**Title**: Towards an automated comparison of OpenStreetMap with authoritative road datasets

**Authors**:

Brovelli M.A. - Politecnico di Milano, Como Campus, DICA, Via Valleggio 11, 22100 Como, Italy

Minghini, M. - Politecnico di Milano, Como Campus, DICA, Via Valleggio 11, 22100 Como, Italy

Molinari, M. - Politecnico di Milano, Como Campus, DICA, Via Valleggio 11, 22100 Como, Italy

Mooney, P. - Department of Computer Science, Maynooth University,Maynooth, Co. Kildare, Ireland

**Keywords**: automation, data quality, GRASS GIS, road network, OpenStreetMap, Volunteered Geographic Information.

## Abstract

OpenStreetMap (OSM) is an extraordinarily large and diverse spatial database of the world. Road networks are amongst the most frequently occurring spatial content within the OSM database. These road network representations are usable in many applications. However the quality of these representations can vary between locations. Comparing OSM road networks with authoritative road datasets for a given area or region is an important task in assessing OSM's fitness for use for applications such as routing and navigation. Comparisons such as these can be technically challenging and no software implementation exists which facilitates such comparisons easily and automatically. In this paper we develop and propose a flexible methodology for comparing the geometry of OSM road network data with other road datasets. Quantitative measures for the completeness and spatial accuracy of OSM are computed including the compatibility of OSM road data with other map databases. Our methodology provides users with significant flexibility in how they can adjust the parameterisation to suit their needs. This software implementation is exclusively built on open source software and a significant degree of automation is provided for these comparisons. This software can subsequently be extended and adapted for comparison between OSM and other external road datasets.

## 1. Introduction and motivation for this work

OpenStreetMap (OSM) is probably the most popular Volunteered Geographic Information (VGI) project on the Internet today. In early 2015 OSM announced that the number of registered contributors to the project had exceeded two million. Several studies (Mooney and Corcoran 2014; Neis and Zipf 2012) have shown that while the number of these contributors actually contributing more than a few perfunctory or exploratory edits is in the tens of thousands the project continues to display incredible growth rates in terms of contributors and the volume of spatial data in the global database. OSM is being used as the source of spatial data for many researchers while the entire project ecosystem itself (the community, motivation of OSM volunteers, etc.) has become the source of increased academic research attention (Arsanjani et al. 2015b).

There have been many concerns raised about the quality, accuracy and general fitness-for-use and fitness-for-purpose of VGI data (Ali et al. 2014). Indeed OSM is the subject and basis for many of these concerns. The use of OSM as a source of spatial data is often justified by highlighting the very high financial cost of accessing and using spatial data collected and produced by National Mapping Agencies (NMA) and Commercial Mapping Companies (CMC) (Arsanjani et al. 2015a) and the fact that it is often more up-to-date (Goodchild 2007). The situation regarding access to these data has changed in the past number of years. Many NMA and CMC are making some or all of their spatial data products available as Open Data. The availability of these authoritative spatial datasets as Open Data has provided many opportunities for researchers to investigate the quality of VGI data such as OSM against authoritative spatial data from NMA and CMC.

Having access to these datasets does not mean that comparisons are easily carried out. Comparing two or more spatial datasets against each other is a challenging geocomputation problem. In the case of comparing a VGI dataset against a NMA or CMC dataset there are technical challenges caused by how the datasets are generated, organised and managed. Usually VGI and NMA/CMC datasets exist for different reasons and are managed, curated, edited, updated etc. under completely different workflows, QA/QC mechanisms and organisational structures and goals. For this reason it is often necessary to develop specific methodologies which are tightly coupled to the data structure, thematic domain and characteristics of the datasets under consideration. We believe that these constraints have contributed to making it more difficult for researchers to compare OSM data with other datasets (in particular NMA and CMC datasets). As we shall discuss there are already methods available which perform these types of comparisons for OSM with different datasets. However researchers must have a very specific skill set to develop software-based implementations of suitable methodologies to undertake comparisons like these. It is also very rare that these comparisons are fully automated.

This is exactly the problem which we address in this work. We propose and introduce a novel and flexible methodology for the automated comparison of road networks in OSM against road networks in NMA and CMC datasets. The software implementation of this methodology can be integrated into Open Source GIS environments such as QGIS which will allow a far wider range of researchers to undertake these comparisons. Details of the precise nature of the comparisons will be outlined in Section 3. While our focus is on NMA and CMC datasets which are available as Open Data we are confident that this methodology can be applied to NMA and CMC datasets which are available under non Open Data license structures and schemes. Researchers will be able to compare OSM street networks against their chosen NMA/CMC datasets. After the comparison has completed they will then make their own assessment and judgement on the relative differences between the two datasets. Quality and accuracy assessment of VGI are subject to more broader research. We, the authors, are involved in two European COST Action Networks: TD1202 'Mapping and the Citizen Sensor' and IC1203 'European Network Exploring Research into Geospatial Information Crowdsourcing: software and methodologies for harnessing geographic information from the crowd (ENERGIC)'. The research outlined in this paper is of great value and interest to the participants in both COST Action networks. Both networks have core themes of development methods and approaches for making the integration of crowdsourced geospatial data and authoritative spatial data more accurate straightforward.

There are a number of assumptions and general guidelines for our work. This paper does not set out to indicate or prove that OSM is better or worse than any particular dataset. Our paper focuses on the description of the technical methodology we have developed to facilitate the automated comparison of OSM to a complementary reference dataset for the purposes of a quality analysis. This technical methodology will compute and generate outputs which will be very helpful to stakeholders and users of OSM to support data quality and fitness for purpose assessments. It will then be the responsibility of the stakeholders or users to interpret these results in light of their application or problem specification for the OSM data. While our technical methodology strives to deliver a generic and automated comparison we assume that users of the methodology have basic knowledge about the reference or comparison dataset. The implementation of the proposed methodology provides users with default values for several important parameters. While users can choose other values for these parameters we attempt to support the user's choice of parameter with the assistance of appropriate supporting explanations and documentation. This methodology only considers the geometry of the two datasets for comparison. We do not consider attribute data or other information. This is subject of ongoing research into the extension of this methodology. Finally it goes without saying that both datasets (OSM and the reference dataset) should be comparable: the reference dataset should have a suitable geographical scale and level of detail to allow comparison with the OSM dataset. It is left to the user to decide if their reference dataset is suitable for comparison.

The remainder of our paper is organised as follows. In the next section we outline the state-of-the-art in similar work reported in the literature with an emphasis on methods which are designed to be generic or reproducible. Section 3 outlines our proposed methodology. Section 4 then provides an evaluation of our proposed methodology by a reference dataset for comparison with OSM on the same region. The paper closes with Section 5 where we outline the main conclusions from this work while offering some directions for the immediate and longer-term future work on this problem.

## 2. Background and Related Work

The task of comparing OSM data to authoritative datasets such as those produced by National Mapping Agencies (NMA) or Commercial Mapping Companies (CMC) has attracted significant attention in the literature over the past number of years. Despite this attention these comparisons display different levels of automation, are carried out for different purposes and focus on different aspects of comparison such as line features, polygons or attributes. In many of these comparisons the authors are attempting to assess or quantify the quality of OSM data against that of an authoritative dataset produced by NMA or CMC. In other studies the purpose is ultimately one of conflation or data fusion. In this section we provide an overview of the most significant and closely related work in the literature with an emphasis on the complexities involved in automating these comparisons.

