

## Quality assessment for building footprints data on OpenStreetMap

Hongchao Fan, Alexander Zipf, Qing Fu & Pascal Neis

**To cite this article:** Hongchao Fan, Alexander Zipf, Qing Fu & Pascal Neis (2014) Quality assessment for building footprints data on OpenStreetMap, International Journal of Geographical Information Science, 28:4, 700-719, DOI: [10.1080/13658816.2013.867495](https://doi.org/10.1080/13658816.2013.867495)

**To link to this article:** <https://doi.org/10.1080/13658816.2013.867495>



Published online: 27 Jan 2014.



Submit your article to this journal [↗](#)



Article views: 3982



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 95 View citing articles [↗](#)

## Quality assessment for building footprints data on OpenStreetMap

Hongchao Fan<sup>a\*</sup>, Alexander Zipf<sup>a</sup>, Qing Fu<sup>b</sup> and Pascal Neis<sup>a</sup>

<sup>a</sup>*Chair of GIScience, University of Heidelberg, Berlinerstr.48, 69120 Heidelberg, Germany;*

<sup>b</sup>*College of Surveying and Geo-Informatics, Tongji University, Siping Road 1239, 200092 Shanghai, China*

(Received 22 May 2013; final version received 8 November 2013)

In the past two years, several applications of generating three-dimensional (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 buildings' footprints and information about their attributes, which are documented as tags in OSM. Therefore, the quality of 3D buildings relies strongly on the quality of the building footprints data in OSM. This article is dedicated to a quality assessment of building footprints data in OSM for the German city of Munich, which is one of the most developed cities in OSM. The data are 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 criteria, and statistical analysis. The results show that OSM footprint data in Munich have a high completeness and semantic accuracy. There is an offset of about four meters on average in terms of position accuracy. With respect to shape, OSM building footprints have a 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 Volunteered Geographic Information (VGI) projects. Currently, there are more than 1.1 million registered members (OSM 2013a), and OSM is growing rapidly. 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 points of interest (POIs) to the database. According to the statistics (the values are derived from our internal OSM database which is updated daily), on 5 May 2013, the number of buildings in OSM was over 77 million. In Germany, there are almost 9 million objects with 'building = yes' to the same time point.

Currently, building footprints data in OSM is mainly used for reconstructing three-dimensional (3D) buildings. At present, there are several projects which generate and

---

\*Corresponding author. Email: [Hongchao.Fan@geog.uni-heidelberg.de](mailto:Hongchao.Fan@geog.uni-heidelberg.de)

visualize 3D buildings from OSM: OSM-3D (<http://osm-3d.org>), OSM Buildings (<http://osmbuildings.org>), Glosm (<http://glosm.amdmi3.ru/>), OSM2World (<http://maps.osm2world.org/>), etc. And the number of applications based on these projects, i.e. 3D navigation on mobile devices, web-based visualization, and simulation, etc., is increasing. Most of the 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 attributes of building footprints or converted from **the number of stories**, while the majority of 3D buildings have random heights. In OSM-3D, many buildings are modeled in LoD2 (level of detail according to CityGML) in case there are indications for their roof types (Goetz and Zipf 2012). Furthermore, Goetz (2013) proposed a concept for generating buildings in LoD3 and LoD4 in CityGML. Besides, buildings in different levels of details (LoDs) from other sources can be uploaded via OpenBuildingModels and visualized in OSM-3D. But the buildings to be uploaded have to be adapted with the corresponding building footprints in OSM (Uden and Zipf 2012).

Since the 3D buildings in the 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 work presented in this article is dedicated to the quality assessment of building footprints data** in OSM within a test area in Munich (Germany) because, firstly, Munich is one of the most representative cities where OSM data is regarded as well developed, and secondly, Munich is the third largest city in Germany and is very densely built up in the center. **Moreover, the geometries of building footprints in Munich reveal a large diversity.**

In this work, four criteria are introduced for the quality assessment of building footprint data in OSM: (1) completeness, (2) semantic accuracy, (3) position accuracy, and (4) shape accuracy. With respect to these four criteria, OSM data are quantitatively assessed by comparison with the reference data from the German ATKIS (Amtliches Topographisch-Kartographisches Informationssystem – Authoritative 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  $\pm 0.5$  m** (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. Based on **this**, parameters are calculated according to the definitions of the four quality criteria. **Then** the differences between the two data sets are analyzed and visualized.

