MSc thesis in Geomatics

Quality assessment and object matching of OpenStreetMap in combination with the Dutch topographic map TOP10NL

Prajnaparamita Bhattacharya

June, 2012

TU Delft Delft University of Technology

Department of GIS Technology, OTB Research Institute for the Built Environment

QUALITY ASSESSMENT AND OBJECT MATCHING OF OPENSTREETMAP IN COMBINATION WITH THE DUTCH TOPOGRAPHIC MAP TOP10NL

A thesis submitted to the Delft University Technology in partial fulfillment of the requirements for the degree of

Master of Science in Geomatics

by

Prajnaparamita Bhattacharya

June 2012

Supervisors: Prof. dr. ir. Peter van Oosterom
Dr. Hugo Ledoux
Co-reader: Dr. ir. Tjeu Lemmens

Prajnaparamita Bhattacharya: *Quality assessment and object matching of OpenStreetMap in combination with the Dutch topographic map TOP10NL,* A thesis submitted to the Delft University Technology in partial fulfillment of the requirements for the degree of, Master of Science in Geomatics, June 2012

LOCATION:

Department of GIS Technology
OTB Research Institute for the Built Environment
Delft University of Technology
The Netherlands

FINAL PRESENTATION:
22 June 2012 at 14:00

EXAMINATION COMMITTEE

Prof. dr. ir. Peter van Oosterom
Ir. Edward Verbree
Dr. Hugo Ledoux
Dr. ir. Tjeu Lemmens

COVER ILLUSTRATION

The cover shows a overlaid image of TOP10NL ans OSM road network of the city Delft.

ABSTRACT

The possibility of an automatic object matching process in combination with a VGI (Volunteered Geographical Information) type of map has been explored in this research. TOP10NL, the topographic map of the Netherlands and OpenStreetMap (OSM), the VGI type of map have been chosen for this purpose. The object matching process in this research became more critical as OSM does not always follow strict rules. Therefore, to overcome the complications special care has been taken into account on the characteristics of these two datasets and have been incorporated in the object matching process. Knowledge from previous studies in this field has been explored and adapted in this study.

As TOP10NL and OSM represent completely different type of data-models, a schema translation had to be done to harmonise them. The average geometric accuracy of TOP10NL is ± 2 meter, whereas the quality of OSM is uncertain over locations. Therefore the quality of the OSM dataset over the study areas was also judged before developing the object matching model by some quality measures of geographic maps. Only the quality measures 'data lineage', 'completeness' and 'positional accuracy' were considered in this study. The data lineage shows that both datasets share quite a large number of objects from the same origin. It was found that the road, railway and building objects were quite complete in the OSM dataset over the study areas whereas the water objects were quite incomplete. Because of the reliable data lineage of the OSM data over the Netherlands the positional accuracy of the OSM data were found to be quite good.

Considering the geographic location and the importance four areas within the Netherlands were chosen for the experiments. The urbanised Delft and Rotterdam, semi-urbanised Dokkum and the rural part of Echt were selected for the experiments. The common features within both TOP10NL and OSM, which are road, building, water and railway, were considered for the experiments. Two different object matching models were developed for line and polygon objects. For both models the decision rule was developed considering the characteristics of the datasets used in this study. For line matching the decision rule was developed to make the list of best matched pairs by accumulating them in a step by step procedure, whereas a straight forward simple rule has been developed for polygon matching process. These models were verified with the different object class data of the different study areas. The accuracy was high and for more than 90% instances the correct match was found.

In the road matching process it was possible to insert the 'street names' from the OSM database as an attribute in the road network database of TOP10NL. A number of matching tables have been formed for different cardinality of matched objects which may be useful in the future in an automatic updating process of TOP10NL and OSM.

ACKNOWLEDGMENTS

I am pleased to get this opportunity to express my gratitude to everyone who helped me directly or indirectly in carrying out my thesis work.

First of all I express my utmost gratitude to my supervisor Dr Hugo Ledoux for his continuous support, valuable comments and discussion to make this study possible. Without his support it would have been impossible to finish this study.

I am very much thankful to my professor Prof Dr Peter van Oosterom for allowing me to carry out my research in his department and for his valuable comments and suggestions to improve my thesis. I am also grateful to Dr Tjeu Lemmens for agreeing to be the co-reader of my thesis. His valuable comments also helped a lot to improve the quality of the thesis.

My special thanks goes to Drs Wilko Quak for allowing me to use his code to start up my project. I am thankful to Mrs. Elfriede Fendel for her help in all practical issues related to carrying out the research. I also acknowledge Ir. Edward Verbree for his support I received.

I am thankful to all my classmates, specially Filip Biljecki and Ken Arroyo Ohori for their friendship, support and inspiration in carrying out my study.

Without blessings of my parents I could not have been able to finish my thesis. My thanks and regards to them also.

Finally, last but not least I am very much thankful to my husband Biswa and my two lovely daughters Mohar and Rai for their patience and support during this period without which I would not have been able to complete my study.

I am fortunate to have all of you in my life.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

TOP10NL, the object-oriented digital topographic map of the Netherlands, which is developed by the Dutch Cadastre [52], has been widely used within the geo-information community of this country. Since 2006 TOP10NL is updated biennially [52]. There were some restrictions in the use of TOP10NL, which has recently been removed by the Dutch Cadastre. The Dutch Cadastre has made an agreement with the Dutch Ministry of Infrastructure and Environment to make the *Key Register Topography* publicly available with effect from 1st January, 2012. This decision was based on the consideration that geographic information is an important foundation for innovative applications. As a result the TOP10NL dataset is now freely available for viewing and downloading in order to be re-used [39].

With this advancement the number of potential users of TOP10NL is anticipated to increase. This anticipation is based on the premise that the demand and use of geographic information in a wide range of areas is increasing exponentially with the advancement of technology and social developments. Besides the institutional purposes, people who are interested in geographic information, will also get chances to use TOP10NL for their own purpose.

Another freely available geographic map 'OpenStreetMap' [16] is getting very popular nowadays. OpenStreetMap (OSM) is a VGI ("Volunteered Geographical Information") type of map [29] [28] [30]. The OpenStreetMap project was started within the framework of Web Mapping 2.0 applications by Steve Coast in England in 2004 with an idea similar to the development of Wikipedia [21], [53] to provide everybody free and legal access to using, creating and developing maps and geographic facts. Since the beginning the popularity and growth of OSM in a short timespan has increased remarkably. People volunteer to contribute location data at free of cost via the project website [16]. Everybody has the freedom to use and correct the OSM data but not the right to sell any product out of it under the license "Creative Commons Attribution-ShareAlike 2.0 license" [44] [27]. The continuous data gathering and updating by volunteers keeps the OSM up to date.

Therefore, currently two popular geographic maps of the Netherlands are available. TOP10NL and OSM both are easily available free of cost, however, they do have some advantages and disadvantages. Users are aware of the quality of the TOP10NL data because it has

(a) OSM

(b) TOP10NL

Figure 1: Visualisation of OSM and TOP10NL dataset of the same area of Delft

been produced by the Dutch Cadastre and the average geometric accuracy of TOP10NL, based on the product report of the first version [3] is ± 2 meter. But the quality of the OSM data is uncertain since it is produced and maintained by volunteers. As volunteers keep the OSM data up to date by continuously providing information of new changes in the real world, users of OSM receive more frequently updated information. On the contrary, TOP10NL users need to wait till the next version is available, which happens once in every two years. Figure 1 is an example of OSM and TOP10NL dataset from the same area of the city of Delft. For better visual comparison only the lines and polygons from both datasets are shown here. It gives an impression on the similarity of both datasets. The white region of the OSM dataset in Figure 1a depicts some incompleteness of the OSM data in that region.

Based on the above mentioned facts, investigating the possibility of using the OSM data to the benefit of the TOP10NL users or, using the TOP10NL data to the benefit of the OSM users seems pertinent and is the motivation of this research. This is only possible by successful combination of TOP10NL and OSM datasets. To investigate that this research focuses on to explore the following research question:

Is an automatic object matching of OSM in combination with TOP10NL possible? If yes, then to what extent this can be performed?

A successful automatic object matching process can be extended to an automatic map updating process as a future follow up work. Both

TOP10NL and OSM users can be benefited by getting more frequent updates if an automatic map updating process is implemented.

1.2 OBJECTIVES

The research has been carried out in three steps to answer the main research question. They are explained here:

- STEP I: *Schema matching* - The TOP10NL and OSM data models represent different model types. The heterogeneity of these two models is caused by the reasons that the data from both sources are available in different formats, different file types and different coordinate reference systems. Therefore, a schema matching was performed to harmonise TOP10NL and OSM data models as a preparatory stage of the object matching process.
- STEP II: *Quality assessment of OSM* - The quality of the OSM dataset over the study areas was judged. The quality of the OSM dataset should be compatible with the TOP10NL data quality for a useful object matching process.
- STEP III: *Object matching of TOP10NL and OSM* - The object matching of both line and polygon features of TOP10NL and OSM was the final stage of this research. The different characteristics of these two data sets were investigated and incorporated in building up decision rules in the object matching process.

1.3 SCOPE

Based on the previous studies a work-flow has been developed for reaching the objective. The work-flow has been tested with a small dataset to make this experiment simple. However, the methodology is suitable for a bigger dataset as well. An algorithm for the object matching process has been presented and applied to the data of TOP10NL and OSM datasets for the pilot areas. Instead of considering the dataset of the entire Netherlands, only four places were chosen to carry out this research. The emphasis was given on the geographic location and the level of urbanisation in the selection of the areas. The city of Delft, the central part of Rotterdam, semi-urbanised Dokkum and the rural part of Echt have been chosen as the pilot areas to get an overall impression of the OSM data over the Netherlands. Only a few number of feature classes were considered for experiments: they are 'road', 'railway', 'building' and 'water'.

Some guidelines were followed for harmonising TOP10NL and OSM which are given by Manoah *et al.*, Horak *et al.*[47] [38]. First, the heterogeneity between the two models was identified and afterwards a schema translation was performed to harmonise these two data models [62] [69]. The OSM dataset has been transformed to the ESRI shapefile format to match the TOP10NL dataset.

Hakley [34] provided some pioneering guidelines to check the quality of the OSM data. He studied the quality of the OSM data in the London areas. Followed by him Girres *et al.*[27] did the quality assessment of the OSM dataset over the region of Hendaye in France. More research on this topic has been carried out for different regions of UK, Germany and Ireland [40] [6] [61] [64] [12]. Based on these research studies a methodology has been developed and applied. In the quality assessment of the OSM data only three quality aspects were checked, which were considered as the most relevant for this study. The history of the data, the data coverage over the study areas and the positional accuracy were taken into account. The results were compared with the previous studies on the quality assessment of the OSM data over different countries.

Several researchers in the past have explored the possibility of partial or complete automation of the object matching process. Most of the research are found on to detect the mismatch between the objects comparing a vector map with recent satellite, aerial or remote sensing imageries [22] [41] [23] [55] [31] [72], whereas some explored matching two vector maps of different versions [32] [9] [58] [57]. There exist some advantages and disadvantages in both techniques. The image based change detection process needs recent high resolution satellite, remote sensing or aerial images which are not always available easily to the wide range of geo-information users. Moreover both TOP10NL and OSM data are available easily at free of cost and both in the vector format. As the motivation of this research is to explore the possibility of using data of one these maps for the benefit of the users of other, the latter technique (comparing two vector maps) was followed in this research. The two vector maps TOP10NL and OSM have been compared for object matching process. In this way the object matching process becomes cost-effective by avoiding the difficulties of getting satellite or aerial imageries.

Two different models were built for the line and polygon object matching purpose. For the line matching process several matching criteria were explored from previous studies [68], [9], [58], [59]. Anand *et al.* [5] used a technique for matching the line objects of the OSM dataset with Ordnance Survey data of UK which is a quite relevant research for this study. Some quality assessment techniques were also used for the object matching process, e.g., the technique described by Girres *et al.* [27] for quality assessment of OSM polygon was found to be a suitable matching criteria for polygon matching process. Finally out of many matching criteria only the relevant ones were selected to fit in the built up algorithm. As an outcome of the object matching process a number of matching tables were formulated, which are classified as 'one-to-one', 'one-to-zero', 'one-to-many', 'many-to-one', 'many-to-many' and 'zero-to-one'.

Due to the time constraints, resolving the mismatches was not done in this study. However, some ideas regarding their implementation have been formulated and elaborated at the end of this thesis in the recommendation. An overview of the work-flow of this study has been shown in Figure 2.

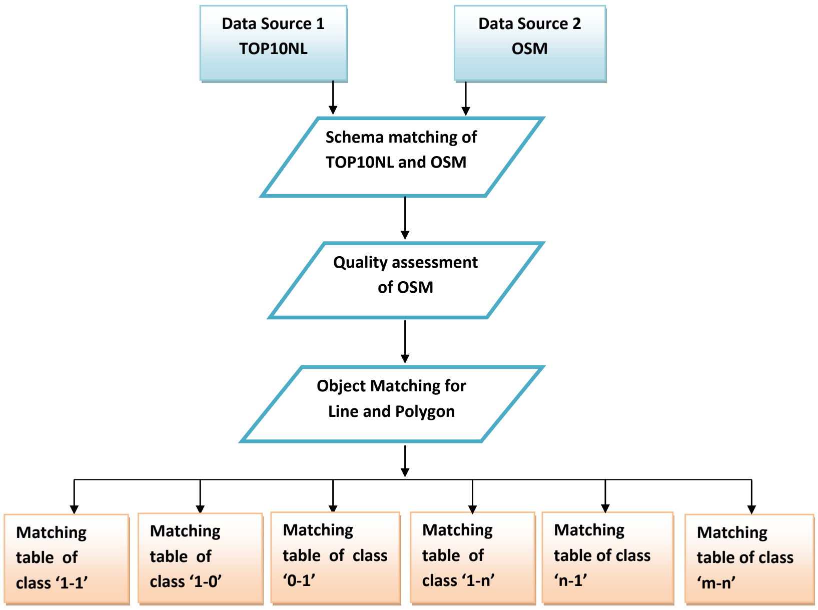


Figure 2: Workflow of this study

1.4 STRUCTURE OVERVIEW

The structure of the thesis is stated below:

- Chapter 2 *Background Information* gives the description of TOP10NL and OSM data models in Sections 2.1 and 2.2. The heterogeneity of these two models are identified. To resolve the heterogeneity between two datasets a schema translation is presented in Section 2.3. The dissimilarities of the data representation of identical objects from different data models are described in Section 2.4. The pilot areas of this study are presented in Section 2.5.

- Chapter 3 *Quality Assessment of OSM* describes the quality assessment of the OSM data over the study areas. The quality aspects 'lineage', 'completeness' and 'positional accuracy' of the linear and polygonal objects were tested and a conclusion was made on the basis of the quality checking results.

- Chapter 4 *Object matching* describes some matching criteria for line and polygon matching found in the literature. The matching criteria were examined and based on their relevance in the study the best matching criteria was selected. Then a decision rule has been developed and implemented for an automatic map matching process for linear and polygonal objects separately. The results are discussed afterwards.
- Chapter 5 *Conclusion and recommendations* summarises the conclusions of the research and present some recommendation for future works.
- Appendix A *Tables and Diagrams* provides some figures and tables.

2 BACKGROUND INFORMATION

To check the usefulness of OSM to the benefit of TOP10NL, both models need to be brought into identical schema because they represent completely different data models. The dissimilarities between the TOP10NL and OSM are caused by the different acquisition methods, different updating period, availability in different data format etc. A schema translation was therefore needed to harmonise these two data models. The purpose of this chapter is to give an overall idea about the differences between the two data models TOP10NL and OSM and how they were harmonised.

This chapter starts with a brief description of the two data models TOP10NL (in Section 2.1) and OSM (in Section 2.2). Then a schema matching process of TOP10NL and OSM has been described in Section 2.3. Some examples of identified dissimilarities of presence of an identical object in both datasets have been shown in section 2.4. Finally the pilot areas have been discussed in Section 2.5.

2.1 TOP10NL

TOP10NL is the successor of previous Top10vector and is a digital topographic file which can be used on a scale level between 1:5,000 and 1:25,000 [52]. The TOP10NL data model contains a collection of topographical base objects, related to a reproduction scale of 1:10,000, which have been included as object classes [52]. In TOP10NL each individual object on land e.g. a building, a plot of farmland, part of a road can be distinguished individually by their unique identification code. Moreover, the object class can also be depicted. A geographical object has certain geometry (a point, line or polygon) and is further characterised by its attributes. In the main structure of TOP10NL, every geographical object is assigned to a specific object class. The current set of object classes consists of 'Road section', 'Railway section', 'Water section', 'Building', 'Land', 'Planimetric feature', 'Relief', 'Registrational area', 'Geographic area' and 'Functional area' [43] [52] [7]. The description of different object classes of TOP10NL is described in Table 1 on page 8. The area of the object classes road, water, land, buildings, functional and registrational area can also be represented in combination for visualisation purpose [52]. In visualisation the sequence of object class layers can play a role for covering the surfaces.

Name	Description	Presence in OSM
FUNCTIONAL_GEBIED_PUNT	*Functional areas, such as camping and sports*	Yes
GEBOUW_VLAK	*The building of any type (high, low etc.) and any purpose (municipality, school, religious building etc.) represented as* polygons	Yes
GEOGRAFISCH_GEBIED_PUNT	*Geographical areas, such as* placesandregions	Yes
INRICHTINGSELEMENT_LIJN	*Row if trees, soundproof wall, railing etc.*	No
INRICHTINGSELEMENT_PUNT	*RD point, tree, tower etc.*	No
RELIEF_LIJN	*Quay, slopes etc.*	No
SPOORBAANDEEL_LIJN	*The tram and train line including the number of tracks presents.*	Yes
SPOORBAANDEEL_PUNT	*Railway lines crossing*	Yes
TERREIN_VLAK	*Orchard, building area, forest, grassland etc.*	Yes
WATERDEEL_LIJN	*All waterline, ditch, channel etc.*	Yes
WATERDEEL_VLAK	*The area (polygon) with water like lake, pond etc.*	Yes
WEGDEEL_LIJN	*The footpaths, bicycle path which are less than 2meter in breadth. And the central lines of the main roads between two junctions.*	Yes
WEGDEEL_PUNT	*All crossings of the roads.*	No
WEGDEEL_VLAK	*The wide roads (highway, motorway), local roads and the traffic free paved area for the pedestrians.*	No

Table 1: Description of TOP10NL object classes

2.1.1 *TOP10NL data model*

TOP10NL datasets are available in GML (Geography Markup Language) and the ESRI shapefile format [7] [52] and maintain the Dutch standard NEN3610 2005 based on ISO and OGC standards. A complete UML (Unified Modeling Language) diagram is shown in Figure 45 (on page 76) in Appendix. The different data models are described below:

- GML is an XML (Extensible Markup Language) encoding for geographic data and developed within the Open Geospatial Consortium, in collaboration with ISO. Similar to XML, GML is also readable by both humans and machines and it is internationally accepted structured information. GML offers the possibility to describe different geographic characteristics such as geometry, topology, coordinate reference systems etc. In GML, the world is modelled using concepts such as features and properties.

- An ESRI shapefile stores the geographical objects as 'simple features' and the attribute information for the features in a data set [25] [26]. The geometry for a feature is stored as a shape comprising of set of vector coordinates. Shapefiles can support point, line, and area features. Area features are represented as closed loops. Attributes are held in a dBASE format file. Each attribute record has a one-to-one relationship with the associated shape record [25]. An ESRI shapefile consists of a main file, an index file, and a dBASE table. They all have the same prefix but different suffices. The suffix for the main file, the index file and the dBASE table are .shp, .shx and .dbf respectively. The main file is a direct access in which each record describes a shape with a list of its vertices. In the index file each record contains the offset of the corresponding main file record from the beginning of the main file. The dBASE table contains feature attributes with one record per feature. The one-to-one relationship between geometry and attributes is based on the record number. Attribute records in the dBASE file must be in the same order of records as in the main file [25]. Table 18 in Appendix on page 77 describes the possible types of shapes. Different object classes of TOP10NL are delivered as different shapefiles. The ESRI shapefile formats of TOP10NL has been used in this study.

2.1.2 *Updating process of TOP10NL*

In the updating process of TOP10NL, aerial images are taken throughout The Netherlands. After comparing the new images with the previous images the differences are identified manually and the updates are incorporated. Different operations such as aggregation of two objects, splitting one object into two objects, changing the shapes of an object, etc. are performed during updating. During updating some predefined rules are followed [8]. The rules have been defined to judge to what extent the object modifications lead up to a new version date. Depending upon the kind of modification such as aggregation, splitting certain rules are applied [8]. A diagram of the production process of TOP10NL is shown in Figure 3.

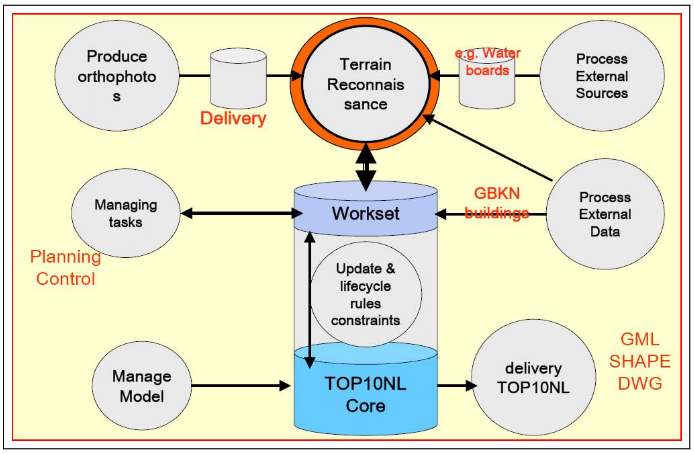


Figure 3: Diagram of the production process of TOP10NL [8]

2.2 OPENSTREETMAP (OSM)

The OSM datamodel is completely different from TOP10NL in all aspects. People volunteer to contribute the location data at free of cost via the OSM website (www.openstreetmap.org). The data is collected from a variety of sources such as recordings from GPS devices, from free satellite imageries or from knowledge about an area. The OSM data model is explained below. The collected information is arranged and stored in a central database and distributed in different digital formats through the Web [21].

2.2.1 *OSM data model*

The OSM follows a different model than the TOP10NL data model in the management of Internet resources. The OSM data model uses the Resource Description Framework (RDF) defined by Manola [27] [48]. In RDF, information is modelled as a tuple ('resource', 'property', 'value'). In case of OSM, the tuple are mentioned as

- the 'resource' or the subject is a geometric primitive with coordinates,
- the 'property' or the predicate is a tag name,
- the 'value' is the tag information.

For example, the tuple follows the form: (way; highway; cycleway). This RDF information is represented in the XML syntax, as shown in Figure 4 (on page 11). The advantage of RDF from a conventional geographic database structure is that RDF allows adding objects which do not belong to a predefined class but certain properties of them are known. However, in the case of OSM, specifications are not as structured as in a RDF model should be.


```
<way id="44431847" user="polderrunner" uid="36080" visible="true" version="1"
changeset="3123085" timestamp="2009-11-15T15:42:43Z">
 <nd ref="44813347"/>
 <nd ref="44812599"/>
 <tag k="highway" v="cycleway"/>
</way>

.........

<node id="44813347" lat="52.0035519" lon="4.3362125" user="AND_fixbot" uid="211771"
visible="true" version="3" changeset="3452869" timestamp="2009-12-26T09:34:39Z"/>

.......
```


Figure 4: A sample of XML format of an OSM file. The first part is the record which explains a cycleway and the second part explains the geographical position

The OSM data is uploaded into the OSM server in terms of tagged geometric primitives. Geometric data primitives in OSM are either nodes, ways or relations. They are described below as per [21]:

- A 'node' is the basic element, the building block, of the OSM data mode. Nodes consist of latitude and longitude of a single geo-spatial point. Nodes can be presented as a group or standalone. A collection of nodes define a way, whereas a single node itself represent any type of point object, such as a building, bus stop or all kinds of other points of interest (POI).The structure of a OSM node is explained in Table 19 on page 78.

- A 'way' is an ordered interconnection of at least 2 nodes that describe a linear feature such as a street, footpath, railway line, river, fence, power line, area or building outline. The polygons are represented as closed ways in which the first and last nodes are identical. Polygons or areas are not separate data primitives, but simply closed ways that are tagged as areas. The structure of a OSM way is explained in Table 20 on page 78.

