Student thesis series INES nr 317

Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network

- A case study in Göteborg, Sweden

Julian Will

2014

Department of

Physical Geography and Ecosystem Science

Lund University

Sölvegatan 12

Julian Will (2014)
Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network- A case study in Göteborg, Sweden
Master degree thesis, 30 credits in *Geomatics*
Department of Physical Geography and Ecosystems Science, Lund University

Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network

- A case study in Göteborg, Sweden

Julian Will

Master thesis in Geomatics, 30 credits

June 2014

Supervisors:

Lars Harrie, Lund University

Daniel Garcia, Kartena AB

Department of Physical Geography and Ecosystem Science

Lund University

Abstract

Julian Will

Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network - *A case study in Göteborg, Sweden*

In the last decade a new and alternative source of geospatial data has become available, so called Volunteered Geographic Information (VGI). Private individuals voluntarily collect large amount of geographic data for a joint project. This data are usually free under certain licence restrictions.

One of the most well known examples of VGI is the OpenStreetMap (OSM) project. Due to its increasing popularity it has developed in less than 10 years to a useful data source. Private, scientific and commercial users are wondering how good is these voluntarily collected data and can it be an alternative to expensive authority and commercial datasets.

This study presents a quality assessment of the OSM road network against a reference dataset from Lantmäteriet (Swedish National Mapping Agency). In order to allow a meaningful evaluation, corresponding features in both datasets have to be identified. This process is commonly referred as matching. Difficulties of automatic matching occur due to the different geometric representations of the two datasets.

An automated matching algorithm is developed to match road network data from OpenStreetMap and Lantmäteriet. The method is adopted from Koukoletsos et al. (2012) and developed further to incorporate feature correspondence. The matching is based mainly on geometric and attribute (road name) constraints. Besides minor pre-processing steps the algorithm works completely automated and can be applied to any region with data coverage in Sweden.

The matching is performed in a case study covering the area of Göteborg, the second largest city Sweden. The created feature correspondence is then used to calculate the quality elements completeness, positional accuracy and thematic accuracy.

The matching algorithm returned good results with an acceptable matching error. The quality assessment revealed that OSM can be seen as a proper data source which has some reservation regarding the attribute accuracy.

Keywords: OpenStreetMap, automated matching, quality assessment, volunteered geographic data, linear features, Physical Geography and Ecosystem analysis

Supervisors: **Lars Harrie, Daniel Garcia.**
Master degree project, 30 credits, in Geomatics. 2014.
Department of Physical Geography and Ecosystem Science, Lund University.
Student thesis series INES no 317

Preface

This study is a master thesis in the master program *Geomatics* at the department of Physical Geography and Ecosystem Science at Lund University. The work was carried out in cooperation with Kartena AB in Göteborg.

I would like to express my gratitude to supervisor Lars Harrie in the department of Physical Geography and Ecosystem Science at Lund University. His helpful read-throughs, comments and suggestions have been very directive and valuable.

I would also like to thank my supervisor Daniel Garcia at Kartena AB for his valuable advice and support which I received during my work. Furthermore, I thank the employees at Kartena AB for encouraging comments and a pleasant work environment.

Table of Contents

1 Introduction

1.1 Background

Geographic data are one of the major data sources for the planning and the development of our society. Nearly everything that happens or what we do is connected to a geographic location. Geographic information is used in all kinds of fields, from more traditional fields such as transportation or environmental services to health care or economics (Longley 2005). The collection of geographical information is usually expensive and done by private and governmental organizations.

During the last 10 years a new type of geographic information has become available, the so called Volunteered Geographic Information (VGI). Goodchild (2007) defines VGI as the volunteer collection of geographic information from a large number of private individuals. VGI can be seen as a special case of User Generated Content (UGC) such as Wikipedia, as it collects information with a spatial component (Neis et al. 2012). The developments of GPS and web 2.0 technologies make it possible to deal with geodata even for non specialist in this field (Haklay 2010).

Due to the high collection cost, geographic data are expensive and has high licensing costs. The high costs often limit the use of geographical data for analysis or products by companies, especially for smaller companies. In contrast to that VGI data are usually free under certain licence restrictions. Therefore, it is interesting to see if VGI, as a no or low cost alternative can compete with traditional geographic information.

One of the most well known examples of VGI is the OpenStreetMap (OSM) project (OpenStreetMap, 2014a; Haklay, 2010). It aims to provide a free and editable map data of the world under certain licence restrictions. Individuals save their collected geographical information, most often gathered with GPS devices, in a database. The data can then be accessed by everyone. More than 1.5 million registered contributors have created a useful data source in less than 10 years. OSM data are nowadays used in several map services (Ludwig et al. 2011; Neis et al. 2012).

A common issue of OSM and VGI data in general, is their trustworthiness (Haklay, 2010; Goodchild and Li, 2012). The data are gathered mostly by amateurs using different collecting methods without strict standards. Therefore, it exists no standard quality assessment of the data. It can be argued that the OSM community itself controls the quality and ensure that it meets a certain quality level. However, it is often required to present stronger quality controls of data in order to use the data for analysis or products. Qualities of geographical data are commonly described by the following elements: completeness, positional accuracy, thematic accuracy, logical consistence and temporal accuracy (ISO, 2002).

A quality analysis of a whole dataset is a extensive and time consuming process. A convenient approach is to investigate the quality of distinct object classes, e.g. road network.

Road objects often build the initial skeleton of geographic dataset and are used as orientation for other features (Ramm et al., 2011). Therefore, it seems suitable to assess the quality of the road network of OSM to obtain a quality assessment for OSM in Sweden.

1.2 Problem statement

Kartena, a geographic IT consult company with the focus on online map-based products, is interested in a quality assessment of OSM data in Sweden. If OSM data meet their quality requirements it could be used instead of authority geographic data in their map services. Furthermore, a quality report of OSM is needed in order to be able to offer the use of OSM data to customers. An automatic quality evaluation method is preferred in order to facilitate the repetition of the calculation and the application of the method in different areas.

The most accurate method of quality assessment is to compare a dataset against its true value. However, to assess the true value is expensive, complex and time consuming. A more suitable method is to compare the quality of a dataset relatively to a dataset which has a documented high quality.

A convenient approach is to identify the same features in the two datasets which represent the same object in reality and then calculate the different quality measurements between the matching pairs. The identifying process is commonly referred to as matching and is also used for different application, for example dataset merging. The processes is usually automated, as manual matching is very time consuming and only for small areas applicable. Difficulties of automatic matching occur due to the different geometric representation of the two datasets. The use of different data models, acquisition methods, geodetic reference systems and uncertainties in measurements leads to different locations of the same feature in different datasets (Goodchild and Hunter, 1997; Walter and Fritsch, 1999).

1.3 Aim

The general aim of this report is to evaluate the quality of the road network of OSM dataset of Sweden relative to a reference dataset.

The specific aims are:

(1) Development of an automatic matching routine to match subsets of the road network of OSM dataset with the real-estate map of Lantmäteriet (Swedish National Mapping Agency).
(2) Calculation of the quality measurements: completeness, positional accuracy and thematic accuracy for the OSM road network dataset based on result of the matching process.

1.4 Method

Fundamental knowledge of OSM, quality assessment of spatial data and matching algorithms is gathered through literature research. The literature studies are conducted using mostly online research databases, such as Google Scholar and library searches. The gathered knowledge is then used to outline a more specific methodology to achieve the aims of this study. Through a comprehensive evaluation of matching routines the routine which is best suitable for this study is chosen. A main part in this study is then the implementation and modification of the chosen matching routine as an automatic process. This matching routine is then used to match OSM and Lantmäteriet data. The matching results are then used to conduct a quality evaluation of the road network of OSM.

1.5 Disposition

The study can be roughly divided into three parts, a theory section, method and implementation and results and discussion.

The theory section starts with an overview of OSM, where the OSM project is described in detail. After this quality aspect, including quality standards, quality measurements and an overview of studies concerning the quality of OSM are presented. The theory section ends with a summarization of relevant matching algorithms for this thesis.

In the beginning of the method and implementation section the study area and the used datasets are presented. The previously described matching algorithms are evaluated before the matching method is described. This is followed by a description of the applied methodology for quality assessment. The method section ends with a summarization of the implementation of the method.

The result method starts with the presentation of the matching and quality assessment results. The results are then evaluated in the discussion section. The conclusion summarizes the result of this study.

2 OpenStreetMap

OpenStreetMap is a VGI project which provides a free and editable map of the world created by volunteers. Anyone is invited to contribute and to improve the OSM project. The data are free to assess and to download under an open data licence.

People, actively contributing to OSM by improving the OSM map, are hereafter refereed as *contributors*. People, only downloading and using OSM data, are hereafter refereed as *users*.

2.1 History of OSM

The OpenStreetMap project was founded by Steve Coast in the United Kingdom in 2004 (Ramm et al., 2011). He started the project in frustration of non-existent free editable data. The initial idea was to create a road network dataset covering the U.K., where the geographic information is voluntarily collected with GPS devices by individuals. The data are uploaded to a database from where the data can be accessed, maintained or downloaded without restrictions.

The OSM project was spreading fast over the internet and soon even people outside the U.K. started to follow and collect data (Ramm et al., 2011) (see figure 1). The community also began mapping all kind of features and not only roads, e.g. Point of Interest (POI), buildings and land use. Besides collecting data the community developed open source tools for uploading, editing and maintaining data as well as to render the data. This results in an increasingly user-friendly interface, making OSM more attractive to people without background in the geographic information field. In less than 10 years nearly one and half million people have contributed to the project (see figure 2).

In 2006 the OpenStreetMap foundation was established to support the OSM project (OSM Foundation, 2014a). It is a non-profit organisation and its purposes are to represent OSM and stimulate the growth of the project. The organisation owns the needed infrastructure as for example server and domains but not the data itself (Ramm et al. 2011). The foundation takes over tasks which are difficult to accomplish for a community, for example to negotiate with data providers and to manage donations.

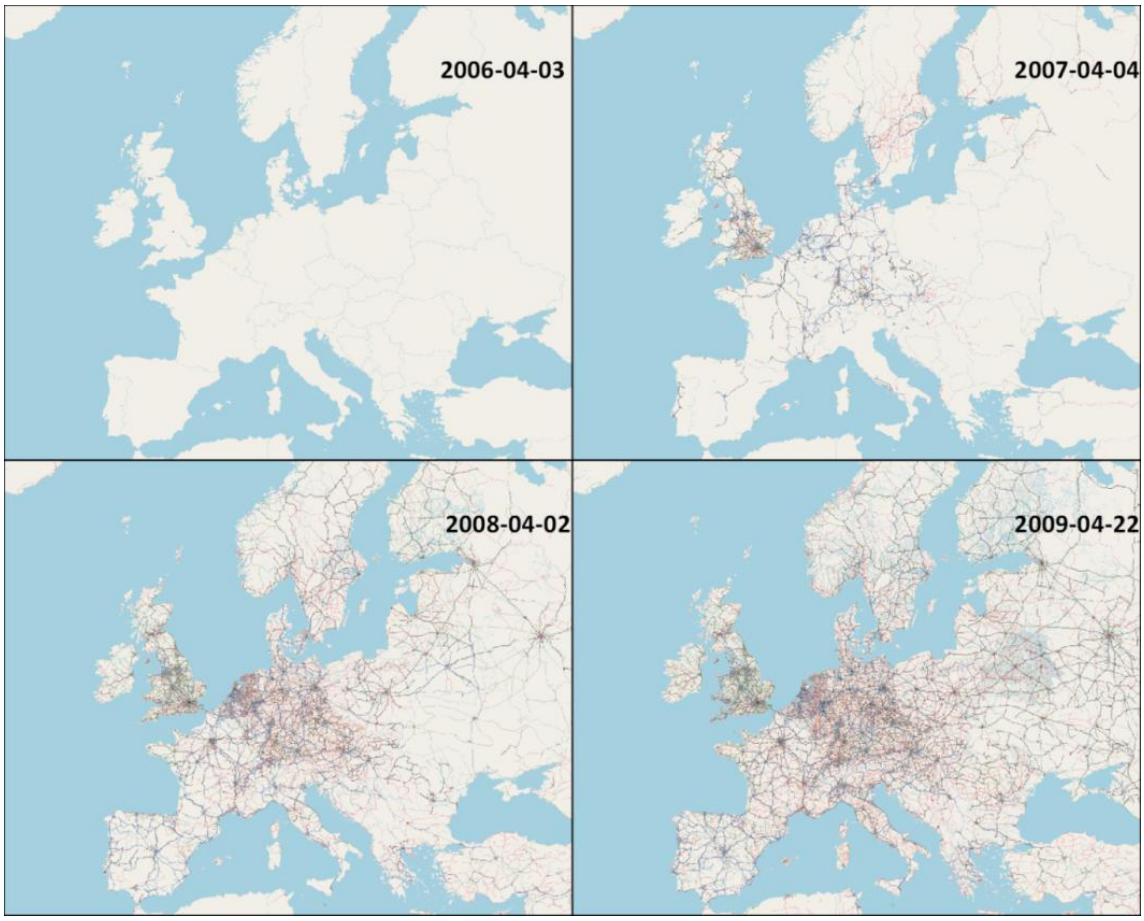


Figure 1: Coverage progress of OSM in Europe between the years 2006 and 2009 (Images from OSM Developer, 2014).

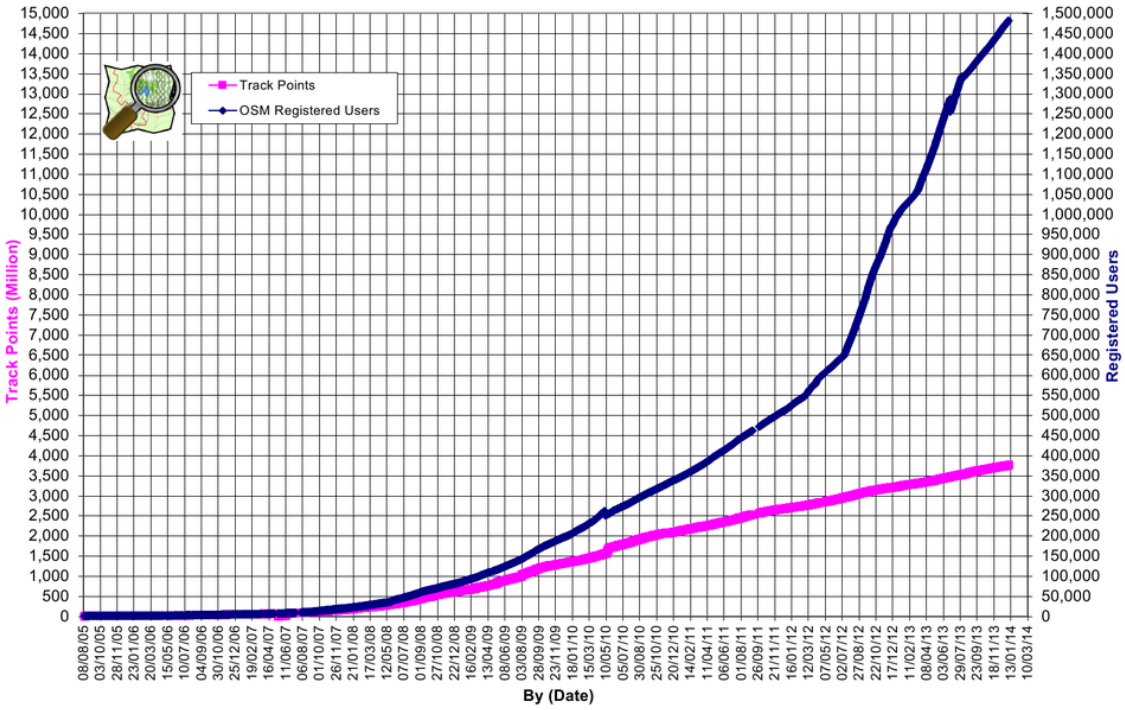


Figure 2: OSM contributors and GPS points upload development from 2005 until 2013 (from OSM Wiki, 2014a).

2.2 Data collection and editing

Anyone is welcomed to participate in the OSM project. A free registration is necessary to be able to add or to edit data. A beginner's guide is available at the Wikipedia page of OSM. It summarizes the basic steps of data collection. Over the years the OSM community developed other collection methods besides GPS tracking. Digitalization from orthophotos is possible after Yahoo in 2006 (until 2011) and Bing in 2010 granted the right to trace from their orthophotos (OSM Wiki, 2014b; OSM Wiki, 2014c). The Walking Papers and Field Papers methods (described below) were introduced to enable mapping without using a GPS device and to simplify attribute mapping (Ramm et al., 2011; OSM Wiki, 2014d). In 2007 OSM was allowed to import the TIGER and AND dataset. The TIGER dataset contain topological data for the entire USA. The AND dataset contain the complete road network of Netherland and major roads of India.

GPS survey is the most popular method of data collection. Equipped with a GPS device the contributor maps features in the reality. The GPS device frequently records the position of the contributor. The coordinates are usually stored in the reference system WGS 84 and they have a positional accuracy of around 5 meter (Ramm et al., 2011). While walking, cycling or driving the mapper notes relevant attribute information. At home the GPS data are uploaded to an OSM server. This is done to have a proof that the actual data are collected from a survey and that other contributors can access the data as well. The GPS traces are then used to create and edit features to the OSM database. Anybody in the community can use the GPS log to edit the map, but it is preferable that the person who made the survey also do the editing.

The OSM community has the permission to use orthophotos from Bing (from Yahoo until 2011) and some other data providers. Digitizing from images is often used to map basic features in unmapped area. Furthermore aerial tracing has been proven to be useful to map features, especially buildings in dense urban environments. However, images contain a positional error and the position should be checked if possible against available GPS tracks. Furthermore, images are seldom up-to-date and the area might have changed since the image was taken. Another major drawback is that attribute information cannot be added to the digitized features unless the contributor has local knowledge about the area.

The Walking-Papers and Field Papers are two similar concepts and are designed to allow simple local and attribute mapping (Ramm et al., 2011; OSM Wiki, 2014d). Field Papers is a continuation of Walking-Papers which has extended functionality, but the main principals are the same. Figure 3 shows an overview over the workflow of Field Papers. Field Papers provide a service to create an atlas over an area which is wished to map. The atlas area is print out as an OSM map. In the field features and attribute information are added to the map. An image or a scan of the edited map is then uploaded to the website. Through certain objects on the printed map, the map is automatically connected to back to its atlas

and hence georeferenced. This map is then used in one of the editors to add features to the OSM database.

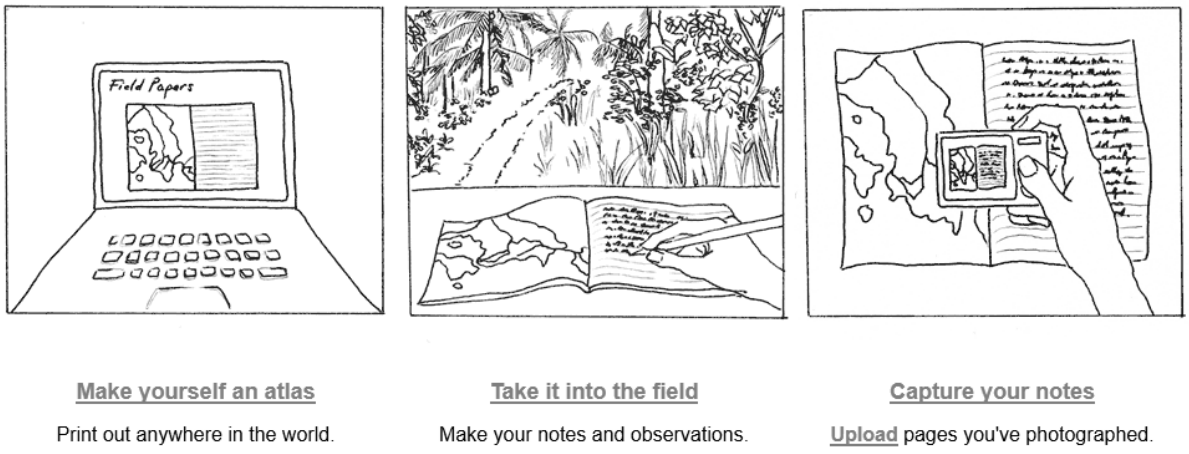


Figure 3: Concept of Field Papers which is an OSM data collection method (from Field Papers, 2014).

Editors are used to make actual changes to the OSM data, by providing functionality for adding or modifying feature. There are several editors, the three main editors are (OSM Wiki, 2014e): iD, Potlatch 2 and JOSM (Java OpenStreetMap Editor). The basic procedure is to digitize feature along GPS tracks, orthophotos and Walking/Field Papers maps and to add attribute information. iD and Postlatch 2 are imbedded within the OSM web page. They have simple user interfaces providing the basic functionality to create, edit and digitize features. Both offer short introduction tutorials which makes it easy for new contributors to understand the editors. JSOM is an external editor with extended functionality, e. g. offline editing, topology check and importing of data. JOSM mainly targets more advanced contributors.

Like in other UGC project, all data in OSM can be edited, modified and deleted by any contributor. However, all changes to data are saved and can be revoked. All contributors have the same rights and there exist no hierarchy, like in other UGC project like Wikipedia. Though, a small group of currently 6 people, called Data Working Group, has extent rights to deal among others with copyright violations, serious vandalism and disputes (Goodchild and Li, 2012; OSM Wiki, 2014j).

2.3 Downloading and viewing

The download and viewing of OSM data are completely free and unrestricted. However, when using the data the licence restriction have to be followed. OSM data can be downloaded from several services.

The homepage of OSM provide the possibility to download a restricted view extent of the map window as XML data. Furthermore can the view extent of any size be exported as raster images (TIFF, JPEG) or Scalable Vector Graphic (SVG) (OpenStreetMap, 2014a)

There exist two Application Programming Interfaces (API), XAPI and Overpass API, which allow download of OSM data in XML format (OSM Wiki, 2014k; OSM Wiki, 2014l). A query map is send through URL to the API which returns the requested data. The query is based on a query language which allows the user to specify its request. XAPI query language is basic and allows the user only to download OSM data over a specific area and with certain tags. The Overpass API has a much more powerful query language which allows the user to download specific OSM data.

An OSM file, called OSM planet, containing the complete OSM data can be downloaded as well (OpenStreetMap, 2014b). The file is updated once a week and is delivered in a compressed XML format. As the planet OSM file is large (about 20 GB), OSM provide also so called Diff files. These files contain all changes made on the OSM data over some period of time. There exist minutely, hourly, daily and weekly diff files. These diff files provide an efficient way of keeping downloaded OSM data up-to-date.