In early work by Mooney et al. (2010) the authors developed an automated quality assessment measures to compare OSM with authoritative datasets containing natural water features. This assessment was based on the mathematical analysis of the shapes of features in both datasets. Haklay (2010) compared the OSM roads and streets dataset with the Ordnance Survey dataset in the UK. This work was extended by Girres and Touya (2010) who compared the quality of the OSM dataset in France with the reference database from the French National Mapping Agency. Both the work of Haklay (2010) and Girres and Touya (2010) indicates that the geometric quality of OSM data compares favourably with these NMA datasets. Ludwig et al. (2011) developed a fully automated approach to matching street objects in Germany contained in OSM and in the commercial Navteq database. The automated approach facilitates repeating the comparison when updates to both datasets dictate. This approach could be extended to other Navteq datasets in different regions but otherwise is tightly coupled to the Navteq data model. Zielstra and Zipf (2010) compared OSM and the CMC TeleAtlas dataset for Germany. The authors indicate that this analysis could be extended to TeleAtlas Europe if an appropriate license to purchase and use this dataset was acquired. Their approach focused heavily on computing comparison metrics such as overall road or street lengths between the two datasets. Koukoletsos (2012) published a PhD thesis in this area and his work includes other additional references and linkages to similar OSM comparison work around this time.

Canavosio-Zuzelski et al. (2013) undertook a photogrammetric approach to determining the positional accuracy of OSM road features using stereo imagery and a vector adjustment model. Their method applied rigorous analytical measurement principles to compute the real world geolocations of OSM road vectors in head-to-head positional accuracy assessment between OSM, the USGS National Map (TNM), and United States Census Bureau’s Topologically Integrated Geographic Encoding Referencing (TIGER) 2007 roads dataset. Fairbairn and Al-Bakri (2013) compared OSM and authoritative large-scale databases in the UK and Iraq to address possible integration of these datasets using a geographical surveying approach. Special software tools and user-interfaces were developed for the comparison which limits the extendability of this approach.

Hecht et al. (2013) considered the comparison of building polygons in OSM against the official building polygon footprint dataset for North Rhine-Westphalia in Germany and the building polygons in the Digital Landscape Model of the Authoritative Topographic Cartographic Information System (ATKIS) for Germany. The authors concluded that differences between the datasets and issues related to how OSM models buildings make these type of comparisons difficult at large scales.

In Jackson et al. (2013) the authors considered comparison of OSM Points of Interest (POI) for schools and educational facilities in the United States against two other point based datasets one of which is an official dataset (US Department of Education list of public and private schools) while the other a VGI dataset (USGS OpenStreetMap Collaborative Project OSMCP). The authors comment that comparison of point-based datasets is much less challenging than polygon or polyline datasets. However automation of the comparison is made difficult due to differences in the classification schemes between datasets.

Siebritz (2014) performed a quantitative and qualitative comparison between OSM and national mapping agency data in South Africa. The author found that uniformity of OSM data across South Africa was poor.

Subsequently it was concluded that OSM data did not have the required accuracy for integration with the South African NMA topographic database. However change detection was possible with OSM.

Graser et al. (2014) developed an open source software toolbox for QGIS for calculating key performance indicators in street network comparison studies between OSM and other authoritative datasets. In their paper Graser et al. make comparisons of OSM and the Austrian reference graph (GIP). The authors emphasise comparing OSM street network turn restrictions, one way streets, completeness of street names, etc. against that of GIP. Other datasets can be compared and the users are able to change some parameters to their preference. The toolbox is aimed at testing street network quality for routing applications.

In Forghani and Delevar (2014) the authors claim to introduce new and innovative metrics for the comparison of OSM data for Tehran against the NMA municipal map of Tehran. Their approach involves computing well known street network metrics for both datasets in combination with the provision of visual overlays of the OSM data on the Municipal map as a base-layer. The authors introduce the concept of an uncertainty map which visualises the uncertainty between the two datasets on a grid-based map. In related comparison work by Arsanjani et al. (2015a) the authors consider the comparison of OSM and authoritative land use and urban atlas datasets. The authors compare OSM and the Global Monitoring for Environment and Security Urban Atlas (GMESUA) for Europe and perform a comparative analysis of land use classifications in OSM and GMESUA. Their analysis found almost 90% accuracy between the two datasets. However the methodology used was developed specifically for OSM and the GMESUA datasets. In recently published work Yang and Zhang (2015) introduce a pattern-mining approach for the conflation of crowdsourced road network datasets. Their proposed methodology builds skeleton graphs to match the two datasets geometrically. When correspondence has been achieved then road-related semantic data can be used to check the data quality of the OSM Points of Interest (POI) and infer the names of the road segments in either dataset.

Koukoletsos et al (2012) proposed an automated feature-based matching method specifically designed for VGI. This approach is based on a multi-stage approach that combines geometric and attribute constraints and is applied to the OSM dataset using the official data from Ordnance Survey as the reference dataset. Data matching proved to be efficient with very low matching errors. The authors suggest their work as a possible first step in a framework for linear VGI quality assessment, including attribute and positional accuracy. Mohammadi and Malek (2014) presented an automated feature matching method for VGI linear data in OSM. The study reports positive results for matching OSM line features to those in reference datasets. However the authors are careful to outline that their methodology is just the first step in a data quality assessment of VGI. In developing quality assessment indicators Mooney and Corcoran (2014) believe that there is significant merit to approaches which consider the development of inherent embedded quality indicators from the OSM data itself. This is partly due to the fact that there may not always be a suitable dataset with which OSM can be directly compared with or comparisons are computationally difficult. This can be dependant on issues such as the thematic area or specific aspects of the quality comparison.

As evidenced by the breadth of the examples of literature on the comparison of OSM with other authoritative datasets there is a strong focus on road network comparisons. While all of these approaches are technically very strong and fit for purpose they are mostly application and dataset specific. Approaches and methodologies are then tightly coupled to the datasets under consideration and not easily extended to other dataset comparisons. In the next section we shall outline our proposed methodology which delivers this flexibility and provides a mechanism through which OSM road networks can be compared with other road network datasets from NMA and CMC.

## 3. Our proposed methodology

In this section we outline our novel methodology to perform automated comparison between OSM and authoritative road datasets in terms of spatial accuracy. Our methodology is rigorous and it provides a high degree of flexibility. In the software implementation users can skip one or more steps (defined as optional) and control the computation by specifying values for the parameters involved. This allows users to consider the precise nature of the authoritative dataset under investigation (e.g. its nominal scale and accuracy) and control the methodology accordingly.

This methodology is currently implemented as a set of three independent GRASS GIS[1] modules written in Python which can be run directly from the command line or invoked from a Graphical User Interface (GUI). The three modules accomplish the following consecutive steps:

1. preliminary comparison of the datasets and computation of global statistics;
2. geometric preprocessing of the OSM road network dataset to extract its subset representing the same road features of the authoritative dataset;
3. evaluation of the spatial accuracy of this OSM subset using a grid-based approach.

Each of these steps are outlined in detail below. For each step of the methodology we indicate the required and optional inputs/operations. The GRASS built-in modules used by the procedure are also highlighted (e.g. *v.buffer*). We use the abbreviation REF (standing for “reference dataset”) to denote the authoritative road network dataset compared to OSM. The GUI screenshots of the three steps are presented in the following for the case study of Erba municipality, located in Como province (Northern Italy), where OSM road dataset is compared to the authoritative road vector dataset, having a scale of 1:2000 and available as open data[2]. Erba will be referred to as the REF dataset.