- A 'relation' is a group of primitives with associated roles. A node or a way might not be a member of any relations. Roles are attributes which can be added to a way, node or relation belonging to a relation. A role consists of a single simple string, which can be empty. Roles describe the part that a particular feature plays within a relation. For example, role=inner and role=outer are used to specify whether a way forms the inner or outer part of a polygon in the multipolygon relation. The structure of a OSM relation is explained in Table 21 on page 78.

All map features are referred to by some predefined tags; however, OSM does not have any content restrictions on tags that can be assigned to OSM elements. Any tag can be used as long as the values are verifiable. However, there exists a core recommended feature set and corresponding tags [21] .

The tags assigned to every OSM element in the OSM database explain the feature by key and value pairs. For example, the 'key=highway' explains the whole category of roads whereas, corresponding 'value= cycleway' categorise the key. "Highway" is a primary tag in this case. Table 2 on page 12 represents an example of different values of "highway" with descriptions. Moreover, the OSM structure is not straight forward to explain the geometry of the object class. For example, in OSM tags the key 'highway' may represent points (e.g. junction), lines (e.g. roads), or surfaces (e.g. parking place). Unlike TOP10NL, OSM does not contain unique identifier for all real-worlds objects.

Key	Value	Comment
Highway	Motorway	*Restricted access major divided highway, normally with 2 or more running lanes plus emergency hard shoulder <tag k="highway" v="motorway">*
Highway	Residential	*Roads accessing or around residential areas but which are not a classified or unclassified highway <tag k="highway" v="residential">*
Highway	Cycleway	*For designated cycleways <tag k="highway" v="cycleway">*
Highway	Motorway_junction	*Indicates a junction (UK) or exit (US) should be set to the exit number or junction identifier <tag k="highway" v="motorway_junction">*
Highway	Road	*A road of unknown classification <tag k="highway" v="road">*

Table 2: Examples of tag value with key in OSM [21]

2.2.2 *Updating process of OSM*

For the updating purpose the OSM website offers an online Flash-based editor, Potlatch, by which the users can add, update, or delete geographical features through a relatively easy-to-use interface. Potlatch also lets users upload and integrate GPX tracks recorded from hand-held GPS units. The Java OSM Editor (JOSM) is another editing suit with an interface to traditional GIS packages. Satellite imagery and out-of-copyright maps are the other important data sources from which users trace features and those are also integrated into the mapping interface. Since the end of 2006, Yahoo granted OSM the right to use its satellite imageries [36]. Microsoft also agreed to provide their Bing Aerial imagery for use in the OSM project [4]. In 2007 the Dutch mapping company AND (Automotive Navigation Data) donated their entire street map of The Netherlands to OSM [15]. The data from TOP10NL may also be used in updating OSM as currently TOP10NL is available at free of cost.

2.3 SCHEMA MATCHING OF OSM AND TOP10NL

To resolve the heterogeneity of TOP10NL and OSM and to make them harmonised a schema translation was done. The heterogeneity between the models is caused due to the difference in acquisition methods using different sensors, difference in data storage in different data-models and presentation in different file types [62]. Some schema matching techniques referred by different researchers were explored. Manoah *et al.* [47] gave some guidelines for schema matching of geographic data. Horak *et al.* [38] mentioned that the data harmonization should start with a description of data sources and mapping of differences among heterogeneous source and target data sets. Therefore, to start this process the heterogeneity of TOP10NL and OSM were explored first. Following the guideline of Goesseln *et al.* [69] different types of heterogeneity have been categorised. They are:

- heterogeneity due to unequal updating periods but stemming from the same dataset
- heterogeneity due to the data acquisition by different operators for the same data-model
- heterogeneity due to the data acquisition from the same sources but stored in different data-models
- heterogeneity due to the acquisition from different sources and stored in different data-models, thematic contents, different updating periods, etc.

However, the cause of the heterogeneity of TOP10NL and OSM does not completely follow one of the above mentioned types. The heterogeneity of TOP10NL and OSM are caused:

- due to the unequal different frequency updating periods,
- due to the data acquisition by different operators,
- due to the data acquisition from different sources and data storage in different data-models,
- due to having different type of files and different reference systems.

As the objective of this research is to compare the objects of the real world present in TOP10NL and OSM, only the heterogeneity in file types and reference systems have been considered to be resolved. The XML format of the OSM data model has been transformed to shapefiles of different object classes, which are equivalent to TOP10NL object classes. The OSM data model has been transformed into shapefiles of different object classes.

As mentioned before the object classes of TOP10NL follow certain strict definitions whereas the object classes of OSM do not follow any strict predefined schema. As there is no restriction, the tagging schema in OSM is increasingly developing into a complex taxonomy of feature classes and objects of real world [36]. So the identification of the map features from the OSM dataset, which are similar object classes in TOP10NL was the preliminary task. This task was performed using thematic semantic mapping. Then the shapefiles of different object classes were formed with the data extracted from the OSM dataset and the coordinate transformation was done afterwards. The schema translation aspects of TOP10NL and OSM are mentioned in Table 3 on page 14.

Harmonising purpose	TOP10NL	OSM	Transformation Goal
Data format	Simple features	RDF	To simple features
File type	ESRI shapefile	XML file	To ESRI shapefile
Reference system	RD_new	WGS84	To RD_new
Object classes	Different shapefiles	Defined in Tags	semantic mapping

Table 3: Schema translation aspects of TOP10NL and OSM

2.3.1 *Semantic mapping of the object classes of TOP10NL and OSM*

By definition, semantic mapping is a strategy that can be used to demonstrate the relationships between ideas in all disciplines. Semantic mappings between the source schema and the mediated schema is one of the main problems in the schema matching process [47]. Therefore semantic mapping was needed to harmonise TOP10NL and OSM. In the present research the task of semantic mapping was to find out and extract the geometric primitive entities from the OSM dataset whose tag contained a word which carried the same meaning of any one of the object classes of TOP10NL.

As this study is aiming to compare OSM and TOP10NL, only some common object classes which are present in both data models were considered. The main four object classes, 'building', 'water', 'road' and 'railway' have been considered for the experiments. Only the object classes in geometric type 'lines' and 'polygons' were tested. The object classes are mentioned in the tags of every OSM element in the OSM database and are assigned by key and value pairs. For example, all types of objects related to road are expressed with "highway" as the 'key' in tag, whereas the type and functionality of the roads are mentioned in 'value', which is shown in Table 2 on page 12. So the class of all geometric primitives with 'highway' in the tag can be considered similar as the object class 'Road section' in TOP10NL because this is a combined object class containing all types of road, parking plot etc. A table of semantic mapping has been shown in Table 4 on page 16. For example, if any OSM feature contains the word 'building' in the values of the 'tag' attribute (in "key" or "value"), it should return as a "building" feature while running the query for searching the object classes. Further division into 'point' and 'polygon' geometry types were done afterwards. In this way the data of different object classes of the OSM dataset were extracted separately following the semantic mapping schema. The whole process was done using Python programming. The steps of this processing is given below.

- Parsing OSM-XML file by XML parser into OSM primitives tables 'nodes', 'ways' and 'relations' separately;
- Identifying them by adding identification number for all 'nodes', 'ways' and 'relations';
- Forming the geometry from the 'ways' and 'nodes' primitives of OSM into 'line' and 'polygon' primitives;
- Classifying the lines and polygons into different object classes by semantic mapping;

2.3.2 *Conversion of OSM data to shapefiles*

After separating the simple features according to the different object classes, separate shapefiles for different object classes were formed. Some attributes were added to all features. The attribute values were

TOP10NL	OSM
GEBOUW_VLAK	*<tag k="building" v="yes">*
	<tag k="building" v="apartment">
	<tag k="building" v="church">
	<tag k="building" v="commercial">
	<tag k="building" v="house">
	<tag k="building" v="industrial">
	<tag k="building" v="school">
WEGDEEL_LIJN	*<tag k="highway" v="motorway">*
	<tag k="highway" v="cycleway">
	<tag k="highway" v="footway">
	<tag k="highway" v="path">
	<tag k="highway" v="pedestrian">
	<tag k="highway" v="primary">
	<tag k="highway" v="residential">
	<tag k="highway" v="road">
	<tag k="highway" v="secondary">
	<tag k="highway" v="unclassified">
WEGDEEL_VLAK	*<tag k="service" v="parking_aisle"> etc*
	<tag k="highway" v="services">
	<tag k="amenity" v="parking_space">
SPOORBAANDEEL$_{L}$IJN	*<tag k="railway" v="rail">*
	<tag k="railway" v="tram">
	<tag k="railway" v="subway">
WATERDEEL_LIJN	*<tag k="waterway" v="canal">*
	<tag k="waterway" v="ditch">
	<tag k="waterway" v="drain">
	<tag k="waterway" v="river">
	<tag k="waterway" v="stream">
WATERDEEL_VLAK	*<tag k="waterway" v="riverbank">*
	<tag k="waterway" v="dock">
	<tag k="waterway" v="dam">
	<tag k="waterway" v="dock">
	<tag k="natural" v="water">

Table 4: A table of semantic mapping of OSM and TOP10NL

extracted from the attribute and tag values of corresponding OSM primitives. Different ESRI shapefiles for different object classes were created afterwards. The OGR simple feature library was used for creating the shapefiles in python. PostGIS was used for storing the data in separate tables of different object classes.

2.3.3 *Coordinate transformation*

The coordinate transformation was done afterwards. The coordinates of the OSM shapefiles were converted from longitude and latitude in WGS84 to local X, Y coordinate in the RD_new system. 'Pyproj', the Python bindings for the PROJ library was used for coordinate transformation. Table 5 shows the specifications of RD_new coordinate system. That can be described as

Ellipsoid	Bessel 1841 $(a = 6377397,155,\ 1/f = 299,1528128)$
Projection center	$lat = 52.1561605\ lon = 5.3876388$
Coordinate shifting	$X : 155000m\ Y : 463000m$

Table 5: The parameters for RD_new coordinate system

The accuracy of the coordinate transformation was checked with the 'Official Software PCTrans' [2]. Randomly a number of points from the OSM dataset was selected from all study locations for testing. The longitudes and latitudes of these points were converted to X,Y of RD_new coordinates in two ways, by PCTrans software and by PROJ using the parameters mentioned in Table 5. The difference between both transformation were calculated. They vary from 8 cm to 61 cm with an average of 40 cm. The difference increases with the distance of the location from the center of the Netherlands in an similar way referred by van Buren *et al.* [65]. The accuracy was also found to be good because it is within the accuracy level tested by van Buren *et al.* [65] over the Netherlands using PCTrans software. In object matching process the accuracy of coordinate transformation should be taken into account in determining the distance thresholds as it might influence the result.

A complete harmosation diagram is shown in Figure 5 (on page 18).

2.4 DATA DISSIMILARITIES FROM TWO SOURCES

By overlaying OSM and TOP10NL dataset an impression of the similarities and dissimilarities can be found. Figure 1 (on page 2) depicts the visualisation of all object classes together from OSM and TOP10NL datasets from the same area of the city of Delft. Apparently they resemble each other, but from a closer view the differences can be better detected. The dissimilarities are caused by the representation of

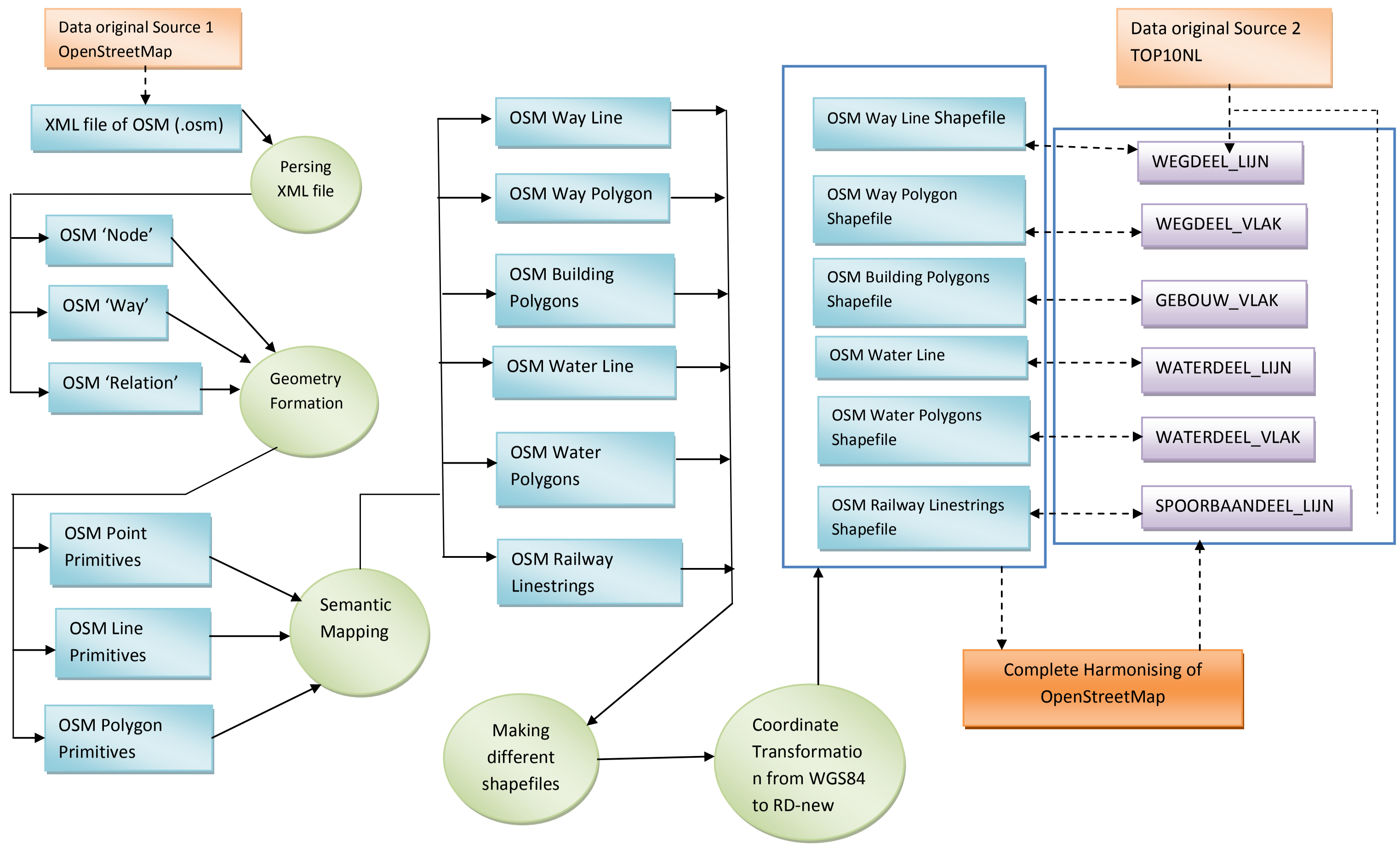


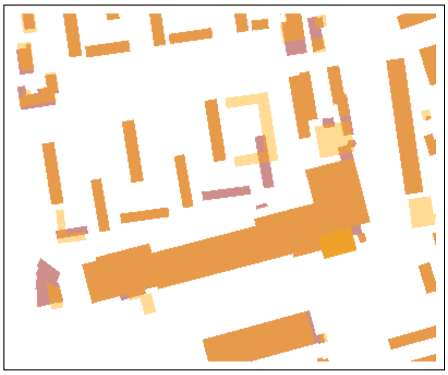
Figure 5: Harmonisation diagram of OSM with TOP10NL

identical objects as different geometry types in different data models or by their presence in both data models. The following subsections describe some different types of dissimilarities.

2.4.1 *Dissimilarities in presence*

Some examples of detected mismatching polygons are shown in Figure 6 (on page 19). In Figure 6a we see that the water polygons from TOP10NL and OSM are different. In Figure 6b we see that in some cases the building polygons from OSM and TOP10NL are quite different.

(a) Differences in shapes of polygons of water bodies(light blue depicts the overlaid area)



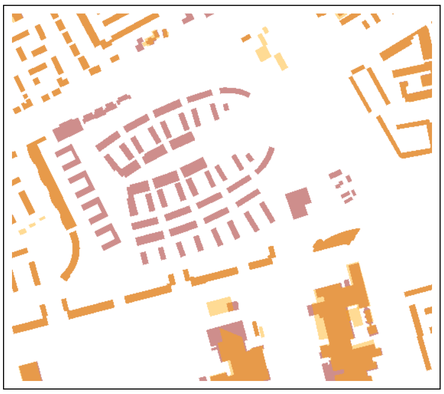
(b) Differences in shapes of polygons of buildings (Yellow depicts OSM, purple depicts TOP10NL and orange depicts overlaid area)

Figure 6: Dissimilarities of Polygons from the TOP10NL and OSM datasets of the same area.

Figure 7a (on page 20) depicts a closer view of the street networks in TOP10NL and OSM. Blue lines represent OSM and yellow lines are for TOP10NL. The figure shows that the street networks from both maps are quite similar though a few new streets are present in OSM whereas a few streets are not recorded in OSM which are present in TOP10NL. Figure 7b (on page 20) represents a situation where the buildings, which are present in the TOP10NL dataset, are not present in the OSM dataset.

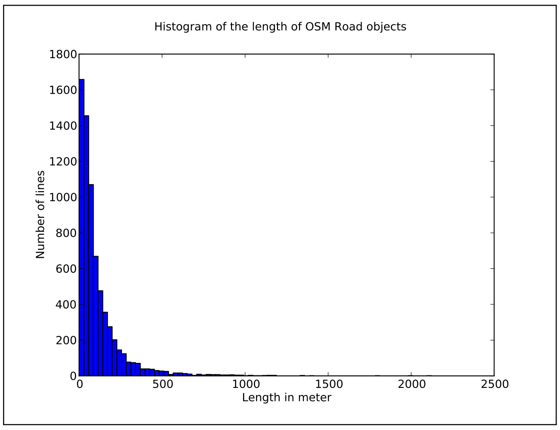
Some more dissimilarities were identified. One of them is the difference between the lengths of the line objects. Most of the TOP10NL road objects terminate in every junction irrespective of small or big junctions, whereas the OSM road objects do not follow such a rule. It was observed that in many cases one OSM road object corresponds to two or many TOP10NL road objects. As a result the number of total road objects in TOP10NL is quite higher than the number of total OSM road objects in the same location. Figure 8 (on page 21) shows the histogram of the length of the road objects from OSM and TOP10NL

(a) Example of overlaying of streets of Delft from OSM and TOP10NL

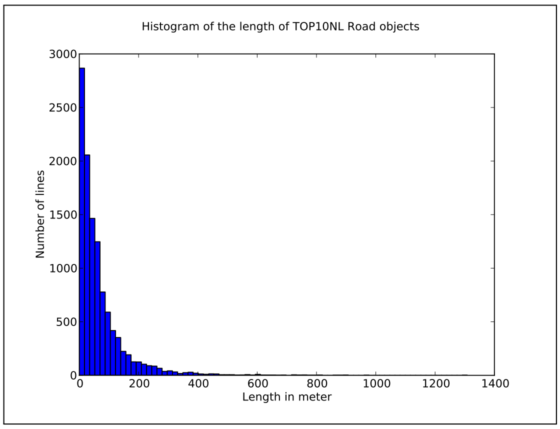


(b) Absence of building polygons in OSM

Figure 7: Dissimilarities by presence in the datasets



(a) Histogram of the length of the OSM road objects



(b) Histogram of the length of the TOP10NL road objects

Figure 8: Dissimilarities of the length of road objects from TOP10NL and OSM datasets of the same area.

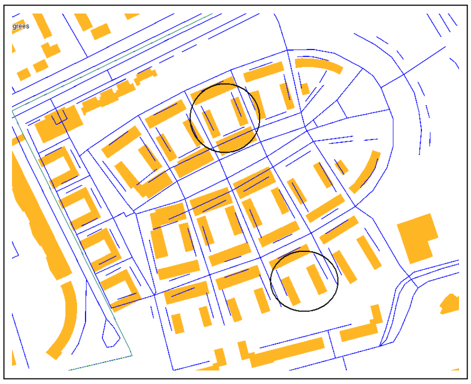
Figure 9: Example of difference in length of roads [Red depicts one OSM line which corresponds to nine TOP10NL lines which are shown in different colors]

of the same location. The histogram shows that the number of road objects with smaller length in TOP10NL is quite higher than in OSM.

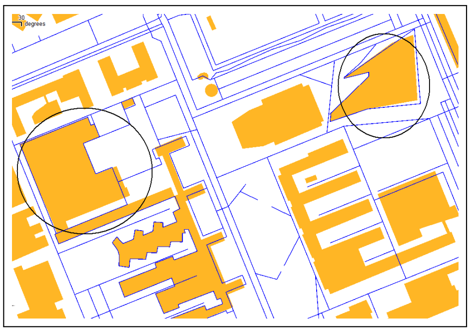
Another dissimilarity is the presence of arc in many OSM road objects. Figure 9 (on page 21) shows one combined example of length and straightness dissimilarities. One curved long OSM road object corresponds to nine TOP10NL road objects.

2.4.2 *Dissimilarities in geometry type*

Some dissimilarities have been detected, which are caused by different geometric type representation of an identical object. Figure 10 (on page 22) depicts some examples where building polygons are represented as lines in OSM. Similar mismatches have been found also in the object class of water. An example is shown in Figure 11 (on page 23) where an identical water object is represented as polygon in TOP10NL and line in OSM.



(a) TOP10NL building polygons are represented as line in OSM (within the circles).



(b) Another example of TOP10NL building polygons which are represented as line in OSM (within the circles).

Figure 10: Dissimilarities of object type of same object from the TOP10NL and OSM datasets.

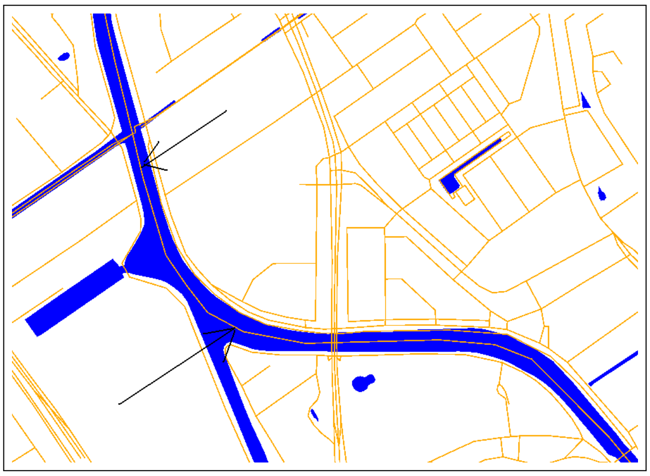


Figure 11: An water polygon in TOP10NL is represented by a water line in OSM

2.5 PILOT AREAS

Four areas within the Netherlands have been selected for carrying out this research. The idea behind choosing the pilot areas was to select some places which are spread over the country and having different characteristics. If the experiment can be verified with the data from different types of locations at different regions of the country, then it can be helpful to get impression over the complete dataset over the country. The locations of the pilot areas within the Netherlands are shown in the map in Figure 12 (on page 24) and the characteristics are described below.

Delft: The city Delft in the Netherlands has been chosen for the experiment. It is a densely populated place within the main economic region of the country. Its historic attraction made it more important. Being a university city, Delft is supposed to have many geo-concerned people who are interested and active in populating geo-information in the VGI map, such as OSM. The urbanised character of Delft was considered in the selection of the study area. An idea of the density of presence of buildings, roads, water and greeneries can be obtained from the satellite picture of the Delft area in Figure 13a (on page 25).

Rotterdam: Rotterdam is a metropolitan city and is located near to Delft. The reason for choosing Rotterdam was to select a place which is even more urbanised and densely populated than Delft. At the same time as it is a port city the presence of plenty of topographic object water makes the area complete for experiments. Figure 13b (on page 25) gives an impression of the area.