Other websites offer the download of OSM data at country and city level and in different data formats (OSM Wiki, 2014i). Some of these websites offer also the possibility to download OSM data back in time.

Another common way to use OSM data are to use OSM data in Web Map Services (WMS) services. WMS is a protocol that is used to render maps on internet sites. There exist various OSM WMS services; two commonly used once are the open-source JavaScript libraries Leaflet and OpenLayers.

2.4 Data model

OSM is using a fairly simple data model. It consists of three element types: node, way and relationship (figure 4). The data are stored as Extensible Markup Language (XML) in combination with the OSM schema. Each element is described with an own XML element (<node>, <way> and <relation>). These XML elements are generally described with the following attributes (Ramm et al., 2011):

- id: individual ID number of the element (required)
- visible: a Boolean value, true if element is visible and false if element is not visible
- version: the version of the element, how often it has been changed
- changeset: id of the changeset, which describe the changes to previous version
- timestamp: time of the last edit
- user: name of user who created or edited the last time the element
- uid: id of user who created or edited the last time the element.

Furthermore, each element can has zero or more tag elements (<tag>) assigned to them. Tag elements have to be created by the contributor.

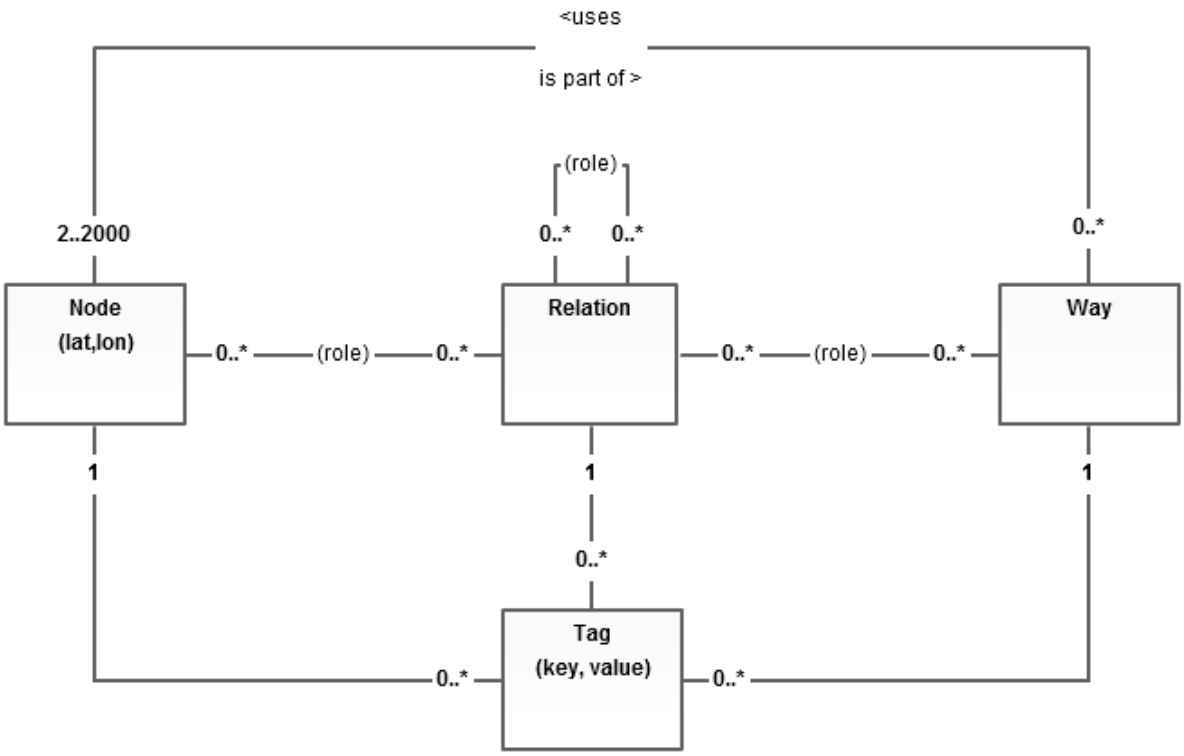


Figure 4: Simplified data model of OSM (modified from Ramm et al., 2011, p. 52).

- **Tag** elements (<tag>) are used to assign attribute information to the element types: Node, Way and Relation (Ramm et al., 2011). A tag element consists of a key and a value. A key usually describes an attribute class, for example highways or names. There exist no complete list of keys and contributors can define their own ones. However, the community has developed a comprehensive list of keys and values for most purposes and contributors are instructed to follow this informal standard (OSM Wiki, 2014f). Tags are from essential importance in the data model of OSM, without tags or wrong tag information data are wrongly displayed and searches cannot return correct results.
  Some common key examples:
  - *Key=name:*
    This key is used for naming of elements.
  - *Key=highway:*
    This key is used to classify any kind of ways which are designed for travel purpose on lands except railroads. Values can be among many others: motorway, tertiary, footpath and cyclepath.
  - *Key=Landuse:*
    Describe the land use of a polygon element.
  - *Key=shop:*
    This key is used to classify any kind of business. Values describe the kind of business.

- **Nodes** are point features and are the only element type which stores information of the geographic position (Ramm et al., 2011). The position is expressed in longitude and latitude in WGS 84 and is stored as compulsory attributes of the node element (lat and lon). Nodes are used to store POI and vertices of lines and polygons.
  XML example code for a node (represents a car park):

```
<node id="301424858" visible="true" version="2" changeset="16837"
        timestamp="2008-10-01T18:30:35Z" user="magol" uid="35601"
        lat="57.7135518" lon="11.9691117">
        <tag k="amenity" v="parking"/>
        <tag k="created_by" v="Potlatch 0.10c"/>
</node>
```

  The example XML node element contains attributes, plus the compulsory attributes lat and lon to store geographic information in WGS 84. Additionally, the example node stores two tags elements. The first tag element has the key amenity, which is used to describe community facilities, with the value parking. The node represents a parking lot. The second tag element describes which software is used to create this node.

- **Ways** represent either a line (e.g. road, river) or the border of a polygon (e.g. building, land cover) feature (Ramm et al., 2011). A way has to consist at least of two nodes. Nodes are assigned to the way element using the element <nd>. If a way represents a polygon the first and last node have to be the same.
  XML example code for a way (represents a bridge for a foot path):

```
<way id="193370533" visible="true" version="5" changeset="16405410"
        timestamp="2013-06-03T12:15:53Z" user="magol"
          uid="35601">
        <nd ref="2038769641"/>
        <nd ref="2038769650"/>
        <nd ref="2038769644"/>
        <tag k="bridge" v="yes"/>
        <tag k="foot" v="designated"/>
        <tag k="highway" v="footway"/>
</way>
```

  The example way element contains attributes. Furthermore, three nodes are assigned to that way using the <nd> element. The way is described with three tag elements. The first tells that the way is a brigde, the second that it is designated to be only used by pedestrians. The last tag element defines the way as a footpath.

- **Relationships** are used to define logical or geographical relationships between elements (Ramm et al., 2011). The element <member> can refer either to a node or a way. Its role attribute defines the function of the member. Relationships are among others used to map multipolygons, routes or driving restrictions. For a multi polygon a role of a member is usually described with “inner” or “outer”.
  XML example code for a relation (represents a bus route):

```
<relation id="2202159" visible="true" version="42" changeset="19924466"
          timestamp="2014-01-10T20:36:35Z"user="tomasy"
          uid="953001">
          <member type="node" ref="2097238852" role="stop"/>
          <member type="way" ref="211442597" role="platform"/>
          <member type="node" ref="2097000221" role="stop"/>
          <member type="way" ref="211442602" role="platform"/>
           <tag k="name" v="Buss Gul express: Partille - Nordstan –
            Torslanda"/>
           <tag k="type" v="route"/>
</relation>
```

  The example relation element contains attributes. Moreover, four member elements are assigned to this relation. Two members refer to nodes which represent stops and two members refer to ways which represent platforms. The relation is described further with two tags elements, the first naming the relation and the second defining the type of relation.

2.5 Licensing

OpenStreetMap is licensed under the Open Database License 1.0 (ODbL). ODbL is a copyleft licence which aims to keep the data always free even when it is combined with other data (Ramm et al. 2011). The license allows to create products from, to share and to adopt OSM data freely as long as two restrictions are followed (OSM Foundation, 2014b):

- The OSM community always has to be attributed as the producer of the data with “© OpenStreetMap contributors” (attribution).
- Public adopted data have to be published under ODbL as well (share-alike).

That means that if OSM data are public distributed any improvements or corrections have to be made available under ODbL. Furthermore it is not allowed to merge OSM data with another database which is not under ODbL on file level. However, it is not restricted to use OSM data with other separated databases. Additionally it is legal to create maps from OSM data and to publish them under whatever license and even demand fees for them.

OSM data are license free if the data extract is insubstantial. The OSM community defined extracts as insubstantial which are less than 100 features, more than 100 features when

extract is done non-systematic and the extracted area is populated by less than 1000 inhabitants (OSM Wiki, 2014g).

OSM data were originally licensed under CC-BY-SA 2.0 (Creative Commons Attribution-Share Alike). This license is similar to ODbL, but has some major difference. CC-BY-SA is based on copyright law while ODbL roots in the European database law. CC-BY-SA is not especially created for data but ODbL is. Therefore, ODbL defines more clear rules what is allowed to do with OSM data and what is not. Under CC-BY-SA publishing of maps derived from OSM data had to be under CC-BY-SA. On the other hand additional data in the map did not have to be made public available, as it has to be now under ODbL. CC-BY-SA required the attribution of each individual member of OSM, which is not feasible for such a large project as OSM.

In September 2012 the OSM licence was changed to ODbL. The license change was not unproblematic and was prepared over several years. The licence change required that each contributor accepted the new licence. If a contributor did not accept the new licence, his data were removed from the database. Additional problems arose from OSM data which were collected from sources which do not allow a licence change. The data providers had to be contacted and asked if OSM can use the data even under ODbL. Data with no permission were removed. In the end only about 1 percent of the data were lost due to the license change (OSM Wiki, 2014h).

2.6 Quality issues

Factors which affect the quality of OSM are mainly depending on the collection method. OSM data are mainly collected with a GPS device or by digitizing orthophotos.

The average accuracy of handheld GPS device (code receiver) is 5 meters (Ramm et al., 2011). However, the quality of a GPS measurement is varying spatially and temporally. Depending on how many GPS satellites that are available and their geometry the position accuracy can be reduced considerably. In areas with high buildings interference and multipath errors also affect the GPS signal negatively. The contributor has a relatively low impact on the quality of the GPS signal, only few things should be consider like not having the GPS antenna in a pocket and not standing still for a longer time when recording a GPS track (Ramm et al., 2011).

Digitized data from orthophotos depend mainly on three factors: the flight height (ground resolution) the georeference quality of the image and the digitizing skills of the contributor. Images have a varying georeference quality and in worst case images can be misplaced by several hundred meters (Ramm et al., 2011). Therefore, it is suggested that the contributor checks the position of the image against GPS traces before starting to digitize. The major editor provides tools to correct such problems. Even if digitizing seems like a simple method it requires some training and common problems are of topological nature. It is

easy to miss to connect roads or to digitize a border of two features twice. Editors provide snapping tools to reduce the errors and digitizing has a steep learning curve.

Quality of attribute information is mainly depending on the effort of the contributor. OSM does not have any restriction on attribute tagging. Contributors can define their own attribute tags and add all kind of information to a feature. As mentioned in section 2.4 the OSM community developed a comprehensive list of keys for most purposes and contributors are instructed to follow this informal standard (OSM Wiki, 2014f). The OSM editors described in section 2.2 suggest a list of these attributes when editing features. The collection of attribute information is often limited by the data gathering method. While walking around with GPS device it is easy to store relevant attribute information. When cycling or driving with a car this become more challenging and time consuming. Orthophotos do not contain any attribute information. Therefore, digitized data often contain only classification information until a contributor with local knowledge add further attributes. The classification is also problematic as it can be difficult to see what the digitized feature represent in the real world.

The OSM community has developed so called quality assurance tools (OSM Wiki, 2014m). The tools aim to point out data that are likely to be wrong so other contributors can check and if necessary correct the data. The OSM homepage incorporates a tool to report map errors created manually by users. Other tools are based on automatic analysis to detect bugs in OSM data. Reported errors are mainly of topological and thematic nature. Some common reported errors are: non-closed areas, intersections without junctions, self intersecting ways and missing or misspelled tags.

As mentioned earlier OSM does not have any standard quality assessment. The idea of OSM is to make the project as open as possible and therefore it does not have restrictive standards. The idea is that the community controls itself. Erroneous data are ideally corrected by other members in the OSM project. This idea is known as Linus' Law: "given enough eyes, all bugs are shallow" (Raymond, 1999, page 29) originally in the context of open-source software development. Goodchild and Li (2012) argued in a theoretical study of quality assurance of VGI that Linus's Law only can be applied to VGI to some extent due to the nature of geographic information. Nevertheless, Hakly et al. (2010) showed that Linus' Law is true for positional accuracy. Girres and Touya (2010) found that completeness of an area is correlated to the number of contributors in the same area. This control mechanism is probably enough for trivial purposes such as using OSM as a city map. But for those who also want to use these data in business, science or decision-making a more certain data quality knowledge is necessary (Koukoletsos, 2012).

3 Quality Aspects

3.1 Quality standards

Quality control has always been an issue of spatial data and has received substantial attention from researches as well as from user and producers of geographical data. Van Oort (2006) presents a comprehensive summary of the research and developments in this field. Since the beginning of the 1990s several quality standards for geographic information have been developed. The U.S. Geology Survey published quality standards as part of their spatial transfer standard in 1992. The international Cartographic Association developed quality standards and evaluation methods (Van Oort, 2006). Comité Européen de Normalisation (CEN) started to develop own standards in 1998 but later approved the ISO/TC211 (see below) as European standards. In 2002 the International Standardisation Organisation (ISO) developed a number of standards concerning spatial data quality within the technical committee 211 (ISO/TC 211).

ISO/TC211 incorporates several international standards concerning spatial data. The two most important for quality issues are: ISO 19113:2002 *quality principals* (ISO, 2002) and ISO 19114:2002 *quality evaluation procedures* (ISO, 2003a). The idea behind the creation of these standards are to ensure that quality measurements of different datasets and accomplished by different people are comparable.

3.1.1: Quality elements:

ISO 19113:2002 (ISO, 2002) defines five quantitative quality elements:

Completeness

Completeness describes the absence or presence of objects in the dataset. There are two sub-elements (ISO, 2002):

- Error of commission is used when a dataset has a feature, which does not exist in reality (reference data).
- Error of omission is used when a dataset miss a feature, which exists in reality (reference data).

This requires knowledge about the features which ideally should be included. If a dataset aims to include all motorways in a country, it is complete when all motorway objects which exist in the real world are included in the dataset. Completeness is related to the purpose and aim of the dataset. Completeness can be evaluated for feature classes as well as for their attribute tags and relationships.

Logical consistency

Logical consistency checks if rules of the data structures are followed. It is divided into four groups: conceptual, domain, format and topological consistency (ISO, 2002).

- Topological consistency is given when the data follow the predefined topological rules. Errors are for example roads that are not linked in the dataset but are connected in the real world (reference data), features which should have the same geometric position are dislocated (a road is at the same time a municipality border) and junctions which not have a node.
- Conceptual accuracy describes how good the data follow their conceptual schema, e.g. each city has to belong to one and only one municipality.
- Domain consistency depict if values are within their value domain, e.g. the name of all European routes in Sweden can only consist of three letters (Wasström et al.,2013).
- Format consistency checks if the data are stored according to physical structure of the dataset, e.g. roads are stored as lines and not as curves.

Positional accuracy

Positional accuracy deals with the exactness of the coordinate position of objects. Coordinates of a geographic feature define its location on the earth in a geodetic reference system. ISO (2002) distinguish between three different positional accuracy elements.

- Absolute accuracy describes the difference to the true coordinate value or the value regarded as true.
- Relative accuracy is a measurement of positional difference between features relative to each other within the dataset.
- The gridded data position accuracy is used for raster data to evaluate the position of the grid against its true value.

Temporal accuracy

Temporal accuracy depicts the accuracy of temporal attributes and relationship. ISO (2002) defines three parameters of temporal accuracy:

- Accuracy of time measurement describes how accurate time information is, the error of a time measurement.
- Temporal consistency depicts if events are correctly ordered, e.g. a road has to be build before it can be expanded.
- Time validity checks if time information is valid, e.g. 31 April is an invalid time.

Thematic accuracy

Thematic accuracy regards the additional information of features besides their position (Van Oort, 2006). Attribute accuracy can be divided into three elements (ISO, 2002):

- Classification correctness depicts if objects are classified correctly. It is commonly measured with a confusion matrix. E.g. a motorway feature should represent a motorway in the real world.

- Quantitative attribute accuracy describes if the value of measurable attributes are correct. E.g. if the area of a real estate is correct.
- Non-quantitative accuracy describes if the value of non-measurable attributes is correct. E.g. if the name of a feature is correct.

Furthermore, the ISO 19113:2002 (ISO, 2002) standard defines three non-quantitative quality elements: *purpose*, *usage* and *linage*.

- *Purpose* specifies if the dataset fits its intended purpose of the creation and its use.
- *Usage* describes if the dataset contains information about its applications and projects for what the dataset has been used.
- *Linage* depicts if information about the origin and development of the dataset is documented. Each feature of dataset should contain information about sources and used software in the creation and maintenance of the data.

Actuality

Actuality is a quality element which is not defined by ISO 19113:2002 (ISO, 2002). Wasström et al. (2013) defines actuality as the time when a feature was the last time, through a control, declared as correct. It is more informative to know the last time a feature was controlled instead of knowing the last time a feature was updated or produced. A feature which was updated or produced a long time ago can still be correct. If actuality information is not given in a dataset, analysis of the update frequency can give an estimation of the actuality of a dataset.

3.1.2: Quality evaluation process

ISO 19114:2003 (ISO, 2003a) specifies standards for quality evaluation processes. It defines five steps as parts of a quality analyses:

1. Selection of quality elements and the geographical extent of the analysis.
2. Choice quality measurements for each quality element.
3. Calculation of quality measurements.
4. Determination of the data quality.
5. Comparison of data qualities against product specifications or user requirements to see if they satisfy the requirements.

A quality evaluation can be divided into internal and external evaluation. Internal evaluation can be performed on the data itself. A common example is the checking of logical consistency. External methods required external reference data, for example for completeness and positional accuracy evaluation. Finally, quality results should be reported as metadata elements following ISO 19115:2003 (ISO, 2003b) or as an evaluation report.

3.1.3 Selection of quality elements:

As mentioned in section 1.3 this study aims to calculate the quality elements completeness, positional accuracy and thematic accuracy for a subset of the road network of OSM compared to the real-estate map of Lantmäteriet. These three elements are chosen because they are among the most desirable characteristics of geospatial data. Jakobsson and Vauglin (2001) reported that European mapping agencies mainly uses these three quality elements.

The quality element actuality was also planned to investigate by this study, as it would be interesting to see if OSM data are more actual than Lantmäteriet. However, actuality is excluded as it is inherently difficult to measure.

3.2 Quality Measurements

In this section some common quality measurements for the three quality elements used in this study are summarized.

3.2.1 Completeness measurements

According, to the definition of completeness, it can be calculated by counting the features of a dataset which are represented or not represented in a reference dataset. For linear features, completeness can be calculated by comparing the summed lengths of all features of the two datasets (Haklay, 2010; Zielstra and Zipf, 2010). This is a more pragmatic approach which is usually used when feature correspondence between the two datasets is not available. A more meaningful evaluation can be achieved when the datasets have been matched. The summed length of matched features divided by the total length of all features in a dataset can be used for completeness measurements (Koukoletsos, 2012). This percentage describes the amount of data which can be also found in the other dataset.

3.2.2 Positional accuracy measurements

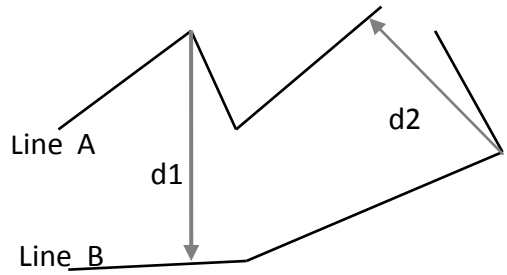
To calculate the positional accuracy for point features is a straightforward process by comparing the coordinates of corresponding points and calculating Euclidian distance and percentile distribution of the distance distribution (Goodchild and Li, 1997). The Federal Geographic Data Committee of the USA developed, based on this method, a standard to assess point positional accuracy (FGDC, 1998). To use this standard to compare linear features, corresponding points in both lines have to be identified. However, this is, except for intersection not possible. Therefore, this standard is not suitable to be applied for linear feature positional accuracy (Airza-López et al., 2011).

The Hausdorff distance is a popular distance measurement between two lines (Mustière and Devogele, 2008). The Hausdorff distance (dH) is defined as (Deza and Deza, 2013):

$$dH(A,B) = \max\ \left\{\max_{P_b \in B}\left\{\min_{P_a \in A}\left\{d\left(P_{a,}P_b\right)\right\}\right\}, \max_{P_a \in A}\left\{\min_{P_b \in B}\left\{d\left(P_{b,}P_a\right)\right\}\right\}\right\} \quad (1)$$

A and *B* are non empty sets of points. *d()* represent the Euclidian distance between two points in A and B. The Hausdorff formula calculates the minimum distance between all

points of set A to B and the minimum distance between all points of set B to A. The maximum minimum distance of both calculations is selected. Finally the Hausdorff distance is the maximum of the both values.



d1: $\max_{P_b \in B}\{\min_{P_a \in A}\{d(P_a, P_b)\}\}$

d2: $\max_{P_a \in A}\{\min_{P_b \in B}\{d(P_b, P_a)\}\}$

Figure 5: Example of the Hausdorff distance between to line objects. The thick arrow (d1) represents the actual Hausdorff distance, as it is larger than d2.

The Hausdorff distance is an asymmetric distance, meaning that the calculation the two components (d1 and d2 in figure 5) have different values (Hangouet, 1995). When calculating the Hausdorff distances between lines all defining points of a line are included in the set of points (Airza-López et al., 2011).

