attribute Accuracy**	Lower, since it is not made professionally.	More accurate, however different OS products have different positional accuracies (as exemplified in Figure 1)
	Temporal Accuracy	Instantly updated	Long updating intervals (6 months or more)

Table 1. *Cont.*

		OSM	VMD
Matching Challenges	**Reference System** (this may also lead to an extra positional drift after transformation from one to another)	EPSG:4326	EPSG:27700
	Road Geometry Structure	Continuously mapped	Segmented between junctions
	Road Classification	Freestyle: the road classification shall be extracted from a number of free text attributes (OSM tags).	Fixed schema: stored in a single field using a defined vocabulary
	Blank Road Names	There are some no-name roads in OSM particularly in rural areas where the user did not find the road name.	It is a part of the VMD specifications that some minor road names have been deliberately left blank.
	File Format	A single XML file including all the thematic layers (to be distinguished by the attributes)	56 shape Files (for each National Grid Reference square). Each shape file has 22 thematic sub-layers.

3. Related Studies

The related studies can be divided into OSM quality research and the studies regarding conflating geographical datasets. After reviewing the related studies, the position of this paper among them will be described.

3.1. OSM Quality

While the volunteers of OSM are highly motivated to generate geospatial contents for social-related reasons [6] and crowd contribution can become a credible source by the "many-eyes" principle, the credibility of VGI in general has been one of the main concerns for authoritative use cases [7]. The success, openness and freedom of OSM has made it a very good examination ground for researchers to study different collaborative mapping characteristics, such as the comparative accuracy and completeness analysis. A statistical comparison of an OSM snapshot in time with official maps was carried out by Haklay in [8]. This study reveals OSM's relative accuracy and/or completeness (also followed up and extended in [9,10]). That research shows that by 2010, 29% of England was covered by OSM, and that 80% of the motorways were mapped within 6 meter accuracy compared to Ordnance Survey's dataset. A dynamic analysis of the changes and the standard geospatial quality evaluation of OSM data has also been performed in [11], where the data for England have been compared among three different instances during 2009. In Ireland, the accuracy of OSM in a number of locations has been manually compared with Google and Bing maps [12] and many obvious examples of inconsistencies in the three studied sources were found.

According to ISO-19157 (Geographic information—Data quality [13]) the geospatial data quality elements are categorized as Completeness, Logical Consistency, Positional Accuracy, Usability, Thematic Accuracy and Temporal Accuracy. This paper does not focus on describing the concept of geospatial data quality, as this has been discussed in many other resources. Examples are the geospatial data quality reviews in [14,15], particularly the OSM-specific quality analysis that has been provided in [16]. The focused quality measure in this paper is the thematic accuracy (finding the missing attributes) and completeness (finding the mismatched attributes) within the road network theme.

In the context of this paper the word "bug" will be used for a difference between OS and OSM road attribution. This word has been defined in the context of the OSM-GB project as any deviation from the defined quality assurance rules. The authors acknowledge that in some circumstances these may not be errors. The aim is for bugs to be fed back to the OSM community for humans to check and fix as appropriate, in line with the OSM community's ethos.

3.2. Spatial Data Matching and Conflating

Conflating geographical datasets is a challenging task owing to the data used being from different domains, standards and schemas. However, due to the increasing number of spatial data providers conflation has become an unavoidable task. The application areas include data integration, change detection, enhancement and updating of one dataset based on the information contained in other datasets [17]. The data sources can be either raster or vector, yet each can also be in different formats and/or delivery standards (*i.e.*, Grid or GeoTiff for raster, and Shapefile or GML for vector).

The work in [17] presents a service-oriented conflation approach that adds the missing features from OSM to a reference map in Germany. The work in [18] also presents a vector-adjustment method for enhancing the OSM positional accuracy of the road network by comparing it with referenced satellite imagery in the US. The method has shown an 86% improvement in positional accuracy between the pre and post-conflation datasets.

Feature matching is a primary task in which geometry and attribution matching techniques can both be employed. The complexity of feature matching is not only due to geometrical or attribution inaccuracy, but also because a single object may be represented differently in different datasets [19]. A method of matching the OSM road network with a commercial map source in the US (NavTeq) is presented in [20] which is mostly designed for business purposes. A method of feature-based matching between OSM roads and ITN (Integrated Transport Network layer of MasterMap from OS, as an example of a Reference dataset) has also been provided in [21]. This method has been used to evaluate the statistics of road network completeness in urban and rural areas based on the lengths of the matching line features. Besides the geometrical feature matching, another approach is to use the ontology-based methods (e.g., in [22]) for merging features from crowd-sourced and authoritative domains.

3.3. The Position of This Paper

The position of this paper within the research domain will now be highlighted. The related studies have mainly focused on the positional accuracy or the completeness of the geometries in OSM when compared to qualified references. Moreover the research outcomes in such studies have provided either the conflation methodologies or the statistical results. The main observable research gap is how the

developed methodologies can potentially act towards an authoritative OSM. Thus the primary contribution of this paper is to integrate the full data sharing within the conflation methodology.

In other words, a more open approach is taken here. To further elaborate on the openness, it will be highlighted that the presented methodology is used to openly share the full details of the daily-updated mismatches as well as the conflated maps within the OSM community. This is partly because both map sources are open, and partly because open web services are utilized. As a result, the community is assisted to making a more authoritative OSM. For example, an OSM user can see the OSM, VMD and the individual mismatched features in a single GIS application (e.g., QGIS) and then easily apply the desired changes in their area of interest. On the other hand, the detailed output can be used to facilitate future research, e.g., examining and analyzing the geographical patterns of OSM shortcomings. In a wider context, the dynamic patterns of all detected bugs in OSM can be analyzed (e.g., in [23]).

In addition to the primary contribution, a number of secondary contributions also arise. Firstly, while the related studies have mainly focused on the geometrical accuracy and completeness, the study presented here focuses on the attributional aspects. Secondly, utilizing a programmable rule/action engine has made the conflation method dynamic and tunable. Finally, the results of the conflation processes reveal some specific geographical patterns as will be discussed in Section 7.

4. Conflating Scenarios

In general, both OSM and VMD can be enriched using each other. While VMD can potentially be enriched by conflating with OSM, the focus of this research is on enriching OSM with VMD. The reason for this direction of enrichment is that the enriched OSM can be instantly and freely digested by the public in order to modify the current OSM data store.

The key OSM/VMD differences and conflation issues have already been summarized in Table 1. Figure 2 shows an example of the relative completeness level of the road networks in urban and rural areas. It is noticeable that in both cases, each data source can potentially be improved using data from the other. Roads such as cycle paths or footpaths, which are extensively mapped in OSM, are by product definition not mapped in VMD. On the other hand, there are some VMD roads that are still unmapped in OSM. Another issue in comparing the two data sources is the updating issue. An example of such issues is shown in Figure 3.

Different scenarios can be designed for enriching OSM. Such scenarios should be designed to have both OSM and VMD details and capacities in mind. Since the focus in this study has been on the road features, the selected scenarios are defined as follows:

(a) Adding/correcting the OSM road names from the VMD road names;
(b) Adding/correcting the OSM's road references from the VMD DFT-Numbers;
(c) Flagging the unmapped roads in OSM which exist in VMD.

In scenario (c) above, an extension could be to add the unmapped roads from VMD to OSM, but the action is currently set to "flagging", *i.e.*, simply highlighting the unmapped road but not adding it to OSM. This can be particularly useful in rural areas where OSM is not rich enough compared to the urban areas. However adding the unmapped roads has been reserved for future work, due to a number of extra challenges involved:

(a) Adding a feature to OSM is normally carried out by OSM users within the central OSM data infrastructure. If a new feature is added to the conflated database, there will be some concurrency issues since a number of attributes including the identification number and the user information will not match between the two datasets.

