Comparison of matching methods of user generated and authoritative geographic data

E. Abdolmajidi[1], J. Will[1], L. Harrie[1], A. Mansourian[1]

[1]Department of Physical Geography and Ecosystem Science, Lund University
Corresponding author: ehsan.abdolmajidi@nateko.lu.se

Abstract

In the last decade a new and alternative source of geospatial data has become available, so called Volunteered Geographic Information (VGI); one example is OpenStreetMap (OSM). To study the progress of the VGI data it is of interest to compare it with authority data. This study presents an assessment of two commonly-used approaches consisting of *Segment-based* and *Node-based* for matching two linear datasets. For that purpose, the OSM road network and a reference dataset from Lantmäteriet (Swedish National Mapping Agency) are chosen to be used in this assessment. The segment-based method is adopted from Koukoletsos et al. (2012) and developed further to incorporate feature correspondence. The node-based method is designed and developed from scratch. The earlier matching method is based mainly on geometric and attribute (road name) constraints while the latter one is also using topological information. The algorithms work completely automated and can be applied to any region with data coverage. The matching is performed in a case study covering the area of Gothenburg, the second largest city in Sweden. Both matching algorithms returned satisfying results with acceptable matching errors, while the node-based method has even slightly better accuracy. The node-based method is also substantially more computationally efficient. We believe that our node-based algorithm is very useful for matching (network) VGI-data with authority data on both local and national level.

Keywords: Geographic data, VGI, OSM, Segment-based matching, Node-based matching,

1. Introduction

The amount of user generated geographic data (so called volunteered geographic information, VGI; Goodchild 2007) has increased substantially during recent years. A common approach to study the increasing VGI data is to compare it with authority data. This comparison could be targeting the quality of the VGI data (Haklay, M. 2010; Girres and Touya, 2010; Ludwig et al., 2011; Neis et al., 2012; Al-Bakri and Fairbairn, 2013) as well as its semantic (Al-Bakri and Fairbairn, 2012). Recently, specific open source tools to perform VGI quality evaluation has been released (Graser et al., 2014). Most earlier quality studies of VGI data was based on an approach of comparing number of features and/or total feature length of objects, while there are some new studies that are based on matching individual features in the VGI and the authority datasets (Ludwig et al., 2011; Koukoletsos et al., 2012).

The general aim of this study is to investigate matching techniques between VGI and authority geographic dataset. More specifically the aim is to match road features (network data) in the VGI dataset *OpenStreetMap* (OSM) with Swedish authority dataset on national level. This implies that the matching techniques must be of high matching quality (i.e., match the correct corresponding features in the two datasets).

There are two main types of matching routines in the literature: segment-based and node-based. The first approach, denoted *segment-based matching*, is based on comparison of segments that create a candidate list; then a selection of the best match in the candidate list is done by e.g. statistical methods. The second approach, denoted *node-based matching*, is based on node matching in the first step and then an evaluation of geometrical and topological properties of the segments/links between the nodes. Both approaches can use semantic information such as road names and road types (if available) in the matching process.

In this study we develop one algorithm for each type (section 3 and 4). The algorithms borrow several ideas from previous work, but include also new techniques especially targeting OSM and authority data. Then we compare the algorithms in a case study (section 5). The paper concludes with a discussion and conclusions.

2. Related work

Java Conflation Suite (JCS) (Vivid Solutions, 2014) is an open source Java library that contains a road network matching tool. The automated road network matching of two datasets is based on a node-based approach. Within a maximum searching distance around each node of the reference dataset, the best matching node of the other dataset, based on distance, is selected. Edges are matched using the Hausdorff distance, edge length and angle measurements. The algorithm splits matched edges when the length differences are too large to create more similar geometries.

Stigmar (2005) adapted the JCS algorithm and increased the matching quality by adding three extensions specifically tailored for her involved datasets. In a pre-processing step the geometry of the topological dataset is simplified. Furthermore she added two steps to the original JCS algorithm to match yet unmatched segments by first looking at their topological relationships, and then putting buffers around the remaining unmatched segments. The best matching pair within the buffer is then chosen depending on a measurement of distance and angular difference.

Volz (2006) proposed another node-based matching algorithm. He started by a rubber-sheeting algorithm to remove the geometric distortions between the datasets. Then nodes that have a high likelihood of correspondence between the datasets were identified; these nodes were denoted seed nodes. By using the seed nodes as starting points a combined node and link matching was performed to identify one-to-one matches. Then, in the next step one-to-many matches were resolved.

