Pre-print version

Quality assessment for building footprints data on OpenStreetMap

Abstract: in the past two years several applications of generating 3D buildings from OpenStreetMap (OSM) have been made available, for instance, OSM-3D, OSM2World, OSM Building etc. In these projects 3D buildings are reconstructed using the building footprints and their attributes information which are documented as tags in OSM. Therefore, the quality of 3D buildings relies strongly on the quality of building footprints data in OSM. This paper is dedicated to quality assessment of building footprints data in OSM for the German City Munich which is one of the most-developed cities in OSM. The data is evaluated in terms of completeness, semantic accuracy, position accuracy, and shape accuracy by using building footprints in ATKIS (German Authority Topographic-Cartographic Information System) as reference data. The process contains three steps: finding correspondence between OSM and ATKIS data, calculating parameters of the four quality criterions, and statistical analysis. The results show that OSM footprint data in Munich has high completeness and semantic accuracy. There is an offset of about four meters in average in terms of position accuracy. With respect to shape, OSM building footprints have high similarity to those in ATKIS data. However, some architectural details are missing, hence the OSM footprints can be regarded as a simplified version of those in ATKIS data.

Keywords: OpenStreetMap, quality assessment, building footprint,VGI

1. Introduction

In the context of Web 2.0, crowd-sourcing has emerged as a new paradigm that leverages community (or crowd) participation to effectively and efficiently accomplish a task traditionally undertaken by a few selected individuals. With a global cast of volunteers, OpenstreetMap (OSM) is considered as one of the most successful and popular VGI (Volunteered Geographic Information) project. For the current state, there are more than 1,1 million registered members (OSM, 2013) who make OSM rapidly growing. Sparked by the availability of high resolution imagery from Bing since 2010, there has been an increase in building information in OSM, proving that volunteers do not only contribute roads or POIs to the database. According to the statistic (the values are derived from our internal OSM database which is updated daily) on May 5th, 2013 the amount of buildings in OSM is above 77 million. In Germany, there are almost 9 million objects with "building=yes" to the same time point.

Currently, building footprints data in OpenStreetMap (OSM) is mainly used for reconstructing 3D buildings. At present there are several projects which generate and visualize 3D buildings from OSM: OSM-3D[1], OSM Buildings[2], Glosm[3], OSM2World[4], etc. And applications based on these projects i.e. 3D navigation on mobile devices, web-based visualization, and simulation etc. are getting increased. The most of 3D buildings in these projects are rendered as polyhedral, extruded footprints with flat roofs, whereby the height information of a number of buildings are directly taken from the attribute of building footprints or converted from the number of stories, while the majority of 3D buildings own random heights. In OSM-3D, many buildings are modeled in LoD2 (Level of Detail according to

[1] http://osm-3d.org
[2] http://osmbuildings.org
[3] http://glosm.amdmi3.ru/
[4] http://maps.osm2world.org/

CityGML) in case there are indications for their roof types (Goetz and Zipf, 2012). In further, Goetz (2013) proposed a conception to generate buildings in LoD3 and LoD4 in CityGML. Besides, buildings in different LoDs from other sources can be uploaded via OpenBuildingModels and visualized in OSM-3D. But the buildings for uploading have to be adapted with the corresponding building footprints in OSM (Uden and Zipf, 2012).

Since the 3D buildings in aforementioned projects are generated mainly by extruding building footprints along the vertical direction, the quality of these buildings strongly relies on the quality of building footprints in OSM. The presented work is dedicated to the quality assessment of building footprints data in OSM within a test area in Munich (Germany), because on the one hand Munich is one of the most representative cities where OSM data is regarded as well developed. On the other hand, Munich is the third largest city in Germany with very dense buildings in the downtown. Moreover, the geometries of building footprints in Munich reveal large diversity.

In this work, four criterions are introduced for the quality assessment of building footprint data in OSM: (i) completeness, (ii) semantic accuracy, (iii) position accuracy, and (iv) shape accuracy. With respect to these four criterions, OSM data are quantitatively assessed by comparing with the reference data from the German ATKIS (Amtliches Topographisch-Kartographisches Informationsystem -- Authorative Topographic-Cartographic Information System). ATKIS is a common project of the Working Committees of the Survey Administrations of the States of the Federal Republic of Germany (AdV) (Grünreich, 2000). It contains information on settlements, roads, railways, vegetation, waterways, and more. The positional accuracy of building data in ATKIS is ± 0.5m (Müller and Seyfert, 1998). The process of quality assessment is composed of three steps. In the first step, correspondences among buildings in two data sets have to be identified. On this base, parameters are calculated according to the definitions of the four quality criterions. Then the differences between the two data sets are analyzed and visualized.

The remainder of this paper is structured as follows: Section 2 gives an overview of the related works to this paper, Section 3 introduces the criterions of the quality, Section 4 describes the algorithm to match building footprints in two data sets, Section 5 firstly gives an overview of the two data set in the test area and presents the results of the test area, and Section 6 discusses the results and concludes the whole work.

2. Related works

2.1 Quality assessment of OSM

In recent years the geo-data provided by the OSM project has been the foundation of a number of scientific publications in a widespread of research fields. In 2008 Haklay conducted a first analysis that investigated the data quality of roads in OSM for England (Haklay 2010). This first approach was followed by further publications about OSM in Germany (Zielstra & Zipf 2010, Neis et al 2012) and France (Girres & Touya 2010) and more detailed investigations about point (Neis et al 2010), line (Helbich et al 2012) and polygon (Mooney et al 2010) objects that can be found in the project's database. As mentioned by Hagenauer & Helbich (2012) nearly all "empirical studies indicate that urban areas are better mapped" in OSM. This it is not surprising since most urban areas with a higher population density inherit larger numbers of contributors, who influenced the quantity and quality of the collaboratively crowd sourced OSM objects (Haklay et al 2010, Girres & Touya 2010, Neis et al 2012).

In contrast to quality assessment of road networks, few works have been made available for evaluating building footprints data in OSM. To the best of the authors' knowledge, only one detailed study

investigating buildings in OSM has been published by Kunze (2012) which applied several methods to assess the completeness of the building information in OSM in comparison to an administrative dataset for two federal states in Germany. As the criterion of quality assessment, the work mainly analyzed the area difference of a group of buildings within hexagon/square instead of individual correspondence. In further, position accuracy and shape characters are not compared.

The most common elements of quality assessment used in the abovementioned research works are position accuracy and completeness. In further, shape similarity is used to evaluate the polygonal objects such as lakes, ponds, and forests (Mooney et al. 2010). In general, the elements for quality assessment can be categorized in three types: elements for geographic data bases, elements for data modeling and for the spatial data. Girres & Touya (2010) did a comprehensive quality assessment for both data and data models of OSM in France. In their work, eight elements are selected from Kresse and Fadaie (2003) and Guptill and Morrison (1995): geometric accuracy, attribute accuracy, semantic accuracy, completeness, logical consistency, temporal accuracy, lineage, and usage. For the building footprints data in OSM, we take four of them, namely, position accuracy, shape accuracy, semantic accuracy, and completeness; because these elements are relevant for the building footprint data while other elements are designed for data modeling. Besides, attributes of building footprints are evaluated in terms of their completeness. Because of the low completeness (see Section 5), the attribute accuracy is not assessed in this work.