### 3.1 Step 1: Preliminary comparison of OSM and REF datasets

The first step of the procedure is aimed at preparing the OSM and REF datasets as well as performing some simple measures of their spatial similarity. The user will have an area of interest for their comparison. Step 1 is delivered using a customized GUI (see Figure 1) allowing users to:

[1] http://grass.osgeo.org

[2] download is available from the geoportal of Lombardy Region at http://www.geoportale.regione.lombardia.it/download-dati

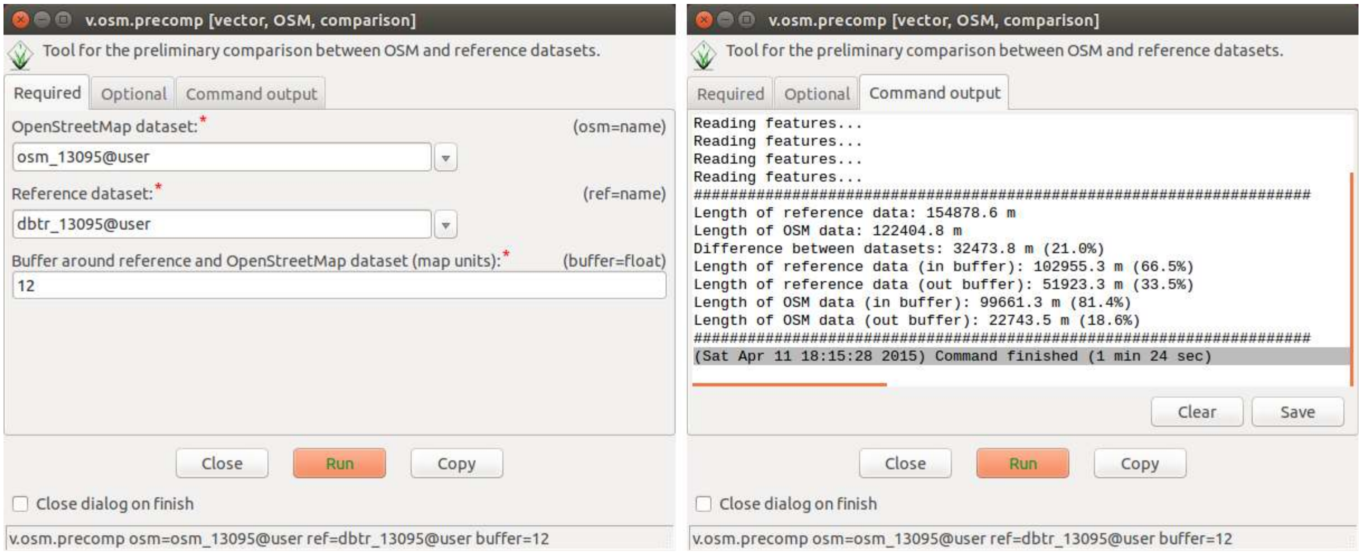
a) Select the OSM and REF datasets to be compared [required]. This assumes they have been already imported into GRASS, e.g. using the module *v.in.ogr*. It is worth noting that it is up to the user to import two datasets which are: 1) actually comparable, i.e. that include the same road feature classes (for instance, if the REF dataset does not include cycleways, these should be preventively excluded also from the OSM dataset imported); and 2) topologically correct, to ensure the success of the following processing;

b) If OSM and/or REF datasets cover a region larger than the one of interest, select a vector layer to be used as the clipping mask *(v.overlay*) [optional];

c) Compute the total length of OSM and REF datasets and their length difference, both in map units and percentage *(v.to.db)* [required];

d) Apply a buffer of user-specified width around the REF and OSM datasets to compute the length and the length percentage of the OSM and REF datasets included in the buffer (*v.buffer*, *v.overlay*, *v.to.db*) [required].

The outputs of sub-steps (c) and (d) are stored in a text file as they are used as inputs to perform further computations.



v.osm.precomp [vector, OSM, comparison]

Tool for the preliminary comparison between OSM and reference datasets.

OpenStreetMap dataset:* (osm=name)
osm_13095@user

Reference dataset:* (ref=name)
dbtr_13095@user

Buffer around reference and OpenStreetMap dataset (map units):* (buffer=float)
12

☐ Close dialog on finish

```
v.osm.precomp osm=osm_13095@user ref=dbtr_13095@user buffer=12
```

v.osm.precomp [vector, OSM, comparison]

Tool for the preliminary comparison between OSM and reference datasets.

```
Reading features...
Reading features...
Reading features...
Reading features...
######################################################################
Length of reference data: 154878.6 m
Length of OSM data: 122404.8 m
Difference between datasets: 32473.8 m (21.0%)
Length of reference data (in buffer): 102955.3 m (66.5%)
Length of reference data (out buffer): 51923.3 m (33.5%)
Length of OSM data (in buffer): 99661.3 m (81.4%)
Length of OSM data (out buffer): 22743.5 m (18.6%)
######################################################################
(Sat Apr 11 18:15:28 2015) Command finished (1 min 24 sec)
```

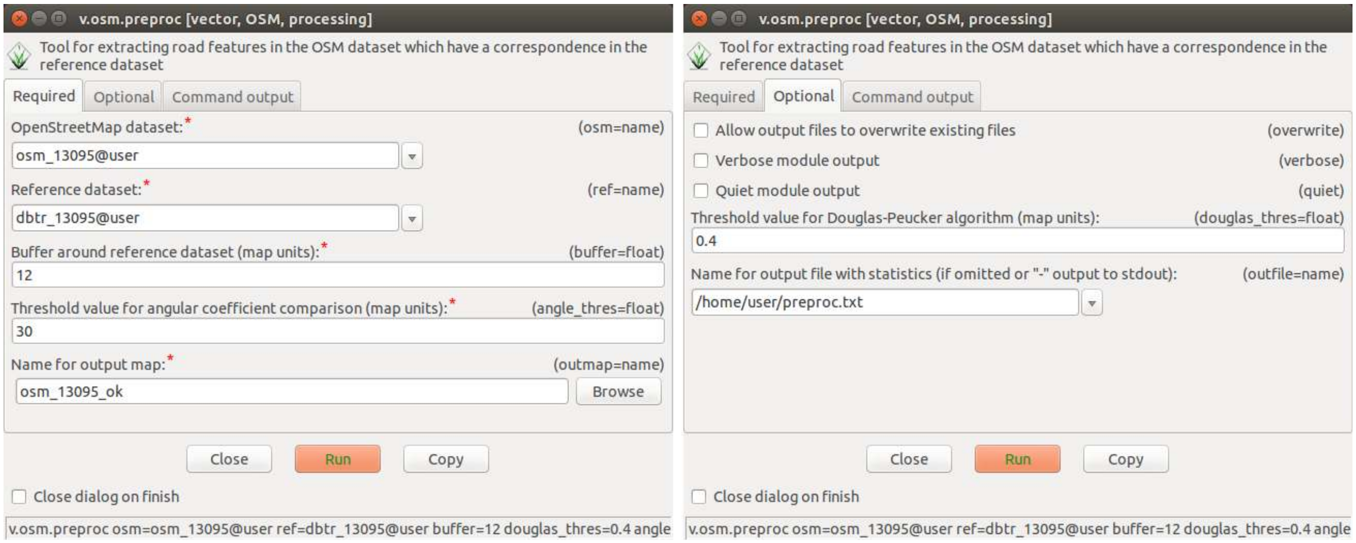
☐ Close dialog on finish

```
v.osm.precomp osm=osm_13095@user ref=dbtr_13095@user buffer=12
```

Figure 1. GRASS GUIs showing required input parameters (left) and outputs (right) for Step 1.