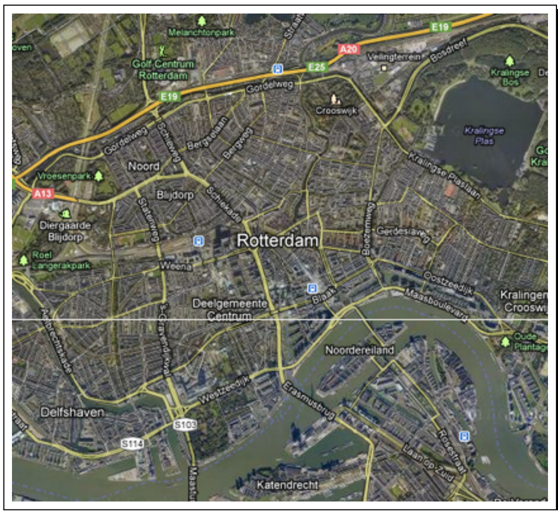
Dokkum: Dokkum is a small town, situated at the north of the Netherlands. Figure 13c (on page 25) shows the satellite image of

Figure 12: The pilot areas for this study

Dokkum. It is seen from this satellite image that Dokkum is comparatively a smaller town than Delft and the number of buildings and roads is less here.

Suburban of Echt The forth area is the suburban rural area of the city Echt, which is situated at the south of the country. The rural characteristics were taken into account in choosing this area because plenty of agricultural plots were found there. The satellite image in Figure 13d (on page 25) gives an impression of that area.

(a) Delft



(b) Rotterdam

(c) Dokkum

(d) Suburban of Echt

Figure 13: Satellite pictures of the pilot areas of same scale adapted from Google map

3 QUALITY ASSESSMENT OF OSM

3.1 INTRODUCTION

The quality assessment of OSM data of different countries within and outside Europe has extensively been carried out by different researchers during the last few years. To meet the objective of this research a good quality OSM data of the Netherlands is required prior to the object matching of TOP10NL and OSM. Therefore, a quality assessment of OSM data over the Netherlands is needed. To get an overall impression of the quality of the OSM data over the Netherlands the assessment has been carried out over the four pilot areas. The criteria for selecting the areas are stated in Section 2.5.

It was described earlier that the OSM data are contributed by the volunteers and there exist many ways of updating OSM. The heterogeneity in the data loggers used by the volunteers makes the OSM quality more critical. The accuracy of a contributor's data logger may vary considerably from other contributors [12]. And also the accuracy of the other sources of OSM data such as Yahoo or Bing map, aerial imagery, GPS track to OSM editor software varies from each other. The accuracy of all sources and tools influences the quality of the OSM data and it also varies from the location to location. Because of all these facts the quality checking of OSM is so important [67]. The quality checking is normally done by comparing with some ground-truth data if it is available. However, in the absence of the ground truth, the methodology developed by Ciepluchuch *et al.* [13] may be used. TOP10NL has been used in this study as a ground-truth for checking the quality of OSM data over the Netherlands.

Two other factors might influence the quality assessment result of the OSM data over the Netherlands. They are the accuracy of the ground-truth TOP10NL and the accuracy of coordinate transformation from WGS84 to RD_new. The average geometric accuracy of TOP10NL is ± 2 meter as reported in the product report at the first version [3]. And the average accuracy of the coordinate transformation from WGS84 to RD_new in this study is found to be 28 cm (described in Section 2.3.3), where it might vary from 25 cm to 100 meters as mentioned in [65].

The quality assessment judgments can be based on a number of factors: number of features, spatial distribution of features and data points, map labelling information, etc. [12]. Hakley [34] provides some pioneering guidelines to check the quality of the OSM data. Some standard quality measures of geographic information are 'lineage', 'positional accuracy', 'attribute accuracy', 'completeness', 'semantic accuracy', 'logical consistency', 'temporal quality' etc. are applicable for VGI type of maps [34] [64] [12]. However, a few of these quality measures were tested in case of OSM by previous researchers. Hakley [34] considered only positional accuracy and completeness for his study in the London area. He compared the roads of the OSM dataset with the adopted different resolutions of the Meridian-2 dataset with 1:1250 in urban areas, 1:2500 in rural areas and 1:10,000 in moorland. The vector dataset 'Meridian 2' provides coverage of Great Britain with complete details of the national road network of Ordnance Survey dataset [34]. Followed by him some more research over the UK region were carried out for the OSM quality assessment [40] [6] [35] [61]. Girres *et al.* [27] did the quality assessment of the OSM dataset with the BD TOPO dataset over the region of Hendaye in France. Their study was found quite relevant to this study.

Only three quality measures 'lineage', 'completeness' and 'positional accuracy' out of the above mentioned quality measures were chosen in this study. Linear and polygonal primitives were tested separately over the study areas. They are described separately in the following sections.

3.2 Lineage

Lineage is the quality aspect which explains the history of the dataset and the way it was collected and evolved [34]. In the OSM dataset the data history can be found in the 'source information'. All OSM elements contain 'user' as an attribute with the source information, contributer's information or software information as values. Moreover, many OSM elements have the source information in the 'tag' attribute. More trusted sources make any VGI more trustworthy. The evaluated result of data lineage has been described below for the four pilot areas separately.

In the Netherlands many renowned datasets such as 'AND' 'AHN', '3dShapes' were donated to enrich the OSM dataset in road network, building, water, etc. The Dutch mapping company 'AND', ("Automotive Navigation Data") has donated their entire street map of the Netherlands [15] to OSM in 2007. However, not all places are updated yet with the 'AND' road data. AHN (Actueel Hoogtebestand Nederland) is a dataset of detailed and precise elevation data of the Netherlands [66]. The dataset '3dShapes' has been released by PBL (Planbureau voor de Leefomgeving) [54] without any restrictions [71]. '3dShapes' is based partly on AHN, LUMOS (Land Use MOdeling System, a toolbox for land use modelling with the Land Use Scanner) [46], and Top10vector, the predecessor of TOP10NL [70]. In the OSM dataset over the Netherlands '3dShapes' is found to be an important data source for many OSM objects [14].

Pilot area Delft: The data lineage of the OSM data of the city of Delft was examined in two parts: lines and polygons. The lines together represent roads, water-line and railways data. It was found that 78 users contributed information in the Delft area. The number of data contributed by a single person ranges from 1 to 1000's. For example, the OSM users 'textlijn', 'polderrunner' and 'cetest' provided 29%, 25% and 11% of the total line data respectively. It was found that in many cases the contributors used the data from the source 'AND'. They are geo-educated amateur or professional users [19] [17] [18].

The polygons are representing the buildings, water bodies, land use, sport terrine, parking places together. Sixty three active users provided polygonal data, out of which 80% data came from '3dShapes' and 6% from 'polderrunner'.

Pilot area Dokkum: Similar to Delft the data lineage of Dokkum was evaluated separately with the line data and polygon data. Eleven users were active in populating the line data in the Dokkum area; 74% of all line data was gatthered by 'Theun' who was busy with the 'WikiProject Nederlandse Fietsroutes'[20] and 'AND' has provided 10% of the total line data.

Seventeen users were active for polygon data gathering. Seventy two percent of total polygon data were provided by'the user '3dShapes' and 36% of the total data were provided by 'theun'.

Pilot area 'Village area of Echt': Fifteen users are active in populating the line data to the OSM of that area. The dataset 'AND' was found directly or through the users as the main contributers of the line data. For 49.3% of the line data the source was 'AND' and similar to several other areas '3dShapes' was also found in this area as the main provider (nearly 98%) of the polygon data.

Pilot area Rotterdam: It was found that 68 individual contributers populated the line data and 37% of the updates were done using 'AND'. It was also observed that 34 contributers populated the polygon data and 90.3% of the updates were done using '3dShapes'.

Table 6 (on page 30) shows the percentage of data provided by two main sources 'AND' and '3dShapes' over the study areas.

3.3 COMPLETENESS

Hakely [34] commented that the completeness of OSM is the most significant aspect of VGI quality. He calculated the difference of total length of all roads from the OSM and Meridian 2 dataset to check the completeness of the OSM dataset. It appeared that the total length

Locations	'AND' Line	'AND' Polygon	'3dShapes' Line	'3dShapes' Polygon
Delft	46.3%	2%	–	83.6%
Dokkum	76%	0.1%	–	98.2%
Echt	49.3%	0.5%	–	97.2%
Rotterdam	37.2%	0.08%	0.5%	90.3%

Table 6: The main data contributers of OSM in the Netherlands

of the OSM dataset covers up to 69% of the Meridian 2 dataset [34]. The same method was applied in this study to check the completeness of all line primitives of the OSM dataset of the study areas. And the completeness of polygon primitives was calculated by taking the difference of total covered area from TOP10NL and OSM. The completeness study was performed for each object class. Road network, water network and railway network were verified for line primities and buildings, way polygon and water polygons were checked for polygonal primitives. An overview of the comparison of total length and total area of different object classes over the study areas are shown in Tables 7, 8, 9 and 10.

Object_class	*Total Covered by OSM*	*Total covered by TOP10NL*	*Percentage covered w.r.t. TOP10NL*	*Remark More covered*
Way_line	786.5714 km	767.0963 km	102.54%	*OSM*
Way_area	0.4167 km^2	5.9985 km^2	6.0%	*TOP10NL*
Building	5.5117 km^2	6.1424 km^2	89.73%	*TOP10NL*
Water_line	77.49 km	295.51 km	26.22%	*TOP10NL*
Water_area	1.9594 km^2	2.5041 km^2	78.25%	*TOP10NL*
Railway	21.1344 km	21.9447 km	96.30%	*TOP10NL*

Table 7: Completeness overview of Delft

The quality measure 'completeness' over an area gives an indication of justification for further experiment on OSM over that area. Object matching, the objective of this study, is not useful if the OSM data is not present sufficiently in the study area. Therefore, checking the completeness of the OSM data over the study area is important prior to further experiments. The observations on the completeness from Tables 7, 8, 9 and 10 it can be summarised as

- The 'road_line' data of OSM over all study areas covered more than 72% of the TOP10NL 'road_line' data. In Delft it was found that the OSM 'road_line' data was covered more than TOP10NL.

Object_class	*Total Covered by OSM*	*Total covered by TOP10NL*	*Percentage covered w.r.t. TOP10NL*	*Remark More covered*
Way_line	477.8402 km	481.8678 km	99.16%	*TOP10NL*
Way_area	0.1487 km^2	4.3043 km^2	3.45%	*TOP10NL*
Building	4.7883 km^2	5.5622 km^2	86.08%	*TOP10NL*
Water_line	23.4095 km	24.4618 km	95.69%	*TOP10NL*
Water_area	1.6159 km^2	6.2221 km^2	25.97%	*TOP10NL*
Railway	125.0664 km	91.06 km	137.34%	*OSM*

Table 8: Completeness overview of Rotterdam

Object_class	*Total Covered by OSM*	*Total covered by TOP10NL*	*Percentage covered w.r.t. TOP10NL*	*Remark More covered*
Way_line	127.7348 km	138.3065 km	92.35%	*TOP10NL*
Way_area	0.0075 km^2	1.00819 km^2	0.70%	*TOP10NL*
Building	0.8305 km^2	0.8602 km^2	96.54%	*TOP10NL*
Water_line	Not present	93.07 km	NA	*OSM absent*
Water_area	0.5115 km^2	0.5078	100.73%	*OSM*
Railway	Not present	Not present	NA	Railways absent

Table 9: Completeness overview of Dokkum

Object_class	*Total Covered by OSM*	*Total covered by TOP10NL*	*Percentage covered w.r.t. TOP10NL*	*Remark More covered*
Way_line	129.76 km	178.05 km	72.87%	*TOP10NL*
Way_area	0.00026 km^2	1.0003 km^2	0.02%	*TOP10NL*
Building	0.226 km^2	0.2265 km^2	99.44%	*TOP10NL*
Water_line	5.7155 km	72.662 km	7.86%	*TOP10NL*
Water_area	0.0226 km^2	0.0227 km^2	99.44%	*TOP10NL*
Railway	Not present	Not present	NA	Railways absent

Table 10: Completeness overview of village of Echt

- The 'road_area' data of OSM was almost absent. Only the 'parking plots' and 'roundabouts' are stored as 'road_area' in the OSM database. From this observation it was decided not to do the experiments of object matching with 'road_area' data.

- Buildings, water polygons and railway networks of OSM were covered less, though comparable to TOP10NL.

- The presence of 'water_line' of OSM was very poor in most of the places. In Dokkum 'water_line' of OSM is completely absent.

- Finally, it can be concluded that though the OSM data over the study areas was not yet 100% complete, they were well presented and the percentage of completeness is different in different locations. From the high presence of some object classes it may be pressumed that in the near future the OSM data of these areas will be complete.

The downside of this quality measure is that it cannot identify the region where the objects are missing. Therefore, it does not reflect the homogeneity of presence of the objects in one specific region. An impression can be drawn from the overlaid picture of one object class data from both datasets. An example of the heterogeneity is the 'road_line' data over Delft. The OSM data is more covered than TOP10NL in this area, but from Figure 14 we see that OSM is completely absent in a small location. Another example of inhomogeneity is the 'building' data of Delft, which is shown in Figure 15. The 'road_line' data of OSM is comparatively homogeneous in the Rotterdam location (Figure 16). Figure 17 shows that the 'road_line' data of OSM in the village area of Echt is poorly present, but they are homogeneously spread over the region. Only main roads are stored in the OSM database.

3.4 POSITIONAL ACCURACY

Haklay [34] used the 'percentage overlapping' method to find out the positional accuracy of the OSM data. The 'percentage overlapping' calculates the percentage of objects from one dataset that is within a certain distance of the same feature in another dataset of higher accuracy. He used buffering for this calculation and concluded that the percentage overlap between Meridian 2 and OSM buffers was pretty high with an average of nearly 80% and the variability from 60% to 89%. He also showed that the positional accuracy of OSM roads was of a reasonable accuracy of about 6 meter. The 'percentage overlapping' was applied in this research to find out the positional accuracy of both linear and polygonal objects of the OSM dataset. Figure 18 (on page 35) shows the principle of percentage overlapping. As the buffer size is related to the positional accuracy of the data, different buffer sizes (e.g. 2, 4, 6, 8 meters) were tested until getting similar results. The

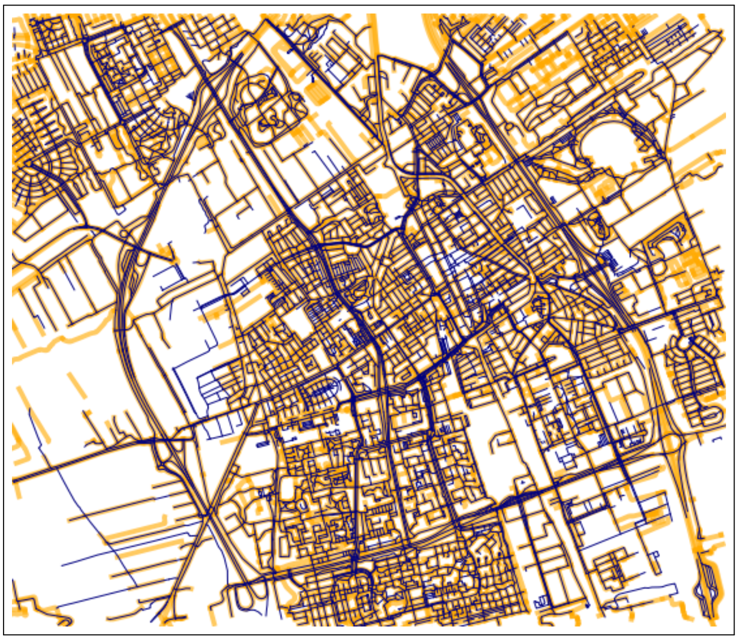


Figure 14: Overlay of highway network of Delft, yellow depicts buffered TOP10NL and blue depicts OSM

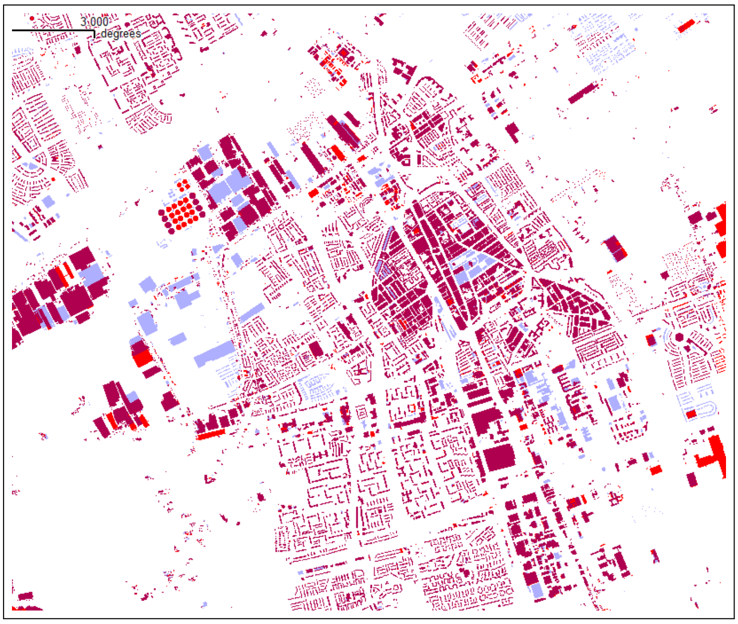
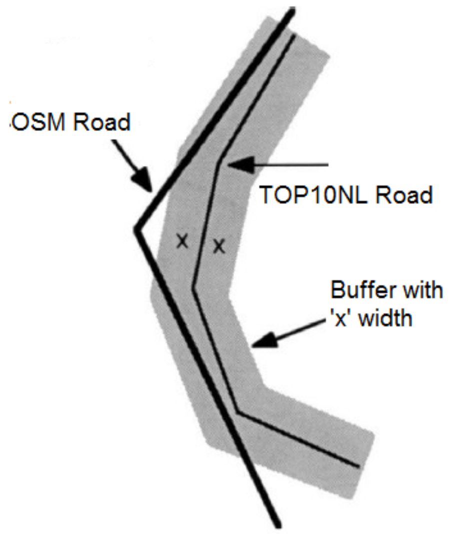


Figure 15: Overlay of Building polygons of Delft, red depicts OSM and blue depicts TOP10NL, purple depicts the overlaid region

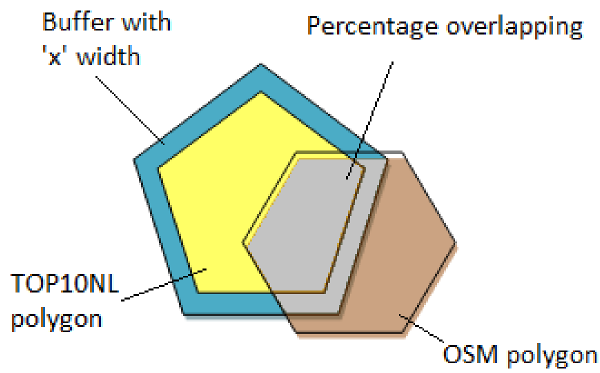
Figure 16: Overlay of highways network of Rotterdam, yellow depicts buffered TOP10NL and blue depicts OSM

Figure 17: Overlay of highways of rural part of Echt, yellow depicts buffered TOP10NL and blue depicts OSM

calculated results of positional accuracy have been shown in Tables 11 , 12, 13 and 14 (on pages 36, 36, 37 and 37). As TOP10NL has been chosen as the ground-truth, it was buffered to check the percentage overlapping of OSM objects.



(a) Percentage overlapping for line



(b) Percentage overlapping for Polygon

Figure 18: Percentage overlapping for line and polygon

From Tables 11, 12, 13 and 14 an overall idea of the positional accuracy over all pilot areas can be observed. The results do not change in case of testing with wider buffer zones. So a buffer zone wider than 6 meter was not considered. Though the positional accuracy according to the 'percentage overlapping' measure varies in different location and object class, it can be summarised that,

- At least 59% of OSM road network placed within 2 meter buffer zone, and upto 122% of OSM road placed within 6 meter buffer zone of corresponding TOP10NL road. This result indicates that there exist more than one roads of OSM within the 6 meter buffer zone of corresponding TOP10NL road. As said in [34], [27] '6 meter' as a good positional accuracy. So OSM road network data in the Netherlands can be considered as quite compatible with TOP10NL.
- The positional accuracy of OSM building data is quite high. Up to 95% of OSM building are situated with 2 meter of buffer zone of TOP10NL building.
- The 'Railway network' and 'water polygon' also found to have good positional accuracy. Up to 90% of both 'Railway network' and 'water polygon' are situated within 6 meter buffer zone of corresponding TOP10NL object.
- The positional accuracy of the 'water line' object class were found quite poor over all pilot areas.

Layer	*Buffer with 2 meter*	*Buffer with 4 meter*	*Buffer with 6 meter*
Way_line	56.72%	91.28%	95.1%
Way_area	23.81%	31.32%	39.49%
Building	94.32%	95.93%	97.9%
Water_line	42.29%	59.78%	67.42%
Water_area	78.49%	83.66%	87.35%
Railway	40.39%	73.23%	90.49%

Table 11: Percentage overlapping results on Delft

Layer	*Buffer with 2 meter*	*Buffer with 4 meter*	*Buffer with 6 meter*
Way_line	60.5%	96.6%	122.8%
Way_area	61.33%	81.67%	104.3%
Building	95.8%	96.93%	98.05%
Water_line	35.5%	38.3%	40.1%
Water_area	85.04%	86.43%	87.77%
Railway	40.5%	69.87%	89.71%

Table 12: Percentage overlapping results on Rotterdam

Layer	Buffer with 2 meter	Buffer with 4 meter	Buffer with 6 meter
Way_line	59.24%	99.22%	122.03%
Way_area	*Hardly present*	*in OSM*	
Building	96.87%	98.44%	101.31%
Water_line	*Not present*	*in OSM*	
Water_area	98.73%	100.33%	101.97%
Railway	*Not present*	*in Dokkum*	

Table 13: Percentage overlapping results on Dokkum

Layer	Buffer with 2 meter	Buffer with 4 meter	Buffer with 6 meter
Way_line	71.86%	86.29%	95.73%
Way_area	*Not present*	*in OSM*	
Building	87.45%	89.65%	93.45%
Water_line	8.07%	24.54%	47.66%
Water_area	99.64%	99.72%	99.89%
Railway	*Not present*	*in OSM*	

Table 14: Percentage overlapping results on village of Echt

The 'percentage overlapping' is a useful quality measure and is widely used for the quality checking of geographic maps. But the downside of the 'percentage overlapping' method is that it also counts the overlapped portion of the adjacent objects of candidate object, therefore it influence the result.

To get more detailed impression of 'positional accuracy' some other tests in addition to 'percentage overlapping' were performed. Some quality assessment tests have been performed with a set of manually selected homologous lines and homologous polygons separately from the TOP10NL and the OSM datasets. Sixty pairs of homologous line and 40 pairs of homologous polygon primitives have been randomly selected. A number of experiments have been performed using different methods proposed by different researchers with these data pairs. This part of the quality assessment are helpful in choosing the parameters for object matching process which has been described later in Chapter 4.

The positional accuracy between the set of homologous lines and polygons was tested using the methods used by Girres *et al.* [27] in the quality assessment of the OSM dataset of France. They compared the OSM dataset with the BD TOPO dataset over the region of Hendaye in France. BD TOPO is a Large Scale Referential (RGE) from IGN (Institut Géographique National) which contains a description of the landscape elements in the form of metric accuracy vectors [51]. They tested the OSM dataset over manually selected pairs of homologous linear and polygonal objects from the study area.

In this present study the following quality measures were tested on the set of homologous objects.

- For linear primitives
  - Euclidean distance between two objects
  - Hausdorff distance between two objects
  - Average distance between two objects
  - Difference of average angle between two objects
  - Difference of angle of the direction of two objects
- For Polygon primitives
  - Compactness
  - Surface distance

The methodologies of testing the above mentioned quality measures are explained in the following sections.

3.4.1 *Euclidean distance*

Euclidean distance is calculated as the minimum distance between the two objects. The Euclidean distances of 60 pairs of homologous lines were calculated. The histogram is showed in Figure 19 (on page 39). It can be seen from the histogram that more than 80% of the line pairs are within 1 meter of Eucledian distance. This is quite a good number for positional accuracy where the width of the roads are between 4 and 10 meter.