2.2 Map matching

Map matching is defined as the process to identify correspondent features between two sets of geospatial data. It is an essential pre-process for data integration, change detection, data updating, and data comparison. The majority of the currently existing approaches for map matching concentrates on road network matching. One of the earlier researches developed a statistical matching algorithm by incorporating the concept of relational matching in their network-matching algorithm (Walter and Fritsch 1999). In the past ten years, most of map matching approaches take features (e.g. distances, angles, shapes and semantics) or structure (e.g. sub-graph and proximity graph) into account for the similarity measurement to identify the corresponding roads (Samal et al. 2004, Xiong and Sperling 2004, Volz 2006, Mustière and Devogele 2007, Min et al. 2007, Olteanu & Mustière 2008, Zhang 2009, Kim et al. 2010, Li and Goodchild 2011). Most recently, Koukoletsos et al. (2012) proposed an automated feature-based matching method specifically designed for OSM, based on a multi-stage approach that combines geometric and attribute constraints. Yang et al. (2013) proposed a heuristic probability relaxation approach to match road networks. Their process starts with an initial probabilistic matrix according to the dissimilarities in the shapes and then integrates the relative compatibility coefficient of neighboring candidate pairs to iteratively update the initial probabilistic matrix until the probabilistic matrix is globally consistent. Then objects correspondences are find out on the basis of probabilities.

In contrast to the road network matching, there are few researches for matching area objects which reveal as polygon objects, such as residential region, water body, forest, meadow etc. The work of Gösseln and Sester (2003) could be deployed to match polygonal objects by using an iterative closet point (ICP) algorithm that detects corresponding point pairs for two point sets derived from each contour of corresponding objects. Huh et al. (2013) developed a method to detect corresponding point pair between polygon object pair with a string matching method based on confidence region model of a line segment. However, these methods are restricted to low density of polygons to be paired. In case that neighboring polygons are located immediately close to each other and similar in shape and size, for instance, polygons of building footprints in dense urban area, there will be error matching.

For this reason, the abovementioned approaches cannot be used to identify corresponding polygons in two building footprints data sets. In the presented paper, area overlapping method is introduced considering the fact that there is not much displacement between OSM building footprints data and the reference data set, namely ATKIS data.

2.3 Similarity measurement by using turning function

Turning function or tangent function was introduced by Arkin et al. (1991) for measuring the similarity of two polygons. Traditionally, there are two ways to represent a closed polygon: (i) by giving a list of vertices or (ii) by giving a list of line segments. Alternatively, a polygon can be represented using a list of angle-length pairs, whereby the angle at a vertex is accumulated tangent angle at this point while the corresponding length is the normalized accumulated length of the polygon sides up to this point. Let C be the polygon on the left of Figure 1. The tangent angle at the starting vertex is $\theta_1 = \varphi_1$. Then θ_i can be calculated as $\theta_i = \theta_{i-1} + \varphi_i$. The right of Figure 1 shows the change of tangent angles (y-axis) along the normalized accumulated length of the polygon sides (x-axis). From this point of view, the tangent angle can be treated as a function of the normalized accumulated length $T_C(l)$. It can be called tangent function or turning function.

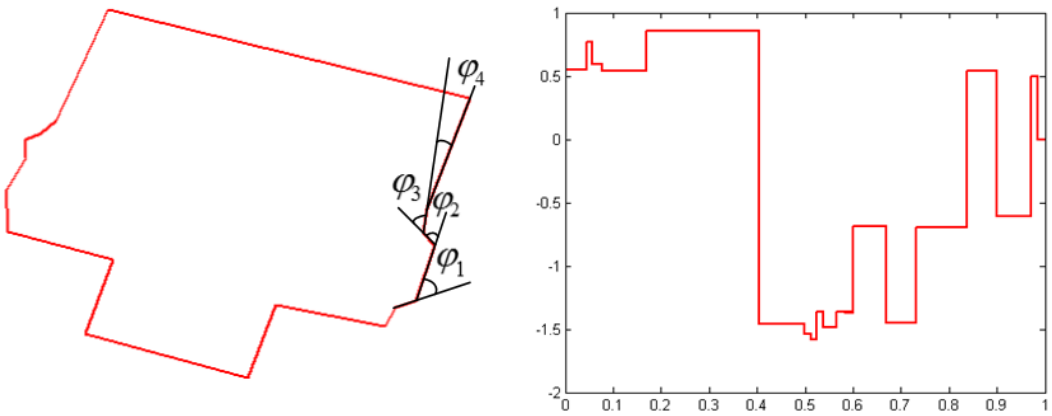


Fig.1. Tangent space representation of polygon

The turning function $T_C(l)$ measures the angle of the counter-clockwise tangent as a function of the normalized accumulated length l. The cumulative angle increases with left hand turns and decreases with right hand turns. This kind of representation is invariant to rotation, because it contains no orientation information. Furthermore, it is invariant to scaling, since the normalized length makes it independent to the polygon size.

The similarity of two polygons (A, B) can be then defined as the distance between their turning functions.

$$S(A,B) = d(A,B) = \|T_A - T_B\|_2 = \left(\int_0^1 \left(T_A(l) - T_A(l)\right)\right)^{\frac{1}{2}} \quad (1)$$

In order to avoid the translation of the tangent angle in relation to the other one, the identical point pair of the two polygons has to be found out and set as reference point for the calculation of the tangent angles. Note that $S(A,B)$ denotes actually the dissimilarity between A and B. The smaller $S(A,B)$ is, the more similar are the two polygons. In the case A is identical to B, there is $S(A,B) = 0$.

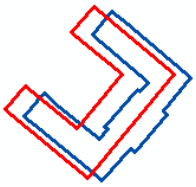
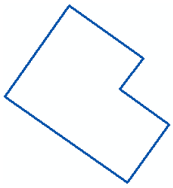
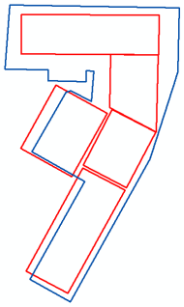
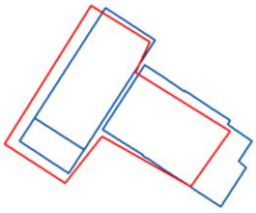
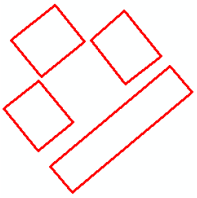
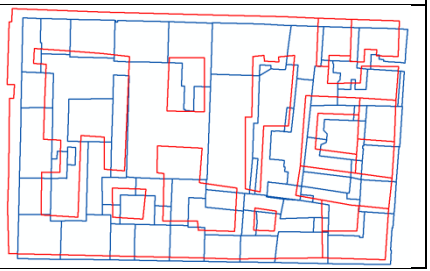
3. The selected elements for quality assessment

As stated previously, four elements are used for the quality assessment in this work, namely, completeness, semantic accuracy, position accuracy, and shape accuracy.

Completeness – this is a measure of the lack of data, which does not record objects that are expected to be found in the database, or excess data that should not be included. Regarding to the data of building footprints in OSM, the completeness is defined as the area difference covered by OSM buildings and ATKIS buildings. In addition, the completeness of the attributive information is given by counting how many buildings in OSM are recorded with attributes such as name, type, height, etc. respectively.