The remainder of this article is structured as follows: Section 2 gives an overview of works related to this article, Section 3 introduces the criteria of the quality, Section 4 describes the algorithm for **matching building footprints in two data sets**, Section 5 firstly gives an overview of the two data sets in the test area and then 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 across a wide spread 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 and Zipf 2010, Neis *et al.* 2012) and France (Girres and Touya 2010), and more detailed investigations about point (Neis *et al.* 2010), line (Helbich *et al.* 2012), and polygonal (Mooney *et al.* 2010) objects that can be found in the project's database. As mentioned by Hagenauer and Helbich (2012), nearly all 'empirical studies indicate that urban areas are better mapped' in OSM. This is not surprising since most urban areas with a higher population density inherit larger numbers of contributors which influence the quantity and quality of the collaboratively crowd-sourced OSM objects (Girres and Touya 2010, Haklay *et al.* 2010, Neis *et al.* 2012).

In contrast to the 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 (Kunze 2012) which applied several methods to assess the completeness of the building information in OSM in comparison to an administrative data set 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 a hexagon/square instead of individual correspondence. Furthermore, position accuracy and shape characters are not compared.

The most common elements of quality assessment used in the aforementioned research works are position accuracy and completeness. Furthermore, 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 into three types: elements for geographic data bases, elements for data modeling, and elements for the spatial data. Girres and 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 10 years, most of the 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, Min *et al.* 2007, Mustière and Devogele 2007, Olteanu and 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 multistage 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 it is globally consistent. Then objects' correspondences are found on the basis of probabilities.

In contrast to the road network matching, there are few researches for matching area objects which are revealed as polygonal objects, such as residential region, water body, forest, and meadow. 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 a corresponding point pair between a polygonal object pair with a string matching method based on a confidence region model of a line segment. However, these methods are restricted to a low density of polygons to be paired. In a case where neighboring polygons are located immediately next to each other and similar in shape and size, for instance, polygons of building footprints in a densely built-up urban area, there will be error matching.

For this reason, the afore mentioned approaches cannot be used to identify corresponding polygons in two building footprints data sets. In this article, an 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: (1) by giving a list of vertices, or (2) 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 the 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  $\theta_i$  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.

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

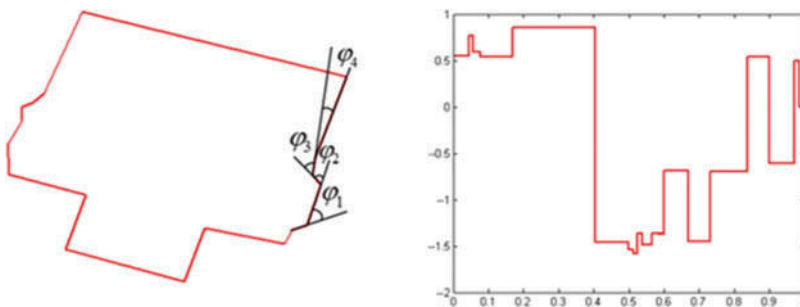


Figure 1. Tangent space representation of polygon.

rotation because it contains no orientation information. Furthermore, it is invariant to scaling, since the normalized length makes it independent of the polygon size.

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

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

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 and set as a reference point for the calculation of the tangent angles. Note that  $S(A, B)$  actually denotes the dissimilarity between  $A$  and  $B$ . The smaller  $S(A, B)$  is, the more similar the two polygons are. In the case where  $A$  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.