(b) As with any other geometry conflation, applying this scenario may need many other adjustments on the existing geometries (including road and non-road features). A common adjustment is to snap the newly created roads to those already present. However, when this process has been initially attempted, the geometrical adjustments have generated a number of false geometries.

Figure 2. Comparing the completeness level of the road network in VectorMap District (brown) and OpenStreetMap (background) in an urban area (Nottingham city, **above**) and a rural area (around Lincoln, **below**).

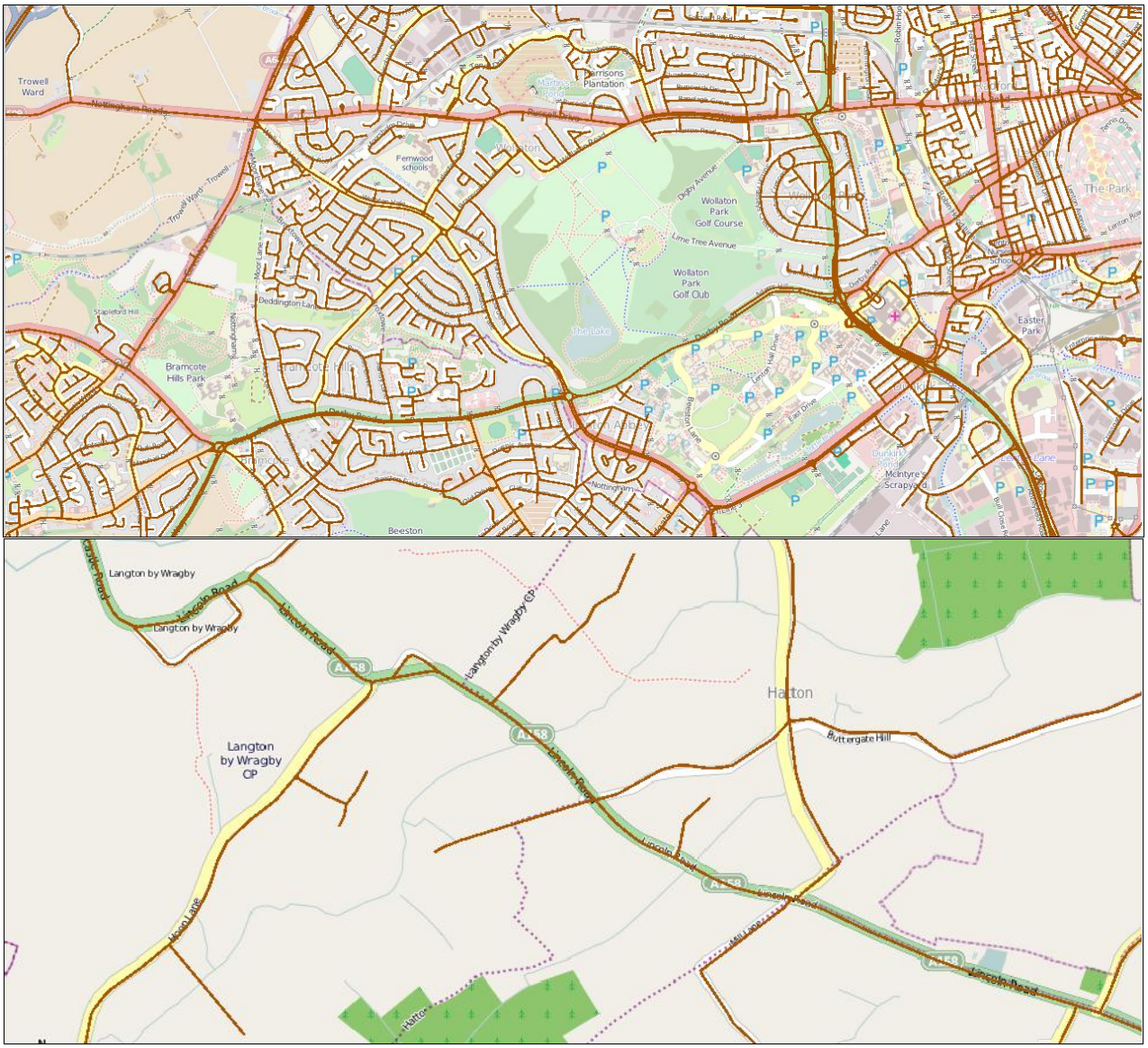
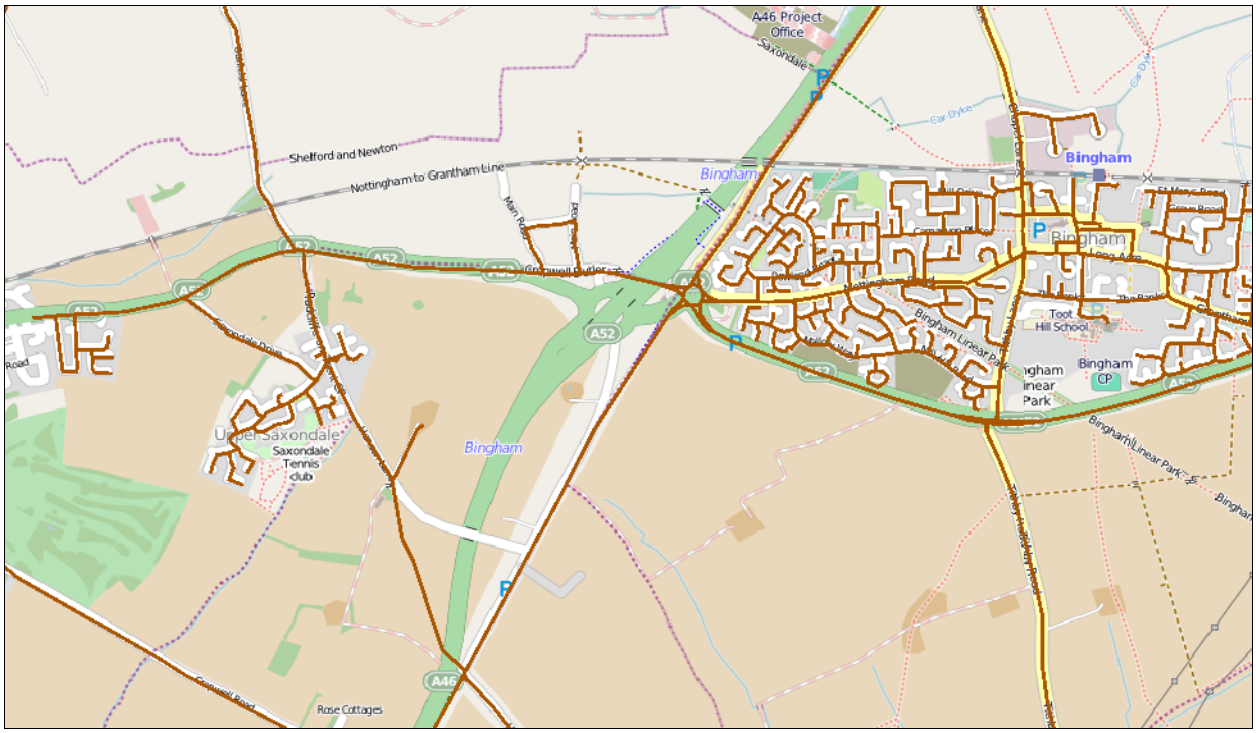


Figure 3. The problems associated with the VMD update cycle: The new road layout (of the A52 by Bingham, Nottingham) that is well mapped in OSM (background) has not yet been updated in VMD (brown).



Another challenge is the trust level to the national maps in that there are cases where the OSM contributors do not believe in the correctness of the national maps. In those cases the question will be how to perform the conflation. The current mechanism allows OSM users to tag a specific road to be ignored in conflation procedure (the key "OSMGB:isbug" shall be set to "no" for the particular OSM road). The tagged road will not be altered, though it is flagged as being mismatched.

5. Methodology

In this section, the details of the methods developed for the conflation scenarios described in the previous section will be presented. The position of this method within a wider project extent, the data management techniques and the developed rules/actions for each scenario will also be described.