Semantic accuracy – this investigates if buildings in the real world are recorded indeed as building objects in OSM on the one hand, on the other hand it measures the percentage of building objects in OSM which are indeed buildings in the real world. Furthermore, it denotes the correctness of inherency between building geometries and their semantic hierarchies. In this work, the semantic accuracy is calculated by analyzing the correspondences among individual buildings in the OSM data and reference data. There might be 1:1, 1:n, 1:0, 0:1, n:1, and n:m relations between OSM building footprints and those in reference data, as shown in Table 1, whereby footprints in two data sets are distinguished in red and blue colors. While footprints in OSM are visualized in red color, footprints in reference data are in blue.

Table 1. Possible relations between building footprints in two data sets

Relation	1:1	1:0	1:n
Illustration			
Relation	n:1	0:1	n:m
Illustration			

According to the OGC standard of CityGML building models (Groeger et al. 2008), semantic hierarchy and geometrical level of details (LoD) relate themselves inherently. Hence, these six kinds of relations denote different semantic accuracy as follows:

- 1:1 relation: a building is semantically correctly recorded.
- 1:n relation: a building in OSM is an aggregation of n buildings in the reference data. Therefore, the building is recorded at higher level on the semantic hierarchy.
- 1:0 relation: a building in OSM is actually not a building (semantically wrong) in the reality.
- 0:1 relation: it is the opposite case of the 1:0 relation.
- n:1 relation: a building in OSM is a part of a building in the reference data. Therefore, the building is recorded at lower level on the semantic hierarchy.
- n:m relation: the buildings are incorrectly recorded with respect to semantic.

In a word, a building is correctly recorded in OSM in terms of semantic only when it has a 1:1 relation with the reference data. This is also indicated in the definition of "building key" in OSM (http://wiki.openstreetmap.org/wiki/Key:building).

Position accuracy – it evaluates how well the coordinate value of a building in OSM relates to the reality on the ground. In the presented work, the corresponding points of a pair of building footprints in two data sets are found at first. Then the position accuracy is calculated as the average distance of these corresponding points.

Shape accuracy – this is a measure of similarity of a building footprint in OSM to the shape of the building footprint in the reality. In this work, the shape similarity between a pair of footprints in two data sets is defined as their turning (tangent) function distance, which is calculated according to (Arkin et al., 1991). The starting points for calculating turning function are selected from the corresponding points whose distance is the shortest of all the corresponding pairs of points.

4. Identification of correspondence

The term of correspondence here has twofold meaning: (i) the relations among building footprints in OSM and ATKIS, and (ii) the corresponding turning points which form the shape of building footprints. In this section, the correspondences among building footprints in OSM and ATKIS are identified at first. For building footprints with 1:1 relation, their corresponding points are found out in the second step.

4.1 Correspondence among building footprints

The six kinds of relations of correspondence can be identified according to the algorithm as follows: Let G_{OSM} be the OSM data set and G_{ref} be the reference data set. For a building footprint $foot_{osm\_i}$ in G_{OSM}, the building footprints in G_{ref} will be checked if they are intersected with the lines of polygon of $foot_{osm\_i}$. In the case that there is intersection by $foot_{ref\_j}$, the intersected area is calculated at first as $Area_{overlap}$. Since the most of building footprints in OSM have been digitalized according to the Bing Map images (http://www.bing.com/maps) (Goetz and Zipf 2012; OSM 2013b, 2013c), there is normally offset between footprints in OSM and the reference data due to the distortion caused by oblique view of the used sensors. Considering this factor, large buildings in OSM have larger percentage of area overlap with their correspondence in the reference data, while small and high buildings might have smaller percentage of area overlap with their correspondence. The threshold of the judgment depends actually strongly upon the parameters of the Bing map images used for digitalization in OSM. In their work, Rutzinger et al. (2009) found out that the correspondence might be caused by their neighboring building if the overlapped area is less than 30%. Therefore, the threshold of the overlapping is set as 30%. If

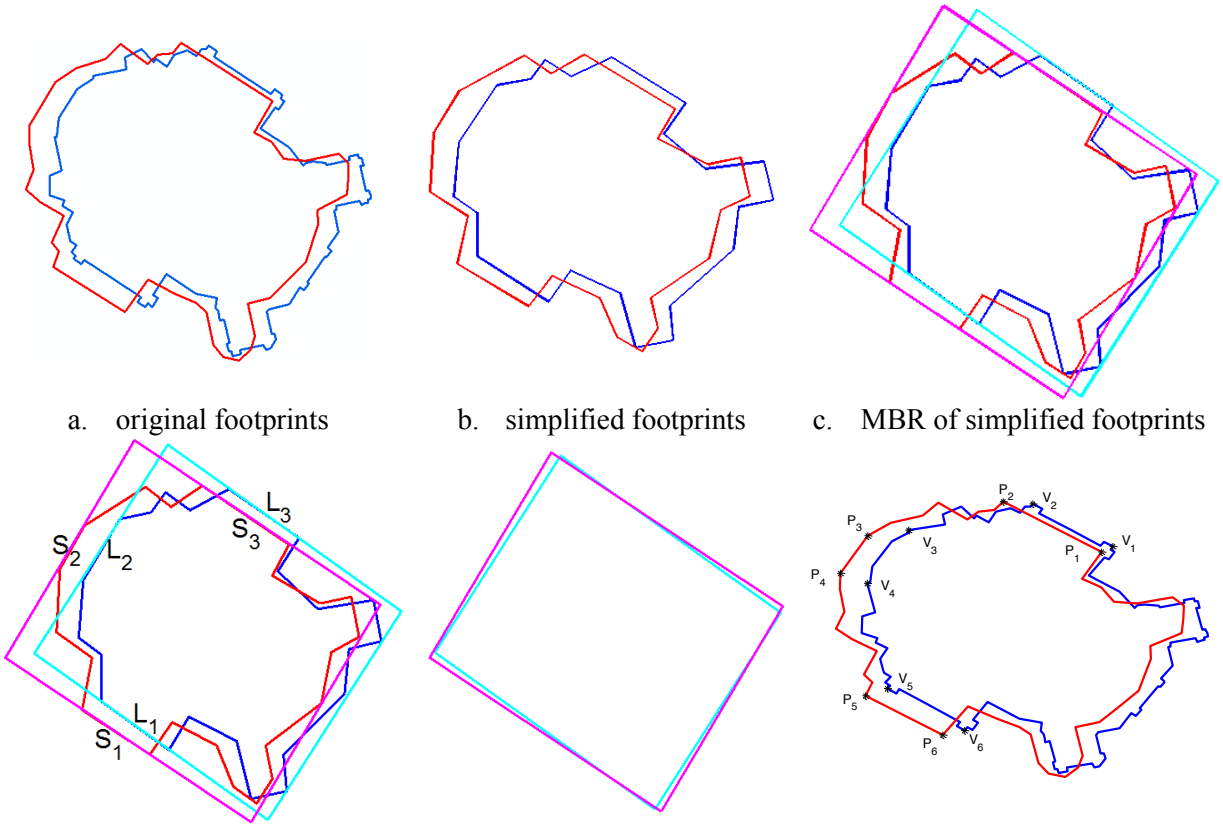
$$\frac{Area_{overlap}}{\min\left(\text{Area}(foot_{osm\_i}),\text{Area}(foot_{ref\_j})\right)} > 30\% \qquad (2)$$

then the footprints $foot_{osm\_i}$ and $foot_{ref\_j}$ are matched. A 1:1 relation is identified when a footprint in G_{ref} can only be matched to one footprint in G_{osm}, while 0:1 or 1:0 relation is indicated to the case that the footprint cannot be matched to those in another data set. If a footprint in G_{osm} can be matched with many footprints in G_{ref}, there might be 1:n or n:m relation. In this case, the matching results will be checked in an inverse way. Namely, for all the n footprints in G_{ref}, their matched footprints in G_{osm} are identified using Eq.2. If all these n footprints are matched to the same footprint in G_{osm}, it is 1:n relation. Otherwise, these n footprints are matched to more than one footprint in G_{ref}, it is then n:m relation.