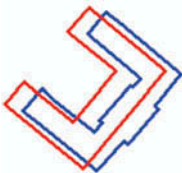
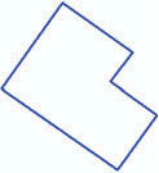
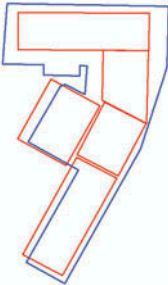
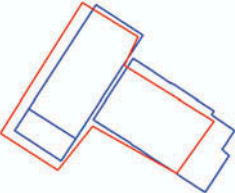
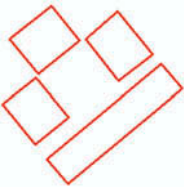
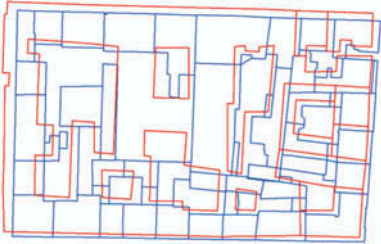
**Completeness** – this is a measure of the lack of data, which does not record objects that are expected to be found in the database, or excess data that should not be included. Regarding the data of building footprints in OSM, the completeness is defined as the difference in the area covered by OSM buildings and the area covered by 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, and height.

**Semantic accuracy** – this investigates if buildings in the real world are recorded indeed as building objects in OSM as well as measuring the percentage of building objects in OSM that 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 the 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 the reference data are in blue.

According to the OGC standard of CityGML building models (Groeger *et al.* 2008), semantic hierarchy and geometrical LoD relate themselves inherently. Hence, these six kinds of relations denote different forms of 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 a higher level on the semantic hierarchy.
- 1:0 relation: a building in OSM is actually not a building (semantically wrong) in 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 a lower level on the semantic hierarchy.
- n:m relation: the buildings are incorrectly recorded with respect to semantic accuracy.

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			

In sum, a building is correctly recorded in OSM in terms of semantic accuracy 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 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 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 ‘correspondence’ here has a twofold meaning: (1) the relations among building footprints in OSM and ATKIS, and (2) 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 first. For building footprints with a 1:1 relation, their corresponding points are found 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  $G_{\text{OSM}}$  be the OSM data set and  $G_{\text{ref}}$  be the reference data set. For a building footprint  $\text{foot}_{\text{osm}_i}$  in  $G_{\text{OSM}}$ , the building footprints in  $G_{\text{ref}}$  will be checked if they are intersected with the lines of a polygon of  $\text{foot}_{\text{osm}_i}$ . In the case that there is an intersection by  $\text{foot}_{\text{ref}_j}$ , the intersected area is calculated first as  $\text{Area}_{\text{overlap}}$ . Since most of the 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 an offset between footprints in OSM and the reference data due to the distortion caused by the oblique view of the used sensors. Considering this factor, large buildings in OSM have a larger percentage of area overlap with their correspondence in the reference data, while small and high buildings might have a smaller percentage of area overlap with their correspondence. The threshold of the judgment actually depends strongly upon the parameters of the Bing map images used for digitalization in OSM. In their work, Rutzinger *et al.* (2009) found 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

$$\frac{\text{Area}_{\text{overlap}}}{\min(\text{Area}(\text{foot}_{\text{osm}_i}), \text{Area}(\text{foot}_{\text{ref}_j}))} > 30\% \quad (2)$$