White (1985) and McMaster (1986) proposed an area displacement measure between two lines to evaluate line simplification methods. It measures the distance between two lines as the ratio of the area between the two lines divided by the length of the line being evaluated (figure 6). It is hereafter referred as average distance (dA) and is defined as:

$$dA(A,B): \frac{Area_{(line\ A, line\ B)}}{lengt\ h(line\ being\ evaluated\ )} \quad (2)$$

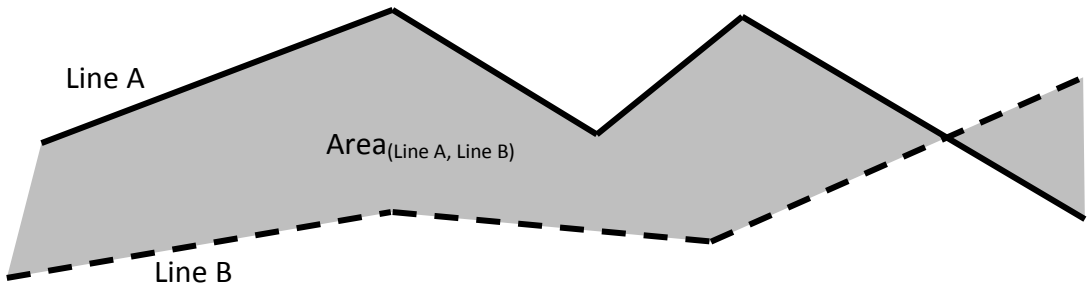


Figure 6: Representation of the area between two lines, which is used in the calculation the average distance between these two lines.

Goodchild and Hunter (1997) introduced a buffering method to calculate positional misplacement of lines. The general idea is to calculate the proportion of the tested line which is within a buffer around the reference feature (figure 7). They propose to iteratively increase or decrease an initial buffer size until a desired overlap percentage is reached. A simpler implementation is to only use certain buffer sizes without using an iterative

process, as used in Haklay (2010). The method is relative insensitive against outliers (Goodchild and Hunter, 1997).

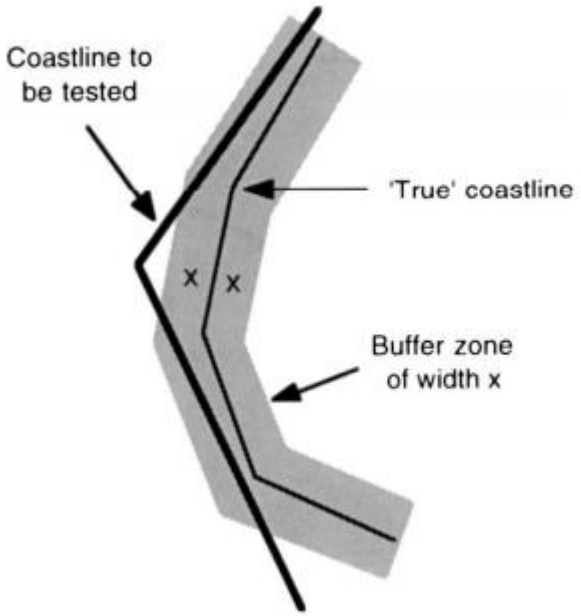


Figure 7: Illustration of the buffer method from Goodchild and Hunter (1997, p. 301)

3.2.3 Thematic accuracy measurement

A straightforward method to assess attribute information is to check if attributes have the same values as the reference attributes. This is often used for classification evaluation, e.g. of land use cover, and the result is presented in an error matrix. In the case of road name accuracy, this divides road names in correct and incorrect names. However, names can be misspelled or contain different abbreviations and still refer to the same road feature, especially in the case of OSM. Therefore, a Boolean measurement is not suitable in the case of OSM. Instead it seems more suitable to measure string similarity of road names.

Text similarity is often used in the field of linguistic and computer science. A common text similarity algorithm is the Levenshtein distance (Navarro, 2001). It calculates the minimum number of operation like insertion, deletion and substitution which are necessary to make two strings identical. Each necessary modification costs one. A value of one to three is regarded to minor misspelling errors (Girres and Touya, 2010). However, Levenshtein distance does not take the length of the two strings into account. A pair of short strings and a pair of long strings can have both the Levenshtein distance of one, while the pair of long strings is still relative similar the pair of short strings could mean something complete different. Serva and Petroni (2007) propose therefore the normalized Levenshtein distance, where the Levenshtein distance of two strings is divided by the number of characters of the longer one. The normalized Levenshtein distance can be defined as:

$$normalized\ Levenshtein\ distance = \frac{Levens\,hein\ \ distance\ \ (string\ 1, string\ 2)}{maximum\ \ \ (string\ 1\ lengt\ ,\ \ string\ 2\ lengt\ h)} \qquad (2)$$

This function returns a value from zero to one. Zero is returned when both strings are identical, while a value of one means the strings are completely different.

Koukoletsos et al. (2012) suggest another name similarity measurement; it is defined by the following equation:

$$name\ similarity = \frac{number\ \ of\ similar\ \ characters}{maximum\ \ \ (string\ 1\ lengt\ ,\ \ string\ 2\ lengt\ h)} \tag{3}$$

The number of similar characters in two strings is counted and divided by the number of characters of the longer one. If the names are completely different this equation return zero and if they are identical one is returned. Koukoletsos (2012) argues that this measurement can cover misspelling and abbreviations.

3.3 Quality studies of OSM

The quality of OpenStreetMap data has been subject of an extensive number of studies over the last years (Neis and Zielstra, 2014). One of the first papers which address this topic was written by Haklay (2010). He assessed the quality of the OSM dataset for England and compared the positional accuracy and completeness of OSM road network against a governmental dataset from Ordnance Survey in 2008. Positional accuracy was assessed using a buffer method based on Goodchild and Hunter (1997). The overlaps of the OSM data with buffers of different sizes around the Ordnance Survey data were calculated. The average overlaps of all motorways in England were 80% using a 20 meter buffer. Furthermore A- and B-roads were investigated in five test areas. A-roads had an overlap of 88% (5.6 meter buffer) and B-roads had an overlap of 77% (3.75 meter buffer).

To estimate the completeness of OSM, Haklay (2010) compared the road length in one by one kilometre tiles covering England between OSM and Ordnance Survey datasets. In 61% of all tiles the Ordnance Survey dataset was more detailed than OSM, in 25% it was contrary and 13% were empty tiles. However, the geospatial pattern showed that OSM had more detail in dense populated areas while the Ordnance Survey dataset had more features in rural areas. He repeated the calculation using only roads of the OSM which were classified as roads. The results were similar but with a slightly higher percentage (64%) where Ordnance Survey data was more detailed.

Zielstra and Zipf (2010) adopted the methodology of Haklay (2010) to evaluate the completeness of the OSM road network in Germany in comparison to a TeleAtlas MultiNet dataset. They extended the analysis by calculating the length differences for entire Germany and for different cities of different sizes. The calculations were based on three OSM datasets from year 2009. Furthermore, they examined the completeness variability within cities by calculating length differences within different buffer sizes around city centres. The results supports the findings of Haklay (2010), OSM data are often more complete in dense populated areas than in commercial or authority datasets, but OSM data lacks coverage in rural areas. There also exists a discrepancy between larger and medium

sized cities; OSM is more complete in larger cities than in medium sized cities. All investigated cities showed a decrease in completeness with an increasing distance from the city center. Moreover the study revealed that the OSM road network in Germany is growing rapidly, the total length difference for Germany decreased from -30% to -7% between OSM and TeleAtlas within 8 months. Neis et al. (2012) showed that the OSM street network of Germany was more complete than a commercial dataset in 2011 (TomTom Multinet).

Girres and Touya (2010) presented a study assessing the quality of the French OSM dataset against BD Topo data from the French national mapping agency (IGN). They used several different methods to assess positional, attribute, semantic and temporal accuracy, logical consistency, completeness, lineage and usage. The study was conducted in several small areas using relative small sample quantities.

Girres and Touya (2010) evaluated positional accuracy for a set of manually matched points, lines and polygons. The geometric difference between matched pairs were calculated using Euclidian distance for points, Hausdorff Distance and Average Distance for lines and Surface Distance, Granularity Measurement and Compactness for polygons. Points were derived from intersections and had an average difference of around 6 meters. Line pairs had an average Hausdorff distance of 13.6 meter and an average distance of 2.2 meter. 106 Lakes were investigated finding that the three measurements for polygons showed small positional differences.

For the lakes pairs the attribute name tag was evaluated showing that only 55% of the OSM lakes had a name tag but then the names were nearly identical to BD Topo data. Furthermore, Girres and Touya (2010) quantitatively checked the attribute accuracy of the complete OSM dataset of France. The main attribute for different feature types is informed nearly 100%, except for road features (only 85%). The second attribute was general very poorly described. Moreover they discovered that a higher number of contributors are related to a higher quantitative attribute quality.

Semantic accuracy was assessed using manual matched road pairs. *Motorways* and *Primary* ways were classified nearly 100% correct, while only 50% of *Secondary* roads are classified correct. The completeness in the test areas was far from complete. Girres and Touya (2010) showed that smaller objects are more likely to be missing than larger and more distinct features. Furthermore, the completeness increases with a higher number of contributors and a higher population density.

Girres and Touya (2010) pointed out that logical consistency is violated within a data theme as well as between data themes. Common problems are that roads features do not end at each intersection, some objects are captured several times or different objects which should have the same location (road and administrative border) have different geometric representations. A comparison of the OSM dataset at two different times showed an

increase of features by 31.7%. Lastly they stated that only 27% of all objects had information about the capture source and only 6% were tagged with information about the used software.

Ludwig et al. (2011) carried out a comparison of the OSM street network in Germany against Navteq dataset. They wanted to assess the quality of OSM for geomatic business applications. The comparison was performed feature-wise. Therefore, they developed a matching routine to find corresponding features in the two datasets. The matching methodology is described in detail in section 4.1. Based on the match they calculated several quality elements. 73% of OSM street objects were within a distance of five meters, 21% between five and ten meters and 6% further away than ten meters. Rural areas had higher deviations than populated areas. Analysis of the attribute quality showed that in urban areas the attribute completeness is better than in rural areas (5.6% vs. 17.5% missing names). The same trend can be seen from important to less important streets. A closer look at the speed limit attribute revealed that most often this information is missing (around 90% of the streets). 80% of all Navteq streets with a road name had a match in inhabited areas, while only 50.8% in uninhabited areas. The completeness decreases also from important to unimportant streets.

For a summary of additional studies of the quality of OSM which also cover other quality elements see Neis and Zielstra (2014). They present a comprehensive summary of the research done on OSM and also discuss potential future trends within OSM research.

4 Matching algorithms and programs

Matching is the identification of corresponding features in two datasets which represent the same object in reality. For example: A road is mapped in *dataset A* as the feature 3 and in *dataset B* as feature 14. The matching process identifies that these two features among all the features in the two datasets represent the same object in reality. This established relationship can be also called feature correspondence.

If a real world object is represented identically in two data sets, it is relative easy to match this two features. However, two geographic datasets have almost always a unique geometric representation of same spatial objects. This makes matching complicated and is the main limitation of automatic matching (Walter and Fritsch, 1999). Differences occur because datasets are often produced independently and using different acquisition methods. Even using the same data collection method, repeated sampling creates different results. Further variations between datasets result from the use of different data models, geodetic reference systems, scale representation and quality requirements. These differences are often not constant and vary spatially.

The research field of automatic matching in geographic information science has received substantial attention and there exist an extensive number of studies related to this problem. Most often algorithms are specifically created for certain datasets and to fit specific purposes. This makes them often not applicable to other datasets or purposes. Section 4.1 summarise some fundamental studies in this field and algorithms related to this study.

4.1 Matching algorithms

Devogele et al. (1996) described a matching algorithm to match road features in a multi-scale database to enable scale transitions relationships between object. Their algorithm is based on semantic, topologic and geographic information of the data. At first semantic information is used to create a match of road pairs. Crossroads are matched under consideration of geometric and topologic attributes. Road section connections are found with the help of Hausdorff distance.

Walter and Fritsch (1999) presented a statistical approach to match spatial data of two road datasets with different data models. Their matching algorithm is divided into five steps. The first step prepared the data for the further analysis; an affine transformation is used to reduce the systematic coordination differences. Next a growing buffer method is introduced to compute a list of possible matching pairs, this method enable to identify 1:M and N:M matching pairs. Step 3 removed unlikely matching pairs, which have certain angle and length differences. Next, each pair is evaluated with a merit function, which takes into account statistical information of angle differences, length, shape, form, position and connection. The statistical information between the two datasets is investigated manually in training sites. In the final step a unique combination of matching pairs with maximum

information about each other is calculated. The algorithm returned a high percentage of correct matches.

The company Vivid Solutions developed the Java Conflation Suite (JCS) (Vivid Solutions, 2014a). It is an open source (under GPL license) Java library of tools for conflation of spatial datasets. Among other conflation tools it contains a road network matching tool. The automated road network matching of two datasets is based on node and edge matching. The dataset which has a considerably higher quality is set as reference dataset. Within a maximum searching distance around each node of the reference dataset, the best matching node of the other dataset, based on distance and topology of adjacent edges calculations, 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. This creates a more similar geometry between the edges and produces a better match result. The output is one dataset with complete coverage of both input datasets. JCS also enable manual matching on the dataset. JCS has not been update since November 2003 according to the homepage of Vivid Solutions (Vivid Solution, 2014a).

Vivid Solution implemented the road network matching tool of JCS as an open source plug-in for the OPENJUMP geographic information system (Vivid Solution, 2014b; OpenJump, 2014). The RoadMatcher plug-in incorporate the road network matching into the OPENJUMP functionality. RoadMatcher was last time updated in 2009 (Sourcefrog, 2014).

Stigmar (2005) tailored a matching algorithm to match routing data with topographic data. Stigmar adapted the JCS algorithm and increased the performance by adding three extensions specific 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. First unmatched segments are match by looking at their topological relationships. This only works on connected unmatched segments, therefore buffers created around the remaining unmatched segments. The best matching pair within the buffer is then chosen depending on a measurement of distance and angular difference. As this step is generally computationally expensive, this step is performed latest to have as few unmatched segments as possible.

Ludwig et al. (2011) created a matching algorithm to perform a quality assessment of OSM for geomatic business applications. They were matching only subsets of OSM and the reference road network dataset which are considered to be important for geomatic businesses. The methodology is adopted among others from Walter and Fritsch (1999). In a preparation step attribute fields between the datasets are adjusted to make them more similar. The next step includes investigation of the data models to build correspondences and relationships between attributes. 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. A visual comparison was used to remove mismatches.

Koukoletsos et al. (2012) presented a feature-based matching algorithm to assess the completeness of OSM against the Intergrated Transport Network (ITN) dataset from Ordnance Survey. The datasets were divided into 1 km^2 tiles, expanded with a 50 m buffer to avoid border artefacts. The algorithm was executed for each tile separately to achieve a faster computation and to get a better representation of the heterogeneity of OSM in the results.

Furthermore each tile was classified as urban or rural depending on the number of features which are presented in a tile. For rural and urban tiles different parameters are used in the algorithm. This considers that in rural areas larger distances between the datasets are more likely than in urban tiles due to in rural areas the image resolution is often reduced (OSM data become more uncertain) and official data has lower accuracy (ITN data become more uncertain). All features are divided into segments, the part between two breakpoints, to enable directional analysis. The algorithm produced robust results with low matching errors, 2% in urban areas and around 3% in rural areas.

The algorithm described in Koukoletsos et al. (2012) consists of 7 stages:

- Stage 1: A candidate list of possible matching pairs is produced. The list contains OSM segments which are within a searching distance and within an angular tolerance from a reference segment. A reference segment can have zero, one or more candidates. The searching distance depends on GPS accuracy, a constant and road width. The angular tolerance is calculated taking the length of the reference segment and GPS accuracy into account. Each reference segment which has only one matching OSM segment which is not three times longer than the reference segment is regarded as a matching pair and is removed from the candidate list.
- Stage 2: In this stage the road name attribute is investigated. If an OSM candidate has exactly the same name as the reference segment, the pair is a match and removed from the list.
- Stage 3: Stage 3 considers the similarity of the road name attribute to account for spelling mistakes. The OSM candidate which has the highest text similarity value above 65% is regarded as a match. Text similarity is calculated using the equation 3 in section 3.2.3.
- Stage 4: The distance between the reference and OSM start-points and end-points of all remaining candidate pairs are calculated. The OSM candidate with the minimum accumulated distance is regarded as match. This was the final step which is performed on segment level.
- Stage 5: The collected matching information at segment level is transferred to feature level. Each OSM and reference feature is classified as matched or not matched. At least

50% of a features total length has to have a corresponding segment match in order to be considered as matched.

- Stage 6: The similarity of the road name attribute of reference features which are within search distance around non-matched OSM features is calculated. The same text similarity calculation as in stage 3 is used (c.f. section 3.2.3). The threshold for being regarded as a match is raised to 75%.
- Stage 7: The final step is focusing on non-matched OSM segments without name attribute. Possible reference features are searched within a search distance. Length of features, unmatched and matched, within the buffer are compared and if they fulfil certain restrictions, a candidate pair is regarded as a match.

5 Material and Methods

5.1 Study area and datasets

5.1.1 Study area

The study is carried out for a region covering Göteborg and surrounding areas (see figure 8). Göteborg is the second largest city of Sweden and is located on the west coast of Sweden. The study area contains areas with dense road network (urban areas displayed as yellow areas) and rural areas with less dense road network (rural areas represented as white and green areas).

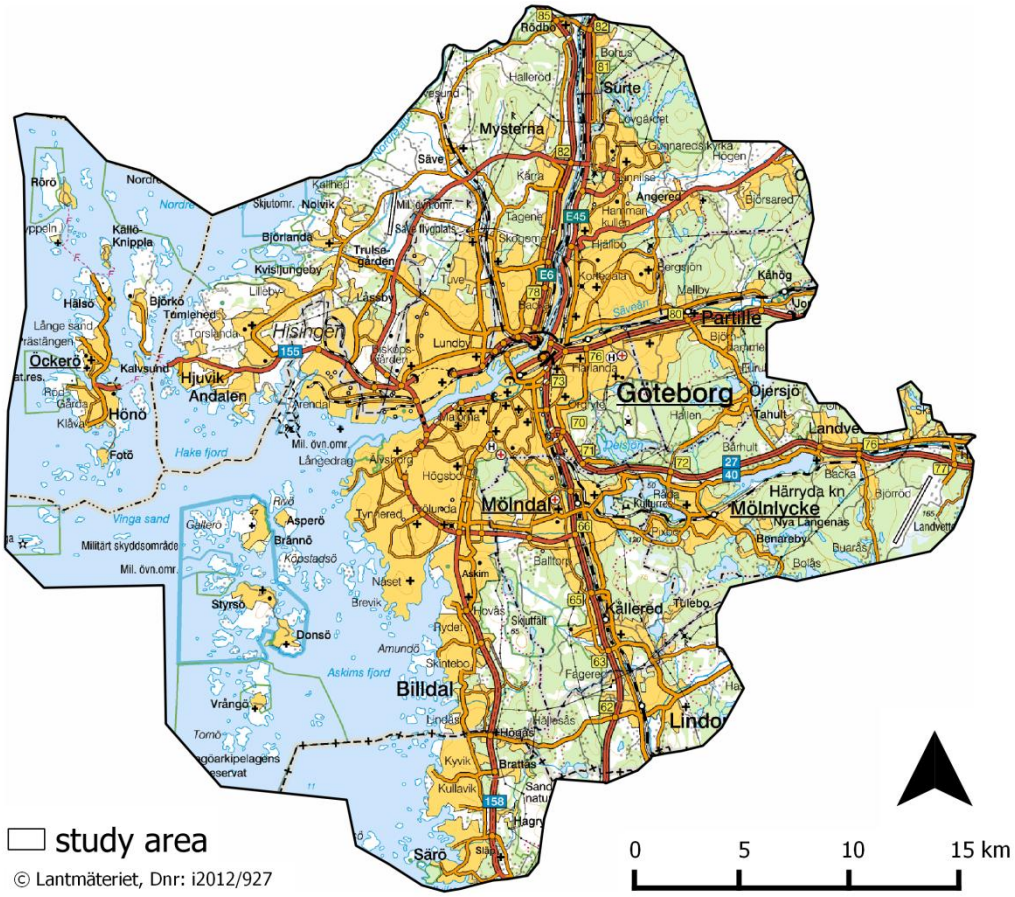


Figure 8: Study area covering Göteborg and the surrounding areas.

5.1.2 Reference data

The aim of this study is to compare OSM data with reference data of high quality. The reference data are supposed to have a similar level of detail as OSM, high positional accuracy and no or a low degree of generalisation.

Lantmäteriet, the Swedish national land survey agency, provides several datasets with varying level of detail and different purposes. The products range from overview maps covering all of Sweden, over specific maps like hiking maps to very detailed maps showing property borders. The most detailed map type is the real-estate map (in Swedish:

Fastighetskartan) (Lantmäteriet, 2014a). Its main purpose is to show properties borders but it also contains detailed information about the topography. The real-estate map is produced for presentation in the scale 1:5,000 – 1:50,000. The positional accuracy is specified with less than two meters standard deviation. Additionally, no cartographic generalisation is applied to the data.

Locality map (in Swedish: Tätortskartan) is another map product with the similar requirements (Lantmäteriet, 2014b). However, the extent of the locality map is limited to areas with a certain minimum population density while the real-estate map covers whole Sweden, except for some mountain areas. With this in mind, the real-estate map of Lantmäteriet is chosen to be used as the reference dataset in this study.

The real estate map is delivered in the coordinate system SWEREFF 99 TM. The data are mainly produced and maintained with photogrammetric techniques. The positional accuracy is less than two meters standard deviation. The positional accuracy is further described individually for each object. The dataset is updated periodically with varying intervals for different regions. The real-estate map contains two road layers, *Roads* and *Other Roads*. These data layers are building networks, meaning that roads are topological structured and have nodes at intersections. The *Roads* layer contains all roads used for car traffic. The *Other Roads* layer includes agricultural road, as well as foot-, cycle- and hiking paths. Only the *Road* layer is used for the quality assessment of OSM. The road type *ferry line* is deleted from this layer. This layer is here after referred to as *LM dataset*. Table 1 presents the different road types in the LM dataset with their assumed width and their corresponding OSM road type.