5.1. The Integrated Approach

As mentioned before, the methodology used in this research forms part of an integrated quality processing framework (OSM-GB) in which a set of rules/actions are developed to check/fix bugs found in the OSM data. The rules/actions in this particular case are designed for comparing OSM and VMD. Once the specialized rules and actions are developed, the rest of the project components are able to apply the rules/actions on the local OSM mirror and serve the results back to the users. The geographical features stored within OSM are checked against each rule. If a feature fails to comply with the rule, the associated action is applied to the feature. For example, a rule may define the criteria to identify whether a feature is matched between VMD and OSM. The associated action will

then define how a missing road is added to OSM when the rule is not satisfied. The software used for this purpose is Radius Studio developed by 1Spatial [24]. The rules/actions are defined using a graphical user interface and/or the specific formal language designed for Radius Studio.

5.2. Data Management

The Radius Studio software requires that the data sources are in an Oracle Spatial database. It also writes the output back to the Oracle Spatial tables. Converting the OSM data into the Oracle Spatial database is done in two steps: the OSM2PGSQL tool [2] is used to load the OSM-XML daily backups (provided by Geofabrik [25]) into a PostGIS database, and OGR2OGR [26] is used to convert it from PostGIS to Oracle Spatial. The main reason behind the implementation of such an indirect conversion process is that the rest of the system in general relies on the PostGIS database. There has currently not yet been identified a different approach which is as efficient as this indirect solution.

To load VMD into Oracle Spatial, a number of additional steps are required. Ordnance Survey allows users to download (or order a CD copy of) VMD, which is a series of shape files for each of the 56 Ordnance Survey National Grid tiles (100 × 100 km per tile). The road network is encoded in one of the 22 sub-layers of each tile, thus a complete road network can be achieved by merging 56 shape files. SHP2PGSQL (an open-source tool in the Quantum GIS package distribution [27]), can be used to both merge and load the shape files into PostGIS. VMD data is needed to be stored in PostGIS due to the system database integration. It is then converted to Oracle Spatial in the same manner described for the OSM conversion. Once both OSM and VMD are converted to Oracle Spatial, Radius Studio applies the developed rules/actions and outputs the results into a secondary Oracle Spatial tables which contains the mismatches details. These tables are converted back to PostGIS using the OGR2OGR tool. A scheduled script applies the corrections to the original data in PostGIS, while backing up the corrected and uncorrected data for analysis purposes.

Having both corrected and uncorrected datasets on the PostGIS, the conflated maps and the individual differences are then served to the public via standard OGC Web Services (e.g., WMS and WFS). GeoServer is used to provide the data stored in the database as WMS and WFS in different coordinate reference systems. For example, a WFS is designed to server the conflated maps while another WFS is designed to serve the detils of the mismatched features. Moreover, the original OSM and VMD are also served as WMS and WFS. Because of the described web service integrity, all the served maps can be accessed consistently on the client side (e.g., using a single desktop GIS application), making the crowd-correction facilitated. More details on the utilized open-source solution can be found in [28].

The daily data transformation cycle discussed does not have any effect on the quality of the OSM data unless the developed actions are invoked. If no action is applied, the daily cycle preserves all the original geometries and attributes consistently. On the other hand, a rule that overwrites the mismatched geometry or attribute will not alter any other geometry or attribute in the dataset.

5.3. Rules/Actions for Adding/Correcting the Road Names/References

The rules developed for the scenarios relating to correcting the presence of road names and references firstly detect all the OSM roads that have an equivalent road in VMD. Checks are then performed against the values stored:

(a) If the VMD road has a name, the OSM road should have the same name.
(b) If the VMD road has a DFT-Number, the OSM road should have the same reference.

If an OSM road does not meet one of the above rules, the following actions will be applied:

(a) If failed by rule (a), the name of OSM road will be replaced by the name of the VMD road.
(b) If failed by rule (b), the OSM road reference will be replaced by the DFT-Number of the VMD road.
(c) In both cases, the detected errors and the changes in the name or reference values are noted in two separate attributes called “bug” and “fix” respectively.

In total, 4 rule/action couples are needed to allow for the correction of road names and references. For example, the following shows the formal language implementation of the rule for adding the missing road names in Radius Studio:

```
Check for OSM_LINE objects that
if OSM_LINE.name equals null
and OSM_LINE.highway does not equal null
and OSM_LINE.highway does not equal "cycleway"
and OSM_LINE.highway does not equal "pathway"
and (there is at least 1 VMD_ROAD object for which
    (VMD_ROAD.geometry is contained within buffer (OSM_LINE.geometry,0.0001)
    Or OSM_LINE.geometry is contained within buffer (VMD_ROAD.geometry,0.0001))
    and VMD_ROAD.name does not equal null
then
    to_lowercase(VMD_ROAD.name) equals to_lowercase(OSM_LINE.name)
```

The rule detects the OSM lines which (a) have no name and no “highway = cycleway” or “highway = pathway” tagging, (b) have a geometrically matched road from VMD, and (c) the VMD matched road has a name. The value of 0.0001 in line 7 (in degrees) is roughly about 7 m in the projected map. This means that the two roads are matched if the VMD road is inside a 7 m buffer of the OSM road. This distance is a heuristic value that has shown an optimized effectiveness of the rule based on the OSM and VMD characteristics. However, future work may be necessary to further refine this distance for an optimized algorithm. Cycle routes and pathways are excluded from the algorithm because firstly these road types do not exist in VMD and secondly they can be too close to other types of roads thus their name can be changed by mistake.

The following source code shows the implementation of the action associated with the roads that do not meet the above rules, in Radius Studio:

```
For OSM_LINE objects
    for the first VMD_ROAD object for which
           (VMD_ROAD.geometry is contained within buffer (OSM_LINE.geometry,0.0001)
           Or OSM_LINE.geometry is contained within buffer (VMD_ROAD.geometry,0.0001))
           and VMD_ROAD.name does not equal null
           and to_lowercase(VMD_ROAD.name) does not equal to_lowercase(OSM_LINE:A.name)
    Create an object of class OSM_LINE_CORRECTED and
    {let OSM_LINE_CORRECTED.{all common attributes} = OSM_LINE.{all common attributes}
    let OSM_LINE_CORRECTED.name = VMD_ROAD.name
    let OSM_LINE_CORRECTED.BUG = "Null road name"
    let OSM_LINE_CORRECTED.FIX = "Name changed to " + VMD_ROAD.name}
```

The object class of OSM_LINE_CORRECTED (line 6 above) is an extension to the OSM_LINE class with two extra attributes called “bug” and “fix”. The system is designed in a way that the outputs are stored in a separate database table called OSM_LINE_CORRECTED. The detected errors can be managed and served independent of the original data using this table before applying the changes to the original data when needed.