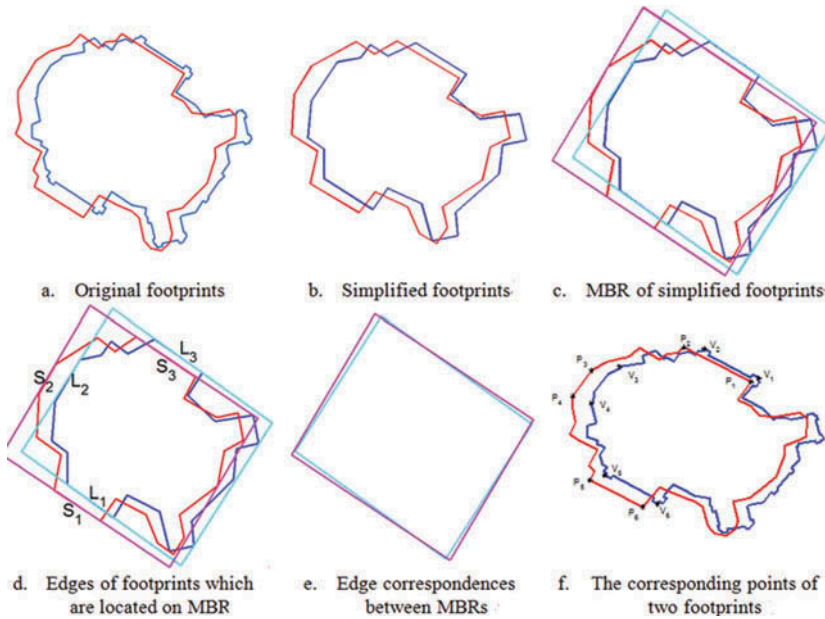
then the footprints  $\text{foot}_{\text{osm}_i}$  and  $\text{foot}_{\text{ref}_j}$  are matched. A 1:1 relation is identified when a footprint in  $G_{\text{ref}}$  can only be matched to one footprint in  $G_{\text{osm}}$ , while 0:1 or 1:0 relations are indicated to the case that the footprint cannot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

#### 4.2. Find identical points of matched building footprints pairs

For the polygon pairs with a 1:1 relation, their corresponding points can be found efficiently by using the following process based on the reality that there is not much difference in shapes, rotation, and scale between OSM building footprints 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 2, whereby a polygon in red stands for building footprints in OSM, while a polygon in blue stands for building footprints in ATKIS.

As shown in Figure 2a, the footprints from different data sets might be formed at a different LoD in terms of geometry. This will lead to a 1: $n$  correspondence among polygon points. To avoid this kind of effect, key points of footprints (Figure 2b) are extracted, first of all, using the Douglas–Peucker algorithm (Douglas and Peucker 1973). Then a minimum bounding rectangle (MBR) is calculated, respectively, for the two polygons (as shown in Figure 2c, rectangle in cyan is the MBR for the OSM footprint and the rectangle in magenta is the MBR for the ATKIS footprint). In the next step (Figure 2d), the edges of the building footprint are marked if they are located on the edges





**Figure 2.** An example of finding identical points of paired footprints.

of its MBR. Then the OSM MBR is shifted to the center of the ATKIS MBR (Figure 2e), so that the edges of these two MBRs can be matched if they are located (almost) at the same place. Finally, edges of footprints can be matched if they are marked to the same edge of MBRs. As shown in Figure 2f, 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  $10 \times 10$  km for almost the entire city of Munich. The OSM data are dumped from our internal database on 10 May 2013. The reference data are ATKIS data in the year 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 close 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  $3 \times 3$  neighborhood of the current cell (where the current footprint is indexed). A sensitivity analysis of the cell size showed that if the cell size is smaller than 1.5 m, there is not much difference in computation time thanks to the high performance of the computer. But, when the cell size is larger than 2 m, the computation time also becomes longer, because the number of buildings within a cell is increased while increasing the cell size. Finally,  $15 \times 15$  grid cells are used both for fast computation and for better illustration of results.

### 5.1. Quantitative assessment of the matching results by using the 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 center 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. Table 2 shows the results of matching using the 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.