4.2 Find identical points of matched building footprints pairs

For the polygon pairs with 1:1 relation, their corresponding points can be found out efficiently by using the following process based on the reality that there is not much difference in shapes, rotation and scale between OSM building footprint and real data, because OSM footprints are created by digitalizing the high resolution Bing images. The algorithm of finding corresponding points of paired

footprints is described by taking a pair of building footprints in Figure 1, whereby polygon in red stands for building footprint in OSM while polygon in blue stands for building footprint in ATKIS.



a. original footprints b. simplified footprints c. MBR of simplified footprints

d. Edges of footprints which are located on MBR e. Edge correspondences between MBRs f. The corresponding points of two footprints

Fig.2. An example of finding identical points of paired footprints

As shown in Figure 1a, the footprints from different data sets might be formed at different level of detail (LoD) in terms of geometry. This will lead to 1:n correspondence among polygon points. To avoid this kind of effect, key points of footprints (Figure 1b) are extracted first of all using douglas-peucker algorithm (Douglas and Peucker, 1973). Then minimum bounding rectangle (MBR) is calculated respectively for the two polygons (as shown in Figure 1c, rectangle in Cyan is MBR for OSM footprint and rectangle in Magenta is MBR for ATKIS footprint). In the next step (Figure 1d), edges of the building footprint are marked if they are located on edges of its MBR. Then OSM MBR is shifted to the center of the ATKIS MBR (Figure 1e), so that edges of these two MBRs can be matched if they are they are located (almost) on the same place. Finally, edges of footprints can be matched if they are marked to the same edge of MBRs. As shown in Figure 1f, three edges are matched. Their ending points are then regarded as identical points of two footprints.

5. The quality of OSM building footprints in Munich

The test data set covers 10km x 10km almost for the whole city of Munich. The OSM data is dumped from our internal database on the May 10th, 2013. The reference data is ATKIS data in the year of 2010 provided by the city of Munich. In order to accelerate the process of finding correspondences in the two data sets, the whole area is divided into a number of grid cells in a preprocessing phase, so that the search area is substantially reduced. A building footprint is indexed to a cell, if its centroid is located in the cell. For some buildings closed to or intersected with the border of a cell, their correspondent buildings could be indexed to the neighboring cells. Therefore, the search area is set as the 3x3 neighborhood of the current cell (where the current footprint is indexed). A sensitivity analysis

of the cell size showed that if cell size is smaller than 1.5m, there is not much difference in computation time thanks to the high performance of the computer. But, when cell size is larger than 2m, the computation time becomes also longer, because the number of buildings within a cell is increased while increasing the cell size. Finally, 15x15 grid cells are used both for fast computation and for better illustration of results.

5.1 Quantitative assessment of the matching results by using method of area overlap

Since the analysis in the following sections is based on the results of matching building footprints in the two data sets, it is essential to know the quality of the matching results. In order to evaluate the matching results quantitatively, building footprints are selected in two data sets by using a rectangular boundary in the downtown of Munich. In the evaluation area, there are 1291 building footprints in OSM, while there are 2470 buildings in ATKIS. The results using the method of area overlap are compared with those from manual matching. Table2 shows the results of the matching using method of area overlap. Both the Recall and the Precision are greater than 99%. Hence, the method of area overlap achieves good and robust matching.

Table 2. statistic of the matching results using method of area overlap

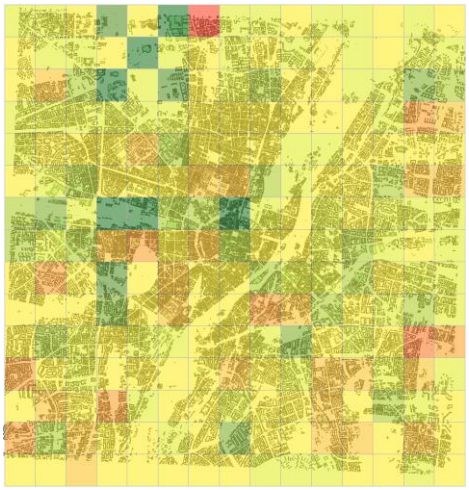
Relation	1:1	1:0	n:1
True matching	569	344	376
False matching	3	2	3
Miss matching	2	0	0
Recall	99.1%	99.4%	99.2%
Precision in total	99.2%		

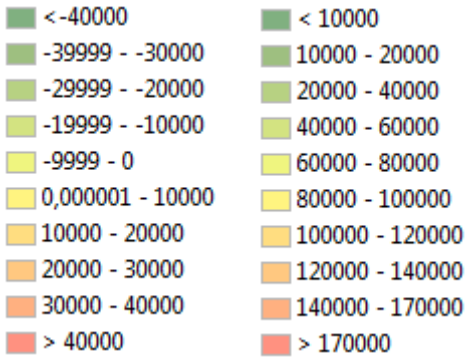
5.2 Quality of data completeness

In terms of area covered by buildings, the city Munich is quite completely mapped, because the total area of buildings in OSM data is even slightly larger than that in ATKIS data. Figure 2 shows cell-based distribution of building areas in both data sets, as well as their differences (ATKIS-OSM) in cells, whereby, the area of each cell is calculated as the sum of area of all buildings located in the cell. Comparing Figure 2a and 2b, the two distributions are almost same. This verifies the fact that almost all the areas covered by buildings have been mapped as buildings in OSM. In most of cells, the differences are between 10000 square meters (Figure 2c).

a. ATKIS data

b. OSM data





Legend for Fig.1c

Legend for Fig.1a and 1b

c. Area difference (ATKIS-OSM)