Mustière and Devogele (2008) developed a node-based matching routine (with inclusion of link matching) that especially targeted matching of datasets with different level of details. The first step was to create graphs of the original networks. Then a prematch of nodes is performed by a distance criterion followed by a pre-matching of links. These prematches result in lists of candidates. Then, for each node in the less detailed dataset a match is performed to the nodes in the other datasets while considering the consistency in the list of candidates.

Toomanian et al. (2013) studied the problem of integrating heterogeneous datasets in a web portal environment. As a sub-problem they matched linear features (in their example it was administrative borders, but it could be any type of network). Their node-based matching method started by a distance-based-node-matching followed by topological and geometrical line segment matching. A special feature of their matching routine was that they handled cases where there was no true correspondence of the datasets.

Walter and Fritch (1999) presented a segment-based method for matching road networks from different datasets and data models. The matching pairs are found by buffering the referent, and from this buffer finding potential matching elements in the other dataset. For the potential pairs parameters such as angular difference, length, shape, and distance between elements are examined. The approach is based on statistical investigations between the two datasets.

Ludwig et al. (2011) developed a segment-based matching algorithm between OSM and a reference dataset; their algorithm is partly based on the work by Walter and Fritsch (1999). An initial list with matching pairs is created by buffering the reference dataset with different sizes. All OSM data within a buffer segment are linked to the reference segment. For each list similarities are calculated and ranked considering name and category attributes. Only the highest OSM ranks are kept as the final candidate list.

Koukoletsos et al. (2012) present another segment-based matching algorithm which is then evaluated for OSM and the Intergrated Transport Network (ITN) dataset from Ordnance Survey. The datasets are divided into 1 km^2 tiles which are computed separately to achieve better performance and to obtain a better representation of the heterogeneity of OSM in the results. The algorithm is based on comparison of segments where the correspondence measures are derived from expected quality of the OSM dataset. The algorithm produces robust results with low matching errors, 2% in urban areas and around 3% in rural areas.

3 Segment-based method

Our method is based on Koukoletsos et al. (2012) and, apart from pre-processing (generalization and segmentation) the data, matching procedure consists of the following steps in two levels (see Will, 2014, for details) (in the description AD stands for an authority dataset and VGI for a VGI dataset):

Segment level:

1) Buffering: for each segment in AD, a list of possible matching segments in VGI located within a buffer around an AD segment is found. These candidates should also have a certain orientation.
2) 1:1 matching: if the created list has only one candidate and its length is not exceeding three times more than the corresponding AD segment, they are considered a matched pair.
3) Exact name matching; the road name of AD segment is matched to the VGI candidates. If only one segment has an exact name then this pair is regarded as a match. If several segments are found, the closest one is chosen as a match.
4) Similar name matching: this looks for the most similar name between VGI candidates and AD segments as the road names are not always correctly spelled or might be written in abbreviations, especially in VGI.

5) Distance matching: AD and VGI segments are matched based on the distances between possible matches regardless of their road name attributes.

Feature level:

6) Feature recomposing: segments along with their matching information are transferred to the feature level. A feature is then considered matched if the matched segments constituted more than half of its length. The corresponding feature/s will be selected based on the length proportion being matched to the feature/s in the other dataset.
7) VGI feature name similarity: this checks the name similarity of the non-matched VGI feature with the AD features located within the buffer twice the GPS accuracy (estimated to be 10 meters) around the VGI feature.
8) Final check: if matching information of non-matched features in one dataset exists in the other dataset, those features are then assigned to their corresponding features in the other dataset where they are listed as a match of

4 Node-based method

The second algorithm is designed by applying the concept of the node-based method. In this algorithm, the network datasets need to be stored in an adjacency list data structure (cf. Sedgewick, 2006). The algorithm involves the following four important checks (after two steps of pre-processing, Figure 1):

1) Node comparison: a list of neighbouring VGI nodes is found for each node in AD dataset. These neighbours are in a range of 10 meters distance from AD nodes.
2) Name check: if there is an AD node with more than one neighbouring VGI node, the node with higher semantics similarity is chosen as the best neighbour. For that purpose, a sequence of names from each neighbouring VGI node is created and compare to the sequence of the AD node.
3) Topology check: all the links connected to the AD node are checked with each link connected to its neighbour.
4) Geometry check: Several conditions must be considered in order to find the best matches. The first measure for considering two links as a possible match is that the azimuth discrepancy of their first segments is in the range of tolerance.