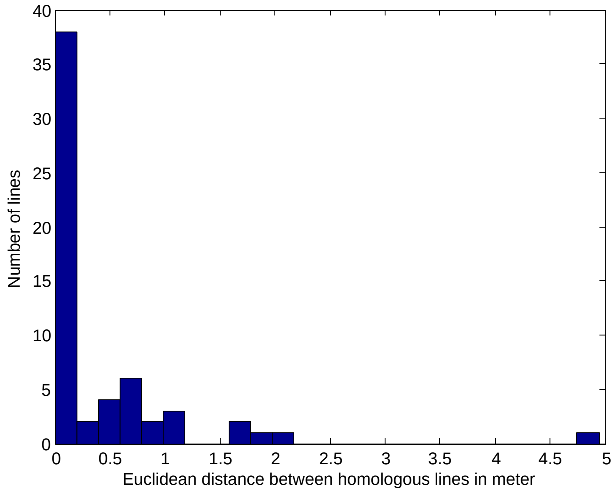


Figure 19: Euclidean distance between homologous lines

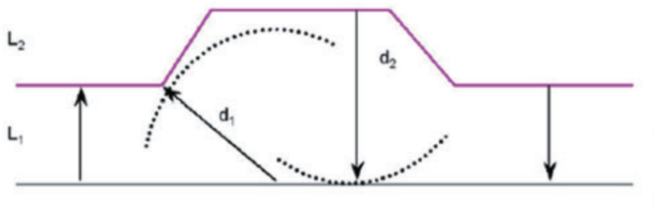


Figure 20: Housdorff distance $H = \max(d_1, d_2)$[27]

3.4.2 *Hausdorff distance*

The Hausdorff distance measures the maximum deviation between two polylines as shown in Figure 20. The Hausdorff distance dH between polylines L1 and L2 is defined as [27]:

$$dH = max(d1, d2) \tag{3.1}$$

where d1 represents the maximum of the shortest distance from L1 to L2 and d2 is the maximum of the shortest distance from L2 to L1.

The Hausdorff distance of the selected pairs of homologous lines were calculated and the histogram is shown in Figure 21 (on page 40). It can be seen from the histogram that nearly 50% instances lie within

20 meter Hausdorff distance and the rest has spread over a wide range upto 275 meter. Almost similar result was found by Girres *et al.* [27] from their study though the range was not that wide. Therefore, it was difficult to judge the quality of the OSM line object from this quality measure for these specific type of datasets. This disperse range of Hausdorff distance in the histogram is possibly the result of the length difference of the lines from two data sources which are explained in section 2.4 (on page 17).

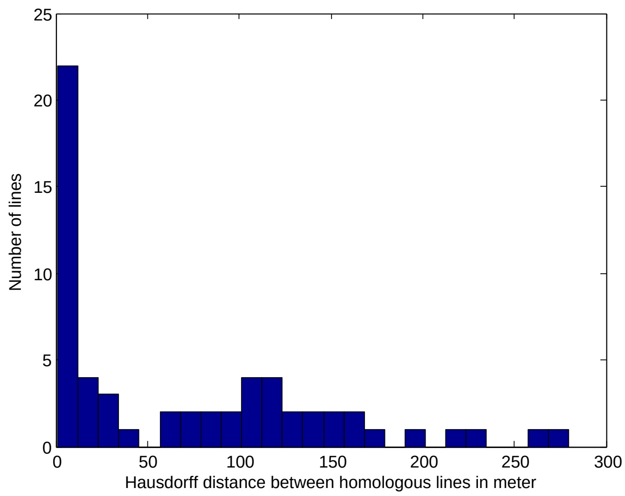


Figure 21: Hausdorff distance between homologous lines

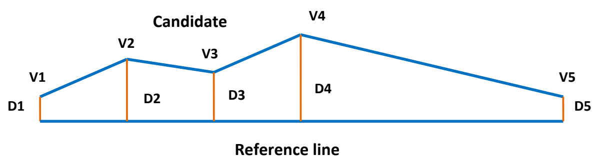


Figure 22: Average distance

3.4.3 *Average distance*

Average distance $\overline{D}$ is determined as the average of the distances of every vertex of candidate line to the reference line. Average distance is widely used in case of object matching process [5] [9]. It has been considered here also as a quality measure to check the acceptability in the object matching process. Figure 22 explains the average distance where,

$$\overline{D} = \frac{\sum D}{\#D} = \frac{D1 + D2 + D3 + D4 + D5}{5} \tag{3.2}$$

Girres *et al.* [27] suggested that an average distance within 6 meter indicates a good accuracy. Figure 23 shows the histogram of average distance between homologous lines. The distribution is an irregular skewed, but all distances fall within 0.9 meter which is within the suggested range by Girres *et al.* [27]. This is a indication of a good positional accuracy of the OSM dataset.

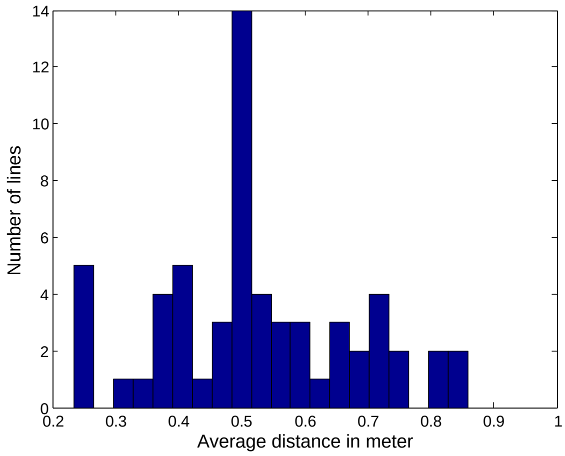


Figure 23: Average distance between homologous lines

3.4.4 *Average angle*

Average angle of any line object can be calculated by taking the average of the angles at every vertex of that line object formed by the horizontal line and the line joining consecutive two points. Figure 24 (on page 41) shows the concept of the average angle. The difference of the average angle of all candidate and reference object lines were calculated. From Figure 25 (on page 42) it can be seen that more than 80% average angles were below 30° while more that 60% were below 10°. The distribution was gradual but the range was a bit wide which is also possible since the set of homologous lines is not always straight lines (explained in Section 2.4.1).

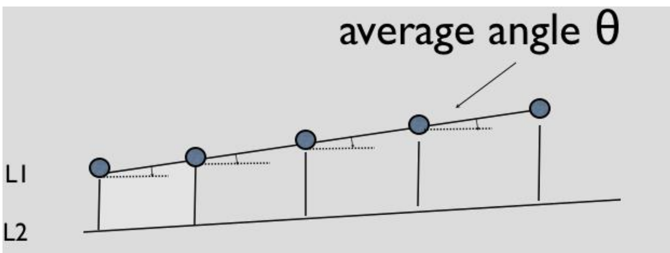


Figure 24: Average angle of a line adapted from [5]

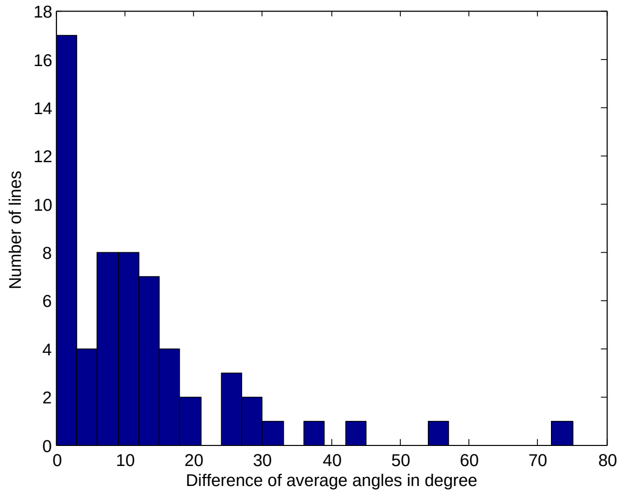


Figure 25: Difference of average angle between homologous lines

3.4.5 *Angle of the direction*

Though the difference of average angle of candidate and reference lines is a good measure of quality assessment, it has been found that often the result is misleading in case of presence of arcs in the lines. In that cases the line pairs having small average angle difference might not be identical because of having completely different directions. Therefore, the direction of the lines was also taken into account for the quality assessment. The direction of every line was calculated by taking the difference of the angles formed by the horizontal line and the line joining the first and the last points [9]. The difference between the angle of direction of the homologous pair was calculated and the histogram is shown in Figure 26 (on page 43). It has been found that the direction of 70% homologous lines are within 5°, which is quite acceptable for identical lines.

3.4.6 *Compactness*

Girres *et al.* [27] used two methods, surface distance and compactness differences, for comparison of positions of polygon primitives. It appeared that more than half of the surface distances among the tested polygons lie between the ratio 0.1 and 0.25. They found that this number is quite small and acceptable for the VGI dataset (OSM here) for polygon matching. They used compactness differences to assess the shape differences of lakes between the OSM and BD TOPO datasets.

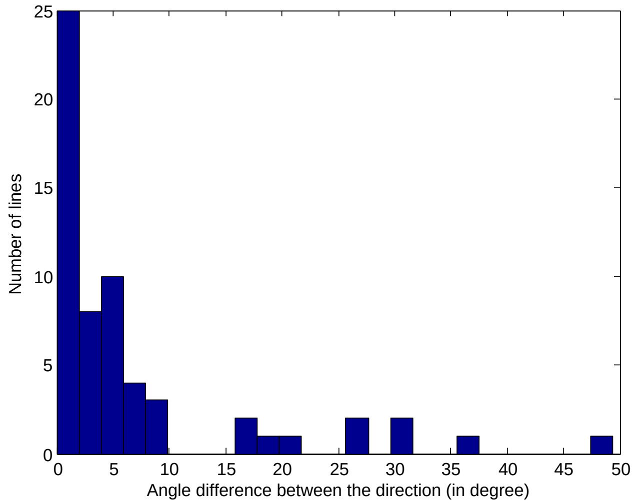


Figure 26: Difference of angle joining start and end point of homologous lines

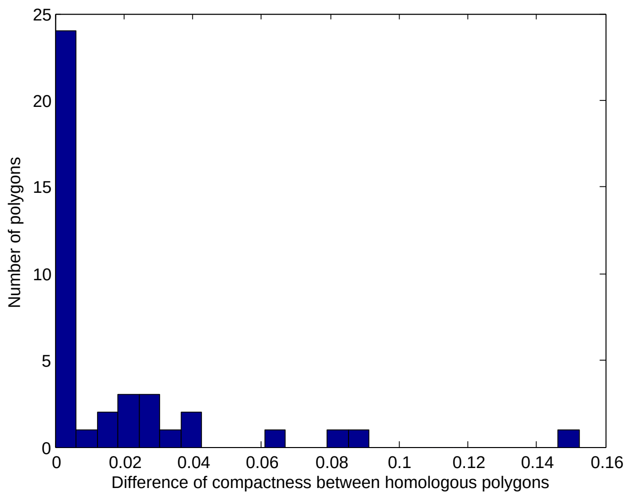


Figure 27: Difference of compactness between homologous polygons

For quality assessment of polygonal primitives the 'compactness' of the polygons was tested. It is defined as the ratio of 2π times area and square of perimeter of each polygon [27], which can be written as

$$C = \frac{2\pi \times area}{perimeter^2} \tag{3.3}$$

The compactness of all selected homologous polygons from candidate and reference set was calculated and the difference of compactness was taken. The histogram in Figure 27 (on page 43) shows that for more than 80% homologous polygons the compactness difference is below 0.04, which is quite small and can be said as a good measure for VGI type of map.

3.4.7 *Surface distance*

Surface distance is another quality measure of polygonal primitives. Surface distance dS of two polygons are computed as [27]:

$$dS = 1 - \frac{S(A \cap B)}{S(A \cup B} \tag{3.4}$$

where dS[0,1] such that
dS = 0 if A = B and dS = 1 if A B = Φ
The surface distance is defined in the interval [0, 1]. The distance is equal to 0 when two polygons are equal and is equal to 1, when two polygons are disjoint. The surface distance of all homologous pair of polygons of candidate and reference sets was calculated. Figure 28 (on page 45) shows the histogram of the surface distances. It can be seen from the histogram that more than 60% polygons have a surface distance less than 0.1 and 85% of polygons have it less than 0.5. The surface distances between the homologous polygons are considered to be reasonably small [27].

3.5 SUMMARY

The quality assessment criteria 'lineage' and 'completeness' were tested with the complete dataset of OSM over the study areas and the 'positional accuracy' was tested with both the complete set of data and also with a set of manually selected homologous object pairs. From the data lineage analysis we see that the sources are quite dependable because the major portion of the data were from authenticated datasets. It was also found that a major part of data from both datasets share same origin, which is Top10vector.

Over the pilot areas the completeness of OSM road data varies from 70% to 102% and building data from 89% to 97% of the TOP10NL

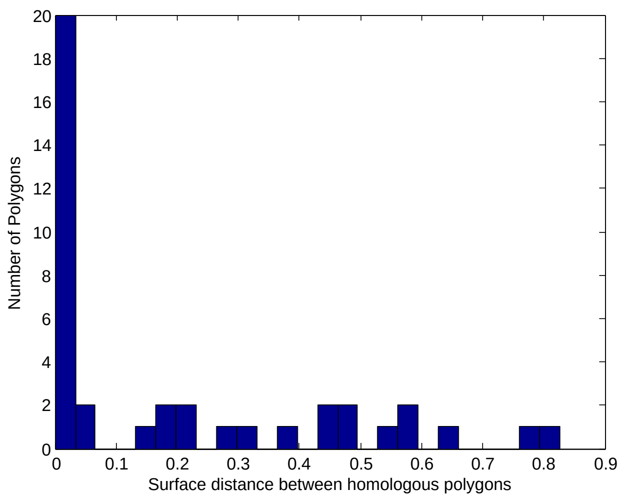


Figure 28: Surface distance between homologous polygons

dataset. The water polygons are also well presented in OSM dataset in most of the places, whereas bit less in the study area Rotterdam, only 25%. In other places the presence of water polygons varies from 79% to 100% of TOP10NL water polygon. In the previous studies it was observed that the completeness of road data over different countries varies from 60% to 80% which was considered as good presentation of VGI data suggested by the previous researchers [64] [74] [27] [34] [45]. However the water line network presentation varies from place to place, somewhere good and somewhere very poor. Overall it can be said from the completeness study that the OSM data is well polulated within the Netherlands.

In finding out the 'positional accuracy' of the OSM data over the study areas it has been observed that up to 100% of total line objects and polygon objects are lying within 6 meters of buffer zone of TOP10NL objects of the same class. From the test results of the different quality measures with the set of homologous object pairs it was found that 'Hausdorff distance' and 'average angle' are not usefull for these type of datasets. However, from other measures it can be said that the OSM dataset of the pilot areas fulfill the criteria of being a good quality data. This confirms the justification of the object matching precess.

4 OBJECT MATCHING

4.1 INTRODUCTION

Changes of geographic data on the surface of the earth has occurred due to two major reasons, natural forces and human actions. Some changes are abrupt and some are gradual. So maintaining a geographic map regular updates are needed. An object matching process is needed in the updating of a geographic map. Haigang *et al.*[33] described a framework to find out the changes in the geographic data of the real world within two time frames. Though the changes can be considered both in thematic and geographic aspects, this research only considered the changes of geographic objects according to their presence in the datasets at the present time. As mentioned in [33] the changes can be categorised into a number of situations which might happen due to the following facts:

- The object may no longer be on the ground;
- The object may be changed;
- The object may be unchanged;
- A new object may appear on the ground.

The purpose of this chapter is to explore the possibility of automatically detecting the mismatches of the objects in TOP10NL and OSM datasets at the present time frame. Unlike other change detection methods in the map updating process, this research concentrates on the detection of differences in two map datasets in the same time frame and consequently 'change detection' has been referred to as 'mismatch detection' in this thesis. In the perspective of this research, while the unidirectional object matching process intended to be carried out from TOP10NL to OSM, the following different cases might be arise. According to the presence of the objects in both datasets a number of matching tables can be formulated after the matching process have been completed.

1. One object might be present in the TOP10NL dataset, but not in the OSM dataset. This can be classified as '1-0'. This might happen due to the removal of that object from the real world after the last update of TOP10NL or due to the fact that the object is not yet recorded in the OSM database by the volunteers.

2. One object might be present in both TOP10NL and OSM datasets in different size and shape. In such a situation the objects might be classified as '1-n', 'n-1' or 'n-m', where 'n' and 'm' defines any integer number. In most cases it may happens due to the difference in the data acquisition techniques in the two datasets (see Section 2.4 on page 17). The 'n-m' classification can only be identified by bi-directional object matching process.

3. One object might be present in both TOP10NL and OSM datasets, which can be classified as '1-1'.

4. One object might be found in the OSM dataset and not in the TOP10NL dataset, which can be classified as '0-1'. This might happen due to the appearance of an object in the real world after the last update of TOP10NL. This case can only be detected by the bi-directional object matching process.

An object matching process is a combination of two matching process, such as 'class matching' process and 'shape matching' process. They are described below:

- The 'class matching' of the objects has been performed in the stage of the schema translation process. It has been carried out by extracting the objects of different classes separately from the OSM dataset by semantic mapping to match the TOP10NL object class. The details of the schema translation of TOP10NL and OSM is explained in Section 2.3 in Chapter 2.

- The 'shape matching' is the most critical component of this study. The main emphasis has been given in this phase to detect the mismatches between the identical objects in both datasets. In doing that the above mentioned mismatches have been considered. As described in Chapter 2, the mismatches between these two datasets are the result of mostly the differences in the data acquisition techniques and partly the recent changes in the real world. Therefore, the differences are mostly due to the difference in representation of an identical object in these two datasets.

  According to the relevance of this study a special care has been taken in defining the mismatch of objects between TOP10NL and OSM. In doing that some differences in the two datasets, which are unimportant for this study, have been ignored. Only the geographic position and the size of the objects were checked rather than the exact shape in the shape matching process. The accurate curves on the polygon object boundaries and the lines were ignored in the matching process. Some examples of line pairs and polygon pairs which are considered as identical in this study are shown in Figure 29.

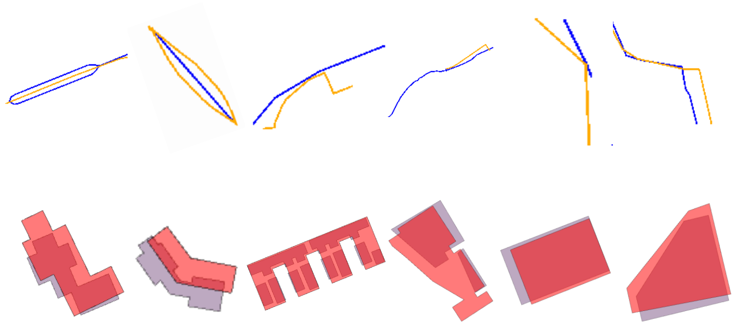


Figure 29: Example of considered identical line and polygon pairs

Several researchers have already carried out research on the geographic object matching in map updating and map integration. Zhang *et al.* [72] [73] mentioned that three modes can be identified for the road database updating. They are

- by ground surveying (using a traditional or a more automatic method)
- by map comparison (using a recent map to update the old road map)
- by image-based road change detection (using new remotely-sensed imagery) methods.

Although as mentioned in [72] that all the three modes have their own strengths and weaknesses, the second method has been chosen in this study, because the main goal of this thesis is to check the compatibility of the vector database OSM with another vector database TOP10NL. Moreover, this is the most economic way of map matching process. Some researches are found on object matching which are GIS based [68], [9], [5], [59], however, most of the object matching research are found on photogrammetry, remote sensing or image-based.

4.2 LINE MATCHING

Some previous studies were explored to get some guidelines for the line matching process. A similar study has been carried out by Volz [68] where he developed a methodology for the object matching of two vector databases. He used some parameters for finding out the best matches separately for point, line and polygon. Basically he followed the way to get the best matches by reducing candidates gradually from the potential matched pairs. By using a buffering technique he first created a list of potential matching candidates and then he filtered the list based on his chosen parameters. Some parameters on line matching are described in Subsection 4.2.1. Almost a similar approach was used by Bont [9] for a similar purpose. However, unlike the method proposed by Volz, [68], this study has explored the possibility of making the list of best matched pairs by accumulating them in a step by step procedure. The detailed method is described in Subsection 4.2.1.

Anand *et al.* [5] used a technique for matching the linear primitives of the OSM dataset with the Ordnance Survey Integrated Transport Network (ITN) data of Portsmouth, UK. Each feature of one dataset was checked with all features of the other dataset for automisation of the matching process. They considered 'average angles' and 'average distances' between the candidate and reference line features to find out the best matched objects.

As the goal of this study is to automate the object matching process, there is no scope for manual checking for the dissimilarities of objects one by one from two different datasets. Therefore some methods have been chosen to carry out the object matching process which consider all objects from the same class, irrespective of relevancy, to find out the best match. Though there are many types of measurement parameters for feature comparison exists, they are very sensitive on the shape of the features, and therefore, they are not universal as mentioned in [63]. As this research aims to compare a topographic map (TOP10NL), with a VGI type of map (OSM) which does not follow any straight forward rule, lots of difficulties arose which needed to be considered in this study in addition to the previous researches. Some problems have been explained in Section 2.4 which are dissimilarities in line length, geometry types, straightness of identical objects from TOP10NL and OSM datasets. Therefore, the matching rules proposed by the previous researcher were tuned with the irregular behavior of the OSM dataset to build up a suitable and simple rule to automatise the object matching process. To make the process simple the complicated calculations has been avoided and simple PostGIS functions have been used in line-matching query.

As mentioned earlier only the 'road network', 'waterline network' and 'railway network' have been considered for this study among the number of linear object classes available over the pilot areas. The 'road network' data of the city Delft has been chosen as the training dataset for building a 'linear object matching model', because the 'road network' was found as the most complicated and dense line network. Then the model was verified with the other line networks from the other pilot areas.

As seen in the 'completeness' study of the quality assessment of OSM (see Section 3.3 in page 29) all roads are represented in the OSM dataset only as lines. But it was shown in Table 1 that in the TOP10NL dataset the roads are represented as polygons as well as lines in different vector files. Polygons represent roads (mainly major roads) and parking plots. The lines represent the roads are the centerline of the major roads, cycleway or foot way. So WEGDEEL_LIJN, the line representation of the road object class of TOP10NL covers all roads of different types and consequently, it was used in this study for comparing with the road_lines of OSM.

4.2.1 *Rule generation for line matching*

A number of matching parameters (described below), proposed by the previous researchers were examined with the training dataset for rule generation of object matching process. Most of these parameters have also been used to check the positional accuracy of the OSM dataset (see Section 3.4 page 32) and found potentially suitable parameters. All the proposed parameters are explained below. The suitability of these parameters have been judged with the characteristics of the TOP10NL and OSM dataset. Only the useful parameters have been selected for generating a matching rule.

Length of lines: Length of lines is a common measure in practice for object matching. In some studies [68],[9] this was considered as a good parameter for the line matching process. But in the present study it has been observed that the length of the road objects from OSM and TOP10NL are quite different (explained in Section 2.4). As many roads in OSM are recorded by the contributers from GPS tracks, the boundaries, where the road starts and ends, are not maintained following any rule. This is unlike most of the road networks in geographic datasets which terminates at junctions. Figures 9 (on page21) and 8 (on page 21) provides the explanation. So comparing the length of the reference line and candidate line is not useful. So 'length of lines' was not considered as a suitable parameter for line matching in this study and has been ignored.

Euclidean distance between two objects: Euclidean distance, which measures the spatial distance between the reference line and candidate line is also widely used as a measure to find out the best match [68],[9]. This is explained in Section 3.4.1 of Chapter 3. From the histogram, (see Figure 19 on page 39) of Euclidean distance of selected homologous line pairs it was observed that more than 80% of homologous line pairs are situated within 2.5 meters. So this measure has been considered as a good measure and 2.5 meter was chosen as a threshold of Euclidean distance for the matching process. The built-in function 'ST_Distance' of PostGIS was used for calculating 'Eucledian distance'.

Hausdorff distance between two objects: Hausdorff distance is explained in Section 3.4.2. It measures the maximum deviation between two polylines. In many object matching process it has been considered as a good measure [59][68]. In quality assessment of OSM dataset it is observed from Figure 21 (on page 40) that nearly 40% of Hausdorff distances lie within 15 meter and the rest has a spread over a wide range upto 275 meter. From this distribution selecting a threshold was meaningless. The differences in length of the OSM and TOP10NL line objects have caused this large spread. Therefore, Hausdorff distance was not considered a usefull parameter for the matching process of these datastes. The PostGIS built-in function 'ST_Hausdorff' was used to calculate the Hausdorff distance.