Fig.3. Cell-based distribution of building areas and their differences

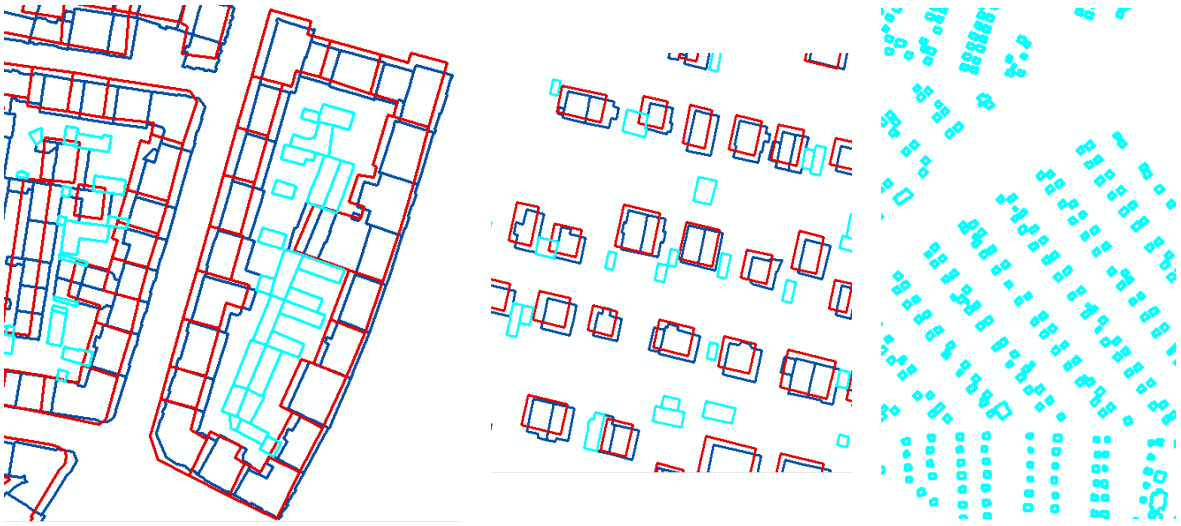
The results of completeness quality are shown in Table 3. In contrast to the high completeness in terms of covered area, there is limited attributive information in OSM footprints data. Only few buildings are recorded with building types, even fewer buildings have attributes of height information and numbers of stories. Both data sets contain few attributes of "building name", because normally only landmarks, commercial and public buildings have a name but most of residential buildings do not have a name. In this context, it can be stated that more than 50% of buildings which have names are tagged with names.

Table 3. Completeness of building footprints in OSM

	Area cover (square meter)	Buildings with types	Buildings with name	Buildings with height	Buildings with numbers of stories
ATKIS	18486805.65	100%	5.24%[5]	100%	100%
OSM	18707108.84	8.46%	2.82%	0.41%	0.06%

In terms of the amount of buildings, there are 33,911 buildings in ATKIS which cannot be matched to the OSM data, while 1,233 buildings in OSM have no correspondent ones in ATKIS. Through a manual inspection in the two data sets and Bing Map, the 33,911 buildings can be classified in three types (Figure 4): (a) most of them are located inside of yard formed by terraced buildings, they are normally occluded by the terraced buildings around them; (b) many smaller building like garages can very difficult to be identified on Bing map images, in addition, their roofs have normally low contrast to the ground and roads; (c) in some regions, villas and other buildings (garages) are small and mostly occluded by trees which make difficulty for the digitalization on Bing Map.

[5] The field of building name in AKTIS is 100% filled. However, only about 5.24% buildings in Munich have individual names, while most of them have a value of "nameless" for the attribute field.



a b c

Fig.4. three types of buildings which have 1:0 relation with OSM data, whereby ATKIS footprints are visualized in blue while OSM footprints are visualized in red, the ATKIS buildings which are not mapped in OSM are highlighted in cyan, the scale is 1:2500.

5.3 Semantic accuracy

As indicated previously, the notion of semantic is defined as what the object is. There are coherent relations between semantic hierarchy and geometrical hierarchy. We assume that all the buildings in the reference data (ATKIS) are semantically correct. That means that every building in ATKIS is corresponded exactly a building in the real world. If a building in OSM is matched only with one building in ATKIS, it is semantically correctly mapped. Otherwise, its semantic is not accurate. The polygon object should be called "building group", if it is matched with several buildings in ATKIS. Or, it should be called "building part", in case that it and its neighboring polygons are matched to an identical ATKIS building footprint.

In total, there are 39,364 buildings in the test bed in OSM data, while there are 100,014 buildings in ATKIS data. As shown in Table 4, almost all the footprints can be matched with ATKIS data except 1,233 buildings with 0:1 relation, because the ATKIS data used in this work is three years older than the OSM data. These buildings are new constructed in the recent three years, according to our local knowledge in Munich. Base on a visual inspection on Bing map image and Google Map, we can state that all these buildings are mapped correctly in semantic. That means that all OSM building footprints are indeed buildings in the real world. Therefore, the semantic accuracy in a broad sense is 100%.

There are 21,775 unique correspondent relation (1:1 relation), which means 21,775 buildings are correctly recorded in OSM data with respect to semantic. 13,131 buildings are semantically coarsely recorded, since they have n:1 relation with building footprints in ATKIS data. 266 buildings are semantically more detailed recorded than ATKIS data, as they have 1:n relation. Then the semantic accuracy of OSM building footprints is calculated as: $\frac{21775+1233}{39364} = 58.45\%$. The value means that each polygon of the 58.45% polygonal objects (with 'building = yes') in OSM is corresponded exactly to a building in the real world. Approximately 40% of polygonal objects (with 'building = yes') in OSM are actually outlines of a group of buildings. According to the definition of semantic in CityGML, they are incorrectly recorded in OSM with respect to semantic.

Table 4. Statistic of relations among building footprints in two data sets (ATKIS:OSM)

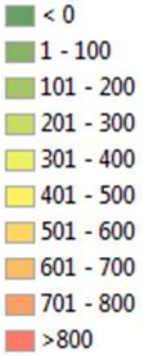
Relation	1:1	1:0	1:n	n:1	0:1
Amount	21,775	33,911	266	13,131	1,233

Figure 5 shows the grid-cell based density map of amount of buildings in ATKIS (Figure 4a) and OSM (Figure 4b) respectively, as well as their difference (ATKIS-OSM) (Figure 4c). Obviously, OSM data has lower building density than ATKIS data. Most of buildings in high densely constructed urban area (red cells in Figure 4a) are semantically wrong recorded (compare Figure 4a and 4c), because they are normally difficult to be distinguished as individual buildings from their roofs on Bing map images. They are normally digitalized as blocks in OSM.

a. ATKIS data

b. data OSM

c. difference of the two data sets (ATKIS-OSM)



d. legend

Fig.5. Distribution of difference of amount of buildings in grid cells

5.4 Position accuracy

The position accuracy is investigated by calculating the average distance among the corresponding points of footprints pair in two data sets. Hence, only the buildings with 1:1 relation are involved in the analysis.

Table 5. Position accuracy of OSM building footprints

	Maximum offset (m)	Minimum offset (m)	Average offset (m)	Standard deviation (m)
Value	14.80	0.002	4.13	1.71

As shown in Table 5, the average offset of OSM building footprints to ATKIS building footprints is 4.13m with the standard deviation of 1.71 meter. The largest offset is near 15m, while the smallest offset is less than a centimeter. The distribution of the offsets is close to normal distribution with $\mu = 4.13$m and $\sigma = 1.71$, as demonstrated in Figure 6.