### 3.2 Step 2: Geometric preprocessing of OSM dataset

The OSM dataset is now geometrically preprocessed to extract its road line features having a corresponding feature in the REF dataset. This preprocessing removes the portions of OSM roads which have no corresponding road in the REF dataset. Hence, this step allows to obtain two fully comparable datasets which will be used to evaluate OSM accuracy in Step 3 (see Subsection 3.3). The GUI (see Figure 2) provides an accessible interface for users allowing them to:



Tool for extracting road features in the OSM dataset which have a correspondence in the reference dataset

OpenStreetMap dataset:* (osm=name)
osm_13095@user

Reference dataset:* (ref=name)
dbtr_13095@user

Buffer around reference dataset (map units):* (buffer=float)
12

Threshold value for angular coefficient comparison (map units):* (angle_thres=float)
30

Name for output map:* (outmap=name)
osm_13095_ok

□ Close dialog on finish

```
v.osm.preproc osm=osm_13095@user ref=dbtr_13095@user buffer=12 douglas_thres=0.4 angle
```

Tool for extracting road features in the OSM dataset which have a correspondence in the reference dataset

□ Allow output files to overwrite existing files (overwrite)

□ Verbose module output (verbose)

□ Quiet module output (quiet)

Threshold value for Douglas-Peucker algorithm (map units): (douglas_thres=float)
0.4

Name for output file with statistics (if omitted or "-" output to stdout): (outfile=name)
/home/user/preproc.txt

□ Close dialog on finish

```
v.osm.preproc osm=osm_13095@user ref=dbtr_13095@user buffer=12 douglas_thres=0.4 angle
```

Figure 2. GRASS GUIs showing required (left) and optional (right) input parameters for Step 2.

a) Perform a generalization of the REF dataset based on the Douglas-Peucker algorithm (Douglas and Peucker 1973) with the threshold value specified by the user (*v.generalize*) [optional]. Reducing the number of line features of the REF dataset decreases the computational time required by all subsequent steps;

b) Split the line features of the REF dataset into segments (*v.split*) [required]; this operation, which clearly increases the number of line features of the REF dataset, is required for the success of sub-step (f) below;

c) Split the line features of the OSM dataset into segments (*v.split*) [required]; this operation, which clearly increases the number of line features of the OSM dataset, is required to improve the result of sub-step (f) below;

d) Compute a centrality measure of degree for each node of the REF dataset (*v.net.centrality*) [required]. The degree of each node of the network corresponds to the number of other nodes the node is linked to (Freeman 1979). This operation, required for sub-step (e) identifies the terminal nodes of the REF network, i.e. those having degree equal to 1;

e) Apply a buffer of user-specified width around the REF dataset to extract the OSM line features included in the buffer (*v.db.select*, *v.extract*, *v.buffer*, *v.overlay*) [required]. The OSM line features which are not included in the buffer are assumed not to correspond to any of the REF line features and therefore are discarded. Besides having a fixed width, the shape of the buffer applied around a line feature is different according to the value of the degree centrality of its starting and final nodes. If this value is 1 (i.e. the node is a terminal node), the buffer is applied without the cap. The result is that the correct portion of OSM dataset is extracted (see Figure 3);

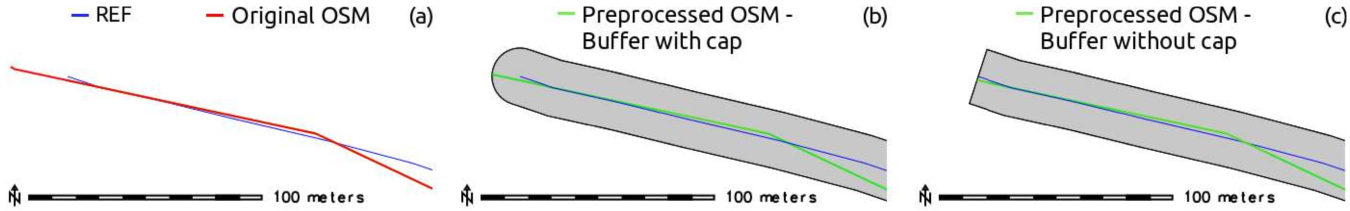

REF
Original OSM
(a)
Preprocessed OSM - Buffer with cap
(b)
Preprocessed OSM - Buffer without cap
(c)
100 meters
100 meters
100 meters

Figure 3. Visual comparison between the REF dataset and: the original OSM dataset (a); the OSM dataset preprocessed using a buffer with cap (b) and the OSM dataset preprocessed using a buffer without cap (c) around a REF line feature with a terminal node: in the last case, the correct part of the OSM line feature is extracted.

f)

- Compute the angular coefficient of each REF segment (*v.to.db*);
- apply a buffer of the same user-specified width of sub-step (e) around each REF line feature (*v.buffer*); compute the angular coefficient of each OSM line feature(s) included in this buffer (*v.to.db*);
- compare the angular coefficient of the REF line feature with the angular coefficients of the OSM line feature(s) included in the buffer: if the difference between them exceeds a user-specified threshold then the corresponding OSM line feature is discarded [required]. This operation is

intended to clean the OSM dataset by removing all line features which, although included in the buffer applied in sub-step (e) do not have a correspondence in the REF dataset (see Figure 4).

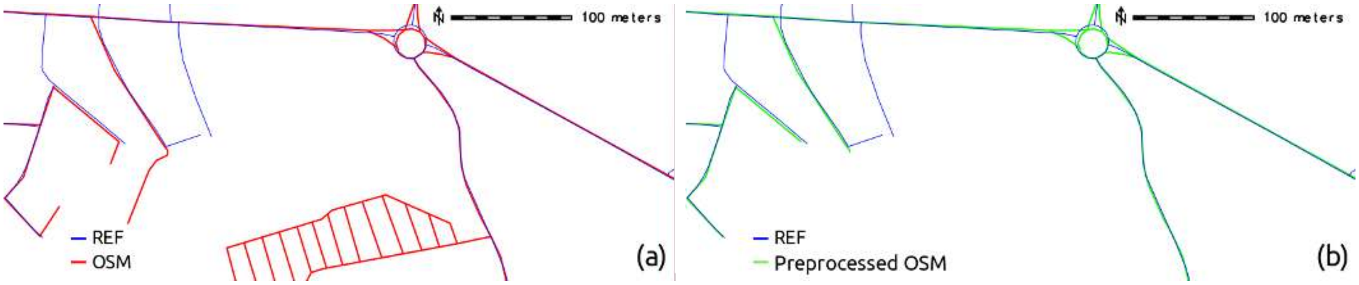
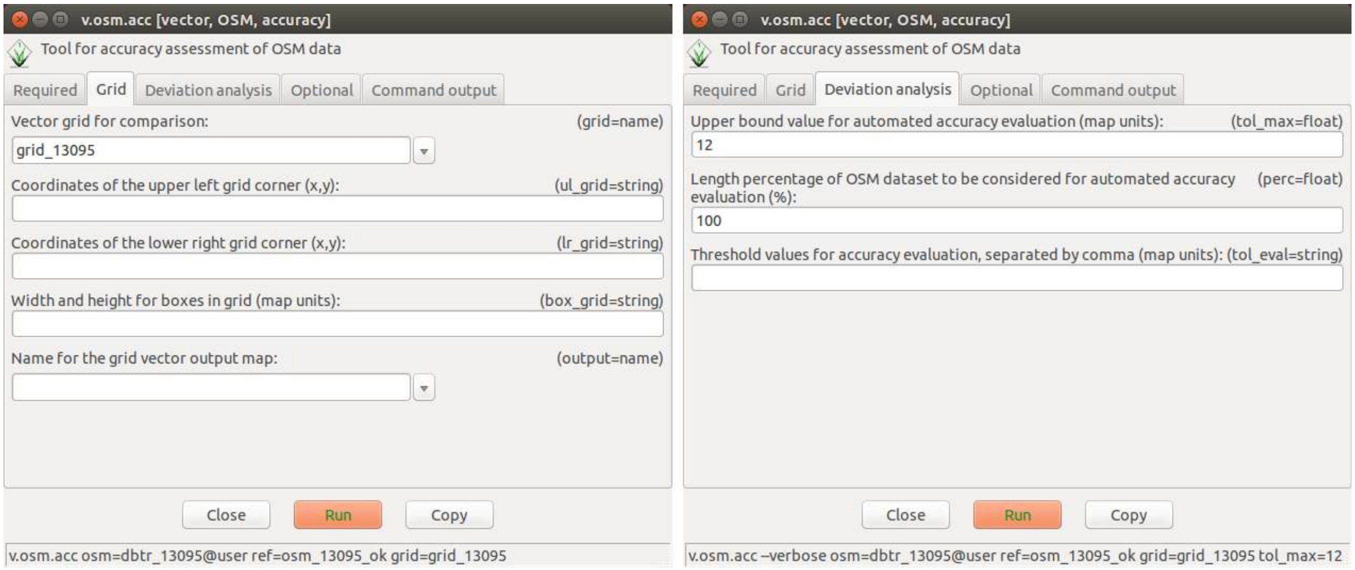


Figure 4. Example of REF and OSM datasets before (a) and after (b) the geometric preprocessing operation: OSM roads without a corresponding in the REF datasets are successfully discarded.

g) Compute the total length of the preprocessed OSM dataset. The difference between the initial OSM dataset and the preprocessed OSM dataset is presented both in map units and as a percentage. The difference between the REF dataset and the preprocessed OSM dataset is also presented both in map units and as a percentage [required]. This information is stored in a text file which can be used as inputs for further computations.