Fréchet distance: Fréchet distance is another measure which is also often used for curve matching process [60]. Fréchet distance of the curves is defined as the minimal leash length necessary for both to walk the curves from beginning to the end [10].

A matching between two curves P and Q is a pair of monotone reparametrizations (α, β) of P and Q respectively, where the point $P(\alpha(t))$ is matched to the point $Q(\beta(t))$. Mathematically, the Fréchet distance between two curves is defined as [24]

$$\delta_F(P, Q) = inf_{\alpha,\beta} max_{t \in [0,1]} \{d(P(\alpha(t)), Q(\beta(t)))\} \tag{4.1}$$

Where $d(P(\alpha(t)), Q(\beta(t)))$ is the Euclidean distance between two points $P(\alpha(t))$ and $Q(\beta(t))$. For every possible function $\alpha(t)$ and $\beta(t)$ at time t, there is the largest distance, and the Fréchet distance should be the minimal one found among these maximum distances.

Fréchet distance has been widely used in several researches as an main criteria of map matching process and has been implemented in many projects, for example 'Track and Trade' project which has been also followed by Tom Tom [1]. Some other research also found on it's use in map matching and coastline matching process by Chen *et al.* [11], Mascret *et al.* [49].

Fréchet distance is a good measure for line matching while the beginning and the end points of the candidate and reference line are same or very closed by. Because of the irregular behaviour of OSM and the length difference of identical line of OSM and TOP10NL it was presumed that this distance measure is not suitable with these datasets. Therefore, this measure has not been included in the decision rule in this study.

Average distance of two objects: Average line distance $\overline{D}$ is the average of the distances of every vertex of the candidate line to the reference line. This parameter is also widely used in the line matching process [5][9]. This is also explained in 'Quality Assessment' chapter in Section 3.4.3 and it indicates the good positional accuracy of the OSM data.

Two disadvantages of using this measure are explained in the following. Firstly, the average distance might be closed to zero even the two lines deviate a lot. This might happen due to the systematic difference in OSM and TOP10NL where one line from one dataset can be partly located above and partly below the corresponding line of other dataset. Secondly, for the same reason the average distance might not be symmetrical in all cases. Average distance from line object A to B might not be the same to the average distance from B to A. In this situation selecting a unique threshold for bidirectional matching process is difficult. Therefore, average distance also has not been taken into account to keep the decision rule simple in line matching process.

Average angle: Average angle, which is described in Section 3.4.4, is another measure for line matching [5] [68] [9]. It was found in this study that average angle is a good measure for straight or nearly straight lines, but not for curved lines. Since the angles at every node changes abruptly for the curved lines the average angle cannot reflect the real trend of the line. This becomes even more critical when the nodes of the identical pairs are not maintained identical distances and they do not maintain the same straightness. This happens because of the different acquisition methods employed for the identical line objects of TOP10NL and OSM datasets. Figure 25 (on page 42) shows that the difference of average angles of a set of homologous pair of lines have a skewed peak distribution and the range is too high. The range of the difference of average angles lies between 0° to 75° though the peak lies within 10°. Therefore selecting a suitable threshold was difficult and consequently this measure was also not considered in this study.

Angle of direction: Angle of direction, which is explained in Section 3.4.5 was also used as a good measure for line matching by previous researchers [9]. Angle of direction is the angle between the line joining the start and end points of the line object and the horizontal line. These were calculated for every candidate and reference line. Figure 26 (on page 43) shows the difference between the angle of the direction of some homologous pair of lines from both datasets. For more than 80% pairs the difference of angle of direction was within 10°. So this parameter was chosen as another measure for line matching in this study and 10° was chosen as a threshold for the matching process.

Out of the above mentioned parameters only two parameters, Euclidean distance and Angle of direction have been found to be suitable for training the dataset and therefore, a decision rule was formed only with these two relevant parameters.

The most common practice in generating a decision rule is to assign weights to parameters and combine them. Bont [9], Anand *et al.* [5] combined the parameters by assigning some weights to them and generated a decision rule. Normally the weights are assigned to the parameters according to their importance or from the experience of previous users. Bont [9] followed the guidelines of Volz [68] and he used all parameters used by Volz. But it was seen that not all the above mentioned parameters are useful in this study. Therefore, the guidelines of Volz [68] could not be followed here. For the same reason, the guideline of Anand *et al.* [5] also could not be followed in this study.

Moreover, the selected parameters were found to be independent of each other and equally important. Therefore, a simple decision rule was formulated for the line matching process. All OSM line objects were checked with all TOP10NL line objects of the same class and the matched pairs were selected if:

- the OSM line object locates within 2.5 meter of Euclidean distance of the TOP10NL reference line object, and
- the difference of angle of direction of the OSM line object and TOP10NL line object is within 10°.

The line matching model was built with the matching rule. Some false positive matching were found which are discussed below.

1. It was found in some matching pairs that one OSM line object was paired with more than one line objects from TOP10NL dataset. That is due to the difference in length of an identical object in TOP10NL and OSM in many cases (explained in Section 2.4). Out of these multiple matches some are found to be wrong. An example of good and bad multiple match is shown in Figure 30 (on page 55). Here one matching pair was found where they match with a very small part of their length. This type of matches have been removed and in doing that another checking criteria was added to the matching rule. That was defined as, if the length of the intersected portion of the candidate line and the buffered reference line exceeds a threshold then it was taken as a good match.

2. When one long straight or nearly straight line corresponds to many small straight or nearly straight lines, then the matching results are perfectly good. But if the set of all reference and candidate lines are not all straight, rather having bents, then the

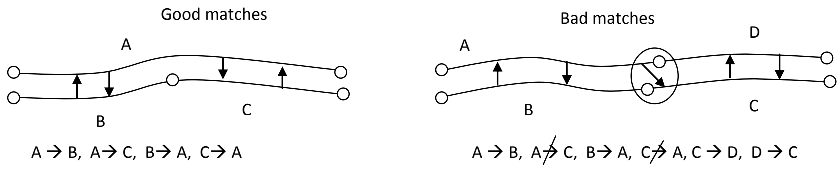


Figure 30: An example of good and bad multiple match

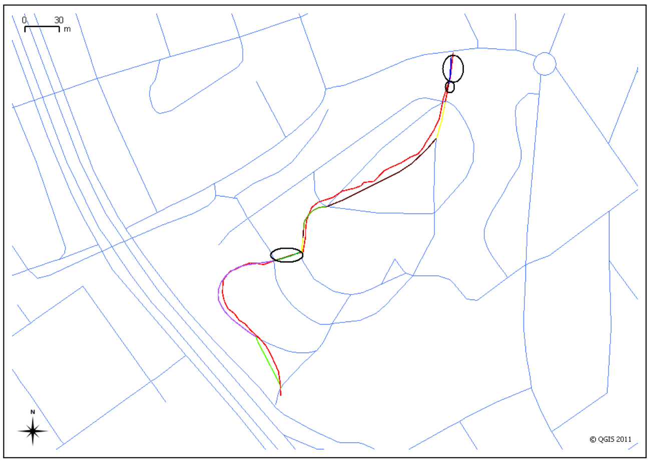


Figure 31: One to many matches with a curved line

model can not match all corresponding lines perfectly. Only the straight lines among the set of small lines turn out as matched line. An example is explained in Figure 31 (on page 55). The rounded portions resulted as the only matched lines. As a result the big line matches with less number of lines and the rest lines remain as unmatched lines. By adding another checking criteria in the decision rule this problem has been solved, that is 'checking the straightness of the lines'.

It was found in [56] that 'the complexity of a line can be identified from the ratio of its length to the distance between the end points'. This ratio was calculated for each line objects from both datasets and denoted as *R* hereafter in this thesis. By visually checking of a number of manually selected examples from the training datasets it was found that the lines with *R* more than 1.06 can be considered as a curved line.

Finally the rule was generated in the following way:

```
if Euclidean_distance (OSM, TOP10NL) < 2.5 meter AND
difference [angle of direction(OSM, TOP10NL)] < 10 deg AND
length [OSM, Buffer(TOP10NL, 2.5 meter)]>=2.5 AND
R(OSM)<1.06 AND
```

```
R(TOP10NL)<1.06
then OSM matches TOP10NL
```

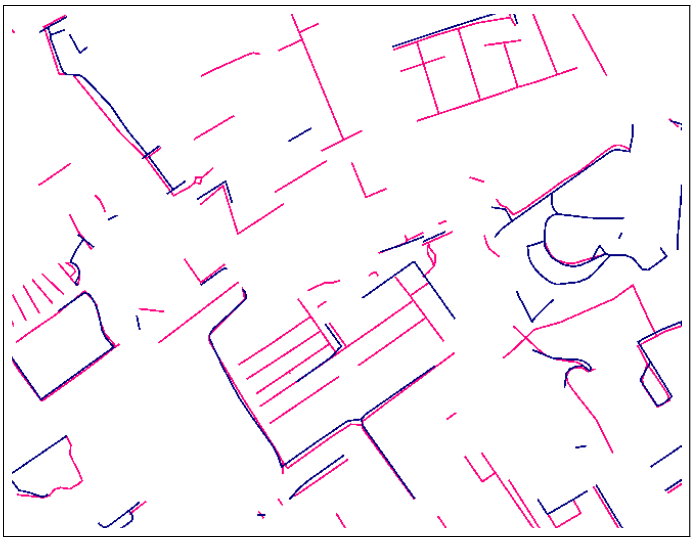


Figure 32: Overlay of unmatched road objects after first matching

In this rule the line objects with curve were excluded. Figure 32 (on page 56) shows the unmatched set of the reference line and candidate line after executing the rule on the training set. It was found from the figure that in some cases two identical objects from different datasets did not match. The unmatched line objects were curved as well as straight. For the curved lines it is understandable, because they were excluded from the previous rule. But the non-curved lines, it has observed that they are bit further situated than the assigned threshold values in previous rule. Therefore, by some relaxation of the threshold values this problem could be overcome. To find out the new threshold values Figures 19 (page 39) and 26 (page 43) were explored. It was found from the histograms of Euclidean distance that all instances lie within 5 meters. By relaxing this distance threshold the errors of coordinate transformation was also incorporated. From the histogram of the difference of angle of directions it was found that 90% of instances were within 25°. Following this information the decision rule was rebuilt for running the model for the second time with the set of unmatched line objects from both datasets.

The matching process for the straight lines were performed two times to avoid the false positive matching. If the matching rule with a relaxed threshold be applied for the first time, more than one match (obviously with wrong match) for one object would be returned in the cases where in reality two or more parallel lines are situated very close to each other. For these cases another filter had to be added to identify the best match. So by rerunning the matching process this confusion was overcome.

Figure 33 (page 57) shows the overlaid picture of the unmatched road objects from both datasets after the second matching process was completed. From the figure it is found that a few identical objects, which are mainly curved, remained unmatched. Therefore, another matching rule was formulated to include the curved lines to make the matching process complete and a third rule was introduced. From a number of manually selected examples it was found that the curved homologous lines were mostly situated within 8 meters from each other. Accordingly, a simple rule for the third matching process was formed for the curved lines: to find out the line objects which were completely within the buffer zone of others. Consequently, the rule was applied to the unmatched set of line objects. The followed rule is mentioned below.

Figure 33: Overlay of unmatched road objects after second matching

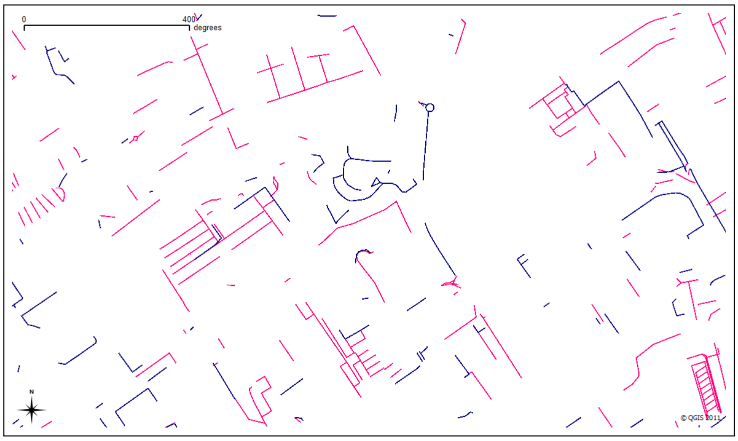
```
if OSM within buffer(TOP10NL, 8 meter) OR
TOP10NL within buffer(OSM, 8 meter)
then OSM matches TOP10NL
```

Summarising, the matching process was carried out in three steps. The steps are

- Step-1: Input–> The complete set of OSM and TOP10NL line objects from similar locations was considered. The straight lines were taken into account.
- Step-2: Input–> The set of unmatched line objects from OSM and TOP10NL after execution of Step-1, was considered. Here also the straight lines were taken into account. The rule was the same as of Step-1, but relaxed with thresholds.
- Step-3: Input–> The set of unmatched line objects from OSM and TOP10NL datasets after the execution of Step-2, was considered. Mainly the curved lines were taken into account in this step.

After running the three steps the complete set of matched pairs was formed by combining the matched pairs from three steps. Afterwards the complete matched and unmatched road objects of Delft location were closely observed. An overlaid picture of matched and unmatched datasets are shown in Figure 34 (on page 58). It is clearly visible from the figure that the unmatched road objects are completely disjoint (what it should be), and the matched road objects visually show a good match. Figure 35 (on page 59) shows the UML process diagram of the complete line matching process.

(a) Overlaying of complete set of matched road objects



(b) Overlaying of complete set of unmatched road objects

Figure 34: Overlaying of complete set of matched and unmatched road objects of Delft

4.2.2 *Verification of line matching*

The built up line matching model was verified with the datasets of 'road network' from the other pilot areas, Dokkum, the rural part of Echt and Rotterdam and then with the datasets of other linear object classes, such as 'railway network' and 'water line' from all four pilot areas. As 'railway network' is absent in the locations Dokkum and

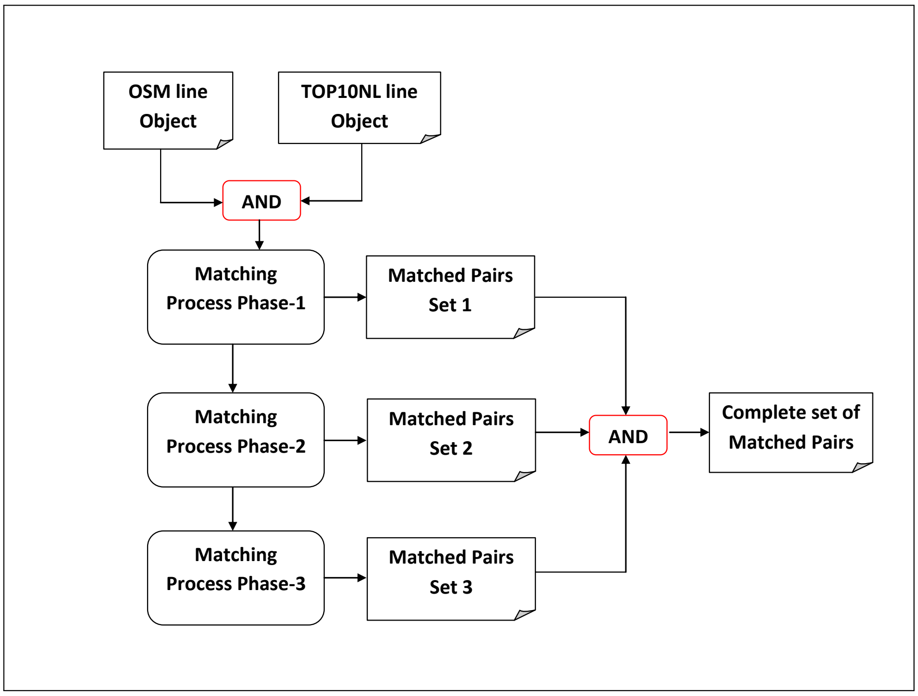


Figure 35: UML process diagram of the line matching process

Echt, the model could not be verified there. The 'water line' object class was rarely present in the Echt OSM dataset and was completely absent in the Dokkum OSM dataset. So the model could not be verified with the dataset of 'water line' from the two study areas.

The matching accuracy was tested from a set of randomly selected pairs of resulted matched lines from all locations. Out of these matched pairs some were found as good matches and some were found as false positive matches. The percentage of good matches over the total matched objects were calculated. Table 15 (on page 59) shows the percentage of good matched for all line objects from all pilot areas.

	%Good Match of Object Classes		
Locations	Road	Waterline	Railway
Delft	87%	96%	94%
Dokkum	89%	–	–
Echt	87%	–	–
Rotterdam	87%	100%	78%

Table 15: The percentage of good matching line objects from different places and different object classes

4.3 Polygonal matching

As mentioned earlier only ’building’, ’water area’ and ’way area’ object classes have been considered in this study. But it was seen in the ’completeness’ study (see Section 3.3 on page 29) in the quality assessment phase of OSM data that ’way area’ was hardly present in the OSM dataset of the pilot areas. Therefore, ’way area’ was not taken account for the polygon matching process. The building dataset of the city Delft was chosen as the training dataset for the ’polygonal object matching model’, since the area is densely populated and changes in buildings happen more often than in other objects. The built up model was verified with the ’building’ datasets of Rotterdam, Dokkum and the rural area of Echt and with the ’water area’ data from pilot areas.

4.3.1 *Rule generation for polygon matching*

Several studies on the building matching process were found. Matikainen *et al.* [50] developed an automatic change detection method for building polygons. They used laser scanner and aerial images to detect changes of building maps. Knudsen *et al.* [42] suggested an algorithm for updating building polygons of the Danish National Topographic Map Database (TOP10DK) using vector and spectral data. Hild *et al.* [37] built an algorithm for automatic change detection of polygonal map objects by integrating satellite imageries with vector map data. But as explained earlier this study concentrated on the object matching process by comparing two vector database, the knowledge from these studies could not be followed completely in this study. Some parameters used for polygon matching by Volz [68], Shen *et al.* [59] were explored to developed a matching rule. Only the suitable parameters for this study were considered. The parameters are explained below:

Area and centroid of the polygon: The area and the centroid of an polygon are common measures for polygon matching. Volz [68], Shen *et al.* [59] used theses measure in their study for polygon matching. The area difference and the distance between the centroids of the candidate and reference polygons could be two good measures for polygon matching where every polygon from one dataset corresponds to only one polygon of the other dataset. But in the case of TOP10NL and OSM datasets this is not the case because of different data acquisition methods. In many cases it was found that one polygon from one dataset corresponds to many polygons from the other dataset. So it was difficult to develop rule with area and centroid which is suitable for an automatic matching process and is valid for all purposes. Therefore, area and centroid of a polygon were not considered in this study.

Surface distance between two polygons: Surface distance between two polygons is a good measure for quality checking of geographic maps as well as for polygon matching. Surface distance [27] is described in Section 3.4.7. This distance measure detects whether two polygons overlap each other or not. If they overlap, it returns the percentage of overlapping area. In order to find out polygons which overlap each other in the polygon matching process, the surface distance has been taken as a good measure in this study. However, in some cases it was detected that this matching criterion also matches the neighbouring polygon which has a small or negligible overlapping area. To filter out these wrong matches one threshold was chosen based on some examples. It was found by manually checking that the surface distance of two polygons which is less than 0.98 can be considered as a matched polygon.

Compactness of the polygon: As explained in Section 3.4.6 the compactness of a polygon is a ratio of 2π times of area and square of perimeter [27]. Compactness differences allow assessing shape differences between reference and candidate polygons [27]. But in this study the position of the identical object has considered as more important factor. Therefore, use of compactness in automatic matching process might results false positive matching. So this measure was not considered in this study for polygon matching process.

Finally a simple rule was generated for carrying out the polygon matching process, which is mentioned below.

```
if surface distance (OSM, TOP10NL)<0.98
then OSM matches TOP10NL
```

The bounding box (bbox) of the polygons was used to speed up the searching process. Figures 36 (on page 62) and 37 (on page 62) show the overlaying picture of resulted matched and unmatched set of building polygons of Delft. The detailed results are discussed in Section 4.4.2.

4.3.2 *Verification of polygon matching*

The built up model was verified with the building polygon data of the other study areas Dokkum, Rotterdam and the village area of Echt. For all the cases the result was satisfactory. The results are provided in Table 16 (on page 63).

Similar to the line matching model, the matching accuracy of the polygon matching model was also verified with a set of randomly selected pairs of resulted matched polygons from all locations. The

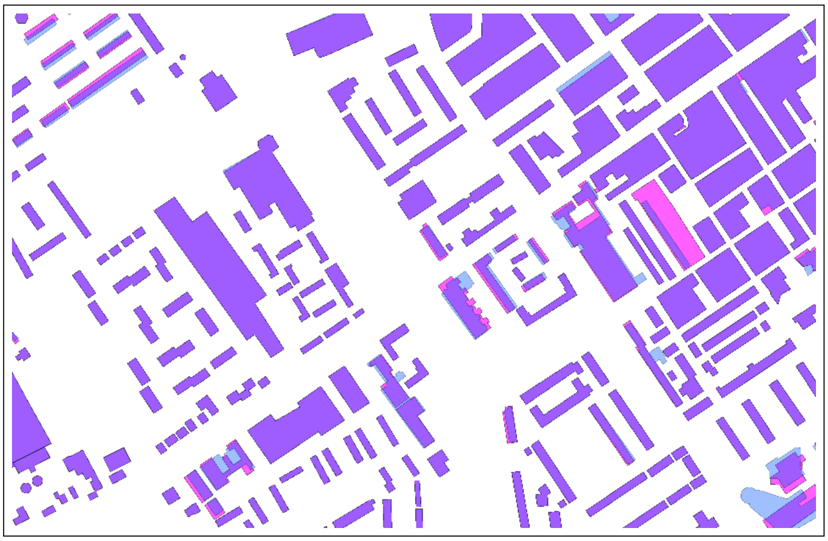


Figure 36: Matched building polygons of Delft. Pink and light blue depict OSM and TOP10NL building respectively, where dark blue depicts the overlaid area

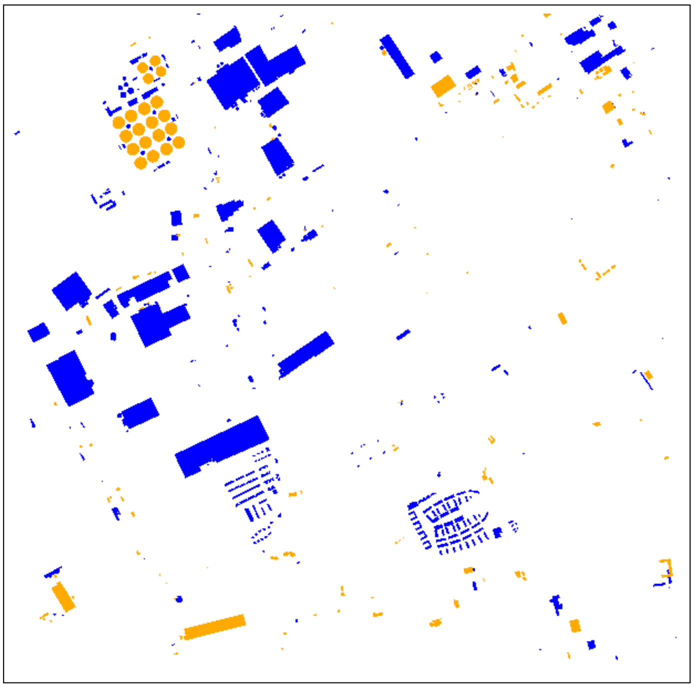


Figure 37: Unmatched building polygons of Delft. Blue depicts TOP10NL buildings and yellow depicts OSM buildings. All are disjoint.

percentage of good matched pairs over the total number of matching pairs was calculated and provided in Table 16 (on page 63) .