5.4. Dealing with the Unmapped Roads in OSM Which Exist in VMD

The rule for detecting road features that are present in the VMD but not in the OSM is very similar to the rule implemented for the identification of features that need correction of names or references. The first stage is to detect the roads in the VMD which do have geometrically matched roads within OSM:

```
Check for VMD_ROAD objects that
    there is at least 1 OSM_LINE object for which
           OSM_LINE.highway does not equal null
           and OSM_LINE.highway does not equal "cycleway"
           and OSM_LINE.highway does not equal "pathway"
           and (VMD_ROAD.geometry is contained within buffer (OSM_LINE.geometry,0.0001)
           Or OSM_LINE.geometry is contained within buffer (VMD_ROAD.geometry,0.0001))
```

Roads that are present in VMD but not in the OSM can then be deemed as “missing”. As described in the conflation scenarios, an action can be designed to actually add the missing road to OSM. This action was found to be challenging since it produced some false geometries, mainly because of the unpredictable positional differences between the two datasets. However a partial adjustment solution is provided here, but it requires further development (the currently implemented action is just to flag it as a missing road):

```
For VMD_ROAD objects:
    if (there are no PLANET_OSM_LINE objects for which
           OSM_LINE.highway does not equal null
           and OSM_LINE.highway does not equal "cycleway"
           and OSM_LINE.highway does not equal "pathway"
           and (VMD_ROAD.geometry is contained within buffer (OSM_LINE.geometry,0.0001)
           Or OSM_LINE.geometry is contained within buffer (VMD_ROAD.geometry,0.0001))
    then create an object of class OSM_LINE_CORRECTED and
           let OSM_LINE_CORRECTED.BUG = "Unmapped road"
           let OSM_LINE_CORRECTED.FIX = VMD_ROAD.name
           let OSM_LINE_CORRECTED.geometry = VMD_ROAD.geometry
           for all OSM_LINE objects for which
                  (start_of(OSM_LINE_CORRECTED.geometry) is within a distance of 0.0001 of
                  OSM_LINE.geometry
                  and it is not the case that OSM_LINE_CORRECTED.geometry equals
                  OSM_LINE.geometry)
                         let OSM_LINE_CORRECTED.geometry =
                         move_vertex(OSM_LINE_CORRECTED.geometry,start_of(OSM_LINE_CORRECT
                         ED.geometry),nearest_point(start_of(OSM_LINE_CORRECTED.geometry),
                         OSM_LINE.geometry,true))
           for all OSM_LINE objects for which
                  (end_of(OSM_LINE_CORRECTED.geometry) is within a distance of 0.0001 of
                  OSM_LINE.geometry
                  and it is not the case that (LINE_CORRECTED.geometry equals
                  OSM_LINE.geometry))
                         let OSM_LINE_CORRECTED.geometry =
                         move_vertex(OSM_LINE_CORRECTED.geometry,end_of(OSM_LINE_CORRECTED
                         .geometry),nearest_point(end_of(OSM_LINE_CORRECTED.geometry),
                         OSM_LINE.geometry,true))
```

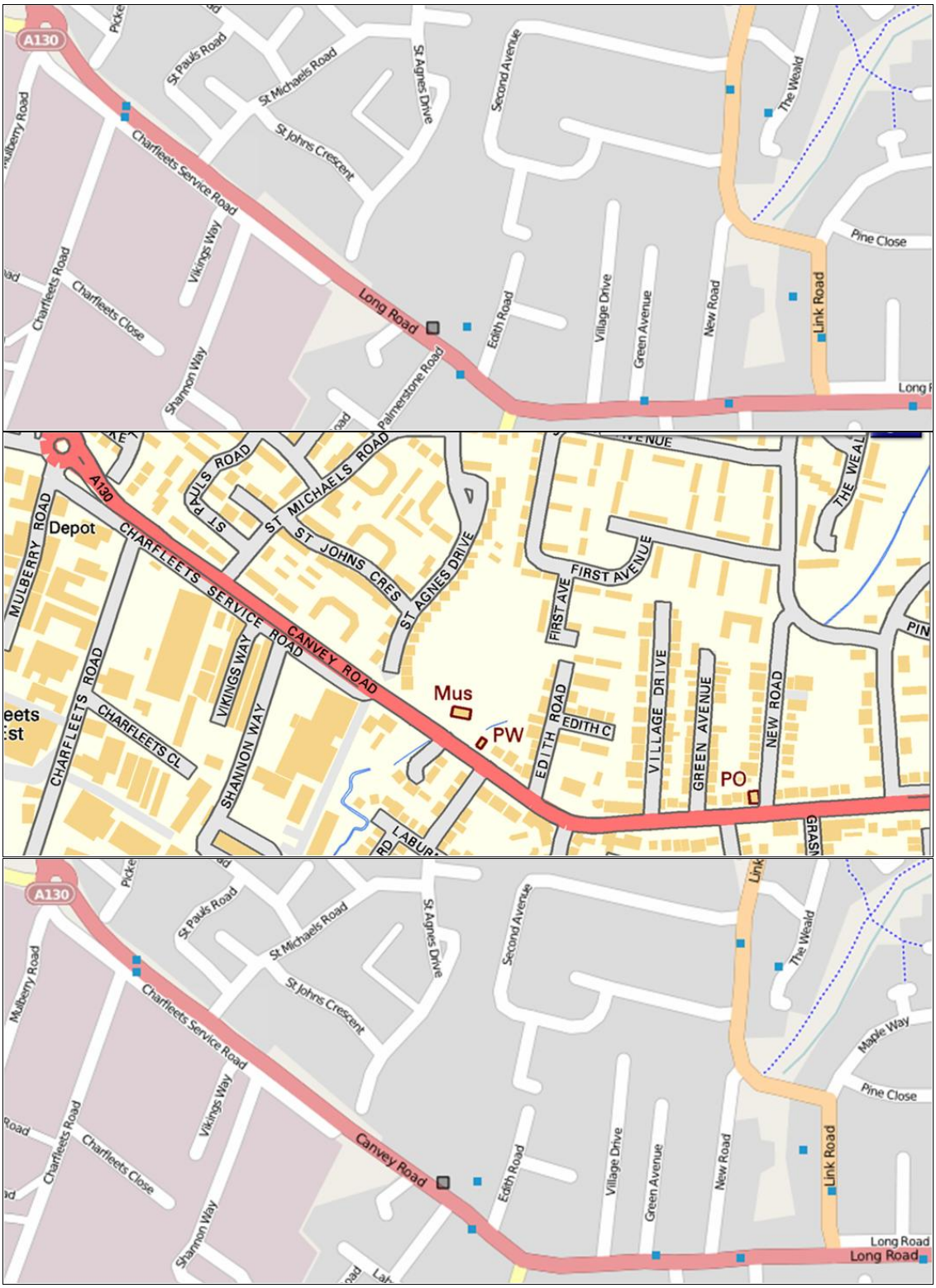
Once the unmapped road is added from VMD, the main issue is snapping the added geometry to the existing road network. If this is not performed the added road may not be geometrically connected to the rest of the road network. This is implemented in the above action after adding the unmapped road using a number of in-built functions including move_vertex(), nearest_point(), start_of() and end_of().

6. Results

6.1. Correcting the Mismatched Road Names

2,471 OSM roads are currently found in the OSM of Britain whos names do not match with VMD. Those OSM road names have been replaced by the VMD road names. A sample conflation is shown in Figure 4.