Table 1: The Different road types in the LM dataset with their assumed width and their corresponding OSM road types.

LM road types (code of type)	**Road width**	**OSM type**
motorway (one direction) (VÄGMO)	12	motorway
larger public road (VÄGAS)	10	trunk, primary
public road (VÄGGG)	8	secondary,
minor public road (VÄGA1, VÄGA2, VÄGA3)	6	tertiary
road (VÄGBN)	5	tertiary, unclassified
residential road (VÄGKV)	4	residential, service, living_street
poor road (VÄGBS)	4	road

5.1.3 OSM data

The most actual OSM data at the time for the study for Sweden was downloaded in XML format (Geofabrik, 2014). The file is an OSM dataset including all features which are mapped within Sweden. All features with a highway tag are imported into a PostGIS database. The data are transformed from the reference system WGS 84 (Lat,Long) to SWEREF 99 TM. This is done to correct the positional offset of about half a meter between

the two reference systems. A shapefile including all corresponding road types of LM dataset and covering the complete study area is exported from the database (see Table 1). Furthermore, all roads with the tag *psv* (Public Service Vehicle) are removed. These roads most represent tram or bus roads which are most often not represented in the LM dataset. The encoding is changed from UTF-8 to Latin1, which is used in the LM dataset. The OSM road data are not restricted to have a node at intersections. To make the geometry between the two datasets more similar, all OSM features are split at line intersection. This dataset is hereafter referred to as *OSM dataset*.

5.2 Evaluation of matching methods

The time frame of this thesis does not allow a developing and testing of a matching method from scratch. Therefore, the methodology which is regarded as most suitable for this thesis of the methods described in section 4.1 is used as a starting point. This method is then modified to fit the aim of this study and the used datasets (see section 5.3).

The matching algorithms developed by Devogele et al. (1996) and Walter and Fritsch (1999) represent two fundamental researches in the field of automatic matching. The process described in Devogele et al. (1996) aims to enable scale-transition relationships between two databases of different level of detail. Besides data matching the algorithm deals also with data schema merging. This study intends to match two datasets with a similar level of detail and has not to deal with database schema problems. Therefore, the approach of Devogele et al. (1996) is regarded as not appropriate.

The matching algorithm proposed by Walter and Fritsch (1999) returned a high matching percentage. But it requires time consuming manual statistical investigation of the datasets. This is the main reason why their algorithm is not used in this study. However, the method of creating a candidate list of possible matches based on a buffer and then finding the best match using different constraints is adopted by several other researches (among others: Ludwig et al., 2011; Koukoletsos et al., 2012) and is also used in this study.

The JCS contains a general matching tool which can be used for different datasets. This matching algorithm is based on node matching while the other presented algorithms match the actual features or segments of features. Stigmar (2005) presents a successful implementation of JCS in a matching project. With some extensions the algorithm returned good matching results. However, the JCS is last time updated more than 10 years ago, which reduces the reliability. Therefore, it is decided not to use the JCS in this thesis. The same applies to the method of Stigmar (2005) and the plug-in RoadMatcher as they are based on JCS.

The matching method of Ludwig et al. (2011) and Koukoletsos et al. (2012) are from special interest as they match OSM road data. Their algorithms are designed to deal with the specialities of OSM compared to commercial datasets. The algorithms differ in the purpose of the matching.

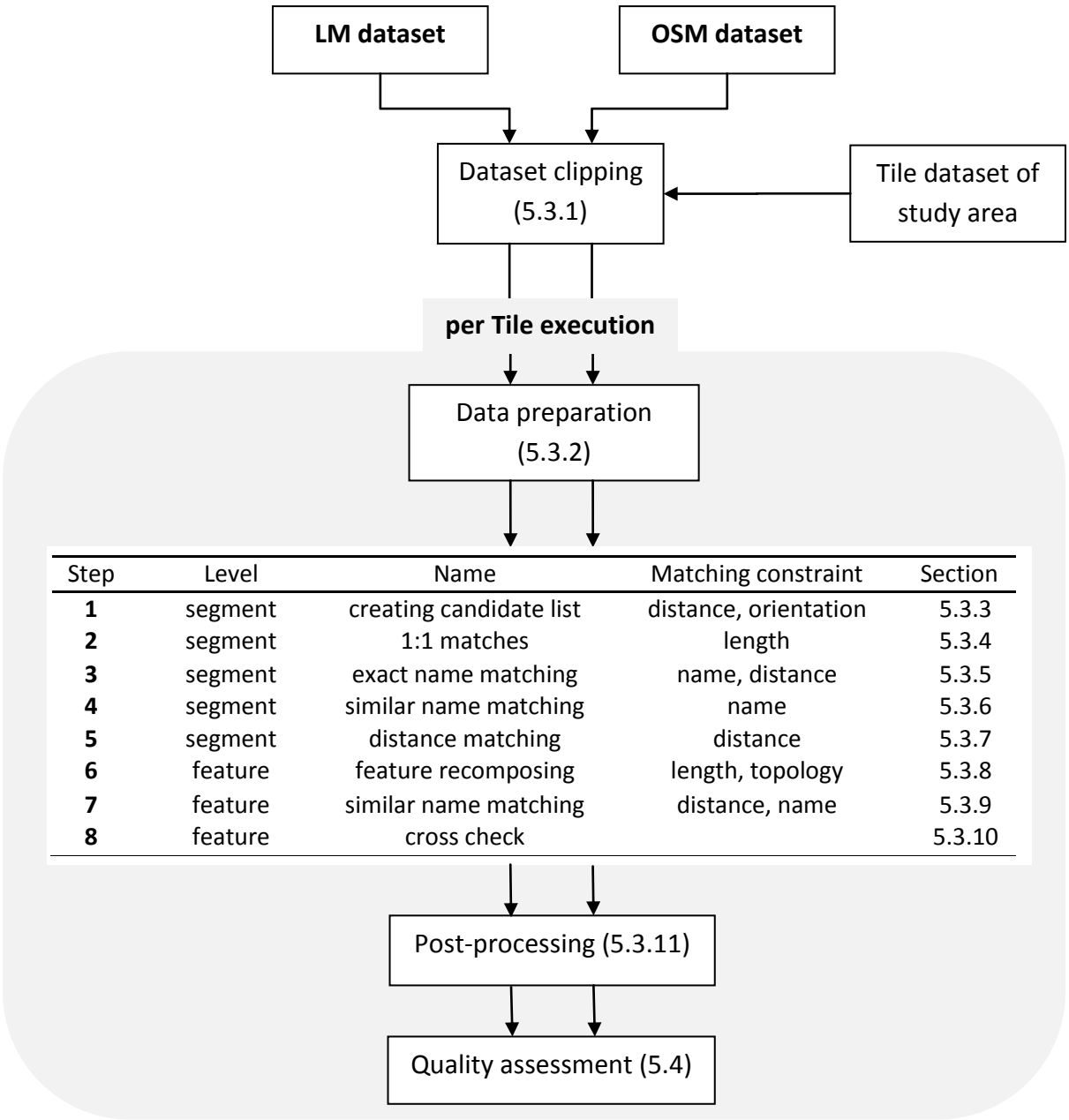
Ludwig et al. (2011) aimed to evaluate the quality of OSM for business geomatic applications. Therefore, they are matching only subsets of OSM and the reference road network datasets which are considered to be important for geomatic businesses. This reduces the application of their method because this study aims to evaluate the general quality of OSM road network and not only for a certain purpose and street types. Furthermore, the method of Ludwig et al. (2011) uses road type information in the matching process. This adds an uncertainty to the matching because OSM roads might be incorrect classified. Another reason why their method is not chosen for this study is that they did not represent an objective matching evaluation. This makes it difficult to decide if their results are reliable.

Koukoletsos et al. (2012) want to evaluate the data completeness of OSM against a national survey agency dataset. The matching algorithm proved to be efficient with low matching errors. For their completeness analysis it was enough to check if a feature in one dataset has a corresponding feature in the other dataset. Exact Information about to which feature a feature of correspond is not collected. Though, in the PhD thesis of Koukoletsos were the algorithm is firstly described, feature correspondence is partly discussed (Koukoletsos, 2012). Therefore, it seems feasible to modify their methodology to implement feature correspondence. Due to these facts it is decided to base the matching algorithm on the methodology from Koukoletsos et al. (2012).

5.3 Matching method

The matching method is adopted from Koukoletsos et al. (2012) and is modified to fit the data sources and aims of this thesis. The main modification is to implement feature correspondence which is not covered by Koukoletsos et al. (2012). The study area is divided into 1 km² tiles and the algorithm is executed separately for each tile. Tiles are not treated differently based on the road density as suggested in Koukoletsos et al (2012).

Figure 9 provides an overview of the main steps of the algorithm, which are explained in detail below. In a pre-processing step the datasets are clipped to the tile extent, geometric simplified and converted into a segment datasets. The matching algorithm works at first at segments and then at feature level.



Step	Level	Name	Matching constraint	Section
1	segment	creating candidate list	distance, orientation	5.3.3
2	segment	1:1 matches	length	5.3.4
3	segment	exact name matching	name, distance	5.3.5
4	segment	similar name matching	name	5.3.6
5	segment	distance matching	distance	5.3.7
6	feature	feature recomposing	length, topology	5.3.8
7	feature	similar name matching	distance, name	5.3.9
8	feature	cross check		5.3.10

Figure 9: Flow chart of the matching routine.

5.3.1 Dataset clipping

The study area is divided into tiles of 1 km² to allow a good presentation of the possible heterogeneity of the OSM datasets. This technique is applied in several OSM quality studies and it has proven to be useful in the interpretation of the results (Haklay, 2010; Zielstra and Zipf, 2010; Koululetsos et al., 2012). The matching is executed separately for each tile. Thereby the number of objects to match at a time is limited which increases the execution speed of the algorithm (Koukoletsos et al. 2012). The 1 km² tiles are extended with a 50 meter buffer to avoid mismatching at the border of a tile. Mismatching can occur when matching pairs are located in different tiles and when clipping produces small parts of features. By applying a buffer these effect are shift to the extended border. The LM and OSM datasets are clipped to the buffered tile and the matching is then executed on the

clipped subsets. After matching, the subsets are clipped to the original 1 km² size of the tile and the effects described above are removed. The buffer size is adopted from Koukoletsos et al. (2012).

5.3.2 Data preparation

Differences in the geometric representation of the datasets make automatic matching more complicated (Walter and Fritsch, 1999). A visual inspection of the datasets showed that the LM features are more detailed represented (include more breakpoints in a feature) than OSM features (see figure 10). Furthermore, features in both datasets contain breakpoints which have no defining function, such as breakpoints on straight parts of a feature. Therefore, a line simplification algorithm is applied to the clipped datasets.

A modified version of the Douglas-Peucker algorithm is used (Douglas and Peucker, 1973). The algorithm goes through each feature in a dataset and removes breakpoints which are within a threshold distance from a straight line segment. A threshold of one meter is used in order to not alter the geometry to much but still remove additional breakpoints. Figure 10 shows an example of features before and after the simplification algorithm.

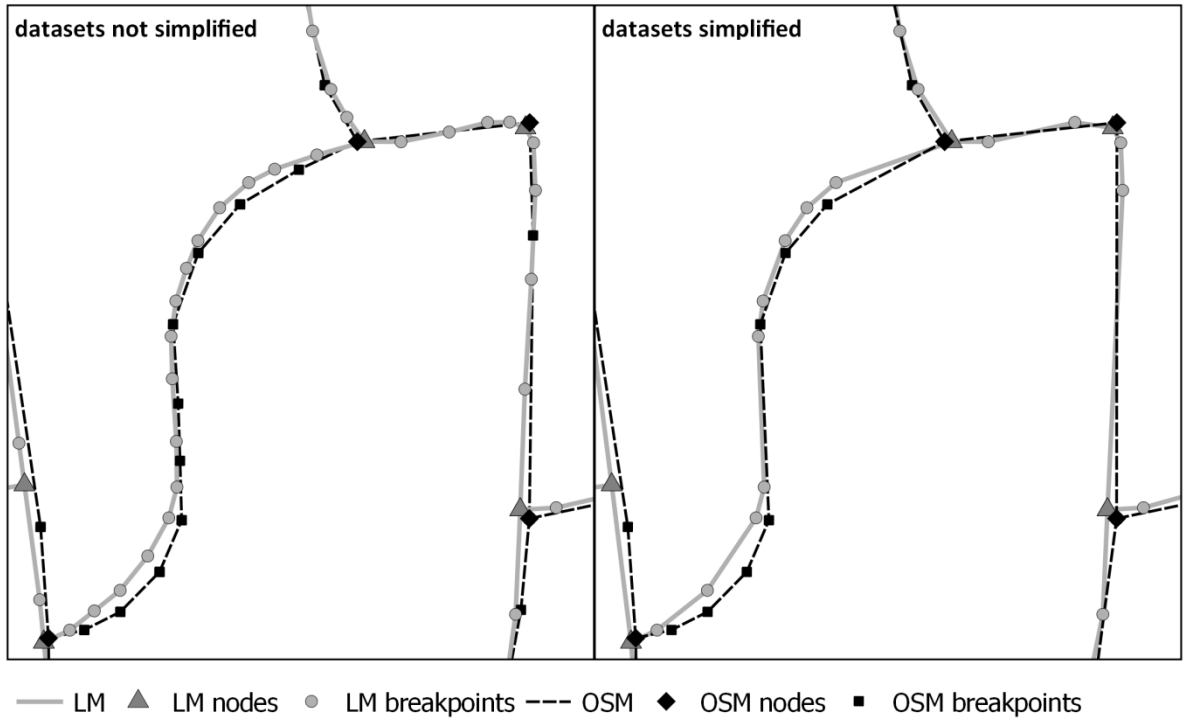


Figure 10: Example of geometric simplification of the LM and OSM dataset using a modified version of the Douglas-Peucker algorithm with the threshold of one meter.

Roads are in both datasets represented as features, which can be straight lines but as well as long polylines with many breakpoints. To uniform the representation, features are divided into segments. A segment is the part between a node and a breakpoint or the part between two breakpoints of a feature (see figure 11). In addition, the matching process uses orientation information which requires segment representation. To calculate

orientation of features is problematic and give non comparable results. Thus, from each dataset a segment dataset is derived. Each segment contains the same attribute information of its initial feature plus its initial feature id to preserve their relationship. The orientation of each segment is calculated and added as new attribute information. Values of orientation are further normalized to values from -90 to 90 degrees because direction is not of importance.

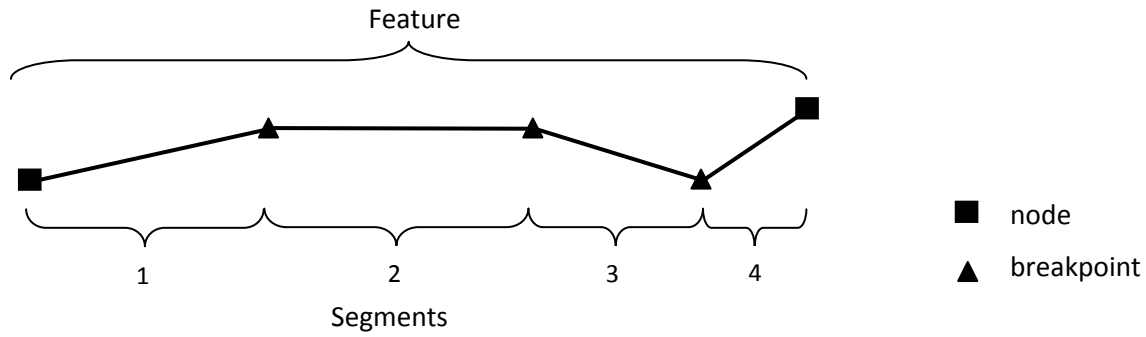


Figure 11: The feature between the two nodes is divided into 4 segments. In the same way OSM and LM features are split into segments in the data preparation step of the matching algorithm.

5.3.3 Step 1 Candidate list

In this step a candidate list is created which contains for each LM segment possible matching OSM segments. This list is the base for step 2 - 5.

All OSM segments which are within a buffer around a LM segments are assigned as possible candidates to the LM segment. The buffer size depends on the road type of the LM segment and is defined as follows (Koukoletsos et al., 2012):

$$Buffer\ size\ (B) = c * \alpha + \frac{w}{2} \quad (4)$$

where *c* is a constant, α is the GPS accuracy and *w* is the road width of LM segment. Values for *c* and α are adopted from Koukoletsos et al. (2012). *c* is set to a value of two, to cover cases with worse GPS accuracy. α is the assumed GPS accuracy and is set to 10 meters. *w* is the road width of the LM segment, it is derived from the different road types. See table 1 for a list of the different road types and their assumed width. The road width is divided by two to cover cases where the OSM contributor walks on the side of the road (Koukoletsos et al., 2012). The buffer size for different road types varied from 20.5 - 26 meter.

To remain in the candidate list an OSM segment has to have similar orientation as the corresponding LM segment. This is done by calculating an angular tolerance for LM segments. Figure 12 explains the calculation of the angular tolerance φ. The black line represents a LM segment with a length of β. The dashed line presents the worst case of mapping it with GPS accuracy α.

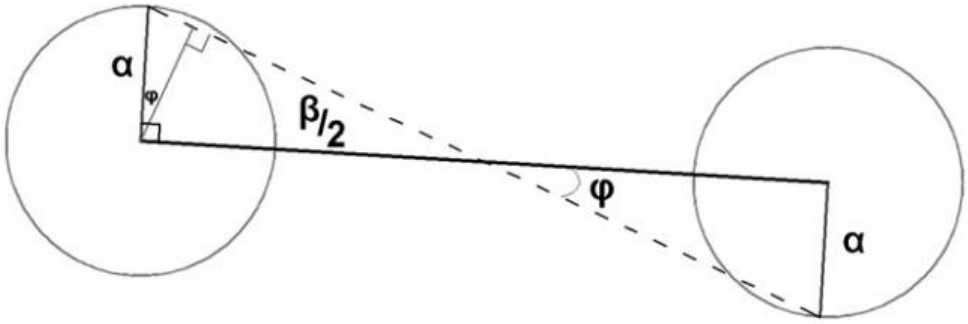


Figure 12: Calculation of angular tolerance (φ) (from Koukoletsos et al., 2012. p. 483). The black line represents a LM segment with a length of β. The dashed line presents the worst case of mapping it with a GPS accuracy of α.

The angular tolerance can be simplified as (Koukoletsos et al., 2012):

$$angular\ tolerance\ (\varphi) = \frac{180}{\pi} * tan^{-1}\left(\frac{\alpha}{\beta/2}\right) \quad (5)$$

For each LM segment the angular tolerance is calculated, using 10 meter as GPS accuracy (α). If the difference of the orientation between the LM segment and an OSM segment found within buffer size is less than the angular tolerance the OSM segment remains as candidate. Otherwise, the OSM segment is removed from the list.

In the following four steps (two to five) the candidates (OSM segments) of an LM segment are tested against different constraints to examine which OSM segment should be the actual matching segment. In these steps only one candidate is accepted as a match to an LM segment. An OSM segment can be a matching counterpart of several LM segments. It is assumed that a LM feature consists of more segments than an OSM feature (see figure 10).

If an OSM candidate is regarded as the match of an LM segment, the segment ID and feature ID of the matching OSM segment is assigned as attribute information to the LM segment. The LM segment and its connected OSM candidates are removed from the candidate list.

The matched OSM segment ID is stored as a key in dictionary with the segment ID and feature ID of the matching LM segments as values. If the same OSM segment matches with other LM segments the LM information is added to the dictionary. After step 5 the dictionary information is assigned as attribute to the corresponding OSM segments. OSM segments which are matched to LM segments belonging to more than one LM feature are not considered.

5.3.4 Step 2 1:1 matching

In this step LM segments are matched when they only have one candidate (OSM segment) and if this OSM segment is only a candidate for this LM segment. Furthermore, the length of the candidate (OSM segment) needs to be less than three times longer than the length

of the LM segment. The length is considered to avoid matching segments which have an unusual length difference.

5.3.5 Step 3 Exact name matching

For each LM segment in the candidate list which has a road name attribute an exact name matching against the OSM candidates is conducted. If only one candidate has an exact name, this pair is regarded as match. If several candidates have the same name the distance between the candidates and the LM segment is investigated.

Distances between segments are calculated between the start points and end points of two segments (see figure 13). The OSM segment with the shortest accumulated distance is regarded as the matching segment.

This technique requires knowing the corresponding start points and end points of two segments. There exist two possible combinations to create two lines between the start and end points of the two segments when a node can only be used once (see figure 13). The combination with the shortest summed length of the two lines represents the desired combination (the green lines in figure 13) and represents the distance measurement.

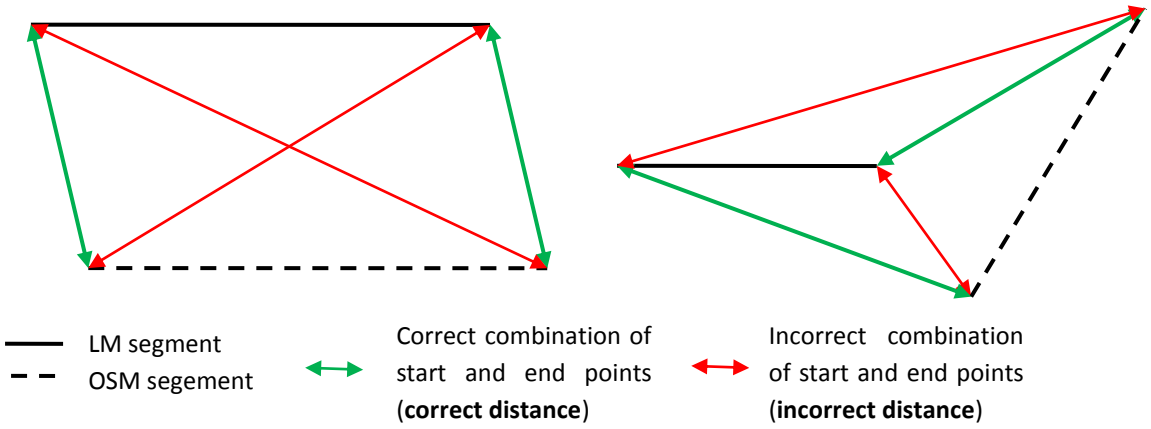


Figure 13: The combinations which are possible to create two lines between the start and end points of two segments when a node can only be used once. The green lines represent the correct combination of corresponding start points and end points of the two segments. The summed length of the two green lines is the distance measurement used in step 3 and 5 of the matching algorithm.