	%Good Match Object Classes	
Locations	Building	Water_Polygon
Delft	99%	100%
Dokkum	99%	100%
Echt	99%	100%
Rotterdam	99%	100%

Table 16: The percentage of good matching line objects from different places and different object classes

4.4 DISCUSSION OVER THE RESULTS

After the whole object matching process was complete the matched pairs were classified according to their cardinality. Different matching tables ('1-1', '1-n', '1-0', 'n-1') were formulated by extracting the matching pairs from the complete set. Afterwards, they were verified by close visual observations. Most of the matched objects were found to be correct in both line and polygonal objects. However, there were some unmatched and false positive matched pairs from both line and polygonal objects. An analysis was performed based on some arbitrarily selected instances and the findings are explained below.

4.4.1 *Line Objects*

Single correspondence: One-to-one match between line objects from OSM and TOP10NL is further divided into two types, namely, straight lines and curved line, and are described below.

- Identical straight line objects from both datasets with the same or almost the same length matched perfectly. For this case 99% of pairs matched successfully. For the other 1% pairs either there exists another line very close in a different side or they are not completely situated side by side. The identical lines with dissimilar lengths sometimes matched more than one line object and are considered as false matching.
- Curved identical line objects with similar lengths from both datasets always matched perfectly.

Multiple correspondence: One-to-many matches were mostly found in case of different length of line objects in both datasets which corresponds to each other. These cases are described in Section 2.4. Especially many OSM line objects were detected which corresponded to more than one TOP10NL line objects. The 'one-to-many' matches were also tested by close visualisation. Though many cases with accurate matching were found, some were found as wrong matching. Basically in the junctions the situation is complicated when many line objects are situate very near to each other. The lines within the circle of Figures 38 on page 64 are the examples of these cases. In this situation the manual matching is also confusing.

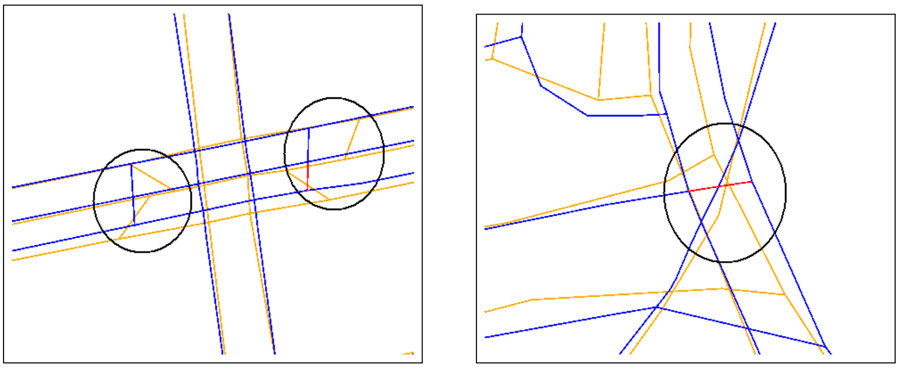


Figure 38: Examples of mismatched line objects

Insertion of 'Street names' in TOP10NL database: There exists an attribute named 'STRNAAMNL' in TOP10NL road dataset whose values are empty almost everywhere. By the line matching process for road networks, it was possible to insert the 'street names' by extracting them from the OSM dataset. This is only possible for those TOP10NL roads which were found as matching objects and the street names were extracted from the corresponding objects from the OSM dataset. This way TOP10NL can be improved with street names. And the accuracy of assigning correct street names for the TOP10NL roads certainly depends on the accuracy of the road matching process.

4.4.2 *Polygonal Objects*

From the verification result of polygon matching process satisfactory results were detected after close observation. The detailed results are given in Table 16 (on page 63). Some matched pairs were found which are a bit confusing about there validity as identical objects. The polygons in Figure 39 (on page 65) shows some example of those cases. Therefore, another study has been performed to check the quality of the polygon matching. It was done only by checking the percentage of overlapped area of OSM polygons over corresponding TOP10NL polygons. The histogram in Figure 40 shows that for more than 85%

of total matched pairs the percentage of overlapped area is more than 90%.

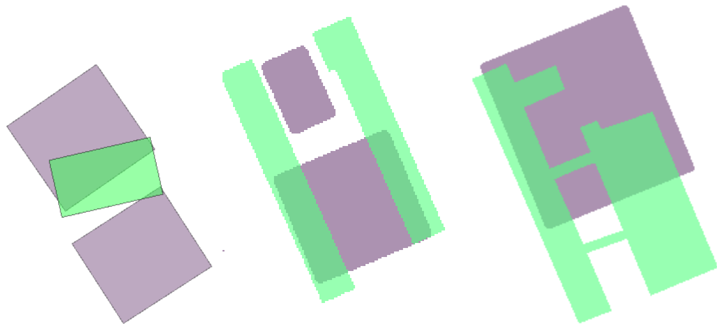


Figure 39: Matched buildings which are not good

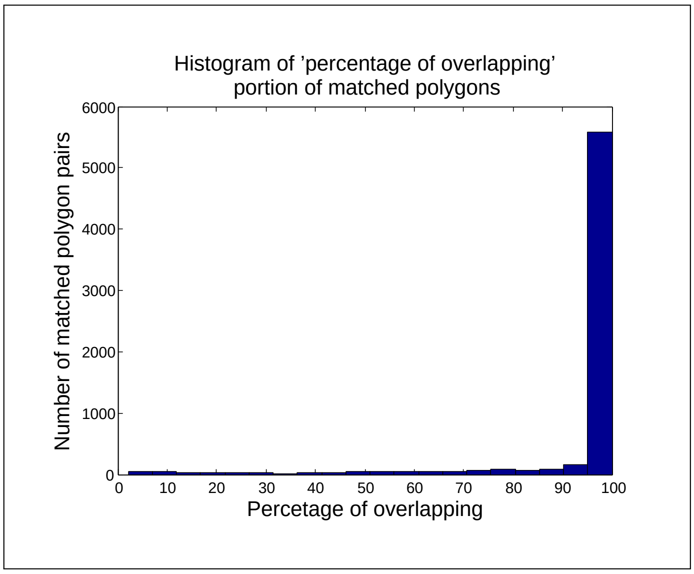


Figure 40: Histogram of percentage overlapping of matched polygon pairs

It was also found in very rare cases that identical water polygons from both datasets, which were a bit shifted and not overlapped, were excluded from the list of matched pairs. One example is given in Figure 41 (on page 66).

All object matching models for both line and polygon were executed afterwards in the opposite direction. The OSM objects were checked with the TOP10NL objects. The accuracy of results was found similar to that of matching TOP10NL objects with OSM objects. The matching correspondence '0-1' and 'n-m' for all objects could be detected by combining the bi-directional matching pairs. Some of the results shown in Appendix. Table 46 (on page 79) shows the matching table of TOP10NL. Here it is seen that the matched TOP10NL road objects got street name as the value of the attribute 'name'. Table 47 (on page

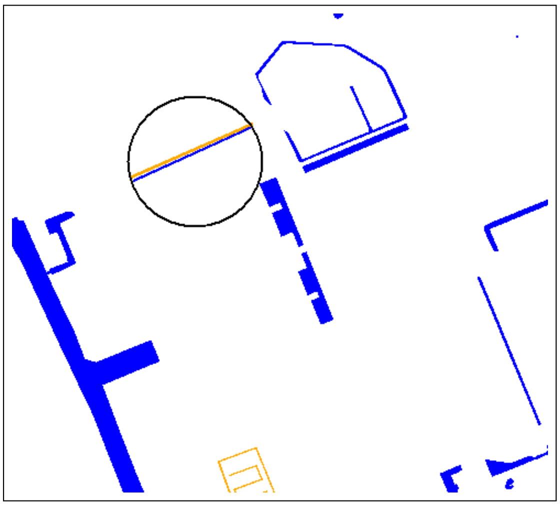


Figure 41: Bad matched water polygon

80) and 48 (on page 81) show one-to-many and many-to-many tables. And the complete updating table is shown in Table 49 (on page 82).

5 CONCLUSION AND RECOMMENDATION

5.1 CONCLUSIONS

The main research question of this thesis is:

"Is an automatic object matching of OSM in combination with TOP10NL possible? If yes, then to what extent this can be performed?"

As the result of this research it can be answered that 'yes, an automatic object matching of OSM in combination with TOP10NL is possible'. The accuracy and the possible level of automatic object matching of TOP10NL and OSM have been explained separately as the findings of the three research steps and are summarised below.

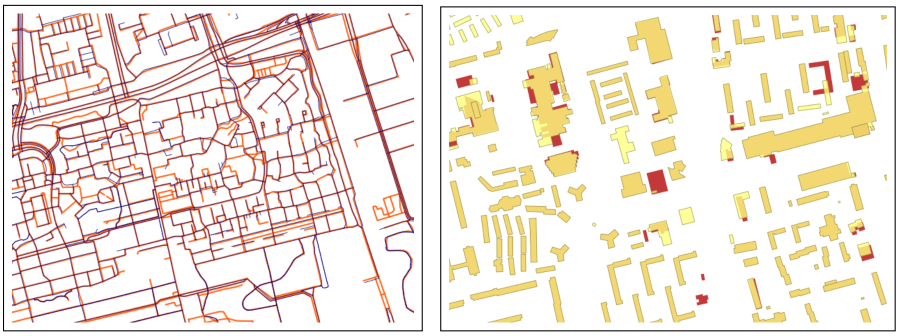
Schema matching: The OSM data model was successfully transformed to the similar data model type of TOP10NL. The ESRI shapefile format of TOP10NL database was chosen for this study. Different shapefiles were prepared with OSM data for different object classes according to match the corresponding TOP10NL shapefiles of different object classes.

The complete schema transformation process was performed in three parts, namely, semantic mapping, converting to ESRI shapefiles and coordinate transformation. The complete process was implemented using Python. The performance of the schema transformation was checked in two ways, by checking the accuracy of the coordinate transformation and by visualisation.

The accuracy of the coordinate transformation was checked by comparing with the 'Official Software' PCTrans [2],see Section 2.3.3 page 17. The average of absolute difference between the two transformation was 40cm, which is within the acceptable range as per van Buren *et al.* [65], .

By visualising the overlaid layers of transformed OSM data on the TOP10NL data of the same object class the performance of the complete schema transformation was verified. Figure 42 page 68 shows an example of overlaying road and building objects of TOP10NL and OSM after the schema translation. No shifts have been identified. The only identified difference was due to the different data acquisition techniques.

Semantic mapping of the object classes from TOP10NL and OSM was found as the most crucial part of the schema translation process. Object classes are mentioned in the tag values of all OSM objects and there are no content restriction on tags. As a result there is a high chance of having a desired OSM object with



(a) Example of overlaying of road network

(b) Example of overlaying of building polygons

Figure 42: Visualisation of the accuracy of schema translation of the OSM data

an unknown tag which is not listed in the semantic mapping list. In such a situation the extraction of the objects from OSM will not be done properly and the completeness result will be misleading.

Quality assessment of OSM: The quality of the OSM data over the pilot areas were checked and found quite good by comparing with the previous research over other different countries. In this research only 'completeness', 'positional accuracy' and 'lineage' of the OSM dataset were evaluated. The result shows that the completeness of the 'road_line', 'railway' and 'building' object classes are quite high though they are not homogeneously spread everywhere. The completeness of the OSM road data over the pilot areas varies from 70% to 102% and building data varies from 89% to 97% of the TOP10NL dataset. The presence of 'water' objects in the OSM dataset is quite poor. In some places the water bodies are not recorded at all. Combining the linear and polygonal water objects, the completeness varies from 25% to 100% of the TOP10NL dataset.

The completeness and positional accuracy of the OSM data was checked by comparing with the ground-truth TOP10NL dataset. The positional accuracy of all OSM objects was quite good in comparison to the guidelines of the previous researchers [34] [27], see Section 3.4,page 32. It was found that up to 100% of total line objects and polygon objects of the different pilot areas are lying within 6 meters of buffer zone of TOP10NL objects of the same class. From the data lineage study it was found that the TOP10NL and OSM share quite a large number of objects from the same origin, which is Top10vector. The reliable data sources such as '3dShapes' and 'AND' influenced the 'positional accuracy' of the OSM data over the Netherlands. From the overall result of quality assessment of the OSM dataset over the pilot areas it can be said that the quality of the OSM data is quite good where the 'completeness' is high. The 'completeness' was found as the most crucial factor in quality assessment of the OSM dataset over the Netherlands.

Object matching: Two different object matching models were built for line and polygon objects. A number of matching criteria from literature were tested and out of them a few were found to be useful for these type of datasets. Only 'Euclidean distance' and 'difference of angle of direction' for linear objects, and only 'surface distance' for polygonal objects were found to be suitable for an automatic matching process. Simple rules were built for both models.

The line matching model was trained and tested with the 'road network' data of the city of Delft and the polygon matching model was trained and tested with the 'building' data of the city of Delft. Finally, the models were verified with the linear and polygonal object classes data over the pilot areas. The detailed results are given in Section 4.4 page 63. In a nutshell it can be said that the percentage of total matched pairs depends on the 'completeness' of the OSM data of the study area. The percentage of correct matching among the matched pairs was reasonably high. For line objects it varied from 78% to 100% whereas for the polygonal objects it varied from 99% to 100% for all the pilot areas. Most of the false matching pairs were found due to their abrupt presence, which were even difficult to detect manually.

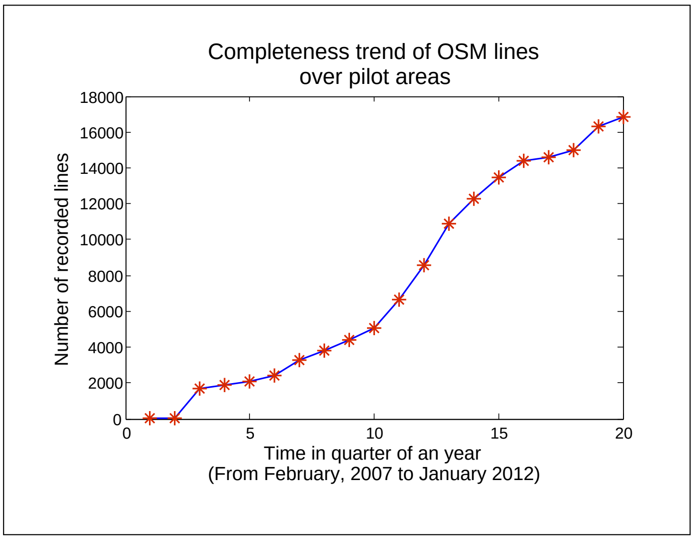
Finally it can be concluded that

1. The accuracy of automatic object matching process of both line and polygon features over the pilot areas are quite high. From the result of the pilot areas the quality of the OSM data over the Netherlands is anticipated to be high. As a result, the possibility of automatic object matching of OSM in combination with TOP10NL is high.

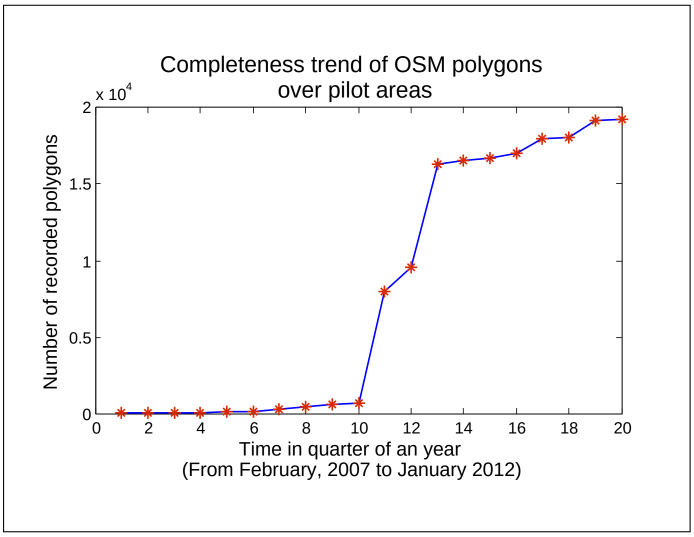
2. The accuracy of object matching process can be improved by more research and some ideas regarding that have been described as recommendations, see Section 5.3 page 71.

3. Future research on automatic updating of OSM and TOP10NL can also be successfully done when the OSM dataset is populated sufficiently. Therefore, higher 'completeness' result of the OSM dataset over the country is needed. During the last 5 years the 'completeness' of 'line' and 'polygon' data over the pilot areas are shown in Figure 43 page 70. This indicates that the OSM data recording will be more complete in the near future if the trend continues.

4. It is observed that the OSM dataset is still not complete over the entire Netherlands. It can be concluded that the OSM can be complete by automatically extracting the missing data of common object classes from the TOP10NL dataset by following this research.



(a) The growth of line objects



(b) The growth of polygon objects

Figure 43: The trend of the growth rate of presence of line and polygon data in OSM dataset over the pilot areas

5.2 CONTRIBUTIONS

The main contributions of this thesis are given below.

- The thesis has explored the possibility of making OSM more useful in the geo-information community. Many problems were identified and stated in Section 2.4 page 17 which show the irregular behaviour of the OSM dataset. This knowledge is helpful in building up a completely automatic object matching model as well as completely automatic updating model in combination with the OSM dataset.

- This thesis shows the way to enrich the TOP10NL 'road' dataset by inserting the 'Street names' automatically from OSM.

5.3 RECOMMENDATION

Some ideas for improving the accuracy of the object matching process and resolving the mismatches have been formulated and elaborated in this section. These might be helpful in future in implementing complete automatic updating process. They are given below.

Split the lines at junctions: Many problems arose during building a line matching model because of the dissimilarity of lengths in line objects of TOP10NL and OSM. It resulted many multiple correspondence (cardinality '1-n', 'n-1', 'n-m') in the matching pairs. The accuracy and usefulness of the matching process of road networks can be upgraded by changing the model to get '1-1' matching pairs. This is possible by breaking the OSM 'road_lines' objects at the junctions. This concept can be explored in two ways:

- the junctions can be identified by searching the nodes having more than two incident lines, or,
- the ground-truth dataset TOP10NL can be used to find out the junction of the OSM dataset. The point objects of 'WEGDEEL_PUNT' dataset of TOP10NL represents the junction nodes of the road networks. The junction nodes of the OSM road networks can be identified by searching the OSM nodes, which are nearest to the TOP10NL junction nodes. Figure 44 page 72 shows the geographical positions of the TOP10NL junction nodes and the vertices of the OSM road networks.

Checking the reliability of OSM with its dynamic behaviour: After completing the matching process all linear and polygonal objects from TOP10NL and OSM were classified by the classes '1-1', '1-0', '1-n','0-1' , 'n-1' or 'm-n' as per their appearance in the TOP10NL and OSM datasets. Most of the multiple correspondences '1-n', 'n-1' and 'm-n' happen due to the reason that TOP10NL and OSM use different data acquisition techniques. These could be resolved by converting them to '1-1'correspondence by splitting the road networks at their junctions. The crucial cases are '1-0'

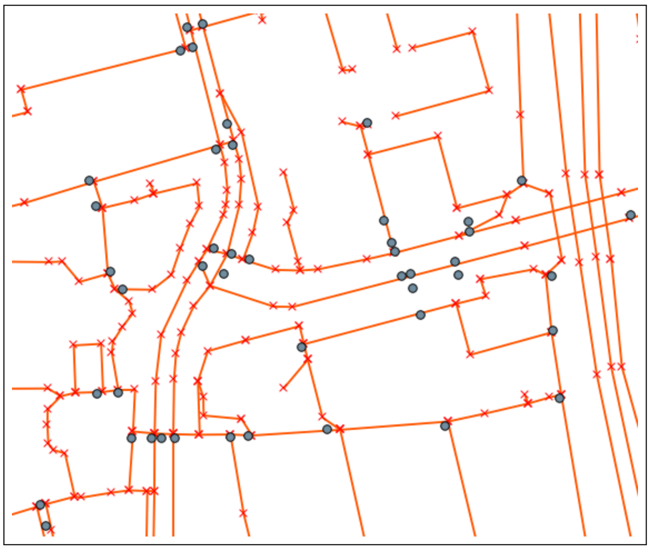


Figure 44: The overlaid picture of TOP10NL junction nodes (the round points) and the OSM road network (the orange lines). The 'x' points depicts the vertexes of the OSM line network

and '0-1', which should be checked for the reliability of the OSM quality. The checking can be done in the following three ways:

- by checking the presence of the object in the real world by going to the field,
- by checking with information from additional data sources maintained by different organisations,
- by checking the the changes of TOP10NL and OSM datasets in consecutive versions.

For checking with the consecutive versions at least 3 current versions of TOP10NL is needed. We can denote them as V_{t-1}, V_t and V_{t+1}, where V_t is the current version, V_{t-1} is the previous version and V_{t+1} is the next version of TOP10NL. The matching process is carried out with the current version of TOP10NL, V_t and at any time between the release of V_t and V_{t+1}. The idea for resolving the correspondences '1-0' and '0-1' are described below.

- While checking the presence of a TOP10NL object in the OSM dataset '1-0' indicates that the object is present in the TOP10NL dataset but not present in the OSM dataset. It might happen due to
  - by mistake in V_t version of TOP10NL or
  - deletion of the object after the last version V_t or
  - it has not yet been recorded in the OSM dataset.

The mistake can be checked with the previous version (V_{t-1}) of TOP10NL. And the deletion or absence can be confirmed by waiting for the next version (V_{t+1}) of TOP10NL.

- The '0-1' correspondence of a TOP10NL object indicates that the object is present in OSM but not in TOP10NL. It might happen due to the insertion of a new object in the real world after the last updates of TOP10NL (V_t), or, due to a mistake in the V_t version of TOP10NL. This situation can be confirmed by the next version of TOP10NL dataset.

Executing the process for resolving the mismatches: The matching process presented in this thesis resulted classified tables described in Section 4.4. Some examples are given in Tables 47, 48 and 49 in Appendix. Due to the time limitation, the mismatches from both datasets have not been resolved. The presence of an object in the real world in current time can be confirmed by the updating/recording dates of the objects. The recording dates are mentioned in one attribute value of all objects of both TOP10NL and OSM datasets. The recording date of TOP10NL objects is almost the same for all objects and is the last updating date of TOP10NL. But for OSM the recording date can be any day. The presence of any OSM object in the real world cannot be guaranteed only by the recording date. For example, if an OSM object is updated recently by the source '3dShapes', then it can be concluded that the data does not contain the latest information of that object. Because it has already been explained in Section 3.2 of chapter 3, that some data of '3dShapes' are acquired from 'Top10vector', which is the predecessor of TOP10NL. So the source of all OSM objects should be taken into account while deciding about the current situation of the OSM data.

Checking the quality of matched pairs for line objects: It is possible to make an initial list of potential matched pairs followed by Volz [68]. Some calculations thereafter may assign a quality indicator to all potential matched pairs without throwing the wrong matches from that list.

Integrating TOP10NL and OSM for building a single dataset might be a good option for getting an upgraded combined dataset.

Building more suitable sophisticated decision rules, for both linear and polygon matching models, may be investigated.

The polygons with hole were not taken into account in this study as they are not present in the datasets. Adding these aspects will help to make this matching model more robust.

The TOP10NL dataset can be enriched by adding some 'Point' object classes, such as rail-station, schools etc. by extracting them from the OSM dataset as OSM has lots of points as 'Point of Interest' (POI).

A

TABLES AND FIGURES

This appendix contains some tables and figures which are relevant to this study.