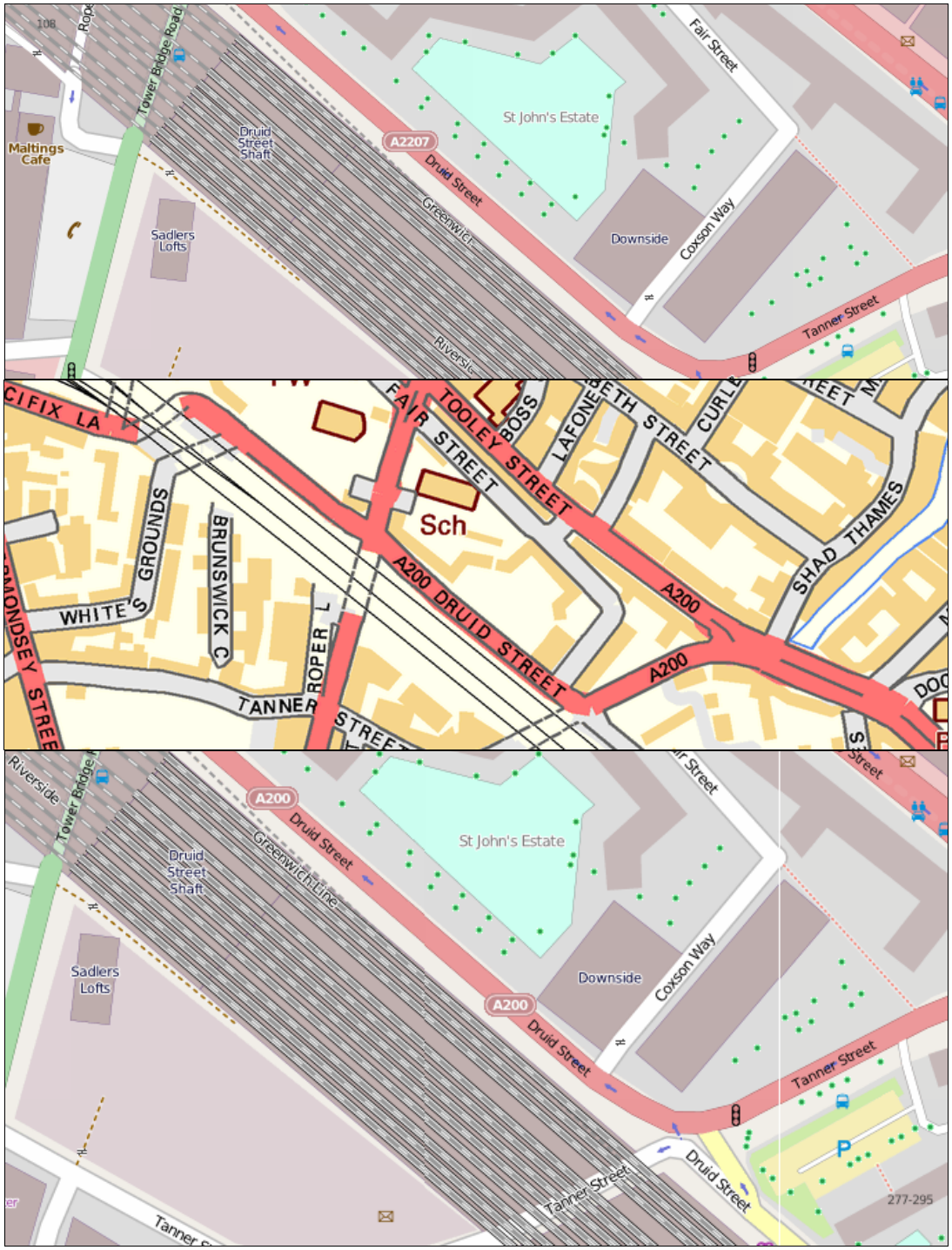
Figure 4. A sample of matched road names (Long road should be Canvey Road, near London). (**Top**): original OSM; (**Middle**): OS Open Data map of the same area; (**Below**): updated in OSM-GB.



6.2. Correcting the Mismatched Road References

Currently there are 377 road references in OSM that are different from the matched roads in VMD. Figure 5 shows an example of those mismatches.

Figure 5. A case of mismatched road references: Druid Street (near Tower Bridge, London) referenced as A2207 in OSM (**Top**) reads A200 in VMD (**Middle**) and updated in OSM-GB (**Below**)—However it is not clear why a parallel street above this is also referenced as A200 in VMD (middle).



6.3. Adding the Missing Road Names

Currently 2,026 missing road names in OSM have been taken from the matched VMD roads. A sample correction is shown in Figure 6.

Figure 6. A sample of added road names (Faversham Road, near Ashford). (**Top**): original OSM; (**Middle**): OS Open Data map of the same area; (**Below**): updated in OSM-GB.

6.4. Adding the Missing Road References

Currently there are 1,121 missing road references in the OSM that have been fulfilled using VMD data. An example is shown in Figure 7.

Figure 8 shows the geographical distribution of the mismatched/missing road attributes.

Figure 7. A sample of added road names (A594, in Leicester). (**Top**): original OSM; (**Middle**): OS Open Data map of the same area; (**Below**): updated in OSM-GB.

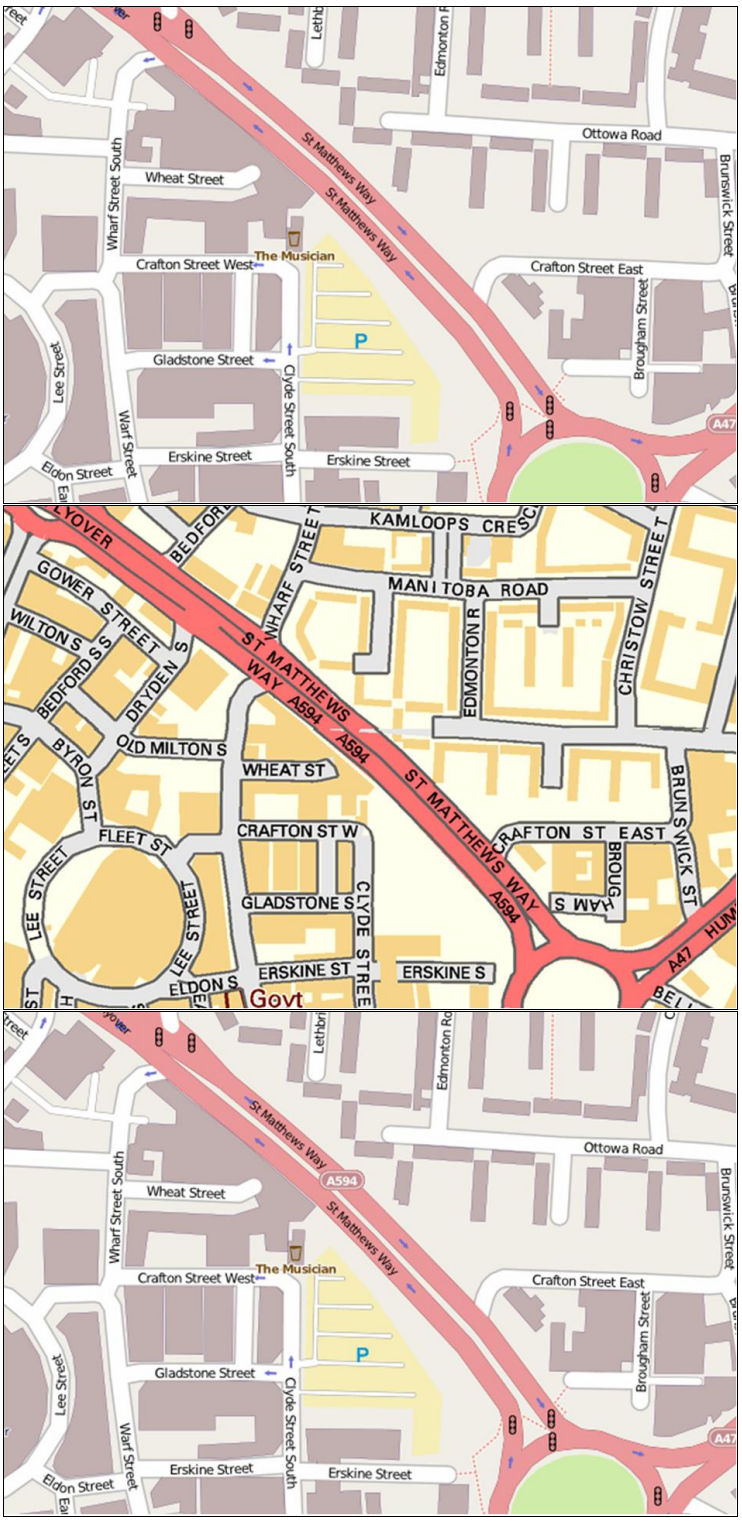
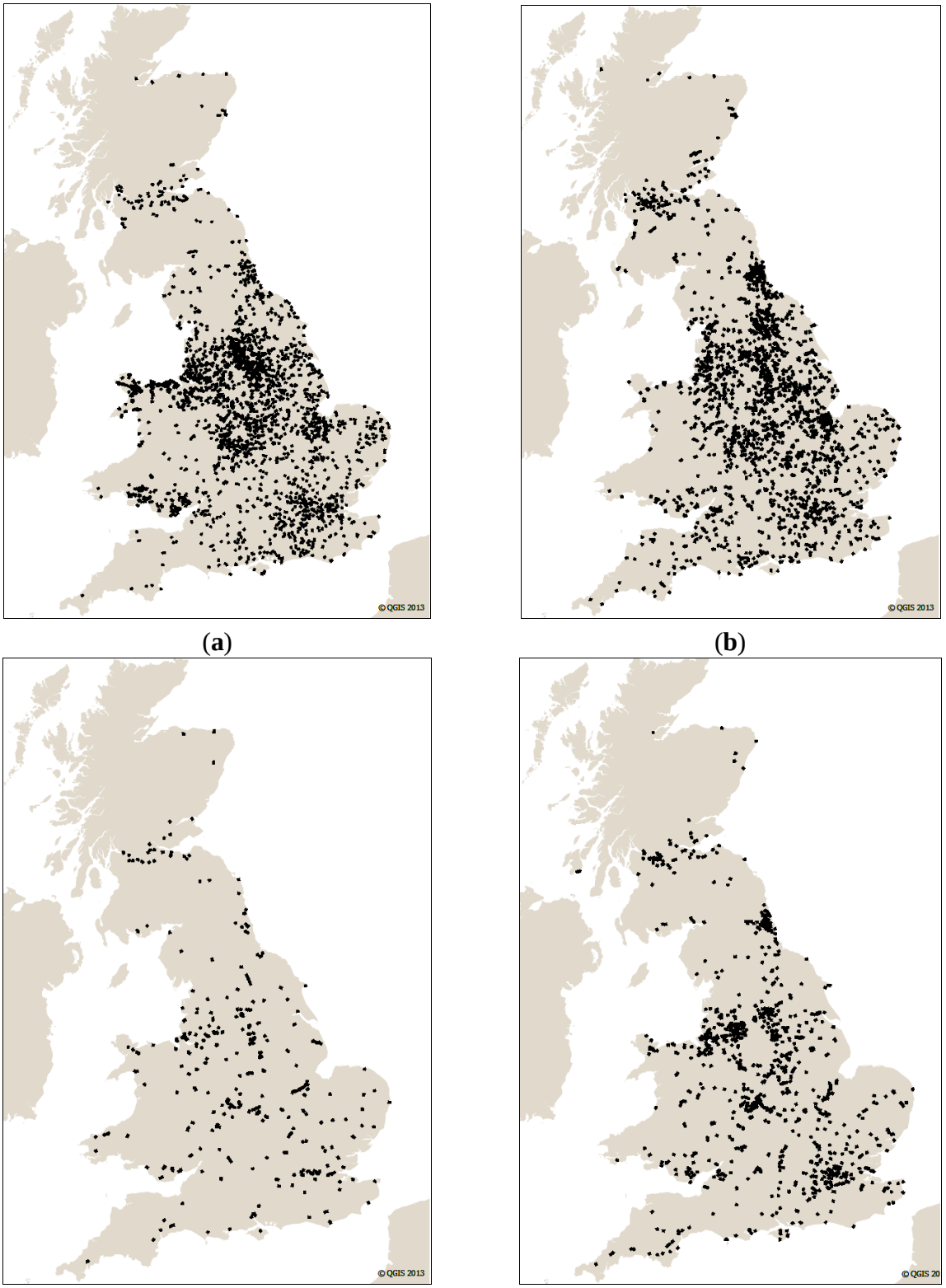