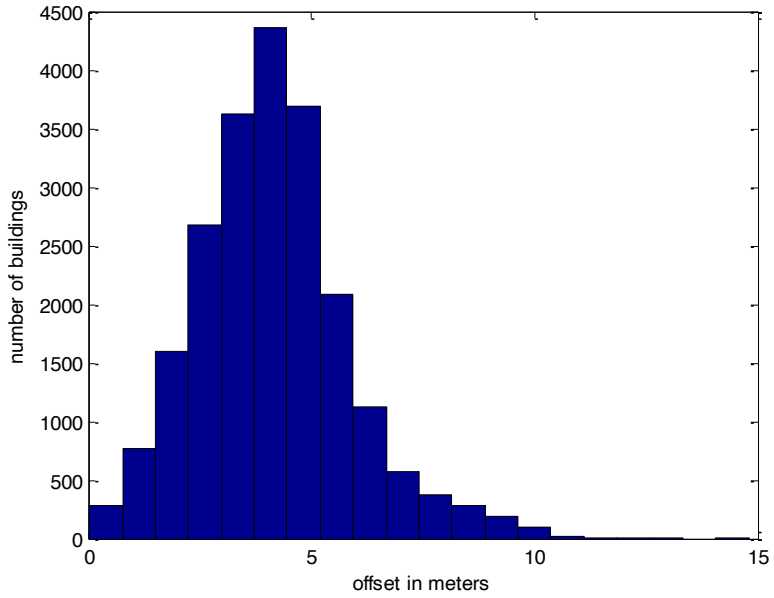


Fig. 6. Distribution of offsets from OSM building footprints to ATKIS building footprints

Note that the precision of building footprints data in ATKIS is $\pm 0.5m$, while the precision of Bing maps imagery in Munich is estimated as 3 to 4 meters by a visual inspection. Comparing with the offset of OSM to ATKIS, the following conclusion can be drawn: the low positional accuracy of OSM building footprint data is caused by the limited resolution of Bing map images.

5.5 Shape accuracy

Similar to the position accuracy, for shape accuracy only footprints which have 1:1 relation are analyzed. The shape accuracy is indicated by the shape similarity between the building footprints pair in the two data sets, whereby the dissimilarity of two polygons can be calculated as their turning function distance. Figure 7 shows the turning functions of the paired building footprints in the example of Section 4.2, the dissimilarity is 1.18 which is calculated using Equation 1. The value is actually the difference of areas covered by the turning function, as shown in Figure 7b.

In fact, Equation 1 is often used to calculate the similarities (dissimilarity) of a set of simplified polygons (with different length thresholds) to the original one. A comparison makes sense, only if the reference polygon for calculating dissimilarity is identical. A cross comparison with the value of similarities is impossible, because the polygon sets are differently. In order to make comparison globally, the similarities have to be normalized by setting the rectangularity of a polygon (polygon A in Eq.3 and 4) equal to the normalized similarity of its MBR to the polygon, because rectangularity is an indicator for polygon shape when comparing with other polygons.

$$\text{rectangularity} = \frac{\text{area(polygon)}}{\text{area(MBR)}} = S_n(A, MBR) \tag{3}$$

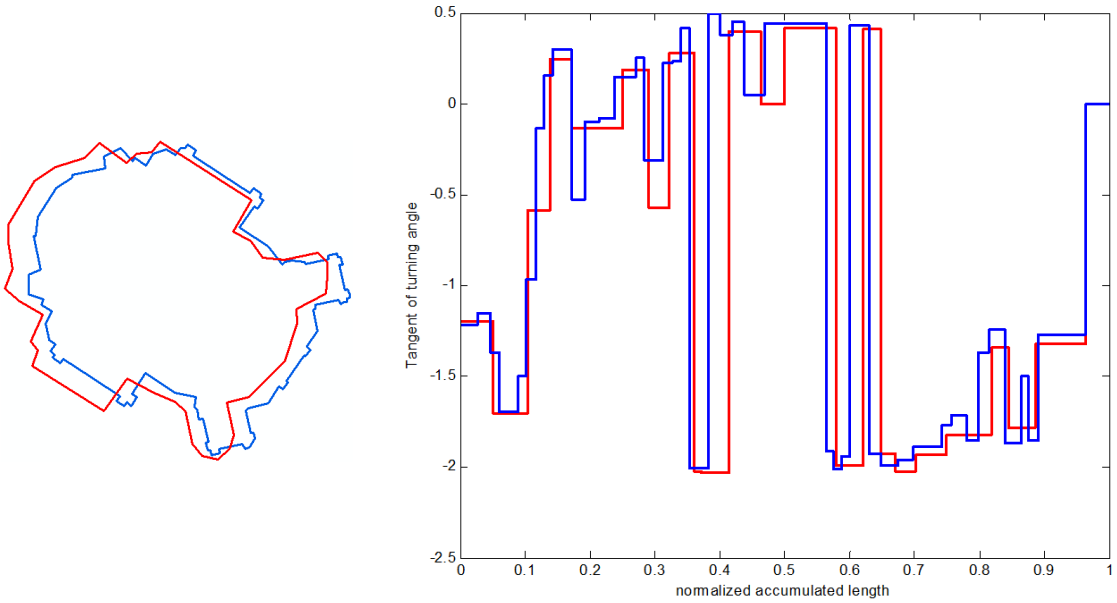
The normalized similarity $S_n(A, B)$ of a polygon B to polygon A can be calculated as:

$$\frac{1-S_n(A,B)}{d(A,B)} = \frac{1-S_n(A,MBR)}{d(A,MBR)} \rightarrow S_n(A,B) = 1 - d(A,B)\frac{1-S_n(A,MBR)}{d(A,MBR)} \quad (4)$$

Whereby: $d(A,MBR)$ is the dissimilarity of MBR to footprint A calculated by Eq.1, $d(A,B)$ is the dissimilarity of footprint B to footprint A calculated by Eq.1.

The principle of the normalization process in Eq.4 can be explained as follows: the ratio of the normalized dissimilarity to the dissimilarity calculated using Eq.1 is a constant value. This value can be calculated by setting the rectangularity of a polygon equal to the normalized similarity of its MBR to the polygon.

Taking the two footprints in Figure 7a as an example, the normalized similarity can be calculated. The dissimilarity of the MBR to the footprint in ATKIS (blue polygon in Figure 7a) is $d(A,MBR) = 1.47$. The rectangularity of the ATKIS footprint is 0.72, which is treated as normalized similarity of the MBR to the ATKIS footprint $S_n(A,MBR) = 0.72$. The dissimilarity of OSM footprint (red polygon in Figure 7b) is calculated using Eq.1, $d(A,B) = 1.18$. Then the normalized similarity of the two footprints can be obtained as 1-1.18x(1-0.72)/1.47 = 0.78.



a. a pair of footprints b. overlap of the turning functions

Fig. 7. Polygon similarity calculated from their turning function distance

Figure 8 shows the distribution of similarities among corresponding building footprints in OSM and ATKIS. Obviously, there is a concentration peak between 0.7 and 1. It means that the most of building footprints in OSM have high similarity (more than 70% similar) to their correspondent ones in ATKIS.

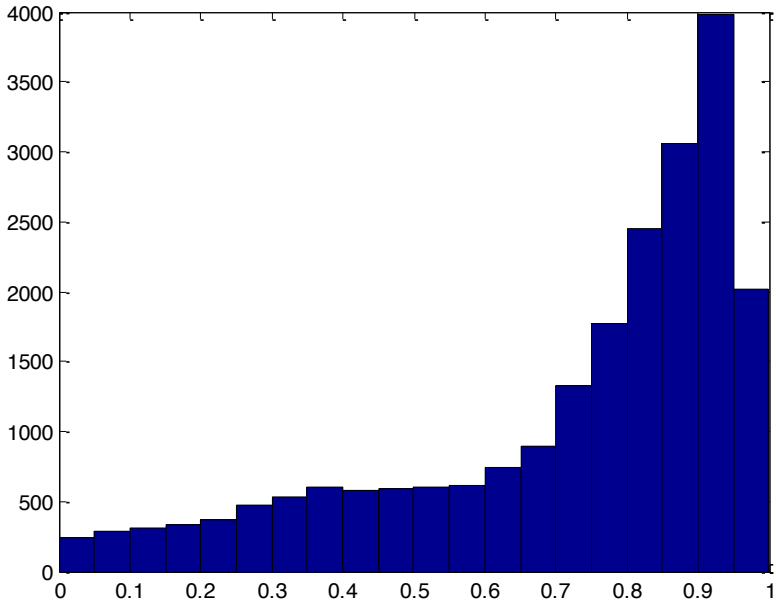


Fig.8. Distribution of shape similarities of corresponding building footprints

In order to find out the reason of the dissimilarity of the corresponding building footprints in the two data sets, the numbers of points which form building footprints are analyzed. The chart diagram in Figure 9 denotes that the most of building footprints in OSM contain up to 10 points less than their corresponding ones in the ATKIS. In other words, OSM building footprints are slightly simplified version of ATKIS building footprints.