Name	Value	Description
type	*"‘node"* *"‘way"*	*type of the member*
ref	*integer*	*id of the member*
role	*string*	*role of the member. This depends on the use of the relation. Popular are multipolygon or route*

Table 17: Structure of OSM way primitive with member [21]

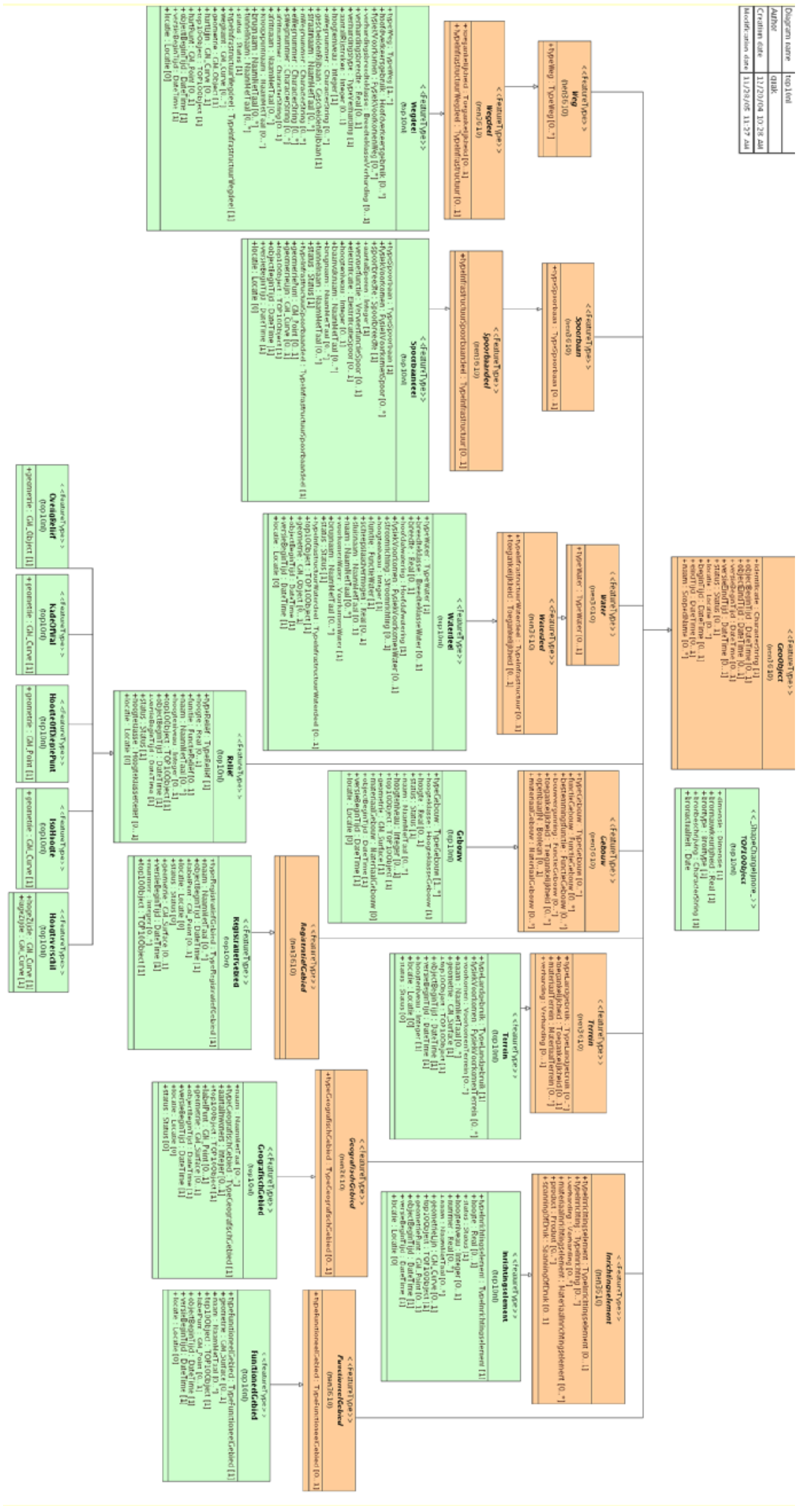


Figure 45: top10uML

Value	*Shape type*	*Field*
0	*Null shape*	None
1	Point	*X, Y*
3	Polyline	*MBR, Number of parts, Number of points, Parts, Points*
5	Polygon	*MBR, Number of parts, Number of points, Parts, Points*
8	Multipoint	*MBR, Number of points, Points*
11	PointZ	*X, Y, Z, M*
13	PolylineZ	*Mandatory: MBR, Number of parts, Number of points, Parts, Points, Z range, Z array* *Optional: M range, M array*
15	PolygonZ	*Mandatory: MBR, Number of parts, Number of points, Parts, Points, Z range, Z array* *Optional: M range, M array*
18	MultiPointZ	*Mandatory: MBR, Number of points, Points, Z range, Z array Optional: M range, M array*
21	PointM	*X, Y, M*
23	PointlineM	*Mandatory: MBR, Number of parts, Number of points, Parts, Points Optional: M range, M array*
25	PolygonM	*Mandatory: MBR, Number of parts, Number of points, Parts, Points Optional: M range, M array*
28	MultiPointM	*Mandatory: MBR, Number of points, Points*
31	MultiPatch	*Mandatory: MBR, Number of parts, Number of points, Parts, Part types, Points, Z range, Z array*

Table 18: The possible shape types, adapted from [26]

Name	value	Description
id	*integer>=1*	*Node id's are unique only between nodes. Editors tend to save these as negative to denote id's that haven't been saved to the server. Node ids on the server are persistent meaning that the assigned id of an existing node will remain unchanged when data is added or corrected. D node id's must not be reused, unless if the same node is undeleted.*
lat	*Float >-90.0 and <90.0 7 decimal places*	*Latitude coordinate. Number between Some applications may not accept latitudes above/below +-85.*
lon	*Float >-180 and <180 7 decimal* places	*Longitude coordinate.*
tags	*A set of key/value pairs, with unique key*	*As per the tagging guidelines*

Table 19: Structure of OSM node primitive [21]

Name	Value	Description
id	*integer >=1*	*Note that the ids are not unique, a way can have the same id as a node. Editors tend to save these as negative to denote ids that haven't been saved to the server.*
nodes	list	*A list of all node id's that make up the way.*
tags	set	*A set of tags (key/value pairs), with no key occurring twice. It follows the tagging guidelines*

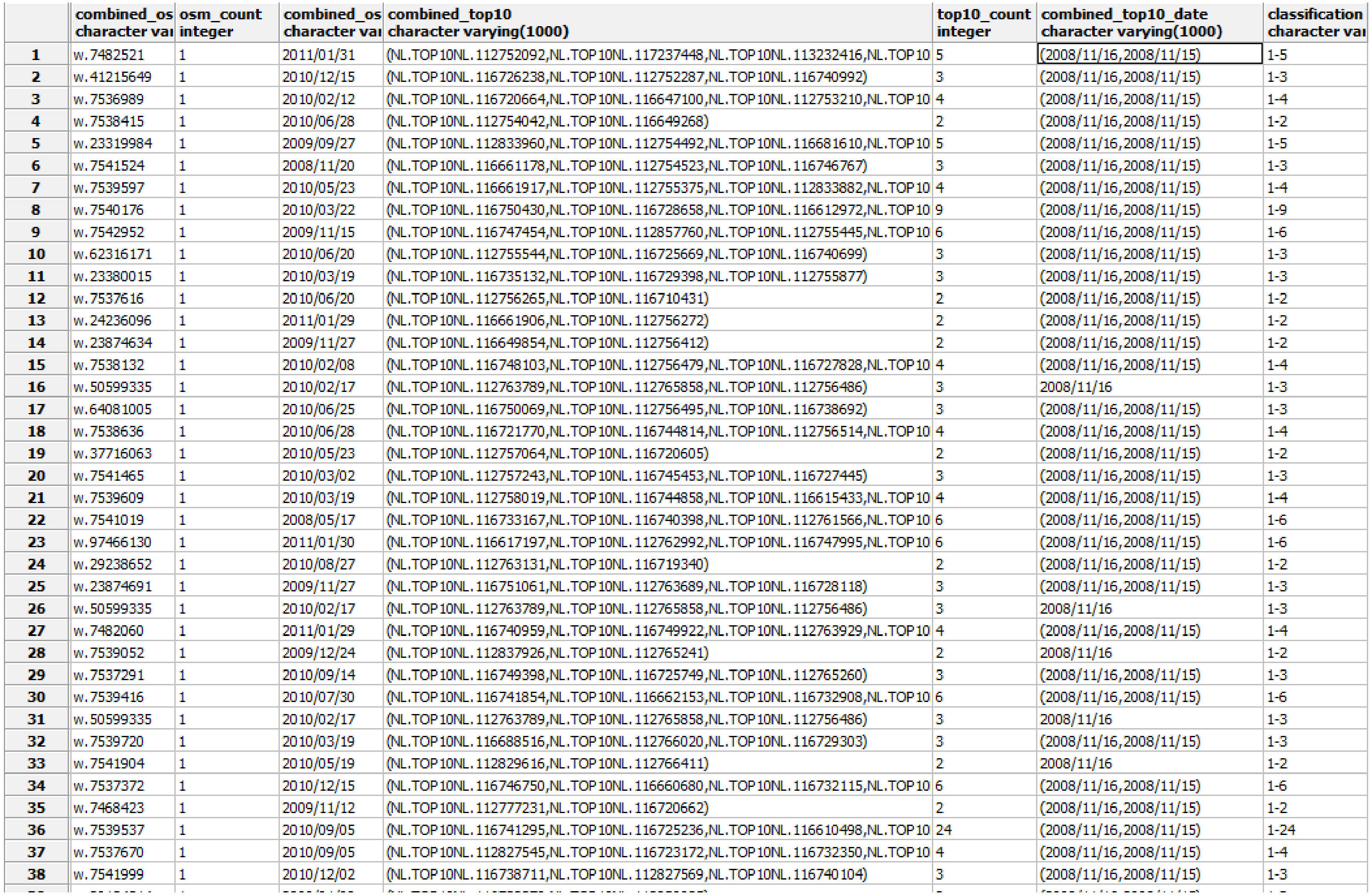
Table 20: Structure of OSM way primitive [21]

Name	Value	Description
id	*integer >=1*	*See above*
tags	|textitlist	*See above*
members	*list(ordered)*	*An ordered list of primitives with associated role attributes (where role may be any text).*

Table 21: Structure of OSM relation primitive [21]

	top10 character varying(30)	osm character varying(80)	name character varying(255)	top10_geom geometry	osm_geom geometry	top10_timestamp character varying(10)	osm_timestamp character varying(1
1	NL.TOP10NL.112752084	w.7539998	['Multatuliweg']	0102000000030	0102000000020	2008/11/16	2009/09/18
2	NL.TOP10NL.112752092	w.7482521	NA	0102000000040	0102000000060	2008/11/16	2011/01/31
3	NL.TOP10NL.112752287	w.41215649	['Delfgauwseweg']	0102000000030	0102000000080	2008/11/16	2010/12/15
4	NL.TOP10NL.112753210	w.7536989	['Pottenbakkerstraat']	0102000000030	0102000000060	2008/11/16	2010/02/12
5	NL.TOP10NL.112754042	w.7538415	['Julianalaan']	0102000000040	0102000000060	2008/11/16	2010/06/28
6	NL.TOP10NL.112754418	w.66137675	['Schieweg']	0102000000030	0102000000020	2008/11/16	2010/07/05
7	NL.TOP10NL.112754492	w.23319984	NA	0102000000030	01020000000A0	2008/11/16	2009/09/27
8	NL.TOP10NL.112754523	w.7541524	NA	0102000000030	0102000000020	2008/11/16	2008/11/20
9	NL.TOP10NL.112755375	w.7539597	['Faradayweg']	0102000000030	0102000000060	2008/11/16	2010/05/23
10	NL.TOP10NL.112755436	w.7540176	['Rietgorsstraat']	0102000000030	01020000000B0	2008/11/16	2010/03/22
11	NL.TOP10NL.112755445	w.7542952	['Boomkwekerij']	0102000000030	0102000000090	2008/11/16	2009/11/15
12	NL.TOP10NL.112755544	w.62316171	NA	0102000000030	0102000000020	2008/11/16	2010/06/20
13	NL.TOP10NL.112755877	w.23380015	['Boliviastraat']	0102000000030	0102000000020	2008/11/16	2010/03/19
14	NL.TOP10NL.112756265	w.7537616	NA	0102000000030	0102000000040	2008/11/16	2010/06/20
15	NL.TOP10NL.112756272	w.24236096	['Rijswijkse', 'Landingslaan']	0102000000030	0102000000030	2008/11/16	2011/01/29
16	NL.TOP10NL.112756412	w.23874634	['Molenwetering']	0102000000030	0102000000080	2008/11/16	2009/11/27
17	NL.TOP10NL.112756479	w.7538132	['Kanaalweg']	0102000000020	0102000000040	2008/11/16	2010/02/08
18	NL.TOP10NL.112756486	w.50599335	NA	0102000000040	0102000000040	2008/11/16	2010/02/17
19	NL.TOP10NL.112756495	w.64081005	NA	0102000000030	0102000000020	2008/11/16	2010/06/25
20	NL.TOP10NL.112756514	w.7538636	['Oosterstraat']	0102000000030	0102000000050	2008/11/16	2010/06/28
21	NL.TOP10NL.112756699	w.7541886	['Reinier', 'de', 'Graafweg']	0102000000030	01020000000C0	2008/11/16	2010/05/19
22	NL.TOP10NL.112756699	w.45830212	NA	0102000000030	0102000000050	2008/11/16	2010/02/17
23	NL.TOP10NL.112757064	w.37716063	NA	0102000000030	0102000000080	2008/11/16	2010/05/23
24	NL.TOP10NL.112757243	w.7541465	['Glenn', 'Millerstraat']	0102000000030	0102000000050	2008/11/16	2010/03/02
25	NL.TOP10NL.112758019	w.7539609	['Tanthofdreef']	0102000000030	0102000000190	2008/11/16	2010/03/19
26	NL.TOP10NL.112761566	w.7541019	NA	0102000000020	0102000000020	2008/11/16	2008/05/17
27	NL.TOP10NL.112762992	w.97466130	['Wiardi', 'Beckmanlaan']	0102000000030	0102000000080	2008/11/16	2011/01/30
28	NL.TOP10NL.112763061	w.7541348	['Mozartlaan']	0102000000030	0102000000020	2008/11/16	2010/08/05
29	NL.TOP10NL.112763102	w.54078017	['Meester', 'Beerninkplantsoen']	0102000000030	0102000000050	2008/11/16	2011/01/29
30	NL.TOP10NL.112763102	w.97285625	['Meester', 'Beerninkplantsoen']	0102000000030	0102000000040	2008/11/16	2011/01/29
31	NL.TOP10NL.112763122	w.7539332	NA	0102000000030	0102000000070	2008/11/16	2010/10/11
32	NL.TOP10NL.112763122	w.7539319	NA	0102000000030	0102000000040	2008/11/16	2010/02/01
33	NL.TOP10NL.112763131	w.29238652	['Duke', 'Ellingtonstraat']	0102000000030	0102000000040	2008/11/16	2010/08/27
34	NL.TOP10NL.112763612	w.48688450	['Delfgauwseweg']	0102000000020	0102000000060	2008/11/16	2010/02/02
35	NL.TOP10NL.112763689	w.23874691	NA	0102000000030	0102000000060	2008/11/16	2009/11/27
36	NL.TOP10NL.112763789	w.50599335	NA	0102000000030	0102000000040	2008/11/16	2010/02/17
37	NL.TOP10NL.112763929	w.7482060	['Meester', 'Beerninkplantsoen']	0102000000030	0102000000070	2008/11/16	2011/01/29
38	NL.TOP10NL.112764851	w.7482974	['Waterbiesweg']	0102000000030	0102000000030	2008/11/16	2010/12/12

Figure 46: TOP10NL to OSM line matching table



	combined_os character vai	osm_count integer	combined_os character vai	combined_top10 character varying(1000)	top10_count integer	combined_top10_date character varying(1000)	classification character vai
1	w.7482521	1	2011/01/31	(NL.TOP10NL.112752092,NL.TOP10NL.117237448,NL.TOP10NL.113232416,NL.TOP10	5	(2008/11/16,2008/11/15)	1-5
2	w.41215649	1	2010/12/15	(NL.TOP10NL.116726238,NL.TOP10NL.112752287,NL.TOP10NL.116740992)	3	(2008/11/16,2008/11/15)	1-3
3	w.7536989	1	2010/02/12	(NL.TOP10NL.116720664,NL.TOP10NL.116647100,NL.TOP10NL.112753210,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
4	w.7538415	1	2010/06/28	(NL.TOP10NL.112754042,NL.TOP10NL.116649268)	2	(2008/11/16,2008/11/15)	1-2
5	w.23319984	1	2009/09/27	(NL.TOP10NL.112833960,NL.TOP10NL.112754492,NL.TOP10NL.116681610,NL.TOP10	5	(2008/11/16,2008/11/15)	1-5
6	w.7541524	1	2008/11/20	(NL.TOP10NL.116661178,NL.TOP10NL.112754523,NL.TOP10NL.116746767)	3	(2008/11/16,2008/11/15)	1-3
7	w.7539597	1	2010/05/23	(NL.TOP10NL.116661917,NL.TOP10NL.112755375,NL.TOP10NL.112833882,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
8	w.7540176	1	2010/03/22	(NL.TOP10NL.116750430,NL.TOP10NL.116728658,NL.TOP10NL.116612972,NL.TOP10	9	(2008/11/16,2008/11/15)	1-9
9	w.7542952	1	2009/11/15	(NL.TOP10NL.116747454,NL.TOP10NL.112857760,NL.TOP10NL.112755445,NL.TOP10	6	(2008/11/16,2008/11/15)	1-6
10	w.62316171	1	2010/06/20	(NL.TOP10NL.112755544,NL.TOP10NL.116725669,NL.TOP10NL.116740699)	3	(2008/11/16,2008/11/15)	1-3
11	w.23380015	1	2010/03/19	(NL.TOP10NL.116735132,NL.TOP10NL.116729398,NL.TOP10NL.112755877)	3	(2008/11/16,2008/11/15)	1-3
12	w.7537616	1	2010/06/20	(NL.TOP10NL.112756265,NL.TOP10NL.116710431)	2	(2008/11/16,2008/11/15)	1-2
13	w.24236096	1	2011/01/29	(NL.TOP10NL.116661906,NL.TOP10NL.112756272)	2	(2008/11/16,2008/11/15)	1-2
14	w.23874634	1	2009/11/27	(NL.TOP10NL.116649854,NL.TOP10NL.112756412)	2	(2008/11/16,2008/11/15)	1-2
15	w.7538132	1	2010/02/08	(NL.TOP10NL.116748103,NL.TOP10NL.112756479,NL.TOP10NL.116727828,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
16	w.50599335	1	2010/02/17	(NL.TOP10NL.112763789,NL.TOP10NL.112765858,NL.TOP10NL.112756486)	3	2008/11/16	1-3
17	w.64081005	1	2010/06/25	(NL.TOP10NL.116750069,NL.TOP10NL.112756495,NL.TOP10NL.116738692)	3	(2008/11/16,2008/11/15)	1-3
18	w.7538636	1	2010/06/28	(NL.TOP10NL.116721770,NL.TOP10NL.116744814,NL.TOP10NL.112756514,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
19	w.37716063	1	2010/05/23	(NL.TOP10NL.112757064,NL.TOP10NL.116720605)	2	(2008/11/16,2008/11/15)	1-2
20	w.7541465	1	2010/03/02	(NL.TOP10NL.112757243,NL.TOP10NL.116745453,NL.TOP10NL.116727445)	3	(2008/11/16,2008/11/15)	1-3
21	w.7539609	1	2010/03/19	(NL.TOP10NL.112758019,NL.TOP10NL.116744858,NL.TOP10NL.116615433,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
22	w.7541019	1	2008/05/17	(NL.TOP10NL.116733167,NL.TOP10NL.116740398,NL.TOP10NL.112761566,NL.TOP10	6	(2008/11/16,2008/11/15)	1-6
23	w.97466130	1	2011/01/30	(NL.TOP10NL.116617197,NL.TOP10NL.112762992,NL.TOP10NL.116747995,NL.TOP10	6	(2008/11/16,2008/11/15)	1-6
24	w.29238652	1	2010/08/27	(NL.TOP10NL.112763131,NL.TOP10NL.116719340)	2	(2008/11/16,2008/11/15)	1-2
25	w.23874691	1	2009/11/27	(NL.TOP10NL.116751061,NL.TOP10NL.112763689,NL.TOP10NL.116728118)	3	(2008/11/16,2008/11/15)	1-3
26	w.50599335	1	2010/02/17	(NL.TOP10NL.112763789,NL.TOP10NL.112765858,NL.TOP10NL.112756486)	3	2008/11/16	1-3
27	w.7482060	1	2011/01/29	(NL.TOP10NL.116740959,NL.TOP10NL.116749922,NL.TOP10NL.112763929,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
28	w.7539052	1	2009/12/24	(NL.TOP10NL.112837926,NL.TOP10NL.112765241)	2	2008/11/16	1-2
29	w.7537291	1	2010/09/14	(NL.TOP10NL.116749398,NL.TOP10NL.116725749,NL.TOP10NL.112765260)	3	(2008/11/16,2008/11/15)	1-3
30	w.7539416	1	2010/07/30	(NL.TOP10NL.116741854,NL.TOP10NL.116662153,NL.TOP10NL.116732908,NL.TOP10	6	(2008/11/16,2008/11/15)	1-6
31	w.50599335	1	2010/02/17	(NL.TOP10NL.112763789,NL.TOP10NL.112765858,NL.TOP10NL.112756486)	3	2008/11/16	1-3
32	w.7539720	1	2010/03/19	(NL.TOP10NL.116688516,NL.TOP10NL.112766020,NL.TOP10NL.116729303)	3	(2008/11/16,2008/11/15)	1-3
33	w.7541904	1	2010/05/19	(NL.TOP10NL.112829616,NL.TOP10NL.112766411)	2	2008/11/16	1-2
34	w.7537372	1	2010/12/15	(NL.TOP10NL.116746750,NL.TOP10NL.116660680,NL.TOP10NL.116732115,NL.TOP10	6	(2008/11/16,2008/11/15)	1-6
35	w.7468423	1	2009/11/12	(NL.TOP10NL.112777231,NL.TOP10NL.116720662)	2	(2008/11/16,2008/11/15)	1-2
36	w.7539537	1	2010/09/05	(NL.TOP10NL.116741295,NL.TOP10NL.116725236,NL.TOP10NL.116610498,NL.TOP10	24	(2008/11/16,2008/11/15)	1-24
37	w.7537670	1	2010/09/05	(NL.TOP10NL.112827545,NL.TOP10NL.116723172,NL.TOP10NL.116732350,NL.TOP10	4	(2008/11/16,2008/11/15)	1-4
38	w.7541999	1	2010/12/02	(NL.TOP10NL.116738711,NL.TOP10NL.112827569,NL.TOP10NL.116740104)	3	(2008/11/16,2008/11/15)	1-3