Figure 8. The geographical distribution of (**a**) Mismatched road names; (**b**) Missing road names; (**c**) Mismatched road references; (**d**) Missing road references.



(**a**) (**b**)

(**c**) (**d**)

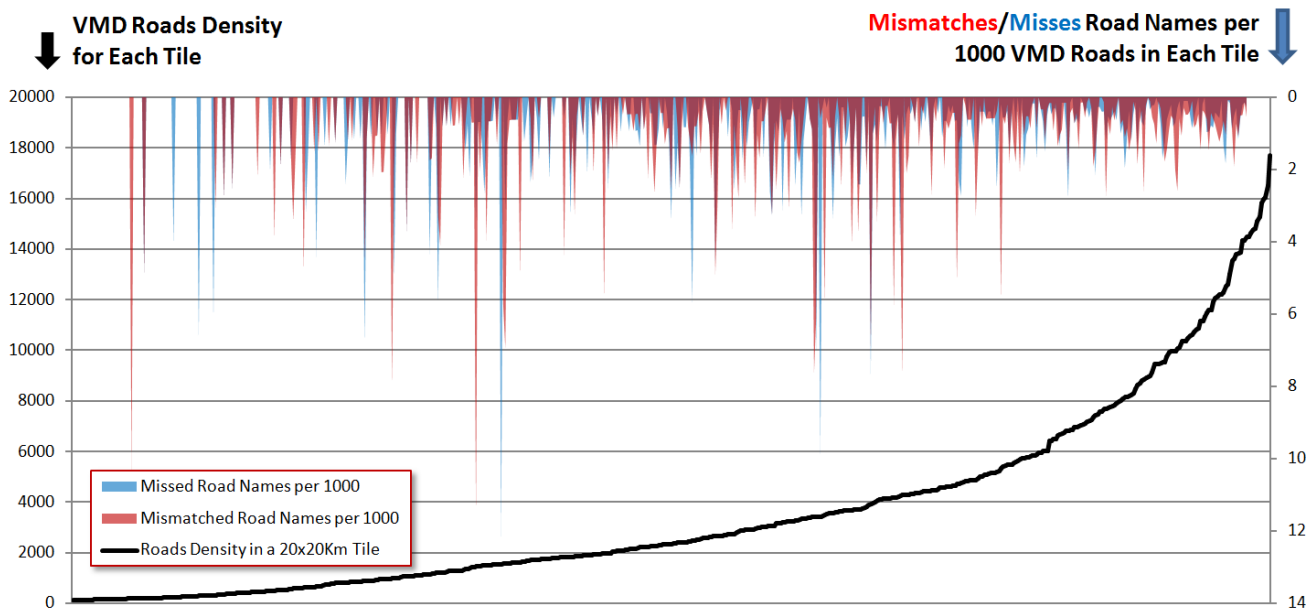
7. Discussion

The distribution of the mismatched and missing road names and references (so called "bugs") illustrated in Figure 8 can be further analyzed to explore the quality of the OSM road attribution. To perform such an exploration, the bug rate needs to be normalized and then the proportion of buggy roads to the total number of roads in an area (which has been termed as bug ratio) examined. This bug ratio inversely indicates the accuracy and completeness of OSM in that area. In that calculation, the total numbers of roads are taken from the VMD (because unlike OSM, the completeness of VMD is supposed to be equally distributed in all areas). It would then expected that the bug ratio may be higher in the rural areas due to less availability of public contributors to the OSM dataset. This hypothesis can be examined for both bug types and attribute types (mismatches and missing values).

To evaluate the validity of the above hypothesis, a zonal analysis was performed. The zone size selection may also have impacts on the analysis result. To differentiate between urban and rural areas, the selected zone size was chosen to cover about a middle to large sized city in each zone. For this reason, the Britain map is divided into approximately 1,200 squares, each 20 × 20 km size. The grids are then sorted by their VMD road density.