Koukoletsos et al. (2012) argue that this distance measurement give better results than comparing line centroid distance or distance between linear segments.

5.3.6 Step 4 Similar name matching

Road names are not always right spelled or have abbreviations, especially in the OSM dataset as contributors have no naming rules. In these cases exact name matching does not work. Therefore, this step investigates name similarity. Name similarity is calculated using the normalized Levenshtein distance (see section 3.2.3, formula 2). Before the calculation the strings are converted to upper letters to only investigate spelling mistakes. The normalized Levenshtein distance value is subtracted from one.

The normalized Levenshtein distance is chosen as it returned more logic values than the name similarity measurement (section 3.2.3, formula 3). This measurement returned high values in cases of names with a large length difference in length even though they were very different.

The OSM candidate with the highest normalized Levenshtein distance is chosen as a match, if it has a value higher than 0.65. The threshold is adopted from Koukoletsos et al. (2012).

5.3.7 Step 5 Distance matching

This step considers only the distance between possible matching pairs. This matches also LM and OSM segments without road name attribute. The OSM candidate with the shortest distance is regarded as the matching segment. The distance is calculated using the distance measurement described in step 3.

5.3.8 Step 6 Feature recomposing

In this step the collected matching information at segment level is transferred to feature level. For each LM and OSM feature, the feature length is compared to the length of the matched segments connected to this feature. If the length of matched segments is more than half the length of the feature, the feature is regarded as matched. The requirement that half of a features length need to have a matching segment can compensate for mismatching in previous steps to some extent. On the other hand it can also remove right matching information. However, it is better to lose matching information than matching features incorrectly.

To ease the explanation of the implementation of feature correspondence, it is explained for the recomposing of an LM feature.

To decide to which OSM feature(s) a LM feature is matched, the OSM feature(s) which has/have matching segment(s) is/are investigated. If matching OSM segment(s) belong(s) to only one OSM feature, this OSM feature is the match of the LM feature. If the matching OSM segments are part of more than one OSM feature, the length proportion of OSM feature to LM feature is calculate. The combination of proportions which is closest to one as well as where the OSM features are adjacent determine the matching OSM feature(s). The same procedure applies when recomposing OSM segments.

A more detailed explanation is given in Figure 14. It shows an example of the recomposing process of an LM feature. The LM feature with a length of 160 meter consists of six segments (1-6).

- LM Segment 1 has no match
- LM segments 2 to 3 are matched to OSM segment A,
- LM segment 4 is matched to OSM segment B,
- LM segment 5 is matched to OSM segment C
- LM segment 6 is matched to OSM segment D.

OSM segments A and B are part of OSM feature I, OSM segment C belongs to OSM feature II and OSM segment D is part of OSM future III.

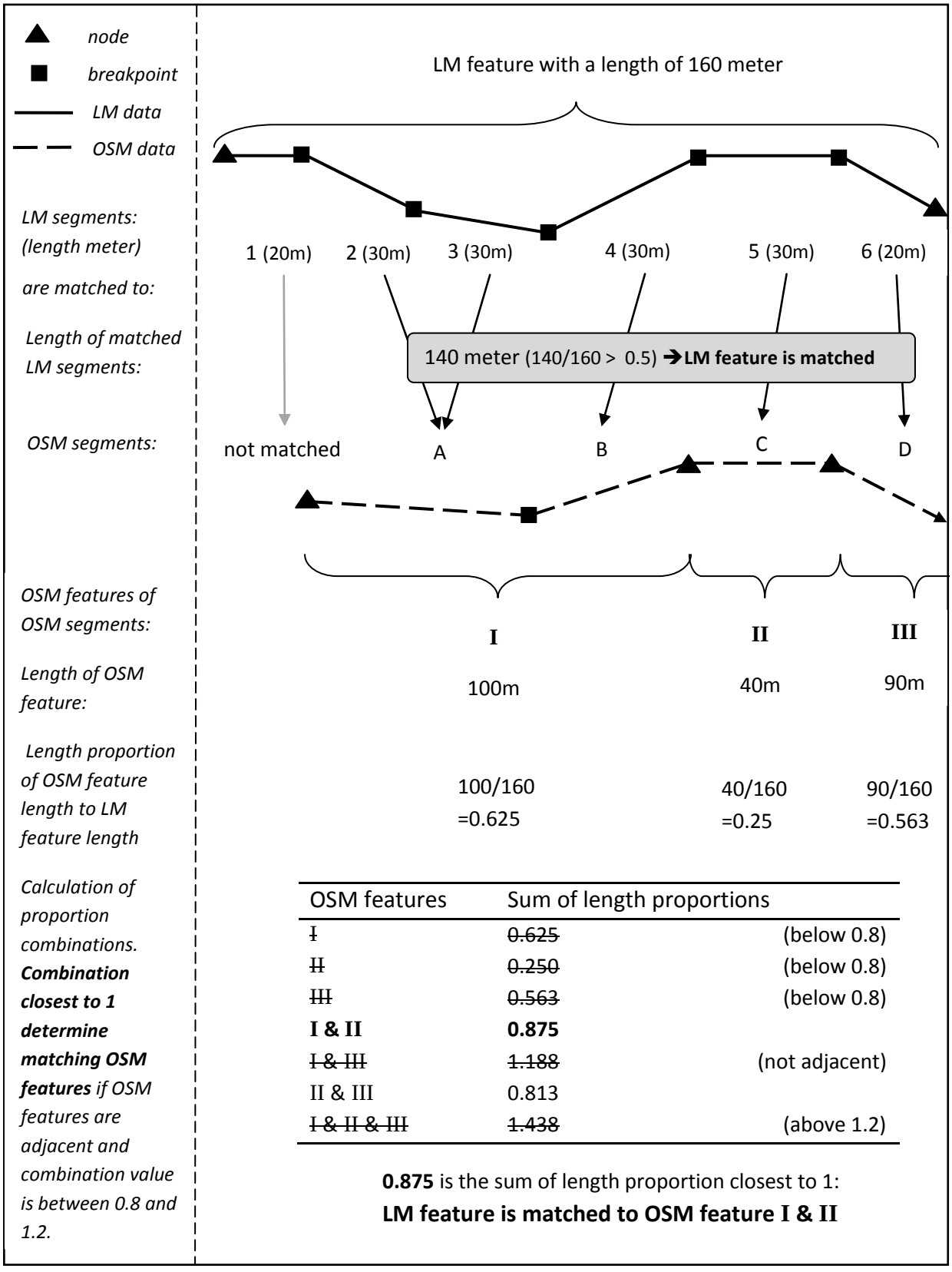


Figure 14: Feature recomposing for an example LM feature.

The five LM segments which are matched have a total length of 140 meters. The requirement that at least half of the features length need to have a matched segment is fulfilled (140/160 > 0.5). The LM feature is regarded as matched.

To decide to which OSM feature(s) the LM feature is matched the OSM features which correspond to the matching OSM segments are investigated. The matching OSM segments belong to three OSM features. For each OSM feature the proportion of its length to the LM feature length is calculated. For all combinations between the OSM features the sum of their length proportions is calculated (table in figure 14).

Combinations of features with a sum of length proportion below 0.8 or above 1.2 and combinations of features which are not connected are not further considered. The sum of length proportion which is closest to one determines the matching OSM features. In this example, 0.875 is closest to one and it represents the length proportion of OSM feature I and II. Thus the LM feature is matched to OSM feature I and II.

5.3.9 Step 7 Feature name similarity matching

Around each OSM feature not jet matched a buffer with the size of the GPS accuracy (10 meters) is created. For all LM segments found within the buffer, similar name matching is applied. The same formula for calculating name similarity as in step 3 is used (see chapter 5.3.6).The value is subtracted from one.

All LM features with a name similarity higher than 0.75 are assigned as matches. A higher threshold compared to step 4 is used following the suggestion from Koukoletsos et al. (2012).

5.3.10 Step 8 Cross Check

A feature can be unmatched in one dataset, but it can be assigned as a matching partner of a feature in the other dataset. Therefore, the final matching step it checkes if matching information about non-matched features in one dataset exists in the other dataset. In more detail explained for the OSM dataset. For all non-matched features in OSM dataset it is check if the LM dataset lists the OSM feature as a match for an LM feature. If this is the case the LM feature is assigned as a match for the OSM feature in the OSM dataset. The same procedure only in the other way around is also applied for the LM dataset.

5.3.11 Post processing

The matching algorithm is performed on the geometric simplified version of the datasets. The quality assessment requires the use of the original data. Therefore, the matching information is transferred to the original datasets. Furthermore, the datasets are clipped to the unbuffered tile extent.

To be able to visualise the matching result matching percentages are calculated per tile and for the complete study for both datasets. The matching percentage describes the ratio of

the summed length of matched features compared to the summed length of all features in the complete study area respective tile.

5.4 Quality Assessment

In this section the applied methods for the quality assessment are described. The evaluation is conducted relative to the LM dataset and is based on the matching results. The quality elements completeness, positional accuracy and road name accuracy are calculated.

5.4.1 Completeness

Matching percentages (cf. 5.2.11) represent the amount of data found in the other dataset (Koukoletsos et al., 2012). Therefore, the two completeness sub-elements (cf. 3.1.1) can be estimate as followed:

- OSM omission: The LM matching percentage can be interpreted as the amount of LM data found in the OSM dataset. A LM matching percentage of 100% means that OSM contain all features of the LM dataset.

- OSM commission: The OSM matching percentage expresses the amount of OSM data found in the LM dataset. Thus, a value below 100% indicate that the OSM dataset contain data which are not included in the LM dataset. To be able to uses this relationship to estimate OSM commission two requirements have to be true:
    - OSM dataset include only road features which are from definition also included in the LM dataset
    - OSM roads are classified correct

  Both requirements are not true. The OSM road type *service* correspond to the LM road type *Residential Road* but it includes more detail, such as car parks which are not represented in the LM dataset specification (c.f. table 1). Furthermore, the OSM classification of road types cannot be trusted 100% which adds further uncertainty to the measurement of OSM commission.

Therefore, completeness quality is only described with the OSM omission which is calculated from the LM matching percentage. The term completeness refers hereafter only to the definition of omission. Completeness is calculated for the complete study area as well as for each tile.

5.4.2 Positional Accuracy,

Only the absolute accuracy of the OSM dataset is evaluated and therefore the term positional accuracy is set equal with the definition of absolute accuracy (see section 3.1.1).

Positional accuracy is calculated using the average distance method described in section 3.2.2 (see formula 2). This measurement is chosen as it returns an average positional error between two lines, while the Hausdorff distance represents the maximal distance between

lines and is therefore more sensitive to outliers. The buffer method is not chosen because it is computationally more expensive than the average distance method.

For each OSM feature which has a match the average distance to its corresponding LM feature is calculated. If an OSM feature is matched to more than one LM feature, average distance of to each matching feature is calculated and the added distance is assigned as positional accuracy.

5.4.3 Road name attribute accuracy

Thematic accuracy is limited to the analysis of the road names. The road name attribute exists in both datasets and the name attribute is among the most important attribute of roads features.

Road name accuracy is calculated using the normalized Levenshtein distance (see section 3.2.3; formula 2). The normalized Levenshtein distance is chosen due to the reasons described in section 5.3.6.

For each OSM feature which has a match the normalized Levenshtein distance with its corresponding LM feature is calculated. If an OSM feature is matched to more than one LM feature, name accuracy with each matching feature is calculated and the worst (highest) value is chosen.

5.5 Implementation

The methodology in 5.3 and 5.4 is implemented using the programming language Python in combination with the geographic information system QGIS (Python, 2014; QGIS, 2014). Both, QGIS and Python are open source projects (see 5.5.3 for detail license description). As this study investigates open data it was natural to use only open source programs. Furthermore, the use of open source software does not restrict the further use and development of the method to licensed software. Both programs support the most common operation systems like Windows, OS X and Linux. In this study a Windows operation system is used.

The QGIS API and all its functionality can be accessed in Python. This makes the combination of the two programs a powerful tool to develop automatic scripts for geospatial tasks. QGIS comes along with a build-in Python console, which allows as straight forward execution of scripts and visual display of results.

5.5.1 Program structure and code

All matching and quality assessment methods described in section 5.3 and 5.4 are implemented as an automatic process. The only manual steps are the preparation of OSM and Lantmäteriet data (see section 5.3.2 and 5.3.3) and the creation of configuration file which stores directory paths and attribute column names of the input data. The required input data besides the OSM and LM data is a polygon shapefile of the extent of the study area.

The algorithm starts by loading the configuration file. Next, the study area is divided into 1km² tiles. After that the matching and quality assessment routines are executed for each tile in the study area.

Basic python functionality and code build the skeleton of the algorithm. Non spatial data are mainly stored in list and dictionaries. Spatial data are stored in shape-files. Shape-files are assessed, edited and created using the QGIS API. Spatial operations, such as buffering, spatial queries and distance calculation are implemented using the QGIS API. The QGIS API commands can be often assessed through only few lines of codes, which make it clean and straightforward to use. In total the code is about 1000 lines long. The code is executed from the QGIS python console (see figure 15).

Some Python QGIS API command examples:

- Opening a shapefile and iterating through all features:
  - for Feature in *QgsVectorLayer(Directory path, Name, Dataprovider).getFeatures():*
- Buffering of a shapefile:
  - *processing.runalg("qgis:fixeddistancebuffer", Input Shapefile, Buffer size, Segment numbers, Dissolve(Boolean),Output Shapefile)*

Variables which are used in the algorithm, such as the buffer sizes or name similarity thresholds are hard coded. Hard coded means, that values are written in the code and are not obtained by user input or external sources.

5.5.2 User input

The user input is limited to a few definitions of variables in the configuration file.

In the configuration file the user has to define:

- directory paths of the input data, output directories and used python files
- the encoding of the two datasets
- the attribute column names of the road name attribute of the two datasets
- the attribute column name of the road type attribute of the LM datasets
- a list of all road types of the LM dataset together with the road width of each road type

The directory path of the configuration file has to be hardcoded in the script. Figure 15 shows a screenshot of the python console of QGIS from where the script is run. The blue circle highlights the required input configuration file path in the script.

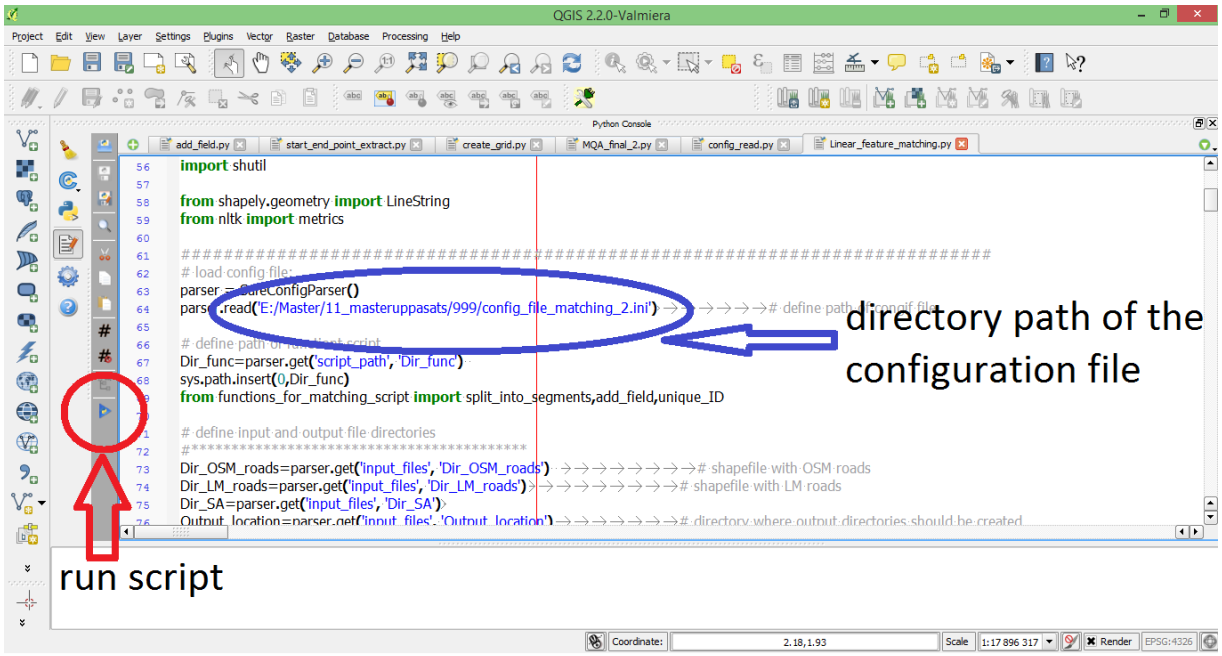


Figure 15: Screenshot of the python consol of QGIS from where the script is run. The blue circle highlights the required input of the directory path of the configuration file. Are all parameters defined within the configuration file, the script is ready to run (red circle).

5.5.3 Licence issues

As mentioned earlier QGIS and Python are open source projects. This means that they are free to use as well as allowed to be modified and to be redistributed under certain license restrictions (Open Source Initiative, 2014).

QGIS is licensed under the GNU General Public Licence (GPL) and Python is licensed under Python Software Foundation License (PSFL) which is a BSD license style (QGIS, 2014; Python, 2014). The main difference is that GPL is a copyleft licence while PSFL is not. Copyleft means that any derived or modified work from the original work has to be published under the same license as the original work. Detailed licence issues of QGIS and Python regard rules for the modification of the source code. Since this study only uses the source code, licence issues are not a restriction in this study. It is only important that the QGIS and Python licenses allow anyone to use their product for free.

The developed matching code in this study is published under GNU General Public Licence (GPL) (see appendix for access information). The author wants that the code is free and that future modifications and redistributions stay free.

6 Results

The method is applied each 1 km² tile in the study area which contain data from at least one of the two data sources. In total 861 tiles covering an area of 861 km² were investigated. The algorithm needed about 5 hours to execute on a machine with a 2.0 GHz dual core processor and 4 GB RAM.

6.1 Matching result

Table 2 presents the matching results for the complete study area. In both datasets around 80% of the data are matched. The LM dataset has a slightly higher match percentage than the OSM dataset.

Table 2: Matching results for the OSM and LM dataset for the complete study area.

Dataset	Total length [m]	length matched [m]
OSM	4596570.04	3550564.09 (77.2%)
LM	4691594.60	3800412.39 (81.0%)

Figure 16 shows the matching percentage per tile for both datasets. Tiles are classed into 4 classes depending on their matching percentage (>80%, 50-80%, <50% and no data). Table 3 shows the distribution of the classes. Both datasets show a similar statistical distribution, more than 50% of the tiles have a matching percentage above 80%. In more than 80% of the tiles at least half of the features are matched. In around 6 % of the tiles no matches are found. However, a spatial difference of the matching percentages is visible. The LM dataset shows an accumulation of tiles above 80% matching percentage in areas with denser road network and tiles with a worse matching percentage are found in areas with less dense road network. The OSM dataset shows the opposite pattern.

Table 3: Number of tiles found in the four classes of matching percentages for the LM and OSM dataset. In parenthesis: the percentage of tiles in one class to total number of tiles (861 tiles).

Matching percentage [%]	LM dataset [number of tiles]	OSM dataset [number of tiles]
>80	467 (54.2%)	527 (61.2%)
50-80	254 (29.5%)	250 (29.0%)
<50	90 (10.5%)	34 (4.0%)
no data	50 (5.8%)	50 (5.8%)

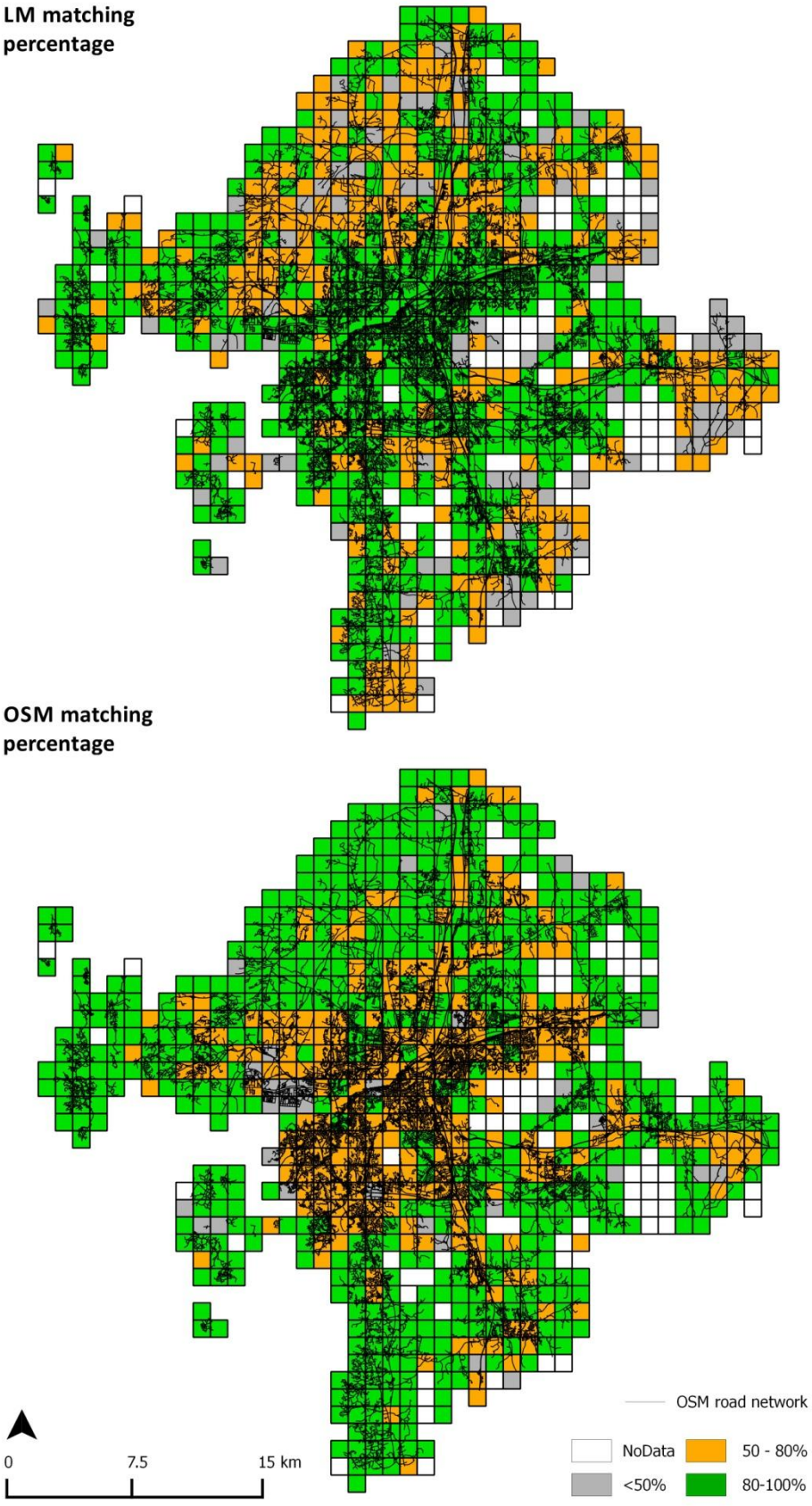


Figure 16: Matching percentages per tiles. Matching percentages are classed into 4 classes. Upper figure, result for LM dataset. Lower figure, result for the OSM dataset.