Figure 47: One to many classification table

	combined_osm character varying(1000)	osm_count integer	combined_osm_date character varying(1000)	combined_top10 character varying(1000)	top10_count integer	combined_top10_date character varying(1000	classification character va
1	(w.45830212,w.7541886)	2	(2010/05/19,2010/02/17)	(NL.TOP10NL.116726655,NL.TOP10NL.116727732,NL.TOP10NL.116749243,NL.T	7	(2008/11/16,2008/11/15)	2-7
2	(w.54078017,w.97285625)	2	2011/01/29	(NL.TOP10NL.116740133,NL.TOP10NL.116671133,NL.TOP10NL.116743095,NL.T	5	(2008/11/16,2008/11/15)	2-5
3	(w.7539332,w.7539319)	2	(2010/02/01,2010/10/11)	(NL.TOP10NL.116736680,NL.TOP10NL.116647937,NL.TOP10NL.116749817,NL.T	4	(2008/11/16,2008/11/15)	2-4
4	(w.7539439,w.7539445)	2	(2010/07/30,2010/02/03)	(NL.TOP10NL.116662679,NL.TOP10NL.116722746,NL.TOP10NL.112776237,NL.T	4	(2008/11/16,2008/11/15)	2-4
5	(w.59619623,w.7541904)	2	2010/05/19	(NL.TOP10NL.112829616,NL.TOP10NL.116726119,NL.TOP10NL.112766411)	3	(2008/11/16,2008/11/15)	2-3
6	(w.7539913,w.7540201)	2	(2009/10/27,2010/01/15)	(NL.TOP10NL.116689990,NL.TOP10NL.116746860,NL.TOP10NL.112830419)	3	(2008/11/16,2008/11/15)	2-3
7	(w.7543244,w.7543240)	2	(2010/08/05,2010/06/12)	(NL.TOP10NL.112834123,NL.TOP10NL.116721166,NL.TOP10NL.116698517)	3	(2008/11/16,2008/11/15)	2-3
8	(w.7540124,w.7539670)	2	2010/03/20	(NL.TOP10NL.116732102,NL.TOP10NL.116715802,NL.TOP10NL.116616971,NL.T	29	(2008/11/16,2008/11/15)	2-29
9	(w.7542951,w.7542898)	2	2009/11/15	(NL.TOP10NL.112834871,NL.TOP10NL.116649402,NL.TOP10NL.116747362,NL.T	5	(2008/11/16,2008/11/15)	2-5
10	(w.7539077,w.7539071)	2	(2008/11/12,2009/10/31)	(NL.TOP10NL.112836674,NL.TOP10NL.116620114,NL.TOP10NL.116751763,NL.T	4	(2008/11/16,2008/11/15)	2-4
11	(w.48688449,w.7537435)	2	(2010/02/02,2010/01/22)	(NL.TOP10NL.116734917,NL.TOP10NL.112836732,NL.TOP10NL.116612721)	3	(2008/11/16,2008/11/15)	2-3
12	(w.7539076,w.7538659)	2	2009/12/06	(NL.TOP10NL.112840959,NL.TOP10NL.116616001)	2	(2008/11/16,2008/11/15)	2-2
13	(w.7540462,w.7540512)	2	(2010/02/17,2010/08/24)	(NL.TOP10NL.112994214,NL.TOP10NL.116662759,NL.TOP10NL.112848651)	3	(2008/11/16,2008/11/15)	2-3
14	(w.7539146,w.28252820)	2	(2010/02/11,2010/02/17)	(NL.TOP10NL.116672415,NL.TOP10NL.112851691,NL.TOP10NL.116648789,NL.T	5	(2008/11/16,2008/11/15)	2-5
15	(w.7467919,w.21059085)	2	(2009/02/19,2010/09/16)	(NL.TOP10NL.113023665,NL.TOP10NL.116750008,NL.TOP10NL.112851742,NL.T	8	(2008/11/16,2008/11/15)	2-8
16	(w.7539279,w.7538486)	2	(2009/11/17,2010/02/08)	(NL.TOP10NL.112854064,NL.TOP10NL.116724628,NL.TOP10NL.112833597)	3	(2008/11/16,2008/11/15)	2-3
17	(w.59619609,w.76254565)	2	(2010/05/19,2010/09/06)	(NL.TOP10NL.112854195,NL.TOP10NL.112852868,NL.TOP10NL.116721161)	3	(2008/11/16,2008/11/15)	2-3
18	(w.7540754,w.7540746)	2	(2007/09/20,2010/06/26)	(NL.TOP10NL.112855750,NL.TOP10NL.116740000,NL.TOP10NL.116730031,NL.T	4	(2008/11/16,2008/11/15)	2-4
19	(w.7541022,w.7540857)	2	2007/09/30	(NL.TOP10NL.112856182,NL.TOP10NL.116745779,NL.TOP10NL.116726562)	3	(2008/11/16,2008/11/15)	2-3
20	(w.7540200,w.7539872)	2	(2009/11/02,2007/09/20)	(NL.TOP10NL.112856694,NL.TOP10NL.116749428,NL.TOP10NL.116731755)	3	(2008/11/16,2008/11/15)	2-3
21	(w.65980009,w.65980008)	2	(2010/07/31,2010/09/21)	(NL.TOP10NL.116659605,NL.TOP10NL.112858354)	2	(2008/11/16,2008/11/15)	2-2
22	(w.7536978,w.7536981)	2	(2010/02/12,2010/12/15)	(NL.TOP10NL.116711534,NL.TOP10NL.116615423,NL.TOP10NL.116729640,NL.T	6	(2008/11/16,2008/11/15)	2-6
23	(w.7468671,w.7468664)	2	(2010/07/31,2009/03/26)	(NL.TOP10NL.116733390,NL.TOP10NL.112859082,NL.TOP10NL.116616480)	3	(2008/11/16,2008/11/15)	2-3
24	(w.23401139,w.59619626,w.28575577)	3	(2010/05/19,2009/10/22)	(NL.TOP10NL.112840977,NL.TOP10NL.112863693,NL.TOP10NL.116729567)	3	(2008/11/16,2008/11/15)	3-3
25	(w.37800697,w.7537770)	2	(2010/05/29,2010/02/02)	(NL.TOP10NL.116743176,NL.TOP10NL.116746221,NL.TOP10NL.112865738,NL.T	4	(2008/11/16,2008/11/15)	2-4
26	(w.7537590,w.7537541)	2	(2010/02/04,2007/11/23)	(NL.TOP10NL.112909771,NL.TOP10NL.116658542,NL.TOP10NL.116725467)	3	(2008/11/16,2008/11/15)	2-3
27	(w.60507482,w.7537971)	2	2010/05/29	(NL.TOP10NL.116723315,NL.TOP10NL.116730028,NL.TOP10NL.112912664,NL.T	5	(2008/11/16,2008/11/15)	2-5
28	(w.7539470,w.7539475)	2	2010/01/07	(NL.TOP10NL.112855699,NL.TOP10NL.112913304,NL.TOP10NL.116742416,NL.T	4	(2008/11/16,2008/11/15)	2-4
29	(w.7536781,w.7536824)	2	(2007/09/20,2009/03/31)	(NL.TOP10NL.116631505,NL.TOP10NL.112842524,NL.TOP10NL.112913658,NL.T	8	(2008/11/16,2008/11/15)	2-8
30	(w.60507464,w.60507496)	2	2010/05/29	(NL.TOP10NL.112914347,NL.TOP10NL.116745726)	2	(2008/11/16,2008/11/15)	2-2
31	(w.7540956,w.7540951)	2	2007/09/20	(NL.TOP10NL.116726102,NL.TOP10NL.112916779,NL.TOP10NL.116736883)	3	(2008/11/16,2008/11/15)	2-3
32	(w.7542715,w.7542713)	2	(2009/10/10,2010/08/05)	(NL.TOP10NL.112917658,NL.TOP10NL.112837521)	2	2008/11/16	2-2
33	(w.7540584,w.7540883)	2	2007/11/13	(NL.TOP10NL.113017908,NL.TOP10NL.116616328,NL.TOP10NL.112954561)	3	(2008/11/16,2008/11/15)	2-3
34	(w.7541976,w.7542181)	2	(2010/12/02,2009/09/27)	(NL.TOP10NL.113006725,NL.TOP10NL.116728383)	2	(2008/11/16,2008/11/15)	2-2
35	(w.7542866,w.7542824)	2	(2009/12/08,2010/02/17)	(NL.TOP10NL.116723884,NL.TOP10NL.113013728,NL.TOP10NL.116751640)	3	(2008/11/16,2008/11/15)	2-3
36	(w.7537223,w.7537221)	2	(2007/09/20,2010/02/02)	(NL.TOP10NL.113014345,NL.TOP10NL.116625075)	2	(2008/11/16,2008/11/15)	2-2
37	(w.7539704,w.53187782)	2	(2009/12/17,2010/03/23)	(NL.TOP10NL.116736767,NL.TOP10NL.113014699)	2	(2008/11/16,2008/11/15)	2-2

Figure 48: Many to many classification table

	top10 character varying(30)	top10_date character vai	combined_os character vai	osm_count integer	combined_os character vai	top10_count integer	combined_top10 character varying(1000)	classification character vai
1	NL.TOP10NL.112752084	2008/11/16	w.7539998	1	2009/09/18	1	NL.TOP10NL.112752084	1-1
2	NL.TOP10NL.112752092	2008/11/16	w.7482521	1	2011/01/31	5	(NL.TOP10NL.112752092,NL.TOP10NL.117237448,NL.	1-5
3	NL.TOP10NL.112752261	2008/11/16		0		1		1-0
4	NL.TOP10NL.112752287	2008/11/16	w.41215649	1	2010/12/15	3	(NL.TOP10NL.116726238,NL.TOP10NL.112752287,NL.	1-3
5	NL.TOP10NL.112753207	2008/11/16		0		1		1-0
6	NL.TOP10NL.112753210	2008/11/16	w.7536989	1	2010/02/12	4	(NL.TOP10NL.116720664,NL.TOP10NL.116647100,NL.	1-4
7	NL.TOP10NL.112753583	2008/11/16		0		1		1-0
8	NL.TOP10NL.112753641	2008/11/16		0		1		1-0
9	NL.TOP10NL.112754042	2008/11/16	w.7538415	1	2010/06/28	2	(NL.TOP10NL.112754042,NL.TOP10NL.116649268)	1-2
10	NL.TOP10NL.112754390	2008/11/16		0		1		1-0
11	NL.TOP10NL.112754417	2008/11/16		0		1		1-0
12	NL.TOP10NL.112754418	2008/11/16	w.66137675	1	2010/07/05	1	NL.TOP10NL.112754418	1-1
13	NL.TOP10NL.112754439	2008/11/16		0		1		1-0
14	NL.TOP10NL.112754492	2008/11/16	w.23319984	1	2009/09/27	5	(NL.TOP10NL.112833960,NL.TOP10NL.112754492,NL.	1-5
15	NL.TOP10NL.112754523	2008/11/16	w.7541524	1	2008/11/20	3	(NL.TOP10NL.116661178,NL.TOP10NL.112754523,NL.	1-3
16	NL.TOP10NL.112754791	2008/11/16		0		1		1-0
17	NL.TOP10NL.112755375	2008/11/16	w.7539597	1	2010/05/23	4	(NL.TOP10NL.116661917,NL.TOP10NL.112755375,NL.	1-4
18	NL.TOP10NL.112755385	2008/11/16		0		1		1-0
19	NL.TOP10NL.112755436	2008/11/16	w.7540176	1	2010/03/22	9	(NL.TOP10NL.116750430,NL.TOP10NL.116728658,NL.	1-9
20	NL.TOP10NL.112755445	2008/11/16	w.7542952	1	2009/11/15	6	(NL.TOP10NL.116747454,NL.TOP10NL.112857760,NL.	1-6
21	NL.TOP10NL.112755466	2008/11/16		0		1		1-0
22	NL.TOP10NL.112755544	2008/11/16	w.62316171	1	2010/06/20	3	(NL.TOP10NL.112755544,NL.TOP10NL.116725669,NL.	1-3
23	NL.TOP10NL.112755604	2008/11/16		0		1		1-0
24	NL.TOP10NL.112755647	2008/11/16		0		1		1-0
25	NL.TOP10NL.112755688	2008/11/16		0		1		1-0
26	NL.TOP10NL.112755877	2008/11/16	w.23380015	1	2010/03/19	3	(NL.TOP10NL.116735132,NL.TOP10NL.116729398,NL.	1-3
27	NL.TOP10NL.112756265	2008/11/16	w.7537616	1	2010/06/20	2	(NL.TOP10NL.112756265,NL.TOP10NL.116710431)	1-2
28	NL.TOP10NL.112756272	2008/11/16	w.24236096	1	2011/01/29	2	(NL.TOP10NL.116661906,NL.TOP10NL.112756272)	1-2
29	NL.TOP10NL.112756412	2008/11/16	w.23874634	1	2009/11/27	2	(NL.TOP10NL.116649854,NL.TOP10NL.112756412)	1-2
30	NL.TOP10NL.112756469	2008/11/16		0		1		1-0
31	NL.TOP10NL.112756479	2008/11/16	w.7538132	1	2010/02/08	4	(NL.TOP10NL.116748103,NL.TOP10NL.112756479,NL.	1-4
32	NL.TOP10NL.112756486	2008/11/16	w.50599335	1	2010/02/17	3	(NL.TOP10NL.112763789,NL.TOP10NL.112765858,NL.	1-3
33	NL.TOP10NL.112756495	2008/11/16	w.64081005	1	2010/06/25	3	(NL.TOP10NL.116750069,NL.TOP10NL.112756495,NL.	1-3
34	NL.TOP10NL.112756514	2008/11/16	w.7538636	1	2010/06/28	4	(NL.TOP10NL.116721770,NL.TOP10NL.116744814,NL.	1-4
35	NL.TOP10NL.112756682	2008/11/16		0		1		1-0
36	NL.TOP10NL.112756699	2008/11/16	(w.45830212,w	2	(2010/05/19,20	7	(NL.TOP10NL.116726655,NL.TOP10NL.116727732,NL.	2-7
37	NL.TOP10NL.112757064	2008/11/16	w.37716063	1	2010/05/23	2	(NL.TOP10NL.112757064,NL.TOP10NL.116720605)	1-2
38	NL.TOP10NL.112757205	2008/11/16		0		1		1-0

Figure 49: Combined table for all types of matching pairs

[1] Creating a data mart for floating car data, project activity report m1-m24. available in textithttp: cordis.europa.eu/documents/documentlibrary/123823881EN19.doc.

[2] Pctrans software for map projections and datum transformations.

[3] Productspecificaties basisregistratie topografie. Technical report, Kadaster The Netherlands, 2007.

[4] Alashgeo. Steve coast joins microsoft. available online at http://slashgeo.org/2010/11/25/OpenStreetMap-Founder-Steve-Coast-Joins-Microsoft.

[5] S. Anand, J. Morley, W. Jiang, H. Du, G. Hart, and M. Jackson. When worlds collide: combining ordnance survey and open street map data. In *AGI GeoCommunity '10: Opportunities in a Changing World 'Innovate - Connect - Succeed'*, AGI GeoCommunity '10: Opportunities in a Changing World 'Innovate - Connect - Succeed', 2010.

[6] A. Ather. A quality analysis of openstreetmap data. Master's thesis, University College London, 2009.

[7] N. Bakker and B. Kolk. Top10nl, a new object-oriented topographical database in gml. In *21st International Cartographic Conference (ICC), Cartographic Renaissance*, pages 841–851, 2003.

[8] N.J. Bakker. Presentations on top10nl. available online at www.gdmc.nl/events/relay4/presentations/TOP10NL.ppt.

[9] T. Bont. Geographic data integration for telecommunication purposes. Master's thesis, Delft University of Technology, 2006.

[10] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *31st international conference on Very large data bases*, Trondheim, Norway, 2005.

[11] D. Chen, A. Driemel, L. J. Guibas, A. Nguyen, and C. Wenk. Approximate map matching with respect to the fréchet distance. In *ALENEX*, pages 75–83, 2011.

[12] B. Ciepluch, P. Mooney, and A. Winstanley. Building generic quality indicators for openstreetmap. In *Proceedings of the 11th International Conference on GeoComputation*, London, England, 2011.

[13] B. Ciepluchuch, P. Mooney, R. Jacob, and A. Winstanley. Sketches of generic framework for quality assessment of volunteered geographical data. In *IGARSS 2011*, Sendai, August 2011.

[14] OpenStreetMap Wiki contributors. 3d shapes for openstreetmap. available online on http://wiki.openstreetmap.org/wiki/3dShapes.

[15] OpenStreetMap Wiki contributors. Details of and data. available online at http://wiki.openstreetmap.org/wiki/AND_Data.

[16] OpenStreetMap Wiki contributors. Openstreetmap. available on http://www.openstreetmap.org/.

[17] OpenStreetMap Wiki contributors. OpenStreetMap user 'cetest'. available online at http://wiki.openstreetmap.org/wiki/User:Cetest.

[18] OpenStreetMap Wiki contributors. Openstreetmap user "polderrunner". available online at http://wiki.openstreetmap.org/wiki/User:Polderrunner.

[19] OpenStreetMap Wiki contributors. Openstreetmap user 'textlijn'. available online at http://www.openstreetmap.org/user/textlijn.

[20] OpenStreetMap Wiki contributors. Openstreetmap user 'theun'. available online at http://wiki.openstreetmap.org/wiki/User:Theun.

[21] OpenStreetMap Wiki contributors. Openstreetmap wiki. available online at http://wiki.openstreetmap.org/wiki/Main_Page.

[22] L. Deren, S. Haigang, and X. Ping. Automatic change detection of geo-spatial data from imagery. *Geo-spatial Information Science (Quarterly)*, 6(3):1–7, 2003.

[23] I. Dowman. Automated procedures for integration of satellite images and map data for change detection: The archangel project. *International Archives of Photogrammetry and Remote Sensing*, 32(4):162–169, 1998.

[24] T. Eiter and H. Mannila. Computing discrete frechet distance. Technical Report 94/64, Christian Doppler Laboratory, Vienna University of Technology, 1994.

[25] ESRI. An esri white paper. available online at http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, July 1998.

[26] Wikimedia Foundation. Shapefile. available online at http://en.wikipedia.org/wiki/Shapefile.

[27] J. Girres and G. Touya. Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14, Issue 4:435459, 2010.

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

[29] M. F. Goodchild. Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2:24–32, 2007.

[30] M.F. Goodchild. Assertion and authority: the science of user-generated geographic content. In *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday. GeoInfo 39*, Vienna University of Technology, 2008. Department of Geoinformation and Cartography.

[31] S. Haigang and L. Deren. A framework for automated change detection system. *Geo-spatial Information Science*, 4(3):29–34, September 2001.

[32] S. Haigang, L. Deren, and G. Jianya. Automatic change detection for road networks from images based on gis. *Geo-spatial Information Science (Quarterly)*, 6(4):44–50, December 2003.

[33] S. Haigang, L. Deren, G. Jianya, and Z. Qing. Analysis and representation of change in change detection. *Geo-spatial Information Scienceo-spatial Information Science (quarterly)*, 5(2):13–16, June 2002.

[34] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. In *Environment and Planning B: Planning and Design*, 2008.

[35] M. Haklay and C. Ellul. Completeness in volunteered geographical information - the evolution of openstreetmap coverage in england (2008-2009). *Journal of Spatial Information Science*, under review, 2010.

[36] M. Haklay and P. Weber. Openstreetmap:user-generated street maps. *Pervasive Computing, IEEE*, 7:12–18, 2008.

[37] H. Hild and D. Fritsch. Integration of vector data and satellite imagery for geocoding. *International Archives of Photogrammetry and Remote Sensing*, 32:246–251, 1998.

[38] J. HORAK, L. JURIKOVSKA, M. UMLAUF, M. HANZLOVA J. JEZEK, and S. ZLATANOVA. Humboldt alignment editor. In *Proceedings - Symposium GIS Ostrava 2011*, 2011.

[39] Dutch Kadaster. Available in http://www.kadaster.nl/window.html?inhoud=/international-english/default.html.

[40] O. Kiunadi. Assessing the quality of openstreetmap data. Master's thesis, University College of London, 2009.

[41] D. Klang. Automatic detection of changes in road databases using satellite imagery. *International Archives of Photogrammetry and Remote Sensing*, 32(4):293–298, 1998.

[42] T. Knudsen and Brian P. Olsen. Automated change detection for updates of digital map databases. *JOURNAL OF THE AMERICAN SOCIETY FOR PHOTOGRAMMETRY AND REMOTE SENSING*, (5):1289–1296, November 2003.

[43] P.G. Lentjes. Evaluatie pionierdataset top10nl. CGI-rapport 06-009 Alterra-rapport 1343, Alterra, 2006.

[44] Creative Commons Attribution 3.0 License. Creative commons. available online at http://creativecommons.org/licenses/by-sa/2.0/.

[45] I. Ludwig, A. Voss, and M. Krause-Traudes. A comparison of the street networks of navteq and osm in germany. In *Tha Association of Geographic Information Laboratories for Europe (AGILE) 2011*, pages 65–84, 2011.

[46] LUMOS. Land use modeling system (lumos). available online at www.lumos.info.

[47] S. Manoah, O.Boucelma, and Y.lassoued. Schema matching in gis. In *AIMSA 2004*, LNAI 3192, pages 500–509, 2004.

[48] F. Manola and E. Miller. Rdf primer. Www document, World Wide Web Consortium (W3C), 2004.

[49] A. Mascret, T. Devogele, I. Le Berre, and A. Hénaff. Coastline matching process based on the discrete fréchet distance. In *12th International Symposium on Spatial Data Handling (SDH)*, pages 383–400, Vienna, Austria, 2006. Springer.

[50] L. Matikainen, J. Hyyppä, and H. Kaartinen. Automatic detection of changes from laser scanner and aerial image data for updating building maps. In *In: Intl Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXV-B2*, pages 434–439, 2004.

[51] Institut Géographique National. Espace professionnel. available online at http://professionnels.ign.fr/.

[52] Kadaster Netherlands. Top10nl. available online on http://www.kadaster.nl/top10nl_engels/.

[53] F. Ormeling. From ortelius to openstreetmap transformation of the map into a multifunctional signpost. *Polish Cartographical Review*, 41(4):319–329, 2009.

[54] PBL. The pbl netherlands environmental assessment agency (pbl). available online at http://www.pbl.nl.

[55] Z. Qiaoping and I. Couloigner. Automatic road change detection and gis updating from high spatial remotely-sensed imagery. *Geo-spatial Information Science (Quarterly)*, 7(2):89–95, June 2004.

[56] Brian G. Schunck Ramesh Jain, Rangachar Kasturi. *MACHINE VISION*. McGraw-Hill, 1995.

[57] Ying S., Li L., X. F. Liu, Zhao H., and Li D. Change-only modeling in navigation geo-databases. In *ISPRS Workshop on Updating Geo-spatial Databases with Imagery & the 5th ISPRS Workshop on DMGISs.*, Urumchi, Xingjiang, China, August 2007.

[58] Y. Shen. Modeling of change-only information and updating of spatial data. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*, volume XXXVII, Beijing, 2008.

[59] Y. Shen, L. Lin, G. Yurong, and M. Yuan. Probabilistic matching of map objects in multi-scale space. In *Proceedings of the 25th International Cartographic Conference ICC 2011*, 2011.

[60] X. Song, V. Raghavan, and D. Yoshida. Open web processing services for improving accuracy of gps tracks by filtering and map-matching. 2009.

[61] H.J. Stark. Quality assurance of crowdsourced geocoded address-data within openaddresses concepts and implementation. Master's thesis, Centre for GeoInformatics (Z_GIS), Salzburg University, June 2010.

[62] J. Stoter, W.Quak, and A. Hofman. Harmonising and integrating two domain models topography. *Nederlandse Commissie voor Geodesie*, 48:89–106, 2009.

[63] H. Sui, D. Li, and J. Gong. Automatic feature-level change detection (flcd) for road networks. In *International Society For Photogrammetry And Remote Sensing*, Geo-Imagery Bridging Continents, pages 459–464, Istanbul, Turkey, July 2004.

[64] V. Talhofer, A. Hofmann, . Hoková Mayerová, and P. Kubícek. Spatial analyses and spatial data quality. In *Tha Association of*

Geographic Information Laboratories for Europe (AGILE) 2011, April 2011.

[65] J. van Buren, E. Hemmatnia, C. Lemmen, and P. van Oosterom. Europees coöordinatensysteem is voor nederland vooral een kans. *Geodata-inwinning; Special of Geo-Info and Vi Matrix*, pages 20–23, October 2008.

[66] N. van der Zon. Kwaliteitsdocument ahn-2. Technical report, Actueel Hoogtebestand Nederland (AHN), 2011.

[67] M. van Exel and E. Dias. Towards a methodology for trust stratification in vgi. In *VGI Preconference meeting at AAG 2011*, pages 2 – 5. Association of American Geographers, 2011.

[68] S. Volz. An iterative approach for matching multiple representations of street data. In *Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data,*, volume XXXVI Part 2/W40, Hannover, Germany., 2006. University of Stuttgart, Institute for Photogrammetry.

[69] G. von Goesseln and M. Sester. Semantic and geometric integration of geoscientific data sets with atkis applied to geo-objects from geology and soil science. In *ISPRS Commission IV Joint Workshop Challenges in geospatial analysis, integration and visualization ii*, page 111116, Stuttgart, Germany, September 2003.

[70] R. Zenger. Details of 3dshapes data. available online at https://rejo-zenger.nl/inzicht/aanvullende-informatie-over-het-3d-shapes-bestand.

[71] R. Zenger. Legal restrictions of 3dshapes data. available online at https://rejo.zenger.nl/inzicht/geografische-bestand-3d-shapes.

[72] Q. Zhang and I. Couloigner. A framework for road change detection and map updating. *International archives of the photogrammetry. Remote Sensing and Spatial Information Sciences*, 35:729–734, 2004.

[73] Q. Zhang and I. Couloigner. Spatio-temporal modeling in road network change detection and updating. In *International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*, China, August 2005. Peking University.

[74] D. Zielstra and A. Zipf. A comparative study of proprietary geodata and volunteered geographic information for germany. In *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal, 2010.