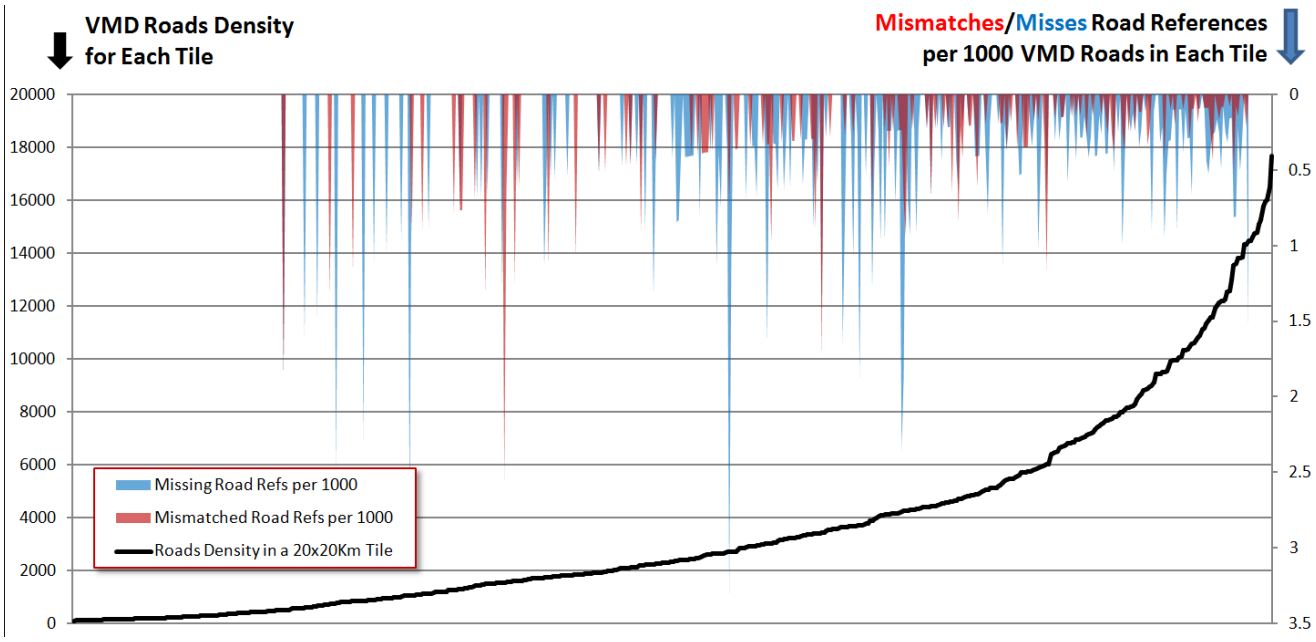
Figure 9 shows the results of the analysis of the bug ratios. This graph firstly shows that the mismatches (red) and the missing (blue) bug ratio peaks mostly lie in lower density areas (an inverse effect of density on the bug ratio). Secondly it shows that the middle-sized cities have more bug ratios than the big cities like London.

Figure 9. The comparison of the patterns of road names bug ratios and the road densities. The x-axis indicates the tiles sorted by the roads density. The dropping down peaks show the two types of bug ratios. Upper plot (**a**): bug ratios for the road name; lower plot (**b**): bug ratios for the road references. Red: mismatched attributes; blue: missing attributes.



(**a**)

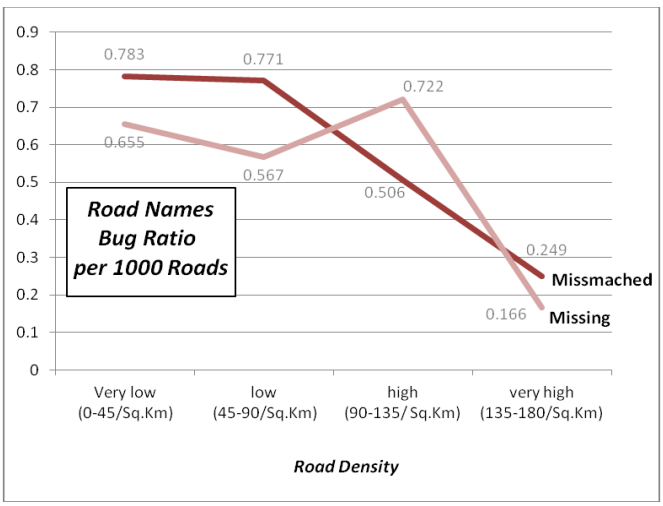
Figure 9. *Cont.*



(**b**)

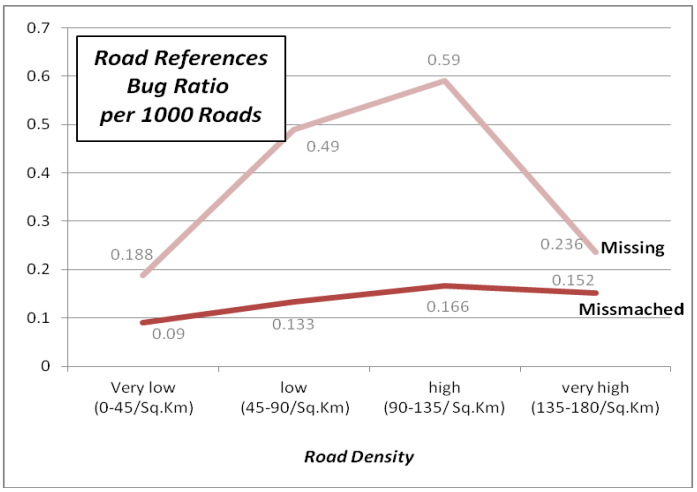
In order to quantify the results, the road density range (0 to 72 K roads per tile) was divided into four equal-interval bands, namely very low (0–18 K), low (18–36 K), high (36–54 K) and very high (54–72 K). Since each tile is a 20 × 20 km square, the intervals are equivalent to road densities of 0–45, 45–90, 90–135 and 135–180 roads/km^2 respectively. Practically, the "very high" band is limited to the London area. Some big cities are in the high band and the other cities are in the low band. The "very low" band is the most frequent one, comprising mostly of rural areas. Figure 10a,b shows the patterns of bug ratios according to the above road density bands.

Figure 10. The graphs of bug ratios per 1,000 roads according to the four density bands. (**a**) The mismatched and missing road names; (**b**) the mismatched and missing road references; and (**c**) the total bugs ratio.

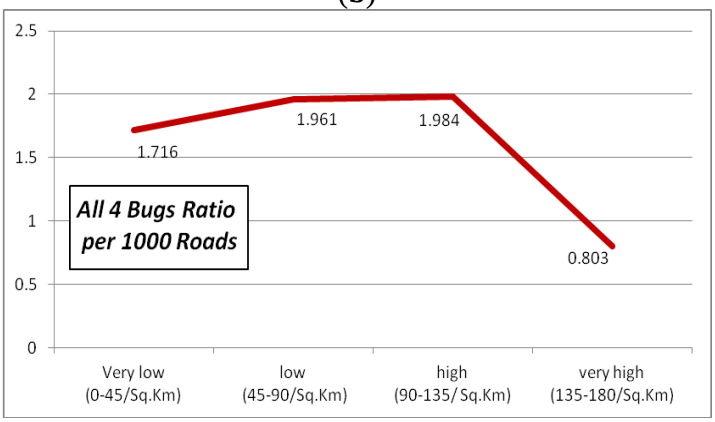


(**a**)

Figure 10. *Cont.*



(b)



(c)

Figure 10a, illustrating the road-name bug ratios, shows two different patterns for the mismatched and missing road names, particularly in the high-density band. This shows a peak in the ratios of missing road names in the relatively big cities excluding London, while otherwise the bug ratios fall with the density. Figure 10b shows that our expectation that the bug ratios should fall with increased feature density is valid only for the road references in the "very high" density band. In the three lower bands, the ratios of missing and mismatched road references increase with road density. The missing road references line shows approximately a 50% drop from the high to the very high band, however the drop in the rate of mismatches is not significant.

Finally when all the bugs are aggregated, as shown in Figure 10c, the density zones can be ordered by general road attribute quality The best quality exists in the very dense areas, then in the very low density areas, and lastly in the middle to large sized cities.

In order to explain the patterns shown in Figure 10, the assumptions about the VMD quality first need to be recalled. If the rate of missing road names or references in VMD is not as spatially uniformed as assumed, it can affect the related bug ratios. Secondly, mismatches (in road names or references) are caused by real differences between ground facts and official maps or by mistakes by the OSM mappers. Missing road names/references may come about by feature tracing in the OSM editors over imagery without ground verification or other sources of attribute data.