Figure 17 displays the example of the matching results for an area in the inner city of Göteborg. The data are classed in matched and unmatched data. It is visible that most features are matched in the two datasets. The OSM dataset has more unmatched data than the LM dataset. Though, most of the unmatched OSM data does not exist in the LM dataset.

Figure 17: Example of the matching result for an area in the inner city of Göteborg. OSM and LM features are classified into matched and non-matched features.

A manual evaluation is performed on 5 percent of the tiles, in total 44 tiles which are randomly selected. Two kinds of errors are distinct.

The first error, called *missing information*, account for features which miss matching information. This error can account for features which are matched as well as for features which are not matched. Therefore, the *missing information* error can be divided into two categories. One category called *additional* describes when a feature is matched but should

be matched to more features. The other category called *new* accounts for features which are not matched but should be matched.

The second error, called *mismatch*, reports features which are matched incorrectly.

Figure 18 shows an example of each error. In the missing information example, the LM feature 1853 should have been also matched to OSM feature 36186 (upper blue circle) (*additional* category) and this OSM feature should have been matched to the LM feature (*new* category).

In the example mismatch error, the LM feature 2795 and 2548 are matched to OSM feature 45241 (lower blue circle) and this OSM feature is matched to these two LM features, which is incorrect.

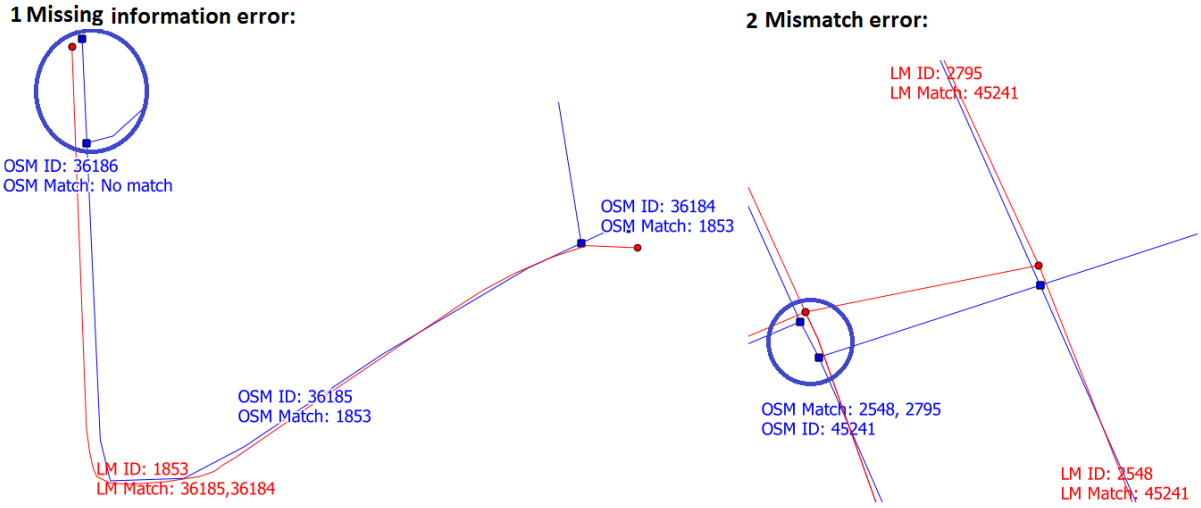


Figure 18: Example of the two kinds of errors which are distinct in the matching evaluation. Blue lines represent OSM data and red lines display LM data. The blue circles highlight features that are not matched (example 1) and mismatched (example 2).

Table 4 presents the results of a manual evaluation of the matching results. The total matching error for OSM dataset is 11.8% and for the LM dataset 14.3%. The *mismatch* error is 1.8% and 2.4% for the OSM, respective LM dataset. The *missing information* error is around 10.0% for the OSM dataset; it is composed of 4.0% belonging to category *additional* and 6.0% to category *new*. The *missing information* error is 11.9% for the LM dataset. The *additional* category has the biggest share with 7.5%, while the *new* category account for 4.4%.

Table 4: Results from the manual evaluation of the matching results. Two matching errors are distinct, missing information and mismatched. Missing information errors can be divided into two categories, new and additional.

Dataset	Total length evaluated [m] (% of total dataset length)	Missing information [m]	Mismatched [m]	Matching error [m]
OSM	279041.84 (6.1%)	27798.74 (10.0%) (4.0%) additional (6.0%) new	5118.63 (1.8%)	**32917.37** **(11.8%)**
LM	286507.24 (6.1%)	34132.10 (11.9%) (7.5%) additional (4.4%) new	6956.21 (2.4%)	**41088.30** **(14.3%)**

Missing information errors occur most often when one of the datasets has an intersection which is not represented in the other dataset, like in example one in figure 18. In the evaluated tiles, OSM data had more additional intersections than LM data.

During the evaluation additional, more subjective impressions about the matching quality are gathered. Features with a similar geometric representation are usually successfully matched. Errors occur mostly when features have a different geometric representation in the two datasets. Different geometric representation can be for example, the same object in reality is represented with different number of features in the two datasets, a road is represented in one dataset as a dual carriage way while in the other as a single road, a crossroads has links in one dataset and in the other dataset it is a simple intersection.

Figure 19 shows an example of roundabouts. In A the roundabouts are represented similar in both datasets, while in the geometries of the roundabouts in B are obviously different. The roundabouts with similar geometry are matched perfectly. Most of the features belonging to the roundabouts with different geometry are not or are incorrectly matched.

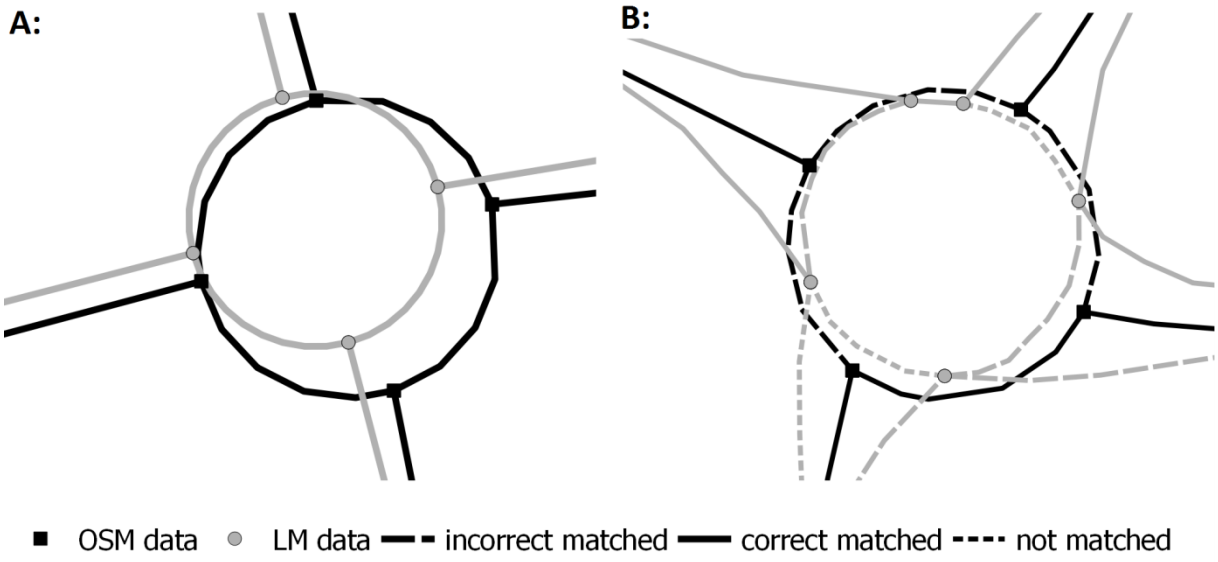


Figure 19: Example of the matching results of roundabouts. In A roundabouts with similar geometric representation in OSM and LM dataset are displayed, while the roundabouts in B is different geometrically represented in the two datasets.

Motorway or dual carriage roads are also often differently geometric represented in the datasets. The differences are mainly that link roads meet the main roads at different locations (see figure 20, black circles). This often leads to matching errors.

Other geometric differences occur at bridges. OSM features are split at bridges, while LM features are not (see figure 20, grey circle). Though, the LM dataset has an extra class for *underbridge* which splits LM features before and after a bridge.

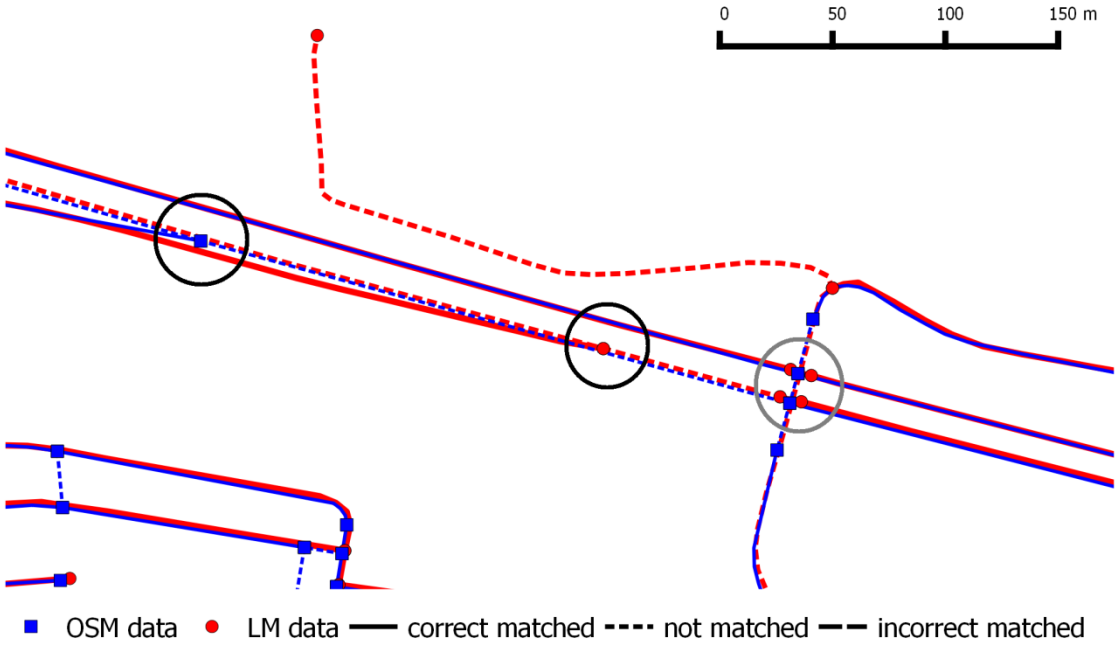


Figure 20: Example of matching a dual carriage road. Blue lines represent OSM data and red lines display LM data. Quadrates respective round points represent node of the two dataset. The features are classified into correct matched (solid line), not matched (dotted line) and incorrect matched (dashed line).

Another matching error which is typical for motorways or dual carriage roads is that a feature is matched to the feature representing the lane in the opposite direction and not to the feature which represents the lane in the same direction.

Features in both datasets are sometimes splitted even though they do not intersect with another feature. This leads as well to matching errors. These problems occur most often on longer roads.

Figure 21 shows the execution time per tile in relationship to the number of OSM and LM features in a tile. The execution time increases with the number of objects to process. A quadratic trend is recognizable.

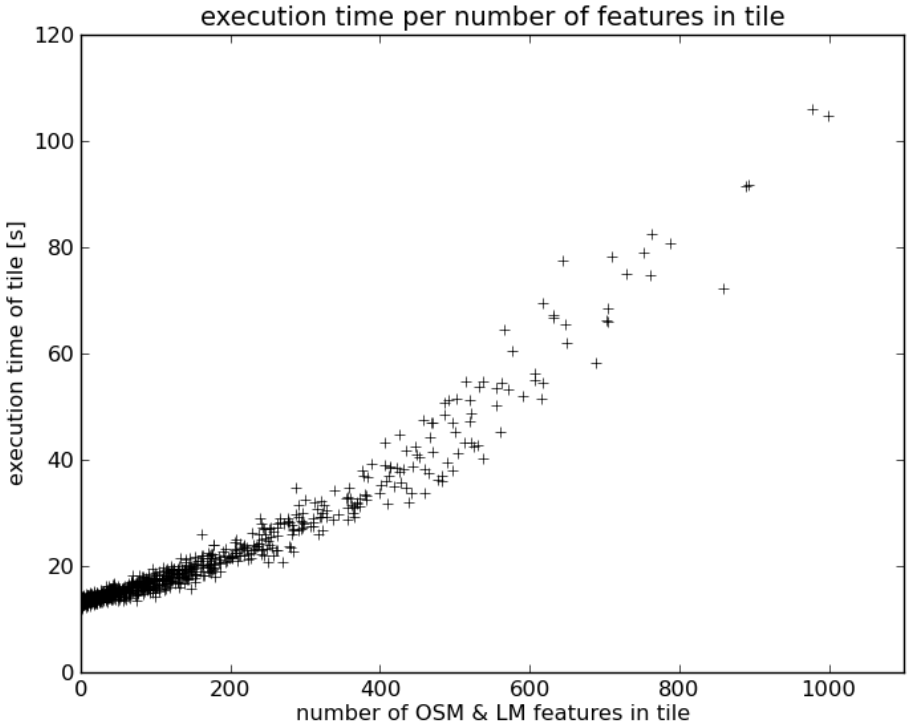


Figure 21: Execution time of the matching algorithm per tile, plotted against the number of OSM and LM features in the tile.

6.2 Quality assessment result

6.2.1 Completeness

As explained in section 3.2.1 OSM completeness can be defined as the matching percentage of the LM dataset. The OSM completeness in the complete study area is 81% (see table 2). Figure 22 shows OSM completeness for each tile. The distribution values of table 3 for the LM dataset are also valid for this figure. Figure 23 shows the distribution in more detail.

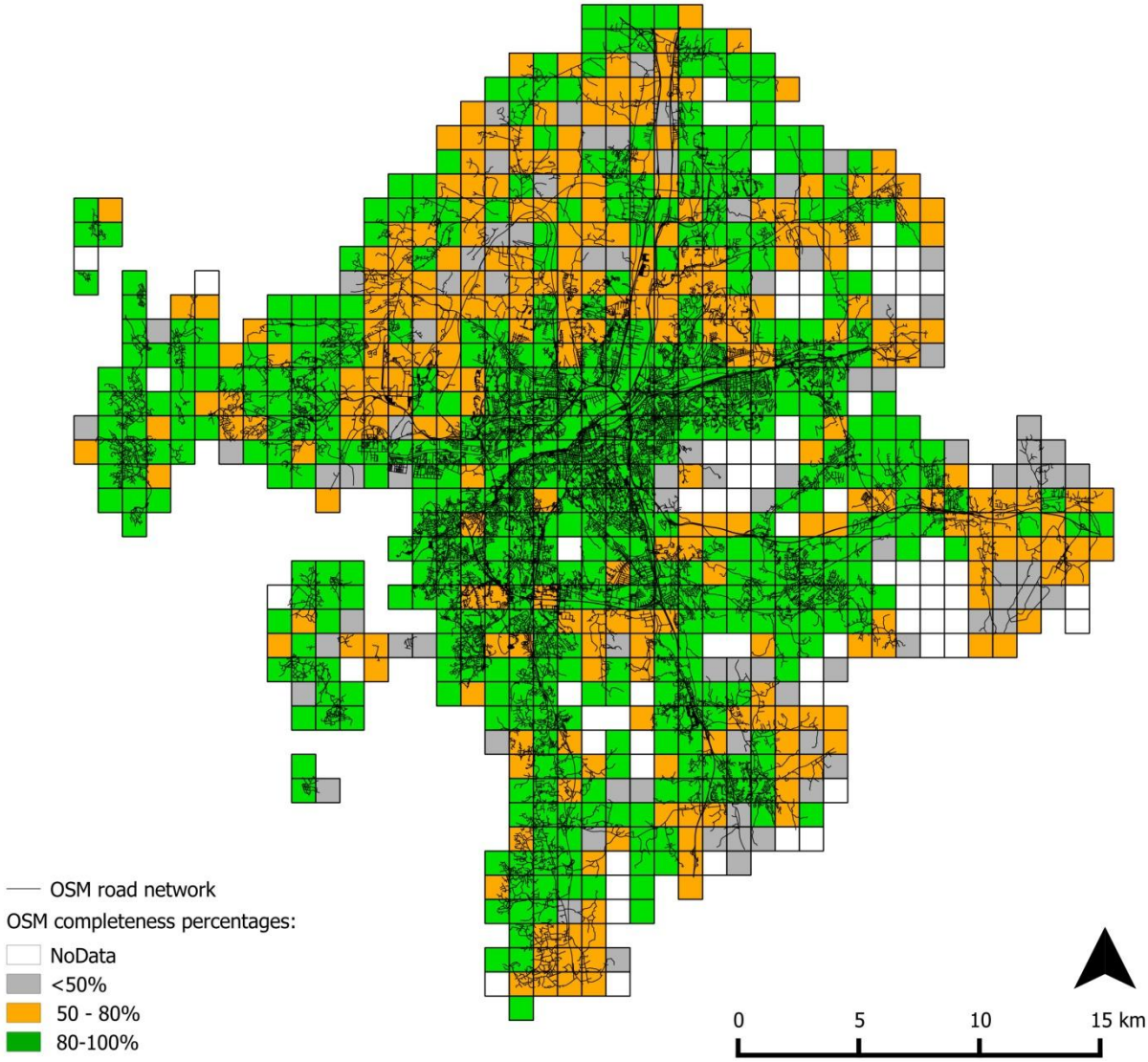


Figure 22: OSM completeness in percentage per tile. OSM completeness percentages are classed into 4 classes.

It is visible that tiles with high OSM completeness are found in areas with denser road network (see figure 22). While in areas with sparse road network the OSM completeness decrease. Tiles with a low completeness mostly contain very little data.

This is confirmed in also figure 23. It shows the OSM completeness per tile plotted against the amount of OSM and LM features in a tile. The frequency of low completeness percentage of a tile decreases with increasing number of features in a tile.

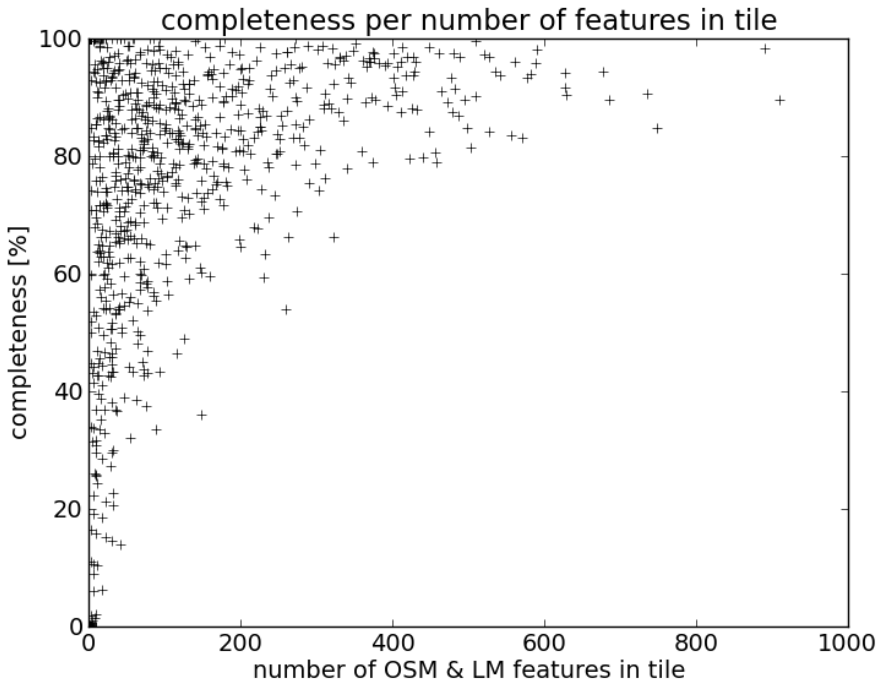


Figure 23: OSM completeness in percentage per tile plotted against the number of features in the tile.

6.2.2 Positional accuracy

Positional error values of above 50 meter are not considered in the results presented below, because they occur most likely due to matching errors. In total 1304 features (3.5% of all matched features) are excluded due to this error. Furthermore, the clipping of the datasets to the original tile sizes created a few multiline strings features for which the average distance could not be calculated. 313 features (0.8% of all matched features) are excluded due to this reason.

The results of the positional accuracy presented below are based on 36002 features.