### 3.3 Step 3: Evaluation of OSM spatial accuracy through a grid-based approach

This step performs the comparison between the REF dataset and the OSM dataset and measures the spatial accuracy of the latter. The OSM dataset was geometrically preprocessed in the second step of the procedure (see Subsection 3.2) and is now fully comparable with the REF dataset. The GUI is presented in Figure 5. This step involves:


v.osm.acc [vector, OSM, accuracy]
Tool for accuracy assessment of OSM data
Required
Grid
Deviation analysis
Optional
Command output
Close
Run
Copy
v.osm.acc [vector, OSM, accuracy]
Tool for accuracy assessment of OSM data
Required
Grid
Deviation analysis
Optional
Command output
Close
Run
Copy

Vector grid for comparison: (grid=name)
grid_13095

Coordinates of the upper left grid corner (x,y): (ul_grid=string)

Coordinates of the lower right grid corner (x,y): (lr_grid=string)

Width and height for boxes in grid (map units): (box_grid=string)

Name for the grid vector output map: (output=name)

```
v.osm.acc osm=dbtr_13095@user ref=osm_13095_ok grid=grid_13095
```

Upper bound value for automated accuracy evaluation (map units): (tol_max=float)
12

Length percentage of OSM dataset to be considered for automated accuracy evaluation (%): (perc=float)
100

Threshold values for accuracy evaluation, separated by comma (map units): (tol_eval=string)

```
v.osm.acc –verbose osm=dbtr_13095@user ref=osm_13095_ok grid=grid_13095 tol_max=12
```

Figure 5. GRASS GUIs showing the definition of the grid (left) and the choice of the analysis to perform (right) for Step 3.

a) Create a grid with user specified coordinates and step size in both the horizontal and vertical directions (*v.mkgrid*) [optional]. The user can also select a polygon vector layer in GRASS which defines the grid [optional]. The use of a grid allows execution of the following analysis on single cells. This takes into account the possible heterogeneous nature of the OSM dataset and subsequently heterogeneous quality. If the user does not create a grid the whole region is treated as a single cell;

b) Perform at least one of the following optional analysis:

- For each cell of the grid:

Find the maximum deviation of the OSM dataset from the REF dataset. The user has to enter an upper bound value for the deviation (this must be less than or equal to the width of the buffer used to preprocess the OSM dataset, see Subsection 3.2); Using a bisection search, the analysis returns the minimum width of the buffer around the REF dataset within which a fixed percentage (specified by the user) of the length of the line features of the OSM dataset is included (*v.buffer*, *v.overlay*, *v.to.db*) [optional]. As suggested by Koukoletsos (2012) by choosing a customized percentage

This is the pre-peer reviewed version of the following article: “Brovelli M. A., Minghini M., Molinari M. and Mooney P. (2016) Towards an automated comparison of OpenStreetMap with authoritative road datasets. Transactions in GIS, DOI: 10.1111/tgis.12182”, which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1111/tgis.12182/full.

different from 100% (e.g. 95%) users can in fact neglect the outliers. These outliers are inevitably introduced by the procedure itself as well as the non-optimal choices of parameters. Neglecting the outliers still allow the user to obtain a satisfying result for decision making purposes. The absence of buffer caps at the terminal nodes when performing the preprocessing of the OSM line features (see Subsection 3.2) was introduced to avoid biasing the result of this buffer-based analysis. For example if the preprocessing was performed using a buffer with caps around the REF dataset then the presence of OSM line features beyond the REF terminal nodes (see Figure 3b) would still return the buffer width as the minimum width of the buffer around the REF dataset including all the line features of the OSM dataset. The output of this analysis is returned as a GRASS polygon vector map corresponding to the chosen grid. For each polygon the maximum deviation of the OSM dataset from the REF dataset is provided as an attribute. If necessary, the module *v.out.ogr* allows the user to export the vector map into one of the supported formats.

- Compute statistics for predefined values of deviation of the OSM dataset from the REF dataset. The user must enter one or more threshold values of deviation; for each cell of the grid, the analysis returns the length and the length percentage of the OSM dataset having a deviation smaller than each threshold value (*v.buffer*, *v.overlay*, *v.to.db*) [optional]. This is very useful when the purpose is to check the accuracy of the OSM dataset against a specified threshold value (a practical example is provided in Subsection 4.3). The output of the analysis is again a GRASS vector map in which the length (and length percentage) of the OSM dataset with a deviation smaller than the predefined threshold value is stored as an attribute for each grid polygon. As before this map can be then exported using the module *v.out.ogr*.

## 4. Evaluation of our Methodology

In this section we provide a practical example of how the methodology described in Section 3 can be used. The example is again referred to the comparison of OSM and REF road network datasets for the case study of Erba municipality. We are not trying to draw conclusions about the spatial accuracy of this specific OSM road network dataset. Rather we describe the different steps of the methodology from the user's point of view. The following discussion will address the three steps of the methodology separately. As already mentioned the key underlying assumption is that the user is familiar with the characteristics of the REF dataset involved.

### 4.1 Evaluation of Step 1: Preliminary comparison of OSM and REF datasets

As stated in Subsection 3.1, this step performs an initial assessment of the similarity between the REF and OSM datasets. It also represents the basis for choosing a suitable value of the buffer width for Step 2. Figure 6 shows the sample REF and OSM road network datasets together with some plots generated from the text files created in Step 1.

A meaningful operation users can perform on the outputs of Step 1 consists of a sensitivity analysis on the value of the buffer width around the REF dataset (see Figure 6b). In other words, by running Step 1 multiple times using different values of buffer width users can ascertain how the length of the OSM dataset included in the buffer around the REF dataset varies with buffer width. The shape of the resulting curve - which clearly must be monotone increasing should suggest a suitable value of buffer width to be used in Step 2. In the example of Figure 6b, a buffer width of 10-12 m (i.e. corresponding to a point located after the change of slope of the curve) is capturing almost all (80%) of the OSM dataset. The remaining percentage (approximately 20%) which is not included in the buffer is due to the OSM road line features which have no correspondence in the REF dataset. These are visually identified in Figure 6a.

Conversely running Step 1 multiple times by varying the buffer width around the OSM dataset allows estimation of the length (or the length percentage) of the REF features which have no correspondence in the OSM dataset (see Figure 6c). This is a first assessment of the OSM spatial completeness against the REF dataset. In the case of Figure 6c, we can infer that there are about 50 km of roads (from a total of about 150 km) which are available in the REF dataset but are not in the OSM dataset. Figure 6a offers a visual assessment of these roads.