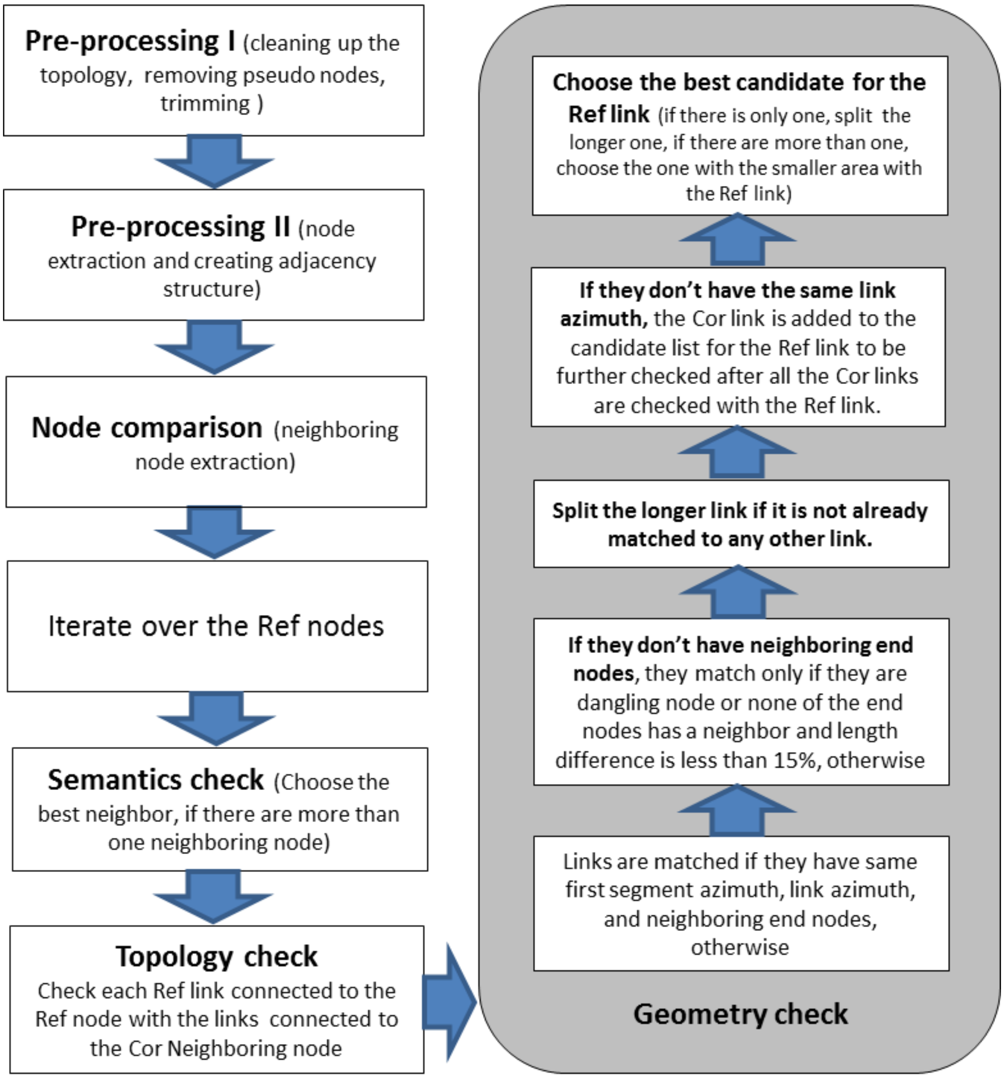

Pre-processing I (cleaning up the topology, removing pseudo nodes, trimming)
Pre-processing II (node extraction and creating adjacency structure)
Node comparison (neighboring node extraction)
Iterate over the Ref nodes
Semantics check (Choose the best neighbor, if there are more than one neighboring node)
Topology check Check each Ref link connected to the Ref node with the links connected to the Cor Neighboring node
Links are matched if they have same first segment azimuth, link azimuth, and neighboring end nodes, otherwise
If they don't have neighboring end nodes, they match only if they are dangling node or none of the end nodes has a neighbor and length difference is less than 15%, otherwise
Split the longer link if it is not already matched to any other link.
If they don't have the same link azimuth, the Cor link is added to the candidate list for the Ref link to be further checked after all the Cor links are checked with the Ref link.
Choose the best candidate for the Ref link (if there is only one, split the longer one, if there are more than one, choose the one with the smaller area with the Ref link)
Geometry check

Figure 1: A general flow diagram of the developed node-based algorithm.

5. Case study

5.1 Study area and data

The study area is Gothenburg, Sweden, and the surrounding region in total 861 km². We use the real-estate map dataset from Lantmäteriet as authority data, and hereafter called LM. The dataset is produced for presentation in the scale 1:5,000 – 1:20,000 and its positional accuracy is specified to be less than 2 meters (standard deviation).