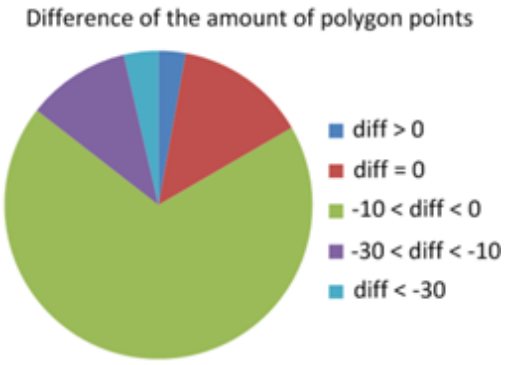
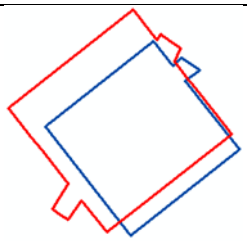
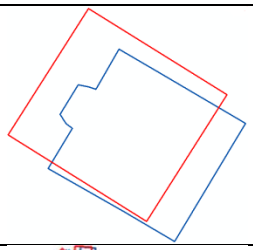
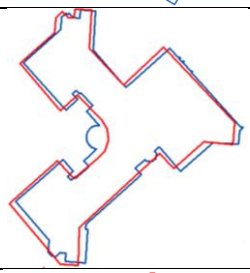
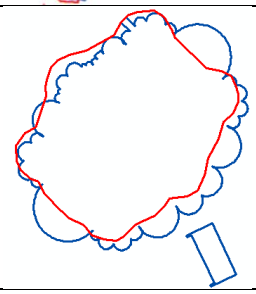


Fig.9. Chart diagram of the differences in terms of number of points

In Table 6, four typical examples of difference in OSM (red lines) and ATKIS (blue lines) are demonstrated. Only in very few cases, footprints in OSM are a little bit complicated than those in ATKIS data. For instance, in Table 6a, the fire escape was digitalized as a part of footprint in OSM, while it is neglected in ATKIS data, because the footprint in OSM is digitalized according to the image of roof in bird view, hence the fire escape cannot be differentiated from the main part of building. In the most case (Table 6b, 6c, 6d), footprints in OSM are simplified. The more complicated a footprint in reality is, the larger difference there is between OSM and ATKIS. There are three major reasons. Firstly, it is difficult to follow the architectural details according to roofs in bird view. Secondly, it is limited by the resolution of the Bing map image used during the digitalization. Thirdly, many volunteers do not have the patience to digitalize a complicated footprint exactly as it is. They normally sketch a simplified polygon with high similarity in terms of shape to the one in the reality.

Table 6. Examples of building footprints in OSM and ATKIS

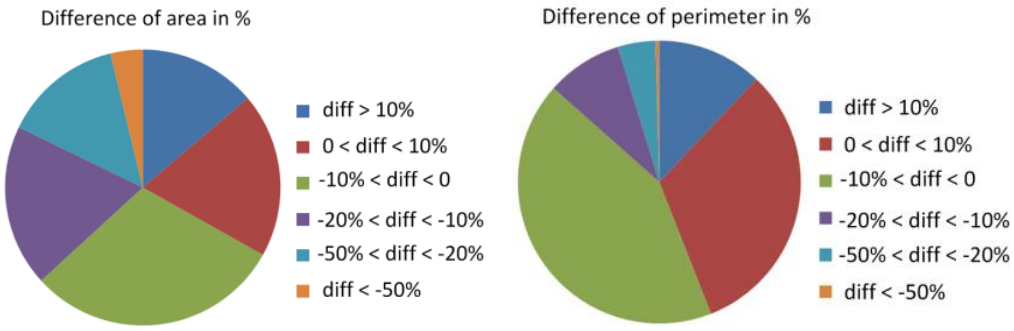
Scenarios	Difference of point amount (OSM:AKTIS)	Building footprint (red: OSM, blue: ATKIS)	Image on BingMap
a	17		
b	-8		
c	-120		
d	-1681		

In addition to the shape similarity, the difference in terms of size is analyzed by comparing the area and perimeter between the corresponding building footprints in the two data sets. Because areas and perimeters of building footprints vary very much, it is senseless to compare them directly. They have to be normalized as follows:

$$\text{Area}_{\text{diff}} = \frac{\text{Area}_{\text{ref}} - \text{Area}_{\text{osm}}}{\text{Area}_{\text{ref}}} \times 100 \quad (5)$$

$$\text{Perimeter}_{\text{diff}} = \frac{\text{Perimeter}_{\text{ref}} - \text{Perimeter}_{\text{osm}}}{\text{Perimeter}_{\text{ref}}} \times 100 \quad (6)$$

The chart diagrams in Figure 10 demonstrate detailed statistics of the difference in terms of the area and perimeters. In terms of area of footprint (Fig.10a), 13% buildings in ATKIS are 10% larger than those in OSM; 20% buildings in ATKIS are slightly larger (less than 10%) than those in OSM; and 30% buildings in ATKIS are slightly smaller (less than 10%) than those in OSM. In terms of perimeter of polygon (Fig.10b), more than 75% footprints have less than 10% difference to their corresponding ones.



a b

Fig.10. Chart diagrams of the differences in terms of area and perimeter

6. Conclusion and future works

This paper presents an approach to assess the quality of OSM building footprints data. A case study in Munich (Germany) is conducted. The results show that OSM building footprints data has high completeness in terms of covered area. Almost all the constructed area in the city is mapped as buildings in OSM. However, OSM building footprints data is still lack of attributes such as name, type, height etc. There are still many buildings which are not mapped on OSM. These buildings can be classified into three types. The buildings of the first type are occluded by their surroundings and hence cannot be visualized on the Bing Map images. The buildings of the second type are mostly small and low garages whose roofs have similar contrast to ground and roads in their surroundings. The third type of these building is referred to villas in forest area. The occlusion by vegetation makes difficulty for the identification on Bing Image. On the other hand, there are new findings on OSM. More than 1200 new constructed buildings are found in OSM which are not recorded in ATKIS data. This shows the preponderance of OSM in terms of the high frequency of data updating.

In a broad sense, the semantic accuracy of OSM building footprint data in Munich is 100%. For the specification of using the data for 3D reconstruction, its semantic accuracy is 58.45%, because semantic hierarchy is considered. In terms of position accuracy, the OSM building footprints have 4 meters offset in average to their corresponding ones in ATKIS. The footprints in OSM are highly similar to those in ATKIS in terms of shape. Most of OSM building footprints are almost identical to those in ATKIS. There is slight difference. Many buildings in OSM footprints consist of less polygon points than ATKIS footprints. Some architectural details are missing, if buildings are complicated in structure. In further, attributive information are not very rich.