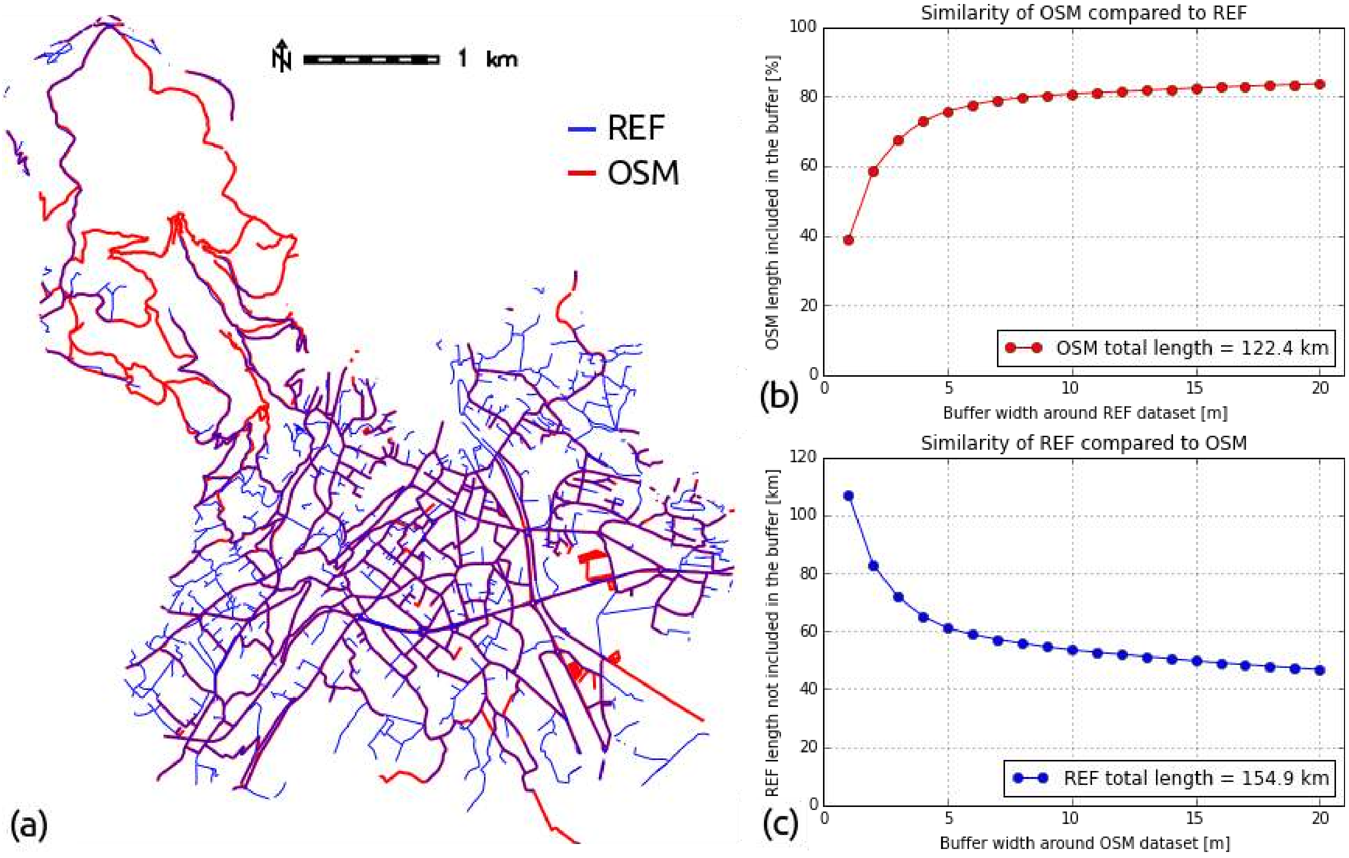


Figure 6. Example of OSM and REF road network dataset (a); plot showing the length percentage of the OSM dataset included in buffers of different widths around the REF dataset (b); a plot showing the length of the REF dataset not included in buffers of different widths around the OSM datasets (c).

### 4.2 Evaluation of Step 2: Geometric preprocessing of OSM dataset

As described in Subsection 3.2, the purpose of this step is to clean the OSM road network dataset to make it suitable for the final comparison with the REF dataset performed in Step 3. Step 2 has a key role in the whole procedure. Due to the heavy geometric processing involved this is by far the most time consuming step. This step uses the choice of the buffer width from Step 1 and, based on the set of parameters chosen by the user, creates the preprocessed OSM dataset suitable for Step 3.

The first parameter users can optionally enter is the threshold value (expressed in map units) used to perform a generalization of the REF dataset using the Douglas-Peucker algorithm. This threshold determines the degree of simplification of the dataset. Users should carefully choose the value of this parameter by taking into account the specific characteristics of the REF dataset being used. The threshold corresponds to the maximum positional difference occurring between the original dataset and the generalized dataset. As such, its value should be set equal to the intrinsic accuracy of the REF dataset, which in turn depends on its scale. This is assumed to be known from the user; if it is not then the user can just leave the field blank which will result in the absence of generalization. In the example of Erba a value of 0.4 m can be chosen as it corresponds to the nominal accuracy of the REF dataset (see Figure 2).

The procedure will then continue by automatically splitting the REF and OSM datasets and assigning a measure of degree to each node of the REF dataset (see Subsection 3.2). The latter is required to correctly isolate the terminal nodes of the REF dataset which have a measure of degree equal to 1.

Step 2 continues then by applying a buffer of user-specified width around the REF dataset to cut out all the OSM line features which fall outside it. As explained in Subsection 3.2, the shape of the buffer around the REF line features ending in a node with degree equal to 1 is without cap. Although users can run Step 2 by choosing a customized value for the buffer width - or again run it multiple times by varying the buffer width they are encouraged to run it using a suitable value resulting from the sensitivity analysis performed in Step 1 (see Subsection 4.1), e.g. 12 m in the example of Erba.

The following stage of Step 2 is to further clean the OSM dataset by discarding all its features who have an angular coefficient which is very different to that of the corresponding REF feature. Users have to define a customized angle threshold to guide this operation. This parameter corresponds to the maximum angular difference users are willing to accept between a REF and an OSM line feature representing the same part of a road. Ideally it must be chosen by looking at the REF and OSM datasets while also using common sense. From our empirical observations and experience we believe that a threshold value of around 30° should be suitable.

Users can finally assess the results of the OSM cleaning from the outputs of Step 2. This consists of both the vector layer of the preprocessed OSM (see Figure 7) and a set of statistics useful to compare the original and the cleaned OSM dataset.