The bugs related to road names might be expected to show different distributions *versus* those related to the road references since their data collection processes are different. The references are specific alphanumeric codes that the OSM mappers can find from official sources or signposts, whilst the names can be collected from other origins. The rate of entering a wrong road reference (assuming the validity of the VMD data) is low and relatively constant across all the density bands—if a reference has been supplied, it is likely to be correct. The chance is higher for missing road references, and this chance has been shown to be even higher in the cities compared to the very low or the very high density zones. The graphs may also be showing the relative importance of the road referencing *versus* the road naming in the different area types. It could be said that the road references are more important attributes of roads outside of conurbations (and indeed, road names may not be clear in the areas), while in towns and cities the road names tend to be more important attributes. There is a further asymmetry where more major routes have references but may not have names while more minor roads are more likely to have names than references. London may be a different environment from other regions of the country not just because of the different road densities (and relative proportions of minor and major roads) but also because of the much greater concentration of contributing OSM users to provide quality assurance.

8. Future Works

The OSM-GB platform has made available a wide range of analytical research on the OSM data and its quality, of which this research is just an example. Regarding the conflation of OSM and OS, the main areas for future works are:

(1) Adding the missing roads and correcting the attributes of OSM roads according to a variety of OS map sources.
(2) Working on other themes, e.g., buildings, POIs, waterways, *etc*.
(3) Making enriched versions of OS maps e.g., adding cycle routes or footpaths from OSM.
(4) Taking into account other attributes and analyzing their effects on the OSM quality. An example is using version_no attribute of OSM (which shows how many times a map feature is edited by the users) and analyzing its correlation with the bug ratios discussed in this paper.
(5) A more in-depth investigation of the bug ratios discussed in this paper to examine how reference data, means of working, real-world coverage of road references and names, zone selection and the varying density of contributors differently contribute to the patterns observed here.

9. Conclusions

In this paper, two map sources of the British road network from two relatively different origins have been matched: OpenStreetMap (OSM) as the crowd-sourced and Ordnance Survey's Vector Map District (VMD) as the officially-sourced datasets. By analyzing the two main road attributes (name and reference) cases where the name or the references are missing in the OSM or mismatched between the two maps (called bugs) have been highlighted. The result is an enriched OSM with added or fixed road names and references which is served via standard Web Services in full details. The observed patterns are:

(1) The higher the road density, the lower the mismatched road names ratios;
(2) The missing road name ratios are high in the large cities, low in the urban and interestingly very low in the capital;
(3) The mismatched road reference ratios are generally higher in the large cities than the low density areas;
(4) The missing road references ratio grows by the road density but significantly drops in the capital;
(5) In total, the best quality is in the very dense areas, then in the very low density areas, and lastly in the middle to large sized cities.

The methodology presented here can be used for different themes in comparing between official and crowd-sourced maps. Analyzing the map matching data can reveal the different patterns of quality in map generation particularly on the crowd-sourced side.

Acknowledgments

This research is based at the Nottingham Geospatial Institute at the University of Nottingham, and is funded and supported by 1Spatial and KnowWhere. The authors also wish to acknowledge the collaborations from Ordnance Survey GB, Snowflake Software and Pitney Bowes. The comments from the anonymous reviewers are highly appreciated too. Finally, the authors would like to thank Adam Rousell who has kindly reviewed and commented on this paper.

Conflict of Interest

The authors declare no conflict of interest.

References

1. OSMGB. *OSM-GB Project Homepage—Measuring and Improving the Quality of OpenStreetMap for Great Britain*. Available online: http://www.osmgb.org.uk (accessed on 1 June 2013).
2. OSM-Wiki. *OpenStreetMap Project Wiki*. Available online: http://wiki.openstreetmap.org/wiki/ (accessed on 1 June 2013).
3. BeyoNav. *OpenStreetMap's Growth Accelerates*. Available online: http://www.beyonav.com/openstreetmaps-growth-accelerates (accessed on 1 June 2013).
4. Pavlenko, A. *Mapnik*. Available online: http://mapnik.org/ (accessed on 1 June 2013).
5. OS. *Ordnance Survey OpenData*. Available online: http://www.ordnancesurvey.co.uk/oswebsite/products/os-opendata.html (accessed on 1 June 2013).
6. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221.
7. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148.
8. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B* **2010**, *37*, 682–703.
9. Ather, A. A Quality Analysis of OpenStreetMap Data. M.Sc. Thesis, University College London, London, UK, 2009.

10. Kounadi, O. Assessing the Quality of OpenStreetMap Data. M.Sc. Thesis, University College of London, Department of Civil, Environmental and Geomatic Engineering, London, UK, 2009.
11. Antoniou, V. *User Generated Spatial Content: An Analysis of the Phenomenon and Its Challenges for Mapping Agencies*; University College London (UCL): London, UK, 2011.
12. Ciepłuch, B.; Jacob, R.; Winstanley, A.; Mooner, P. Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences, Leicester, UK, 20–23 July 2010.
13. ISO. *ISO/CD 19157 Geographic Information—Data Quality*; ISO: Geneva, Switzerland, 2010.
14. Devillers, R.; Jeansoulin, R. *Fundamentals of Spatial Data Quality*; Wiley Online Library: London, UK, 2006.
15. Oort, P.A.J. Spatial Data Quality: From Description to Application. Ph.D. Thesis, Wageningen University, Wageningen, The Netherland, 2006.
16. Mooney, P.; Corcoran, P.; Winstanley, A.C. Towards Quality Metrics for Openstreetmap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; ACM: New York, NY, USA, 2010.
17. Wiemann, S.; Bernard, L. Conflation Services within Spatial Data Infrastructures. In Proceedings of 13th AGILE International Conference on Geographic Information Science, 2010 AGILE, Guimarães, Portugal, 10–14 May 2010.
18. Canavosio-Zuzelski, R.; Agouris, P.; Doucette, P. A photogrammetric approach for assessing positional accuracy of OpenStreetMap© roads. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 276–301.
19. Stankutė, S.; Asche, H. Improvement of Spatial Data Quality Using the Data Conflation. In *Computational Science and Its Applications-ICCSA 2011*; Springer: Berlin-Heidelberg, Germany, 2011; pp. 492–500.
20. Ludwig, I.; Voss, A.; Krause-Traudes, M. A Comparison of the Street Networks of Navteq and OSM in Germany. In *Advancing Geoinformation Science for a Changing World*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 65–84.
21. Koukoletsos, T.; Haklay, M.; Ellul, C. Assessing data completeness of VGI through an automated matching procedure for linear data. *Trans. GIS* **2012**, *16*, 477–498.
22. Du, H.; Anand, S.; Alechina, N.; Morley, J.; Hart, G.; Leibovici, D.; Jackson, M.; Ware, M. Geospatial information integration for authoritative and crowd sourced road vector data. *Trans. GIS* **2012**, *16*, 455–476.
23. Pourabdollah, A.; Morley, J.; Feldman, S.; Jackson, M. Studying the Dynamic Patterns of OpenStreetMap Bugs in Great Britain. In Proceedings of 16th AGILE International Conference on Geographic Information Science, Leuven, Belgium, 14–17 May 2013.
24. 1Spatial. *1Spatial Website*. Available online: http://www.1spatial.com (accessed on 1 June 2013).
25. Geofabrik. *Geofabrik*. Available online: http://www.geofabrik.de/ (accessed on 1 June 2013).
26. GDAL. *ogr2ogr*. Available online: http://www.gdal.org/ogr2ogr.html (accessed on 1 June 2013).
27. QuantumGIS. *Quantum GIS Desktop Project*. Available online: http://www.qgis.org/ (accessed on 1 June 2013).

28. Pourabdollah, A.; Morley, J.; Feldman, S. OSM-GB: Using Open Source Geospatial Tools to Create OSM Web Services for Great Britain . In Porceedings of the 6th Conference on Free and Open Source Software for Geospatial (FOSS4G), Nottingham, UK, 17–21 September 2013; in press.