### 5.2. Quality of data completeness

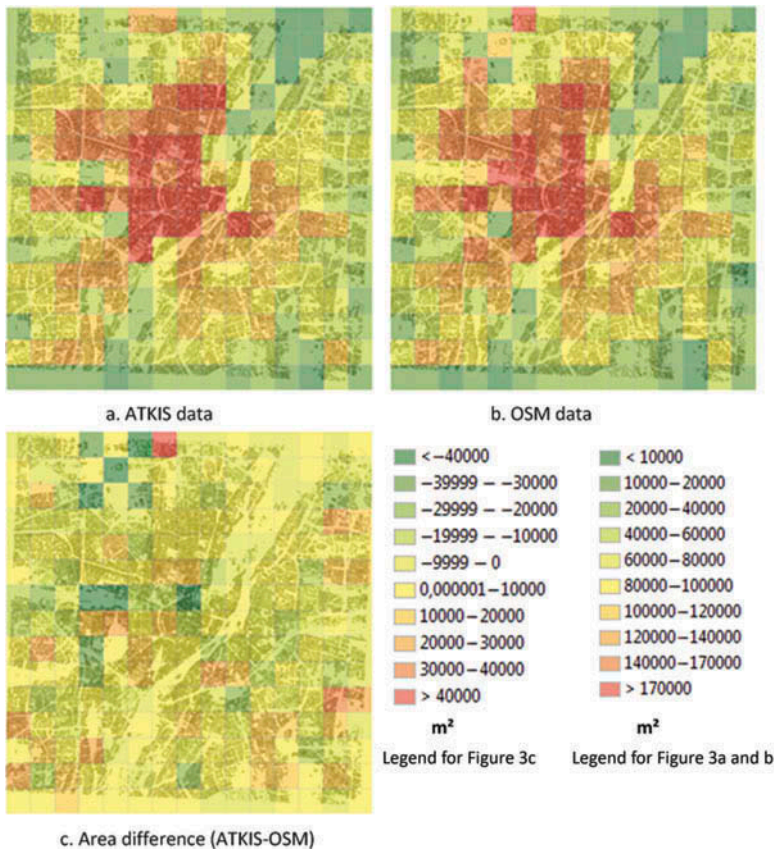
In terms of area covered by buildings, the city of Munich is almost completely mapped, because the total area of buildings in OSM data is even slightly larger than that in ATKIS data. Figure 3 shows the cell-based distribution of building areas in both data sets, as well as their differences (ATKIS-OSM) in cells, whereby the area of each cell is calculated as the sum of area of all buildings located in the cell. Comparing Figure 3a and b, the two distributions are almost the same. This verifies the fact that almost all the areas covered by buildings have been mapped as buildings in OSM. In most of the cells, the differences are within 10,000 m<sup>2</sup> (Figure 3c).

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 a few buildings are recorded with building types, even fewer buildings have attributes of height information and the numbers of stories. Both data sets contain few attributes of ‘building name’, because normally only landmarks and commercial and public buildings have a name, whereas most residential buildings do not. In this context, it can be stated that more than 50% of the buildings which have names are tagged with names.

Although the mappings of the built-up area in ATKIS and OSM are almost the same, it does not mean that all the buildings in ATKIS are mapped in OSM. Therefore, the completeness has to be evaluated by analyzing the matching results. In terms of the number of buildings, there are 33,911 buildings in ATKIS which cannot be matched to the OSM data, while 1233 buildings in OSM have no correspondent ones in ATKIS. From this point of view, the completeness of OSM is  $66.1\% (100,014 - 33,911) / 100,014 = 66.1\%$ , note that 100,014 is the total amount of buildings in ATKIS in the test bed).

**Table 2.** Statistics of the matching results using the 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%	

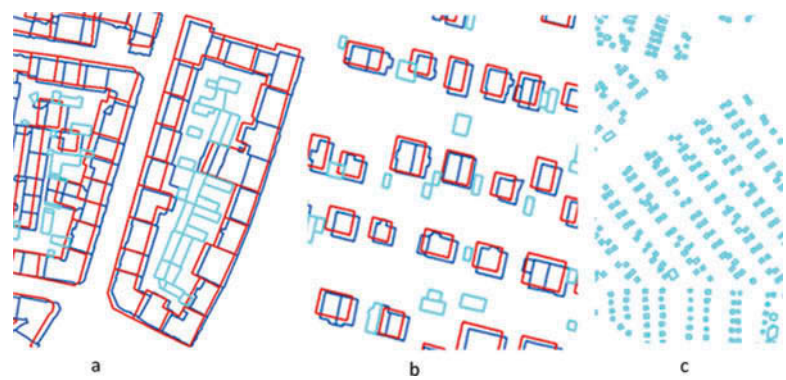


**Figure 3.** Cell-based distribution of building areas and their differences.

**Table 3.** Completeness of building footprints in OSM.

	Area cover (m <sup>2</sup> )	Buildings with types	Buildings with name	Buildings with height	Buildings with numbers of stories
ATKIS	18486805.65	100%	5.24% <sup>1</sup>	100%	100%
OSM	18707108.84	8.46%	2.82%	0.41%	0.06%