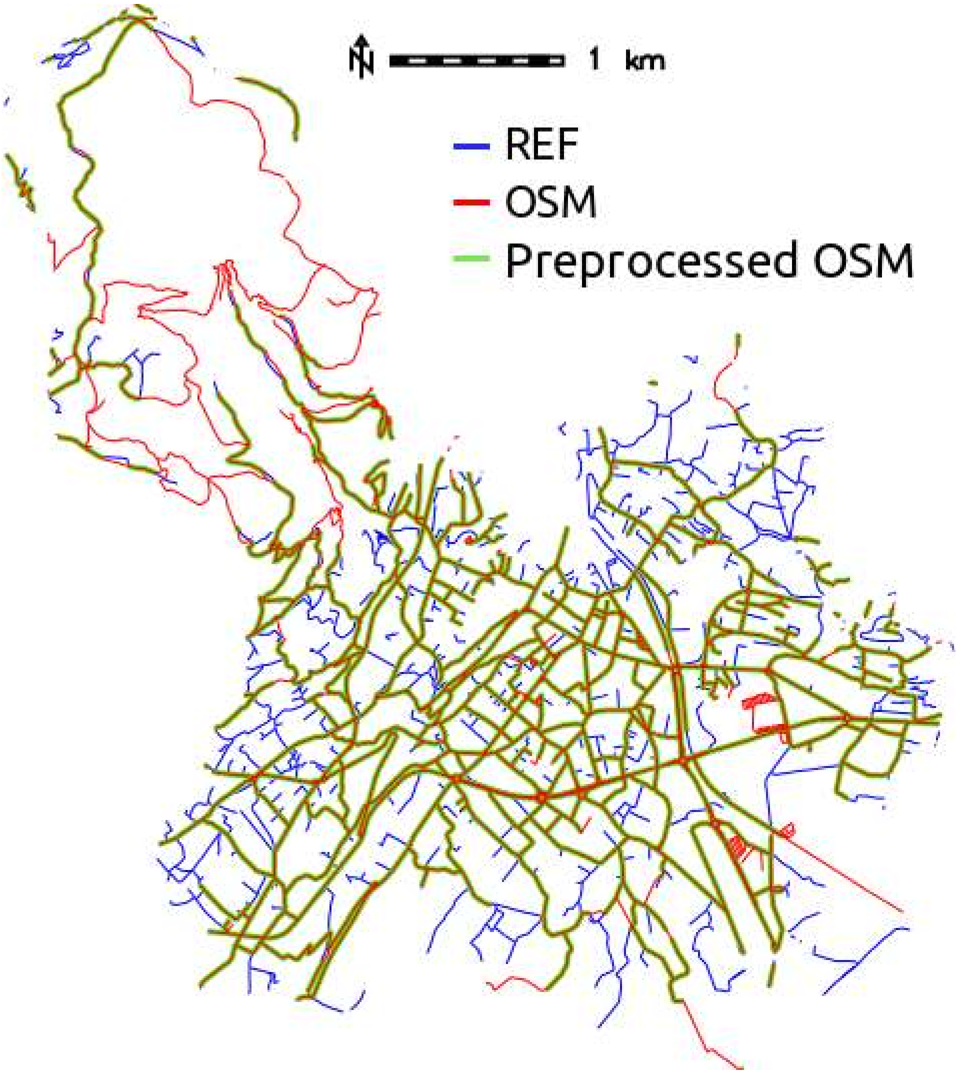


Figure 7. An example of REF, original OSM and preprocessed OSM road network datasets.

### 4.3 Evaluation of Step 3: Evaluation of OSM spatial accuracy through a grid-based approach

Once the OSM road network dataset has been preprocessed users can run Step 3 to perform the actual comparison between the REF and the OSM datasets (see Subsection 3.3). This is the final and most important stage of the whole methodology and depends upon the outcomes of Step 2.

As mentioned in Subsection 3.3 users can define a customized grid in order to separately run the following analysis on different sub-areas and capture the (typically) different spatial accuracy of the OSM dataset. Users can import and select an external polygon vector layer representing the grid. Alternatively users can create the grid in GRASS by defining an origin and the step sizes in the horizontal and vertical directions. In this case the grid should be created according to the level of detail of the results users wish to achieve: a dense grid (i.e. made up of many cells) allows to highly differentiate the measures of OSM accuracy in space; on the opposite, a sparse grid returns a measure of OSM accuracy which is averaged on large areas. The creation of the grid should also guarantee that OSM accuracy patterns due to spatially-diversified areas are properly detected. If the dataset under investigation includes both urban and rural areas, where OSM accuracy can be lower (Hecht et al. 2013), these should be contained in different cells of the grid.

Using the created grid users can then perform at least one of two possible analysis. The first analysis produces an output vector map corresponding to the grid where each polygon has an attribute specifying the maximum deviation of the OSM preprocessed dataset from the REF dataset. This is calculated according to the percentage of the length of OSM line features chosen by the user (see Subsection 3.3). It should be restated that the upper bound value for the deviation that users must define should correspond to the buffer width used to clean the OSM dataset during Step 2 (12 m in the case of Erba, see Figure 5). Styling the output vector data according to the values of the maximum deviation returned from Step 3 allows the user to get an immediate estimate of the OSM dataset’s similarity to the REF dataset for the study area (see Figure 8a).

The second analysis is useful for testing the OSM accuracy (meant as the deviation from the REF dataset) against one or more specific target accuracies which the user might be interested in. This could include accuracies required to use OSM data for some specific purpose or use-case. The analysis returns a new vector map corresponding to the grid. The attributes store the length and length percentage of the OSM dataset with a deviation from the REF dataset less or equal than the target accuracy for each target accuracy specified by the user. Besides quantitatively analysing the values returned, this vector map can be used to visually detect the spatial patterns of OSM accuracy throughout the study area (see Figure 8b). From both the analysis performed in Step 3 on the case study of Erba, a progressive decrease in accuracy can be detected when moving from the south area (corresponding to the city center) to the north-west area (a mountainous, rural area) (see Figure 8).

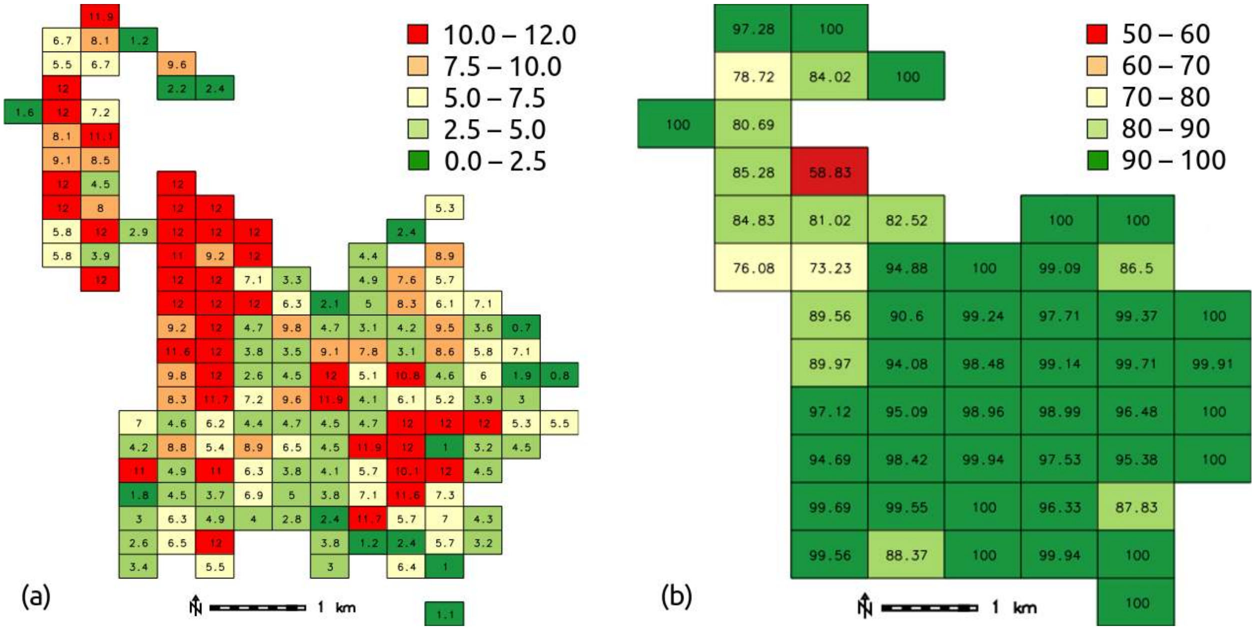


Figure 8. Vector data returned from Step 3 showing the OSM similarity to REF, expressed as the maximum deviation (in map units) of the OSM dataset from the REF dataset, for each grid cell (a); and the spatial distribution of OSM accuracy, expressed as the length percentage of the OSM dataset included in a predefined buffer around the REF dataset (b).