The OSM data are from 16 April 2014. All roads with the tag “psv” (Public Service Vehicle) are removed; these roads most often represent tram or bus roads. The OSM road data is not restricted to have a node at intersections. To make the geometry between the two datasets more similar, all OSM features are split at intersections. For the segment-based method the OSM data of the study area are split to 861 tiles of 1 km^2 each.

The OSM data are transformed from the reference system WGS 84 (Lat,Long) to SWEREF 99 TM (which is a UTM 33 projection of the Swedish implementation of ETRS 89). The LM data set is provided in SWEREF 99 TM.

5.2 Implementation

The segment-based algorithm is implemented as a plugins to PyQGIS using Python and the QGIS API. The node-based algorithm is chiefly implemented using ArcPy package. The other package utilized for creating the KD-tree index for the extracted nodes is the spatial module of the SciPy package.

5.3 Comparison measures

The comparison is based on: (1) percentage of correct matches and (2) computational performance.

5.4 Result from the segment-based method

5.3.1 Matching quality

Table 1 shows the results of the segment-based matching method for the whole study area with almost 80 percent matched features in both datasets. Figure 2 also demonstrating the distribution of matching in the study area for each dataset.

Table 1: Matching results for total study area.

Dataset	Total length [m]	Length matched [m]
OSM	4596570	3550564 (77%)
LM	4691594	3800412 (81%)