Figure 24 shows the cumulative frequency of the positional accuracy values for the complete study area. 75% of all matched OSM values have a positional accuracy below 3.6 meter. 95% of the features have a position accuracy of less than 17 meters. The mean positional error is 3.8 meter

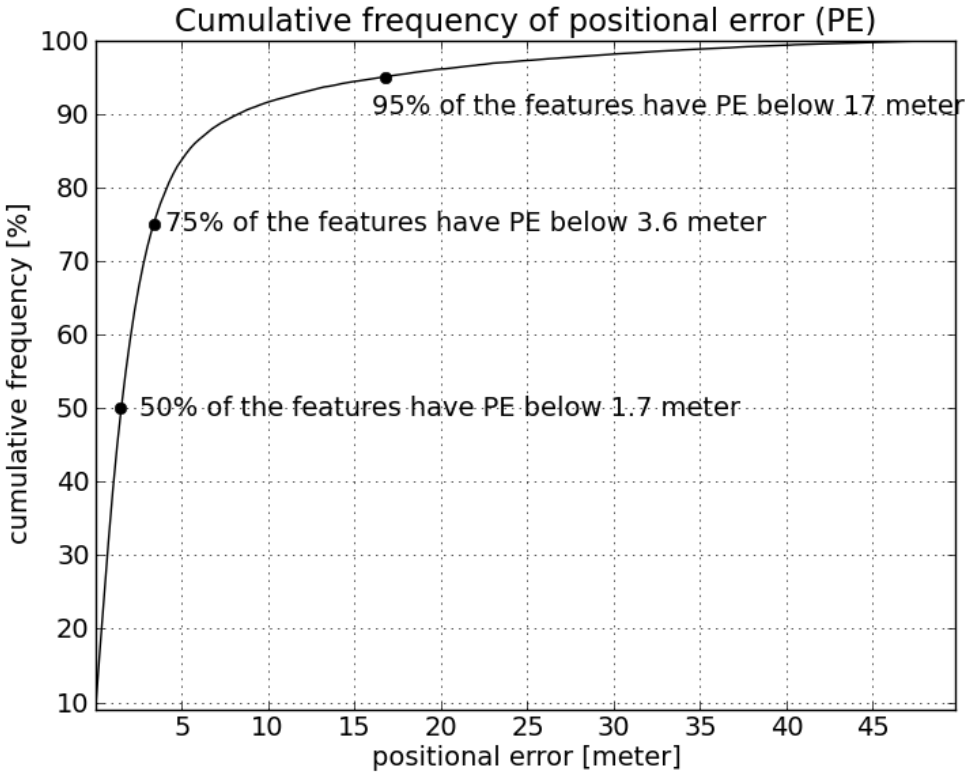


Figure 24: Cumulative frequency in percentage for positional error values in meter of the complete study area.

Figure 25 shows the average positional accuracy per tile divided into 4 classes (<5, 5-10, >10 meters and no data). Table 4 shows the number of tiles in each class. Nearly 70% of the tiles have an average positional error of less than 5 meters. 4% of all tiles have an average positional accuracy above 10 meters.

Table 4: Number of tiles found in the four classes of positional accuracy of the OSM dataset. In parenthesis: the percentage of tiles in one class to total number of tiles (861 tiles).

Positional accuracy [m]:	Number of tiles:
< 5	591 (68.64%)
5-10	183 (21.25%)
>10	34 (3.95%)
No data	53 (6.16%)

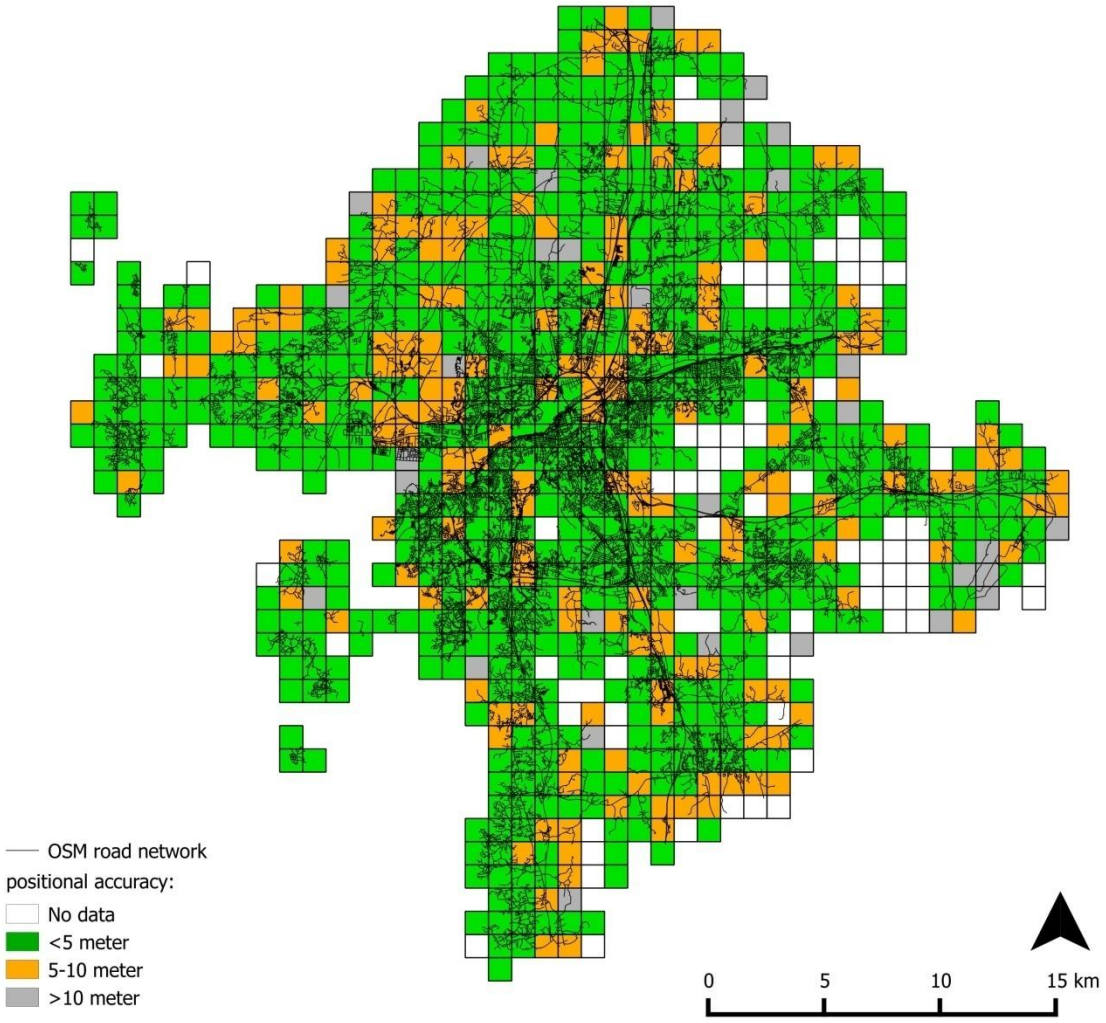


Figure 25: Positional accuracy in meter per tile classified into for classes. The values represent the mean "average distance" of features per tile.

Figure 26 displays two examples of matches between OSM and LM features. In the example A one OSM feature is matched to two LM features (1:M relationship). In the example B two OSM feature are matched to the same LM feature (M:1 relationship). These cases create problems for the calculation of the positional accuracy measurement.

In the case A the average distance is calculated once between OSM feature 43175 and LM feature 24163 and once between OSM feature 43175 and LM feature 24648. The positional accuracy for element for OSM feature 43175 is then the sum of these two calculations. However, this does not represent the true value.

In the case B the positional accuracy for the both OSM feature 43174 and 43173 are calculated as the average distance from each feature to the LM feature 24650. Though, these values do not represent the true positional error for the features.

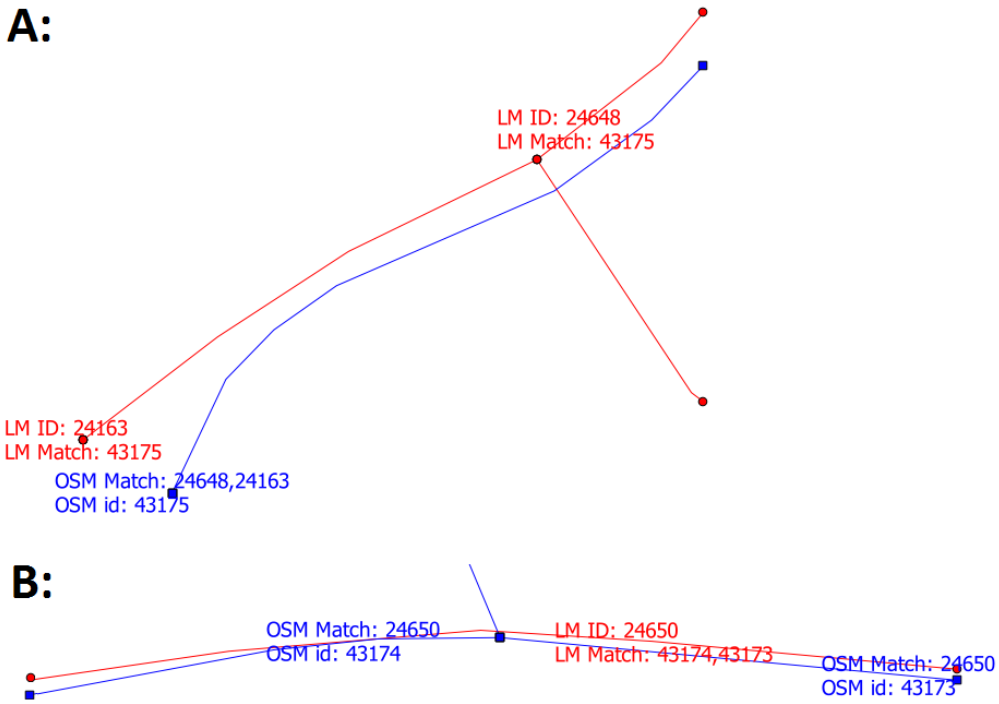


Figure 26: Example of M:1 and 1:M matching relationship. These relationships create problems for the calculation of the positional accuracy measurement.

6.2.3 Road name attribute accuracy

Figure 27 represents the distribution of road name analysis. 67% of all matched OSM features have a correct name, 4.5% are misspelled. 25% of the matched OSM features do not have a name and in around 3% the matching LM feature(s) do not have a name.

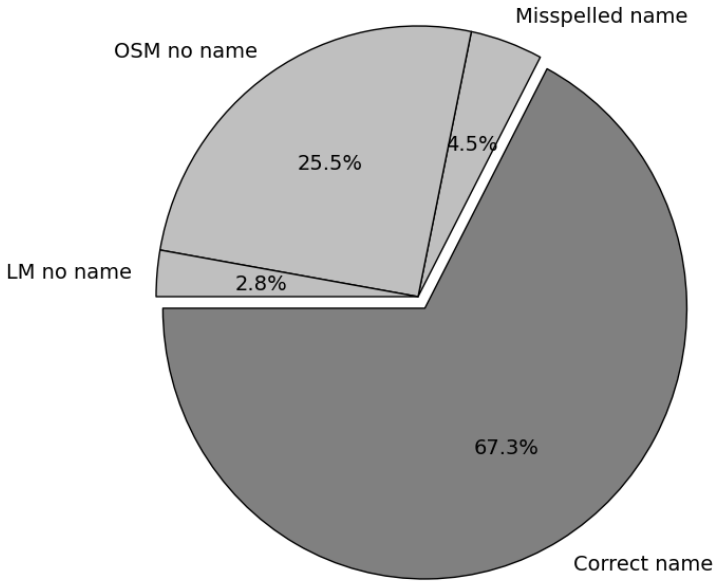


Figure 27: Percentage of matched OSM features with a correct name, misspelled name, no name and no name of the matching LM feature compared to the total number of matched OSM features (37254) in the study area.

Figure 28 shows the name accuracy per tile. For each tile the percentage of matched OSM features with a correct name and matched features with a misspelled name which is still readable (a normalized Levenshtein distance less than 0.2) compared to the total amount of matched features is calculated. The tiles are classified in four classes (>80, 50-80, <50% and no data). Table 5 shows the distribution of the tile classification.

It is visible that no clear pattern of the spatial distribution is recognizable. Tiles with dense as well as with sparse road network have both high and low percentages.

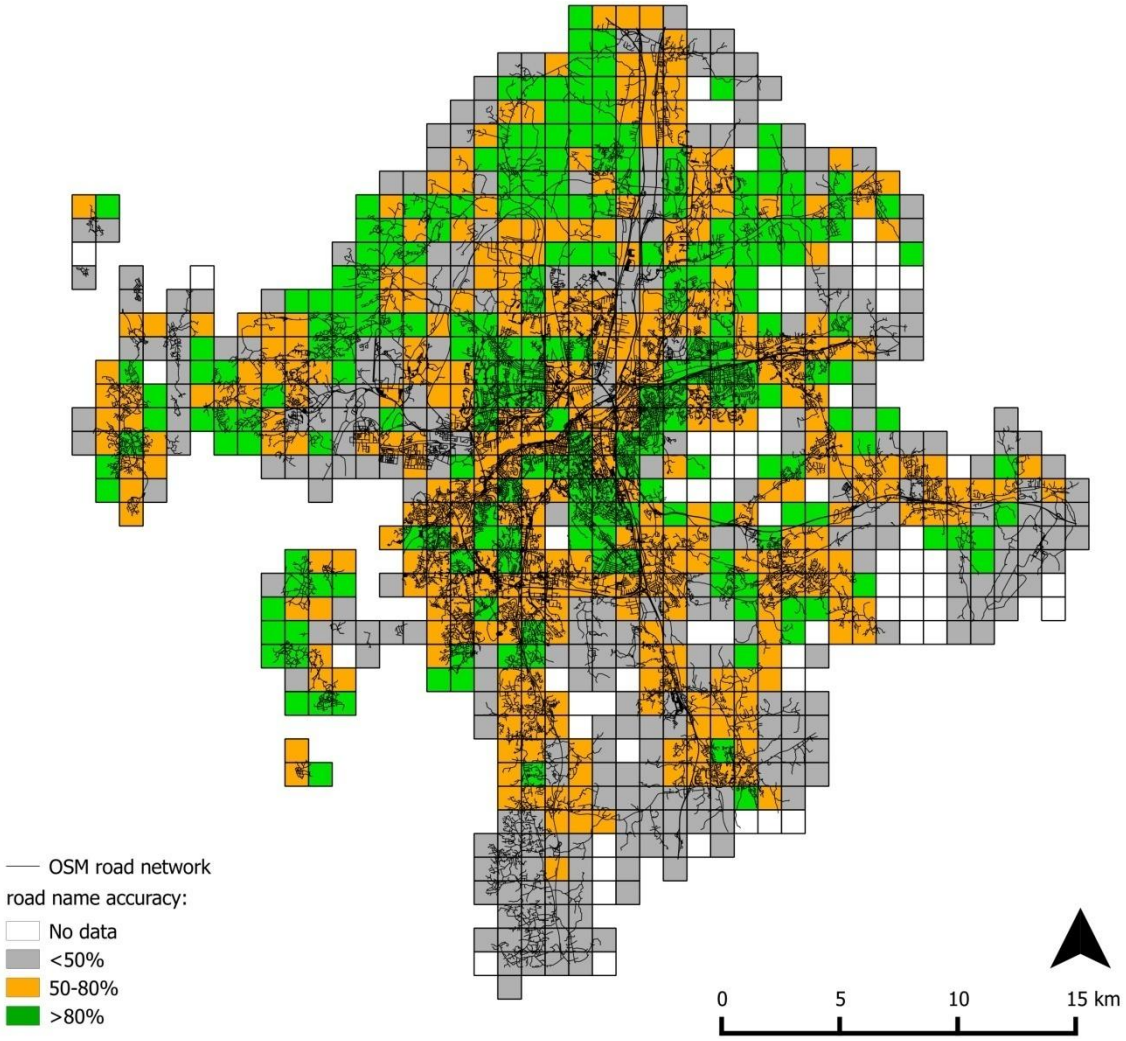


Figure 28: Road name accuracy classified into four classes per tile. Percentage of matched OSM features with a correct name and matched OSM features with misspelled names which are still readable compared to the total amount of OSM features per tile.

Table 5: Number of tiles found in the four classes of road name accuracy of the OSM dataset. In parenthesis: the percentage of tiles in one class to total number of tiles (861 tiles).

Name accuracy [%]:	Number of tiles:
>80	231 (26.83%)
50-80	325 (37.75%)
<50	255 (29.62%)
No data	50 (5.88%)

7 Discussions

In this study an automatic matching algorithm is developed to match OSM and LM road datasets. The results of the matching process are then used to evaluate the quality of OSM, regarding the completeness, positional accuracy and road name accuracy of the OSM road network.

The first part of the discussion concerns the developed matching algorithm and in the second part the quality assessment of OSM is discussed.

7.1 Matching process

7.1.1 Implementation

As explained in section 5.3 the matching algorithm, used in this study is an extension of the methodology from Koukoletsos et al. (2012). Their methodology is developed to match OSM data and has a low matching error. Though, their algorithm only classifies features into matched and unmatched features and does not consider feature correspondence. This part therefore had to be developed in this study. Two main problems needed to be solved during the implementation:

- Implementation problem: The translation of the methodology of Koukoletsos et al. (2012) into code form
- Method problem: The development of feature correspondence.

A limitation of the implementation of the matching algorithm is the insufficient testing of the different threshold and variables used in the algorithm. Most of the thresholds, among others the buffer size, angular tolerance and similarity thresholds are directly adopted from Koukoletsos et al. (2012). Others, such as the threshold for geometric simplification are specified based on tests with a limited amount of features

Robust testing of different thresholds is a very time consuming process as it involves a manual evaluation of the results. Furthermore, thresholds have to be tested in a larger area to include the spatial varying representations and properties of the datasets. The timeframe of this study did not allow an extensive testing of the thresholds, even though this would be preferred and might improve the matching results.

The spatially changing representations and properties of the datasets created the main problems during the implementation of the matching routine. A successful execution of the matching algorithm in one test area did not guarantee that the algorithm produced good results in another test area.

The main problem of implementations of feature correspondence was how to deal with M:1 and 1:M relationships in step 6 (see section 5.3.8). If segments of a feature were matched to segments belonging to the same feature of the other dataset, feature correspondence can be set between the two features. However, when segments of a

feature were matched to segments belonging to more than one feature of the other dataset, it has to be decided which are the correct matching features.

This is done by investigating the lengths of the features. The described method in section 5.3.8 assumes that the length of a feature is similar to the length of its matching feature(s). Figure 26 shows two examples where this assumption is true. This assumption is also in general true, as matched features should represent the same object in reality and thus should approximately have the same length. Furthermore, matching features need to be adjacent. This constraint is added to assure that the result is topologic correct. A limitation of this approach is that it fails to matches 1:M relationships with different lengths. These cases are rare but can occur, especially with OSM data. An end of a road can be defined differently by an OSM contributor and a LM cartographer. This mainly applies for minor roads which end in e.g. a forest or field.

Another approach would have been to simply match a feature to all the possible features. However, this solution would transfer all matching errors from the segment level to the feature level.

In Step 2-5 OSM segments can have more than one matching partner (1:M relationship) while LM segments only can have one match. The reason for this is that it is assumed that a LM feature consists of more segments than an OSM feature (see figure 10). However, this is only based on a visual comparing of datasets and no robust tests have been performed. Therefore, it would be interesting to see which effect 1:M matching at LM segment level has on the matching results.

7.1.2 Matching results

The developed matching routine matched around 80% of the features in both datasets. A relative clear pattern is visible when looking at the matching results for each tile. The LM datasets has the highest matching percentages in areas with denser road network while OSM dataset has the highest matching percentages in areas with less dense road network.

Koukoletsos et al. (2012) reported similar results for their study area covering London. Their matching percentage of the reference dataset was about 10% higher than in this study. The higher percentage can be explained as their study area contains a denser road network than the study area in this study. Their study reported the same matching percentage patterns of the two datasets depending on the road density.

The matching returned subjectively good results. The algorithm succeeded to match most of the features correctly. On the other hand, features which have a different geometric representation in the two datasets are frequently incorrectly matched.

The manual evaluation of the matching revealed a matching error of 11.8% for the OSM dataset and 14.3% for the LM dataset. The main proportion of the total error belongs to *missing information* errors. The *mismatch* error is less than 2.5%. This means that the

matching algorithm has a low mismatching error but on the other hand need improvements to reduce the number of features which are not matched.

The error percentages are considerably higher than the errors (around 3%) reported by Koukoletsos et al. (2012). However, it should be noticed that the matching routine from Koukoletsos et al. (2012) classify features into matched and unmatched features and does not considered feature correspondences. Feature correspondence adds an additional error source to the matching.

Errors belonging to the category *additional* from *missing information* errors cannot be identified in the results of Koukoletsos et al. (2012). Subtracting theses percentages from the total matching error, OSM and LM dataset have a matching error of 7.8%, respective 6.8%. These values are still higher than the errors reported in Koukoletsos et al. (2012). The differences can be further explained with the use of Koukoletsos threshold and the insufficient testing of them. The thresholds are developed to fit their input data and might not be directly transferable to other datasets.

7.1.3 Error discussion and match improvement

The most frequent matching error *missing information* errors occur most often when one of the datasets has an intersection which is not represented in the other dataset, like the example 1 in figure 18. It has to be stressed that the algorithm still succeeds to match most of these features (see figure 26). No clear pattern has been found why the algorithm failed in some cases. The most likely cause is that the features of the two datasets are split into segments with length difference above the buffer size in step 1 (cf. 5.3.3) and therefore the correct matching segments cannot be found.

An additional matching step might be able to match non-matched features and so reduce the *missing information* error. To consider topological relationships have been proven to be useful in matching algorithms (Stigmar, 2005). Therefore, an additional matching step should use topological properties. A possible additional matching step could check for each unmatched feature whether the neighbouring features contain matching information and then uses this information to select the matching feature.

Furthermore, the manual evaluation revealed that matching errors arise most often when corresponding features are represented geometric differently in both datasets. If the differences become too large correct automated matching becomes very difficult. The roundabout in B in figure 19 is an example of this case. No direct relationship can be established between the features of the datasets. A solution could be to treat the features belonging to the roundabout as one and match theses. In order to implement this in automatic matching routine, problematic roundabouts have first to be automatically identified.

Other matching errors connected to different geometric representation of the datasets can be solved easier. One example is that OSM features are cut at bridges while LM features

are not. This error is introduced in the data preparation of OSM. OSM data are splitted into features at each line intersection. This step is of great importance, because otherwise the geometry between OSM and LM data would be very different. However, this step has to be modified so it does not split features at bridges and other line intersection which not meant to have a junction.

The LM dataset has another issue concerning the representation of bridges; the parts of a road under a bridge are classified as *underbridge*. OSM data does not include this classification. This results in features of considerably different lengths which makes matching difficult. To solve this problem LM features belonging to the class *underbridge* should be classified to its road type and merged to their neighbouring features.

Features in both datasets are sometimes spitted even though they do not intersect with another feature. This causes that the same object in reality is differently geometrically represented in the two datasets which create problems for the matching algorithm. An additional data pre-processing step, which merges these features, should be able to remove these problems.