The main reason for the abovementioned differences is that OSM footprints were digitalized using the base map of Bing images while ATKIS footprints are based on cadastral data. The offset is resulted by the distortion of buildings due to oblique aspect of the used sensor, while the fact of missing geometrical detail is caused by the limited resolution of Bing map images. The semantic accuracy of OSM in dense urban area is rather low, because many buildings in high densely constructed area are digitalized together with their neighbors as large blocks, since they cannot be distinguished on the Bing map images. But OSM data will be improved quite soon thanks to power of VGI: a huge number of volunteers for contribution and high frequency of the data updating.

So far, regular cell grids are used to reduce the computation cost and for a better overview of illustration and visualization of results in the quality assessment. In the future, this will be compared with the partitioning based on geographical zones i.e. city center, commercial area, industrial area, rural urban area etc. Besides, building footprints on OSM of large region (i.e. Baden-Wuerttemberg) containing large cities, middle and small cities, as well as rural area will be evaluated against the authority data.

References and Notes

Arkin E.M., Chew L.P., Huttenlocher D.P., Kedem K. and Mitchell J.S.B. 1991. An Efficiently Computable Metric for Comparing Polygonal Shapes. In: IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 13, No. 3, March 1991.

Budhathoki, N.R. and Haythornthwaite C. 2012. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. American Behavioral Scientist 2012.

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

Girres, J.F. and Touya, G. 2010. Quality assessment of the French OpenStreetMap dataset. Trans. GIS 2010, 14, 435–459.

Goetz, M. and Zipf, A. 2012. Towards defining a framework for the automatic derivation of 3D CityGML models from volunteered geographic information. Int. J. 3-D Inf. Model. 2012, 1, 496–507.

Goetz, M. 2013. Towards generating highly detailed 3D CityGML models from OpenStreetMap, In: International Journal of Geographical Information Science (IJGIS), Volume 27, Issue 5, pages 845-865.

Gösseln, G., Sester, M., 2003. Semantic and geometric integration of geoscientific data sets with ATKIS-applied to geo-objects from geology and soil science. In: Proc. ISPRS Commission IV Joint Workshop 'Challenges in Geospatial Analysis, Integration and Visualization II´, Stuttgart, Germany, 8–10, September (on CDROM).

Gröger, G., Kolbe, T. H., Czerwinski, A. and Nagel, C. 2008. OpenGIS City Geography Markup Language (CityGML) Encoding Standard – Version 1.0.0. OGC Doc. No. 08-007r1.

Grünreich D. 2000. Spatial Data Infrastructures and Geoinformation Engineering – Germany's Approach and Experiences. In: Proceedings of o the United Nations Regional Cartographic Conference for Asia and the Pacific. Kuala Lumpur, Malaysia, April 11 to 14, 2000.

Guptill S and Morrison J (eds) 1995. Elements of Spatial Data Quality. Oxford, Pergamon

Hagenauer J. and Helbich M. 2012. Mining urban land use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. International Journal of Geographical Information Science, 26:6, 963-982.

Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design 37(4) 682 – 703

Haklay, M., Basiouka, S., Antoniou, V. and Ather, A. 2010. How many volunteers does it take to map an area well? The validity of Linus' Law to volunteered geographic information. Cartographic Journal 47: 315-22

Hays, J. and Efros, A.A. 2007. Scene completion using millions of photographs. InACM SIGGRAPH 2007 papers (SIGGRAPH '07). ACM, New York, NY, USA, Article 4.

Helbich, M., Amelunxen, C. and Neis, P. 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. Int. GI_Forum 2012. Salzburg. Austria.

Kim, J.O., Yu, K.Y., Heo, J., Lee, W.H., 2010. A new method for matching objects in two different geospatial datasets based on the geographic context. In: Computers and Geosciences 36 (9), 1115–1122.

Koukoletsos T., Haklay M. and Ellul C. 2012. Assessing data completeness of VGI through an automated matching procedure for linear data. In: Transactioins in GIS, Volume 16, Issue 4, pages 477-498.

Kresse W and Fadaie K. 2003. ISO Standards for Geographic Information. Berlin, Springer-Verlag

Kunze, C. 2012. Vergleichsanalyse des Gebäudedatenbestandes aus OpenStreetMap mit amtlichen Datenquellen. Student research project at the Technical University of Dresden, online avaiable: http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-88141

Li, L., Goodchild, M.F., 2011. An optimisation model for linear feature matching in geographical data conflation. In: International Journal of Image and Data Fusion 2 (4), 309–328.

Li X., Wu C., Zach C., Lazebnik S. and Frahm J. 2008. Modeling and reconstruction of Landmark image collections using iconic scene graphs. In: ECCV, pages 427-440.

Min, D., Zhilin, L., Xiaoyong, C., 2007. Extended Hausdorff distance for spatial objects in GIS. In: International Journal of Geographical Information Science 21 (4), 459–475.

Mooney, P., Corcoran, P. and Winstanley A.C. 2010. Towards quality metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. Pages 514-517

Müller W. and Seyfert E. 1998. Quality assurance for 2.5D building data of the ATKIS DLM 25/2. In: D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS – Between Visions and Applications, Stuttgart, Germany.

Neis, P., Zielstra, D., Zipf, A. and Struck, A. 2010. Empirische Untersuchungen zur Datenqualität von OpenStreetMap – Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. AGIT 2010. Symposium für Angewandte Geoinformatik. Salzburg. Austria.

Neis, P., Zielstra, D. and Zipf, A. 2012. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet. 2012; 4(1):1-21.

Olteanu A. and Mustière S. 2008. Data matching – a matter of belief. In: Headway in Spatial Data Mining, A. Ruas and C. Gold eds., Lecture Notes in Geoinformatioin and Cartography, pp. 501-519.

OSM. 2013. Stats - OpenStreetMap Wiki. Retrieved 05/08/2013, from http://wiki.openstreetmap.org/wiki/Statistics

OSM 2013b. Bing - OpenStreetMap Wiki. Retrieved 10/04/2013, from http://wiki.openstreetmap.org/wiki/Bing

OSM 2013c. Buildings - OpenStreetMap Wiki. Retrieved 10/04/2013, from http://wiki.openstreetmap.org/wiki/Buildings

Rutzinger, M., Rottensteiner, F. and Pfeifer, N. 2009. A Comparison of Evaluation Techniques for Building Extraction From Airborne Laser Scanning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2(1): 11-20.

Uden, M. and Zipf, A. 2012. OpenBuildingModels - Towards a platform for crowdsourcing virtual 3D cities. 7th 3D GeoInfo Conference. Quebec City, QC, Canada.

Walter V. 1996. Zuordnung von raumbezogenen Daten - am Beispiel der Datenmodelle ATKIS und GDF. Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften : Reihe C, Dissertationen ; 480. Dissertation at the University of Stuttgart, Germany.

Walter V. and Fritsch D. 1999. Matching spatial data sets: a statistics approach. International Journal of Geographical Information Science, 13 (5), 445–473.

Yang B., Zhang Y. and Luan X. 2013. A Probabilistic Relaxation Approach for Matching Road Networks, In: International Journal of Geographical Information Science, 27(2):319-338

Zielstra, D. and Zipf, A. 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In Proceedings of 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 10–14 May 2010.