## 5. Discussion and Conclusions

A number of automated and semi-automated procedures have been developed over the past number of years to assess the quality of OSM data in comparison to authoritative datasets from National Mapping Agencies (NMA) and Commercial Mapping Companies (CMC). Our paper outlines a methodology which delivers a flexible and generic means by which OSM road network datasets can be compared to NMA or CMC road network datasets. Our approach addresses a gap in the knowledge in this area where previous developed procedures, while rigorous and scientifically valid, are typically very specific and tightly coupled to the characteristics of the authoritative road datasets involved. As a result, we believe they are not easily replicable or extensible to a broader range of authoritative datasets. In this paper we have developed an automated and flexible procedure to perform comparisons between OSM and authoritative road datasets in terms of geometrical accuracy. We outlined the methodology in detail and then provided an example of its implementation.

The purpose of our paper is not to carry out an investigation into the accuracy of the OSM road network dataset compared to a reference dataset for a specific area or to make specific statements about the comparative quality. The goal of the paper is to present and technically illustrate the capability, flexibility and potential of our procedure as a means of supporting other researchers and practitioners in carrying out their own accuracy and data quality comparisons between OSM and a reference dataset. Our methodology is implemented as a customized module for the open source GRASS GIS software. It is written in Python and consists of three separate steps which guide users as they perform comparison of OSM road data and an NMA or CMC road network dataset for a specific area. Users can fully customize each step by specifying appropriate values for all the parameters involved. While making the procedure flexible and in principle extendable to any reference dataset comparable with OSM we make the assumption that users are familiar with the reference dataset involved. Choice of suitable parameter values should be linked to the reference dataset's intrinsic characteristics (e.g. the nominal scale and accuracy). The statistical outputs of the procedure allow also users (by running the procedure multiple times varying the parameter values) to obtain deeper insights of the analysis as well as perform sensitivity analysis on one or more parameters. The output vector data can be reused or further analysed in any GIS software. Due to the flexibility and adaptability of our methodology we believe that the methodology can be extremely useful to many users and for a number of use-cases. The methodology allows users to obtain a measure of the OSM deviation from a known dataset which in turn can inform whether OSM is suitable or not for a specific application.

There are also some weaknesses to our approach. Firstly, the computational time required to run the whole procedure is rather long. The majority of run time is required by Step 2, which, involving heavy geometric processing operations (see Subsection 3.2), is by far the most time consuming stage of the implementation. The overall execution time is dependent upon the number of features of the REF dataset after its generalization and split; in the case of Erba, this dataset has 9563 features and the time required to run Step 2 is about 15 hours on an Intel Xeon E5620 processor with 2.4 GHz and 24 GB of RAM (Linux Ubuntu server). Reducing this computational time is the major focus of our immediate and long term future work. Parallel processing approaches provide the potential for a major speedup in overall execution time. We are also considering the possibility of providing this implementation as a Web Processing Service (WPS). A WPS would make the procedure available as an interoperable service through the Web. This could provide users with an opportunity to compare OSM road network datasets with other datasets for small spatial areas without the need to install GRASS and other software on their local system. Clearly, with the aim of reducing the computational time for Step 2, it is very likely that WPS results for Step 3 (which depend on Step 2) will have to be sent by e-mail to the user after offline processing is complete. In terms of alternative implementations of the procedure we know that typically a small number of researchers and users are familiar with GRASS GIS. Hence, we also plan to develop the extension of the procedure as a QGIS plugin making it usable by a larger number of users. At present the methodology has been tested on several road datasets. These are mainly Italian datasets with which the authors are highly familiar. Our future work will include an extensive test of the procedure on an additional number of authoritative datasets ideally from different countries and with heterogeneous characteristics in order to maximize the validation of results. This will be reported in a future paper. Work is also underway to extend the methodology to include the attributes of the OSM road network dataset in addition to the geometry.

## Acknowledgements

The authors gratefully acknowledge support from COST Actions TD1202 “Mapping and the Citizen Sensor” and IC1203 “European Network Exploring Research into Geospatial Information Crowdsourcing: software and methodologies for harnessing geographic information from the crowd (ENERGIC)”. Specifically the Short Term Scientific Mission (STSM) COST-STSM-TD1202-24885, supported by COST Action TD1202, was where this work was partially started. Finally the authors would like to sincerely thank the anonymous reviewers for their helpful suggestions which assisted us in improving the overall quality of this paper.

## References

Ali A L and Schmid F 2014 Data Quality Assurance for Volunteered Geographic Information. In Duckham M, Pebesma E, Stewart K, and Frank A U (eds.) *Geographic Information Science.* Springer International Publishing: 126–141

Arsanjani J J, Mooney P, Zipf A, and Schauss A 2015a Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets. In Arsanjani J J, Zipf A, Mooney P, and Helbich M (eds) *OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography*. Springer International Publishing: 37–58

Arsanjani J J, Zipf A, Mooney P, and Helbich M 2015b An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications. In Arsanjani J J, Zipf A, Mooney P, and Helbich M (eds.) *OpenStreetMap in GIScience*. Springer International Publishing: 1-15

Canavosio-Zuzelski R, Agouris P, and Doucette P 2013 A Photogrammetric Approach for Assessing Positional Accuracy of OpenStreetMap Roads. *ISPRS International Journal of Geo-Information* 2(2): 276–301

Douglas D H and Peucker T K 1973 Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization 10*(2): 112-122

Fairbairn D and Al-Bakri M 2013 Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS International Journal of Geo-Information* 2(2): 349–370. doi:10.3390/ijgi2020349

Forghani M and Delavar M R 2014 A Quality Study of the OpenStreetMap Dataset for Tehran. *ISPRS International Journal of Geo-Information* 3(2): 750–763

Freeman L C 1979 Centrality in social networks conceptual clarification. *Social networks* 1(3): 215–239

Girres J F and Touya G 2010 Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14(4): 435–459

Goodchild M F 2007 Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211–221

Graser A, Straub M, and Dragaschnig M 2014 Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph. *Transactions in GIS* 18(4): 510–526

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

Hecht R, Kunze C, and Hahmann S 2013 Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information* 2(4): 1066–1091

Jackson S P, Mullen W, Agouris P, Crooks A, Croitoru A, and Stefanidis A 2013 Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information* 2(2): 507–530

Koukoletsos T., 2012. A Framework for Quality Evaluation of VGI linear datasets, PhD, University College London, UK http://discovery.ucl.ac.uk/1359907/11/1359907.pdf last accessed August 2015

Koukoletsos, T., Haklay, M. & Ellul, C. (2012), Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data, Transactions in GIS 16(4), 477–498.

Ludwig I, Voss A, and Krause-Traudes M 2011 A Comparison of the Street Networks of Navteq and OSM in Germany. In Geertman S, Reinhardt W, and Toppen F (eds) *Advancing Geoinformation Science for a Changing World*. Berlin, Springer Berlin Heidelberg: 65-84

Mohammadi, N. and Malek, M. (2014), VGI and Reference Data Correspondence Based on Location-Orientation Rotary Descriptor and Segment Matching. Transactions in GIS. doi: 10.1111/tgis.12116

Mooney P, Corcoran P, and Winstanley A C 2010 Towards Quality Metrics for OpenStreetMap. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, USA: 514–517

Mooney P and Corcoran P 2014 Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. *Transactions in GIS* 18(5): 633–659

Neis P and Zipf A 2012 Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information* 1(2): 146–165

Siebritz L A 2014 Assessing the accuracy of openstreetmap data in south africa for the purpose of integrating it with authoritative data. PhD Thesis, Department of Architecture, Planning and Geomatics, University of Cape Town, SA

Yang B and Zhang Y 2015 Pattern-Mining Approach for Conflating Crowdsourcing Road Networks with POIs. I*nternational Journal of Geographical Information Science* published online. doi: 10.1080/13658816.2014.997238

Zielstra D and Zipf A 2010 A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal: 1–15