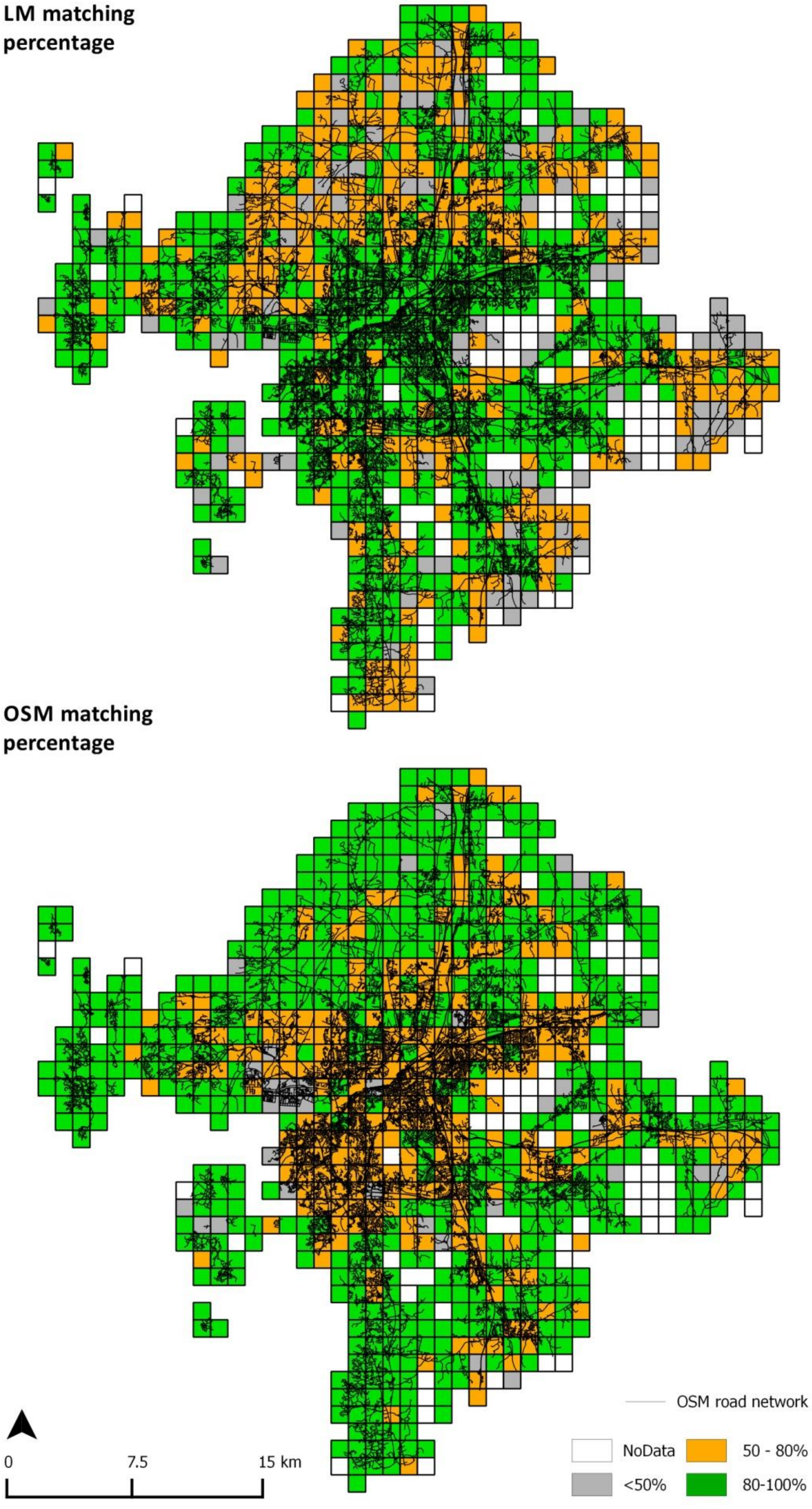
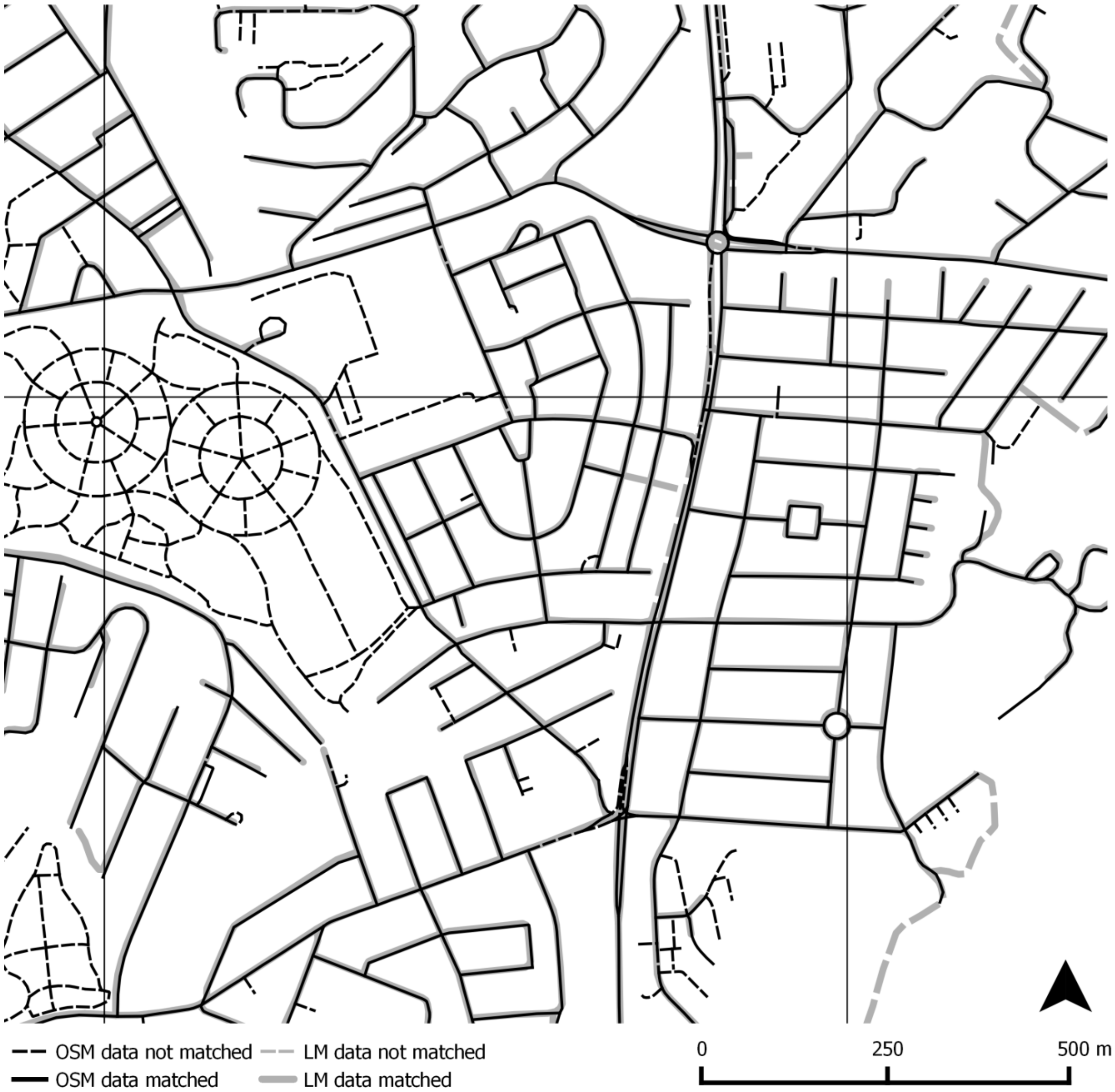


Figure 2: Matching percentages per tiles.