Link roads of the two datasets often meet the main road at different positions. In the example at figure 20 (black circles) the OSM link connects to the main road at a position around 180 meters away from the connection point of the LM link. The links in this example are correctly matched to each other. Though, the different connection points split the main roads in features of different lengths which causes that the features are not matched (see figure 20).

A matching error which is typical for motorways and other dual carriage roads is that a feature is matched to the feature representing the lane in the opposite direction and not to the feature which represents the lane in the same direction. These errors are caused most likely during the creation of the candidate list in step 1 (section 5.3.3). For an LM segment no OSM segment belonging to the same lane but an OSM segment belonging to the other lane is found within the buffer size. The segments are then matched in the following steps. The segmentation matching approach might not be optimal for matching dual carriage roads. An approach to match this road class only on feature level can be also problematic due to the problems caused by links. Further investigations have to be done in order to improve the matching results of dual carriage ways and for solving matching problems related to links.

As mentioned earlier thresholds are insufficient evaluated. A sensitive analysis of the thresholds could probably improve the matching results by better adopting the thresholds to the used datasets. Though, testing is a very time consuming process due to the necessary manual evaluation of the matching results.

Nevertheless, the thresholds used for similarity matching (step 4 and 7) should be tested more detailed. A different method has been used in this study to calculate name similarity

than in Koukoletsos et al. (2012). It was intended to use their method, but the method return incorrect results and was therefore replaced with the normalized Levenshtein distance. Whether the method failed because of language difference between English and Swedish or simple was incorrect programmed is not clear. Due to time limitation the initial thresholds are kept. A very limited test (20 names) conformed that the thresholds in combination with the normalized Levenshtein distance returned appropriated results. However, these thresholds need to be validated against more misspelled names.

The threshold used for the geometry simplification should be evaluated as well more detailed. The requirement for the threshold was to not alter the geometry to much but still remove additional breakpoints. The threshold is then chosen based on visual inspections. A detailed analyse of the effect on the matching results using different thresholds is not performed.

The shape and size of the used tile dataset can have an effect on the matching results. This is not tested in this study. However, Koukoletsos et al. (2012) tested the robustness of their matching algorithm for different tile sizes and shapes. They stated that the matching results remained practically the same. Due to the fact that the applied method is based on Koukoletsos et al. (2012) it can be assumed that the tile dataset has no major effect on the matching results. Figure 21 show that the execution time per number of features in a tile has a quadratic trend. Thus, executing the matching process per tile reduced the total processing time.

7.2 Quality assessment

7.2.1 Completeness

The completeness analysis of OSM produced results similar to findings of other studies. The OSM completeness found in this study is 81%. Koukoletsos et al. (2012) reported for the greater area of London OSM completeness of 93%.

A analyse of spatial distribution of completeness per tile revealed that higher OSM completeness values are found more frequent in areas with a dense road network (urban areas) than in areas with less dense road network (rural areas). This relationship is discovered by all OSM quality studies presented in section 3.3.3 and is one of the most well known characteristics of OSM. Girres and Touya (2010) reported a positive relationship between number of contributors active in an area and its OSM completeness.

Only 10% of all tiles in the study area have less than 50% completeness, while around 85% of the tiles have OSM completeness above 50%. To decide if OSM data can replace other datasets in geospatial application depend on the exact requirements of the application. However, it might be more of interested to know how complete certain road types are presented in OSM. This analysis has not been done in this study. Though, the matching routine delivered all necessary data to calculate OSM completeness also for certain road types.

As the OSM completeness is defined as the LM matching percentage, the quality of the matching has a direct influence on the completeness measurement. *Mismatch* errors lead to an overestimation of the OSM completeness while *missing information* error of the category *new* underestimate the OSM completeness. The matching error of the category *new* (4.38%) is larger than the *mismatch* error (2.43%). Thus, OSM completeness is probably slightly better than 81%.

7.2.2 Positional accuracy

The average positional accuracy is 3.8 meter for the complete study area. 75% of all considered features for the calculation have an average distance of less than 3.6 meters. 90% of all tiles have an average positional accuracy below 10 meter.

The results are nearly identical with the values for positional accuracy reported by Ludwig et al. (2011). Haklay (2010) presented somewhat higher values.

The decision to not include average distance above 50 meters in the results of positional accuracy has been made after a visual control of these values. It was visible that high average distances can be related in most cases to *mismatch* errors. The threshold of 50 meters was set quite subjectively based on the impressions of the visual control. This might add some errors to the positional accuracy assessment. However, the threshold was set relative high to even include high average distances between correct matches in the assessment. 3.5% of all matched features are not included due to this reason; the OSM *mismatch* error is 1.8%. This indicates that the chosen threshold is acceptable but could be probably higher as more features are excluded than features are mismatched.

Another issue with the positional accuracy assessment is that the method calculates incorrect average distances for 1:M and M:1 matches (see figure 26). In both cases the calculated distance are larger than the correct distance. Thus, this limitation results only in an underestimation of the position accuracy. The problem is not actual distance measure but the implementation of it. In order to correct this, 1:M and M:1 matches have to be pre-processed before the average distance is calculated. This could be implemented by temporally merging OSM and LM features which are matched to the same LM respective OSM feature and then calculating the average distance between them. Furthermore, the method has to be extended to solve problems related to multiline string features.

Nevertheless, it can be concluded that the OSM data in this study area have a good positional accuracy.

7.2.3 Road name accuracy

Matched OSM features have in 67% a correct name, in 5% the name is misspelled, 25% have no name and in 3% the matching LM feature(s) has/have no name. It is visible that if contributors map the name of roads they do it mostly correct. However, in one of four features they do not map the name of a road. No relationship of areas with dense and less dense road network to road name accuracy per tile has been recognized.

Ludwig et al. (2011) reported better road name accuracy and they also discovered a decrease of accuracy in rural areas compared to urban areas. Girres and Touya (2010) also discovered that if an OSM feature has a name, then this name is most often correct.

The analysis shows that the name attribute in OSM is most often correct if it is included. However, the chance that a feature has road name is only 25%. This is most likely to be less than required for most geospatial applications.

Like for the completeness element it would be interesting to conduct further analysis to see for example if the name accuracy depends on road types.

8 Conclusions

The matching algorithm, developed in this study, succeeded to match volunteered geographic data from OpenStreetMap (OSM) and authority data from Lantmäteriet (LM) automatically for a larger area. The overall matching result is good with an acceptable matching error (matching percentages of around 80% with a matching error of about 14%). The matching returned very good results when corresponding features in the two datasets had a similar geometric representation. When the geometry of corresponding features was different, the matching performance decreased. With that the first aim of this study was reached.

The evaluation of the matching results showed that the matching algorithm produced a low mismatching error but on the other hand failed to match a considerably amount of features. Therefore, improvements for the future development of the matching algorithm are suggested. The first is to extend the algorithm with an additional matching step, based on topological relationships, to decrease the amount of unmatched features. Second, improvements in data pre-processing, to make the topology to the datasets perfect as well to align the geometry of the two datasets. Furthermore, proper sensitive test of the used thresholds are suggested.

The second aim regarded the quality assessment of OSM. This is achieved using the feature correspondence, which is created during the matching routine, to compute completeness, positional accuracy and road name accuracy of OSM relative to the LM dataset.

OSM can be seen as a proper data source which has some reservation. Completeness (omission) and positional accuracy of OSM are good, while the road name accuracy is of poorer quality. OSM completeness is higher in urban areas compared to rural areas; positional accuracy and road name accuracy have not shown this pattern.

From a business geospatial point of view, the assessed quality of OSM is good enough for simple application, e.g. a background map. For more advanced applications, such as address searches and logistic services, the quality of OSM is inadequate. However, more detailed analysis of the quality of OSM, which are tailored to an actual application of OSM data, are required to decide if OSM data can be used in business geospatial applications.

OSM data are continuously growing and improving (cf. figure 1 and 2). Hence, it is necessary and interesting to repeat OSM quality evaluations. This can be achieved easily with this method due to its automatic implementation.

The approach to base the quality assessment on matching results has proven to be useful, as it due to the establish feature correspondence enable the possibility of more detailed quality analysis of OSM, for example to calculate quality elements for certain road types.

The matching routine can be moreover the foundation for further applications of OSM data which require feature correspondence, e.g. the integration of other free data into OSM.

References

Ariza-López, F.J., A.T. Mozas-Calvache, M.A. Ureña-Cámara, V. Alba-Fernández, J.L. García-Balboa, J. Rodríguez-Avi, and J.J. Ruiz-Lendínez. 2011. Influence of sample size on line-based positional assessment methods for road data. *ISPRS Journal of Photogrammetry and RemoteSensing* 66: 708-719. doi: 10.1016/j.isprsjprs.2011.06.003

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.

Deza M.M., and E. Deza. 2013. *Encyclopaedia of Distances* (Second edition). Berlin-Heidelberg: Springer.

Douglas, D.H., and T.K. Peucker. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. In *Cartographica: The International Journal for Geographic Information and Geovisualization* 10: 112-122. doi: 10.3138/FM57-6770-U75U-7727

FGDC. 1998. Geospatial Positioning Accuracy Standards Part 3: National Standard for Spatial Data Accuracy. Federal Geographic Data Comittee. Reston, USA.

Field Papers. 2014. Field Papers. Retrieved 27 April 2014, from http://fieldpapers.org/.

Geofabrik. 2014. OSM data for Sweden. Retrieved 16 April 2014, from http://download.geofabrik.de/europe/sweden.html.

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., and G. J. Hunter.1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science* 11: 299-306. doi: 10.1080/136588197242419

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

Goodchild, M. F., and L. Li. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics* 1: 110-120. doi: 10.1016/j.spasta.2012.03.002

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

Haklay, M., S. Basiouka, V. Antoniou, and A. Ather. 2010. How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic Information. *The Catographic Journal* 47: 315-322. doi: 10.1179/000870410X12911304958827

Hangouet J.F. 1995. Computation of the Hausdorff distance between plane vector polylines. *Auto Carto* 12: 1-10.

ISO, 2002. ISO 19113:2002 Geographic information - Quality principles. 29 p.

ISO, 2003a. ISO 19114:2003 Geographic information - Quality evaluation procedures. 63 p.

ISO, 2003b. ISO 19115:2003 Geographic information - Metadata. 140 p.

Jakobsson, A., and F. Vauglin. 2001. Status of data quality in European national mapping agencies. In *Proceeding of the 20th International Cartographic Conference*, Volume 4, pp. 2875–2883. Beijing, China.

Koukoletsos, T. 2012. A Framework for Quality Evaluation of VGI linear datasets. Phd Thesis. London, United Kingdom: UCL (University College London).

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: 477-498. doi: 10.1111/j.1467-9671.2012.01304.x

Lantmäteriet. 2014a. Fastighetskartan – Produktbeskrivning (in Swedish). Retrieved 01 April 2014, from http://lantmateriet.se/Global/Kartor%20och%20geografisk%20information/Kartor/produktbeskrivningar/fastshmi.pdf.

Lantmäteriet. 2014b. Tätortskartan – Produktbeskrivning (in Swedish). Retrieved 01 April 2014, from http://lantmateriet.se/Global/Kartor%20och%20geografisk%20information/Kartor/produktbeskrivningar/tatoshmi.pdf.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2005. *Geographic Information Systems and Science*, 2nd Edition. ISBN: 978-0-470-87002-0. New York: Wiley.

Ludwig, I., A. Voss, and M. Krause-Traudes. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. 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. doi 10.1007/978-3-642-19789-5_4

McMaster, R. 1986. A statistical analysis of mathematical measures for linear simplification. *The American Cartographer* 23: 103–17.

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

Navarro, G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* 33: 31-88. doi: 10.1145/375360.375365

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

Neis, P.,and D. Zielstra. 2014. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet* 6: 76-106. doi: 10.3390/fi6010076

OSM Developer. 2014. Coverage progress - Europa on OSM. Retrieved 17 Mars 2014 , from http://random.dev.openstreetmap.org/progress/?region=europe.

OSM Foundatation. 2014a. FAQ - What is OpenStreetMap and what is the OpenStreetMap Foundation. Retrieved 20 February 2014, from http://wiki.osmfoundation.org/wiki/FAQ#What_is_OpenStreetMap.3F_And_what_is_the_OpenStreetMap_Foundation.3F.

OSM Foundatation. 2014b. License. Retrieved 24 February 2014, from http://wiki.osmfoundation.org/wiki/License.

Open Source Initiative. 2014. Open Source Initiative homepage. Retrieved 23 Maj 2014, from http://opensource.org.

OpenJump. 2014. OPENJUMP GIS. Retrieved 14 Februray 2014, from http://www.openjump.org/.

OpenStreetMap. 2014a. Main page. Retrieved 24 April 2014, from http://www.openstreetmap.org.

OpenStreetMap. 2014b. Planet OSM. Retrieved 26 April 2014, from http://www.openstreetmap.org.

OSM Wiki. 2014a. Stats Rigistered users. Retrieved 17 Mars 2014, from http://wiki.openstreetmap.org/wiki/Stats#Registered_users.

OSM Wiki. 2014b. Yahoo. Retrieved 17 Mars 2014, from http://wiki.openstreetmap.org/wiki/Yahoo.

OSM Wiki. 2014c. Bing. Retrieved 17 Mars 2014, from http://wiki.openstreetmap.org/wiki/Bing.

OSM Wiki. 2014d.Field Papers. Retrieved 17 Mars 2014, from http://wiki.openstreetmap.org/wiki/Field_Papers#How_to_Upload_a_Snapshot.

OSM Wiki. 2014e. Editor. Retrieved 17 Mars 2014, from http://wiki.openstreetmap.org/wiki/Editor.

OSM Wiki. 2014f. Map features. Retrieved 24 February 2014, from http://wiki.openstreetmap.org/wiki/Map_Features.

OSM Wiki. 2014g. Open Data License -Substantial – Guideline. Retrieved 24 February 2014, from http://wiki.openstreetmap.org/wiki/Open_Data_License/Substantial_-_Guideline.

OSM Wiki. 2014h. Open Database License. Retrieved 24 February 2014, from http://wiki.openstreetmap.org/wiki/Open_Database_License.

OSM Wiki. 2014i. Planet.osm. Retrieved 26 February 2014, from http://wiki.openstreetmap.org/wiki/Planet.osm.

OSM Wiki. 2014j. Data Working Group. Retrieved 28 April, from http://wiki.openstreetmap.org/wiki/Data_working_group

OSM Wiki. 2014k. XAPI. Retrieved 28 April, from http://wiki.openstreetmap.org/wiki/XAPI

OSM Wiki. 2014l. Overpass API. Retrieved 28 April, from http://wiki.openstreetmap.org/wiki/Overpass_API

OSM Wiki. 2014m. Quality Assurance OSM. Retrieved 19 May, from http://wiki.openstreetmap.org/wiki/Quality_Assurance.

Python. 2014. Python homepage. Retrieved 7 May 2014, from https://www.python.org/.

QGIS. 2014. QGIS Homepage. Retrieved 7 May 2014, from http://qgis.org/en/site/.

Ramm, F., J. Topf, and S. Chilton. 2011. OpenStreetMap - Using and enhancing the free map of the world. Cambridge: UTI Cambridge.

Raymond, E.S. 1999. Cathedral and the Bazaar. *Knowledge, Technology & Policy* 12: 23-49. doi: 10.1007/s12130-999-1026-0

Serva, M., and F. Petroni. 2008. *Indo-European languages tree by Levenshtein distance. Europhysics Letters* 81: 68005. doi: 10.1209/0295-5075/81/68005

Sourcefrog. 2014. OpenJUMP Plugins – Roadmatcher. Retrieved 27 April 2014, from http://sourceforge.net/projects/jump-pilot/files/OpenJUMP_plugins/More%20Plugins/Roadmatcher%201.4%20for%20OJ/.

Stigmar, H., 2005. Matching Route Data and Topographic Data in a Real-Time Environment. In *ScanGIS'2005 - Proceedings of the 10th Scandinavian Research Conference on Geographical Information Sciences, e*d. Hauska, H. and H. Tveite. pp. 89-107. Stockholm, Sweden. doi=10.1.1.131.1077

Van Oort, P.V. 2006. Spatial data quality: from description to application. PhD Thesis. Wageningen, Netherlands: Wageningen University.

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

Vivid Solutions. 2014b. RoadMatcher . Retrieved 14 Februray 2014, from http://www.vividsolutions.com/products.asp?catg=spaapp&code=roadmatcher#down_1_4.

Walter, V. and D. Fritsch. 1999. Matching spatial data: a statistical approach. *Journal of Geographical Information Science* 13: 445–473. doi: 10.1080/136588199241157

Wasström, C., G. Lönnberg, and L. Harrie. 2013. Kvalitetsaspekter (in Swedish). In: *Geografisk informationsbehandling – teori, metoder och tillämpningar* (in Swedish), 6th edition, ed. L. HARRIE. pp. 263-282. Lund: Studentlitteratur.

White, E.R.1985. Assessment of line-generalization algorithm using characteristic points. *The American Cartographer* 12: 17-27. doi**:** 10.1559/152304085783914703

Zielstra, D., and A. Zipf. 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *AGILE 2010: 13th AGILE International Conference on Geographic Information Science*, ed. Painho. M., M. Y. Santos, and H. Pundt. Springer Verlag. Guimaraes, Portugal.

Appendix 1 - Matching code

The developed matching algorithm presented in this study can be downloaded from:

- https://github.com/JulianWill/Linear_Feature_matching.git

To access the latest version or to contribute to the development of the matching algorithm visit this web page.

For question contact the author: julianwill88@web.de

Observe that the matching algorithm might change due to future development and so differ from the describe methodology in this study.

Seminar Series

Institutionen för naturgeografi och ekosystemvetenskap, Lunds Universitet.

Student examensarbete (Seminarieuppsatser). Uppsatserna finns tillgängliga på institutionens geobibliotek, Sölvegatan 12, 223 62 LUND. Serien startade 1985. Hela listan och själva uppsatserna är även tillgängliga på LUP student papers (www.nateko.lu.se/masterthesis) och via Geobiblioteket (www.geobib.lu.se).

The student thesis reports are available at the Geo-Library, Department of Physical Geography and Ecosystem Science, University of Lund, Sölvegatan 12, S-223 62 Lund, Sweden. Report series started 1985. The complete list and electronic versions are also electronic available at the LUP student papers (www.nateko.lu.se/masterthesis) and through the Geo-library (www.geobib.lu.se)

288 Emma Li Johansson (2013) A multi-scale analysis of biofuel-related land acquisitions in Tanzania - with focus on Sweden as an investor

289 Dipa Paul Chowdhury (2013) Centennial and Millennial climate-carbon cycle feedback analysis for future anthropogenic climate change

290 Zhiyong Qi (2013) Geovisualization using HTML5 - A case study to improve animations of historical geographic data

291 Boyi Jiang (2013) GIS-based time series study of soil erosion risk using the Revised Universal Soil Loss Equation (RUSLE) model in a micro-catchment on Mount Elgon, Uganda

292 Sabina Berntsson & Josefin Winberg (2013) The influence of water availability on land cover and tree functionality in a small-holder farming system. A minor field study in Trans Nzoia County, NW Kenya

293 Camilla Blixt (2013) Vattenkvalitet - En fältstudie av skånska Säbybäcken

294 Mattias Spångmyr (2014) Development of an Open-Source Mobile Application for Emergency Data Collection

295 Hammad Javid (2013) Snowmelt and Runoff Assessment of Talas River Basin Using Remote Sensing Approach

296 Kirstine Skov (2014) Spatiotemporal variability in methane emission from an Arctic fen over a growing season – dynamics and driving factors

297 Sandra Persson (2014) Estimating leaf area index from satellite data in deciduous forests of southern Sweden

298 Ludvig Forslund (2014) Using digital repeat photography for monitoring the regrowth of a clear-cut area

299 Julia Jacobsson (2014) The Suitability of Using Landsat TM-5 Images for Estimating Chromophoric Dissolved Organic Matter in Subarctic Lakes

300 Johan Westin (2014) Remote sensing of deforestation along the trans-Amazonian highway

301 Sean Demet (2014) Modeling the evolution of wildfire: an analysis of short term wildfire events and their relationship to meteorological variables

302 Madelene Holmblad (2014). How does urban discharge affect a lake in a recreational area in central Sweden? – A comparison of metals in the sediments of three similar lakes

303 Sohidul Islam (2014) The effect of the freshwater-sea transition on short-term dissolved organic carbon bio-reactivity: the case of Baltic Sea river mouths

304 Mozafar Veysipanah (2014) Polynomial trends of vegetation phenology in Sahelian to equatorial Africa using remotely sensed time series from 1983 to 2005

305 Natalia Kelbus (2014) Is there new particle formation in the marine boundary layer of the North Sea?

306 Zhanzhang Cai (2014) Modelling methane emissions from Arctic tundra wetlands: effects of fractional wetland maps

307 Erica Perming (2014) Paddy and banana cultivation in Sri Lanka - A study analysing the farmers' constraints in agriculture with focus on Sooriyawewa D.S. division

308 Nazar Jameel Khalid (2014) Urban Heat Island in Erbil City.

309 Jessica, Ahlgren & Sophie Rudbäck (2014) The development of GIS-usage in developed and undeveloped countries during 2005-2014: Tendencies, problems and limitations

310 Jenny Ahlstrand (2014) En jämförelse av två riskkarteringar av fosforförlust från jordbruksmark – Utförda med Ekologgruppens enkla verktyg och erosionsmodellen USPED

311 William Walker (2014) Planning Green Infrastructure Using Habitat Modelling. A Case Study of the Common Toad in Lomma Municipality

312 Christiana Marie Walcher (2014) Effects of methane and coastal erosion on subsea-permafrost and emissions

313 Anette Fast (2014) Konsekvenser av stigande havsnivå för ett kustsamhälle- en fallstudie av VA systemet i Beddingestrand

314 Maja Jensen (2014) Stubbrytningens klimatpåverkan. En studie av stubbrytningens kortsiktiga effekter på koldioxidbalansen i boreal barrskog

315 Emelie Norhagen (2014) Växters fenologiska svar på ett förändrat klimat - modellering av knoppsprickning för hägg, björk och asp i Skåne

316 Liisi Nõgu (2014) The effects of site preparation on carbon fluxes at two clear-cuts in southern Sweden

317 Julian Will (2014) Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network - A case study in Göteborg, Sweden