More detailed results show that OSM has lower percentage of matching in the urban area. This is due to the more completeness and higher density of OSM in the urban areas. This finding is in line with other research outputs (cf. Koukoletsos et al. 2012). Figure 3 demonestrates the density of OSM and LM in an urban area. It is clear that OSM has features which do not exist in LM and will remain unmatched. A test area was evaluated in terms of correctly matched/unmatced features whose result has 86% accuracy.



OSM data: © OpenStreetMap contributors; LM data: © Lantmäteriet, Dnr: i2012/927

Figure 3: Example of the segment-based matching result.

5.3.2 Computational efficiency

The average execusion time for the segmen-based algorithm was investigated for each step to assess the computational efficiency of this approach. The following list shows the running time of each part of the algorithm in seconds:

1- Pre-processing: 10,959
2- Step 1: 2,500

3- Step 2: 38
4- Step 3: 396
5- Step 4: 141
6- Step 5: 300
7- Step 6: 1,401
8- Step 7: 252
9- Step 8: 1,514
10- Step 9: 612

This algorithm takes more than 4 hours to be accomplished.

5.4 Result from the node-based method

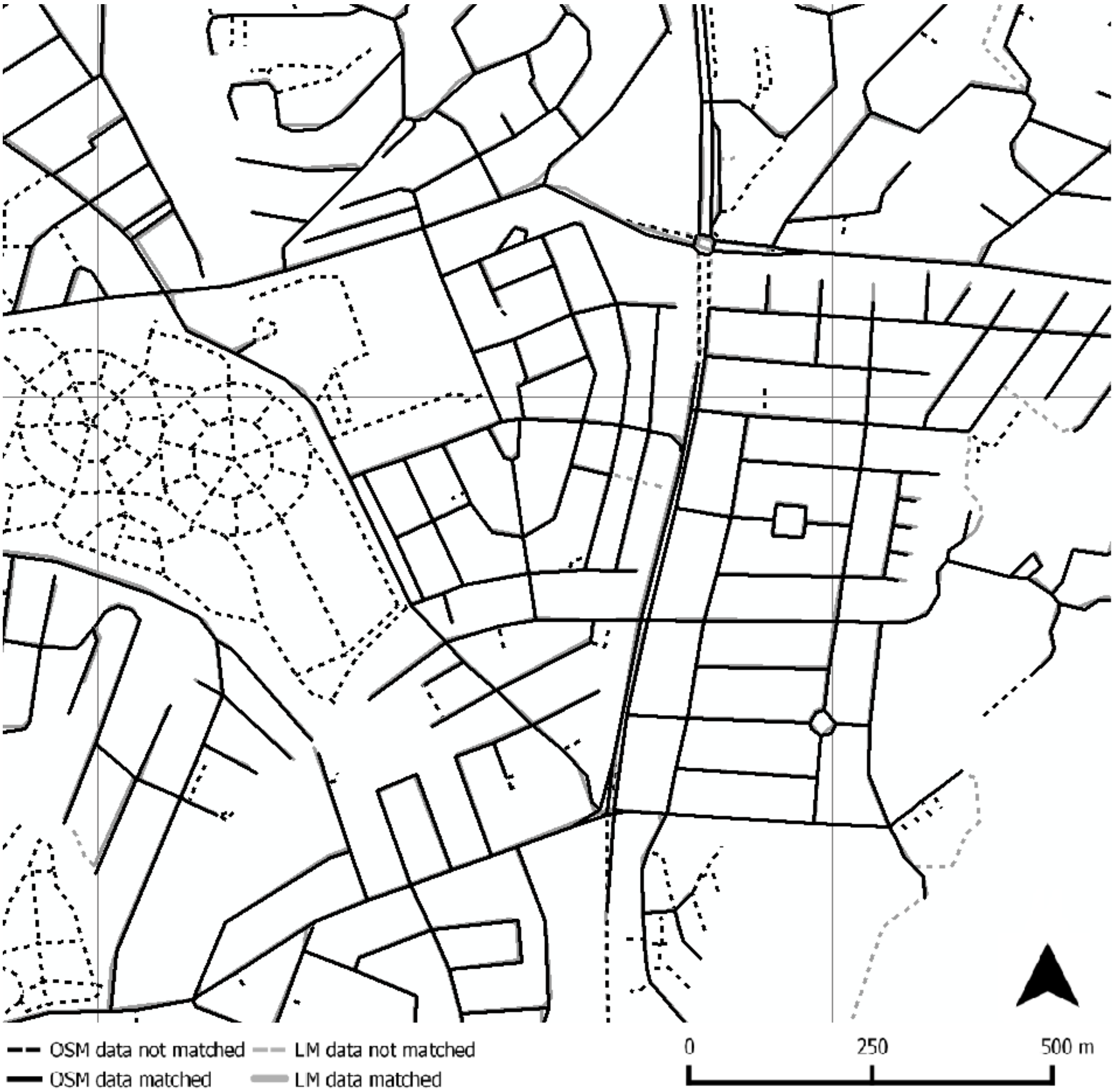
5.4.1 Matching quality

Table 2 shows the results of the node-based matching algorithm for the whole study area. This algorithm also matched almost 80% of features in both datasets. In this algorithm, the tiling approch is not used, hence the result of accuracy per tile is not producesd. Nevertheless, 10% of the features of the study area was randomly chosen as the evaluation features. They were then manually investigated to evaluate the overall accuracy of the results. The evaluation showed that 92% of the features are correctly matched. This estimation is calculated regarding the proportion of the length of the features which are correctly matched/unmatched to the total length of the evaluation features.

Table 2: Matching results for total study area.

Dataset	Total length [m]	Length matched [m]
OSM	4489797	3561441(79%)
LM	4542694	3594120(79%)

The same area is shown in Figure 4 as in Figure 3. Comparing the figures can reveal that the matched features are nearly the same; however, the segment-based method is better in some cases while the node-based is more accurate in some other.



OSM data: © OpenStreetMap contributors; LM data: © Lantmäteriet, Dnr: i2012/927

Figure 4: Example of the node-based matching result.

5.4.2 Computational efficiency

In order to evaluate the computational efficiency of the algorithm it was run 10 times and the average running time (in seconds) for each part is as follows:

1- Pre-processing I: 37
2- Pre-processing II: 161
3- Node-comparison: 50
4- The rest of algorithm: 180

The name, topology and geometry checks are used in the body of the algorithm for measuring the similarity level of features. In the other words, they are combined and it is not possible to measure the execution time for each of them. It could be possible to modularize each of the checking sections in the next versions of the algorithm. Presently, the whole algorithm takes almost 7 minutes to be executed on the test datasets.

6. Discussion

6.1 Comparison of the matching methods

In the matching procedure both algorithms are suffering from almost same barriers. The heterogeneous geometrical representations, varying positional accuracy across the study area, different representations of complicated structures such as roundabouts and bridges, multi-carriage roads, and data errors are the most important barriers. In the node-based algorithm another problematic case is where a link makes a ring. A ring is a link which has the same start- and end-nodes. These cases are excluded from the current version of the algorithm as their link azimuth is not defined. There are, however, various solutions for this problem. Both datasets are also lacking accurate descriptive information such as name.

The segment-based method is benefiting from buffer for creating the candidate list. The buffer is built around each segment of a link to make the candidate list. On the other hand, the node-based method is focusing on the neighbouring nodes of two datasets. This approach is computationally simpler than buffering around a link or segment. Moreover, this can be supported by an efficient spatial indexing such as KDTree. The segment-based method splits the study area into tiles in order to consider heterogeneity of urban and rural areas and also to increase the time-efficiency of the algorithm. The tiling can be considered as a B-Tree indexing with depth of one which may not be the most efficient one. As the segment-based method requires splitting the links into their segments, the node-based approach also needs to extract the nodes of two datasets.

The results for both matching algorithms are satisfying with a slightly better quality in the node-based approach. The accuracy of node-based and segment-based algorithms is respectively 92% and 86% in the evaluation area. Regarding the computational efficiency, the node-based method was completed in almost 7 minutes while the segment-based algorithm took nearly four and half hours to be executed. However, different programming packages and indexing techniques have been used for the implementation of the algorithms and therefore the results are not fully comparable. Nevertheless it is still apparent that the node-based algorithm is substantially more efficient than the segment-based method.

6.2 Advantages and disadvantages

The segment-based approach selects the segments which are within the buffer around a reference segment. This will cause to have limited number of candidates which are expected to be highly similar. On the other hand, the node-based approach chooses the neighbouring nodes without considering the condition of the links connected to them. This makes the node-based approach to be more complicated in detecting the correct matches. Despite this fact, the number of candidates in this approach is still reasonably few. As the node-based algorithm is focusing on the neighbouring nodes, the result can be affected if there are clusters of nodes. For example, in the pre-processing I (See Figure 1), the bridges along the multi-lane carriageways are split in the clean-up as they cross over another multi-lane carriageway. This causes many short links to be created whose nodes neighbour each other. On the other hand, the node-based algorithm benefits from the topological information coming with the nodes.

This information enhances the matching results and decreases the mismatches by tying the matched links at their shared nodes. The other essential advantage of the node-based algorithm is its high computational efficiency which enables us to consider more complicated conditions to increase the accuracy of the results.

6.2 Complete versus incremental matching

In an operational mode it would be useful to perform a matching routine a couple of times each year. This is e.g. necessary to monitor the quality evaluation of OSM. To perform such a monitoring it could be interesting to utilize an incremental matching approach. In this approach, result from previous matching is stored and only changes to the two datasets are considered. The incremental approach is of special interest if there is a labour-extensive manual step after the automated matching. If the matching routine is completely automatic then an incremental approach is not of interest.

7. Conclusions

In this paper we have investigated two main approaches for matching network dataset: segment-based and node-based. The problem studied was the matching road features in *OpenStreetMap* (OSM) and the real-estate map dataset from Lantmäteriet. To perform, one algorithm of each type was developed and implemented. A case study was performed in an area around Gothenburgh, Sweden which includes both urban and rural regions. The case study reveals that both the segment-based and node-based algorithms provided good matching results (both had a matching error around 10%). The main difference between the methodologies lies on the computational side. The node-based approach is computationally much more efficient than the segment-based approach. Our recommendation is therefore that a node-based approach should generally be used for matching OSM with authority datasets.

References

Al-Bakri, M. and Fairbairn, D. 2012. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *Int. J. Geographic Information Science*, 26*:*1437–1456, doi:10.1080/13658816.2011.636012.

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

Devogele, T., J. Trevisan and L. Raynal. 1996. Building a Multi-Scale Database with Scale--Transition Relationships. In *Advances in GIS Research II*, ed. M -J. Kraak, M. Molenaar and E. M. Fendel, pp. 337-351. London: Taylor & Francis.

Girres, J-F., and G. Touya. 2010. Quality assessment of the french OpenStreetMap dataset. *Transaction in Gis* 144: 435-459. doi: 10.1111/j.1467-9671.2010.01203.x

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69: 211-221. doi: 10.1007/s10708-007-9111-y.

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 volunteered geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B* 37: 682–703. doi: 10.1068b35097.

Koukoletsos, T., M. Haklay, and C. Ellul. 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. Transaction in GIS 16(4): 477-498. doi: 10.1111/j.1467-9671.2012.01304.x

Ludwig, I., A. Voss and M. Krause-Traudes. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. doi 10.1007/978-3-642-19789-5_4. In A*dvancing Geoinformation Science for a Changing World*, Lecture Notes in Geoinformation and Cartography, vol. 1, ed. Geertman S., W. Reinhardt and F. Toppen, pp. 65-84. Berlin Heidelberg: Springer-Verlag.

Mustière, S., and T. Devogele. 2008. Matching networks with different levels of detail. Geoinformatica 12 (4): 435–453. doi: 10.1007/s10707-007-0040-1

Neis, P., D. Zielstra, A. Zipf. 2012. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet 4: 1-21. doi:10.3390/fi4010001.

Sedgewick, R., 2002. *Algorithms in C – Part 5 Graph algorithms*. Addison Wesley.

Stigmar, H., 2005. Matching Route Data and Topographic Data in a Real-Time Environment. In Hauska, H. and Tveite, H. (eds.), *ScanGIS'2005 - Proceedings of the 10th Scandinavian Research Conference on Geographical Information Sciences*, Stockholm, Sweden, pp. 89-107.

Toomanian, A., Harrie, L., Mansourian, A., Pilesjö, P., 2013. Automatic integration of spatial data in viewing services, *Journal of Spatial Information Science*, 6:43-58, doi:10.5311/JOSIS.2013.6.87.

Vivid Solutions. 2014. Java Conflation Suite. Retrieved 14 Februray 2014, from http://www.vividsolutions.com/JCS/.

Volz S., 2006. An iterative approach for matching multiple representations of street data. *Proceedings of the ISPRS workshop on Multiple Representation and Interoperability of Spatial Data*, pp. 101–110, Hanover (G).

Walter, V. and Fritsch, D., 1999. Matching Spatial Data Sets: A Statistical Approach, *International Journal of Geographical Information Science*, Vol. 13, No. 5, pp. 445-473.

Will, J., 2014. Development of an automated matching algorithm to assess the quality of the road network of OSM in Sweden. MSc thesis. Department of Physical Geography and Ecosystem Science, Lund University, Student thesis series INES nr 317.