Through a manual inspection of the two data sets and the Bing map, the 33,911 buildings can be classified in three types (Figure 4): (1) most of them are located inside of yards formed by terraced buildings, and they are normally occluded by the terraced buildings around them; (2) many smaller buildings, like garages, can be very difficult to identify on Bing map images; in addition, their roofs normally have a low contrast to the ground and roads; (3) in some regions, villas and other buildings (garages) are small and mostly occluded by trees, which creates difficulty for the digitalization on the Bing map. In Figure 4, ATKIS footprints are visualized in blue, while OSM footprints are visualized in red, and the ATKIS buildings which are not mapped in OSM are highlighted in cyan; the scale is 1:2500.



**Figure 4.** Three types of buildings that have a 1:0 relation with OSM data.

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 corresponds exactly to 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, it is not semantically accurate. The **polygonal** object should be called the ‘building group’ if it is matched with several buildings in ATKIS. Or, **it should be called the ‘building part’ in a case in which 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 the OSM data, while there are 100,014 buildings in the ATKIS data. As shown in **Table 4**, almost all the footprints can be matched with ATKIS data except 1233 buildings with a 0:1 relation, because the ATKIS data used in this work are three years older than the OSM data. These buildings are newly constructed in the last three years, according to our local knowledge in Munich. Based on a visual inspection of the Bing map image and Google Maps, 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 relations (1:1 relations), which means 21,775 buildings are correctly recorded in the OSM data with respect to semantic. A total of 13,131 buildings are semantically coarsely recorded, since they have an  $n:1$  relation with building footprints in the ATKIS data. And, 266 recorded buildings are semantically more detailed than the ATKIS data, as they have an  $1:n$  relation. Then, the semantic accuracy of OSM building footprints is calculated as:  $\frac{21,775+1233}{39,364} = 58.45\%$ . The value means that each polygon of the 58.45% polygonal objects (with ‘building = yes’) in OSM corresponds exactly to a building in the real world. Approximately 40% of polygonal objects

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

Relation	1:1	1:0	1: <i>n</i>	<i>n</i> :1	0:1
Amount	21,775	33,911	266	13,131	1233





**Figure 5.** Distribution of difference of number of buildings in grid cells.

(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.

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

#### 5.4. Position accuracy

The position accuracy is investigated by calculating the average distance between the corresponding points of footprints pair in two data sets. Hence, only the buildings with a 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

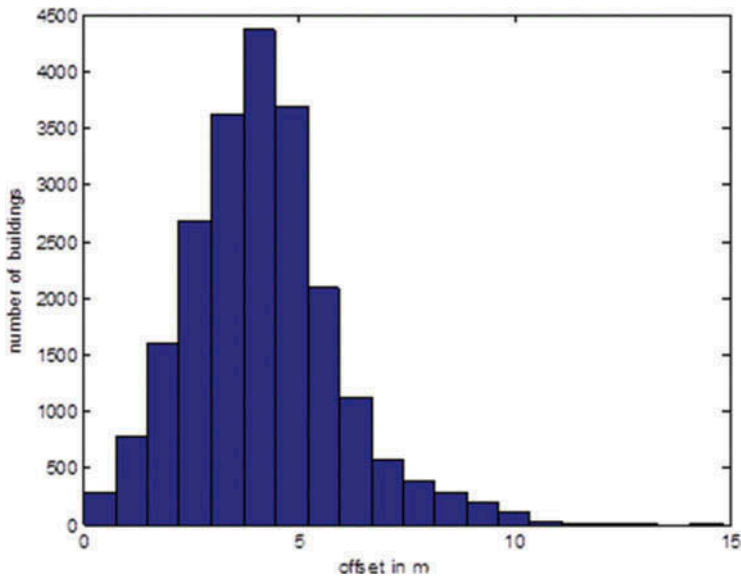


Figure 6. Distribution of offsets from OSM building footprints to ATKIS building footprints.

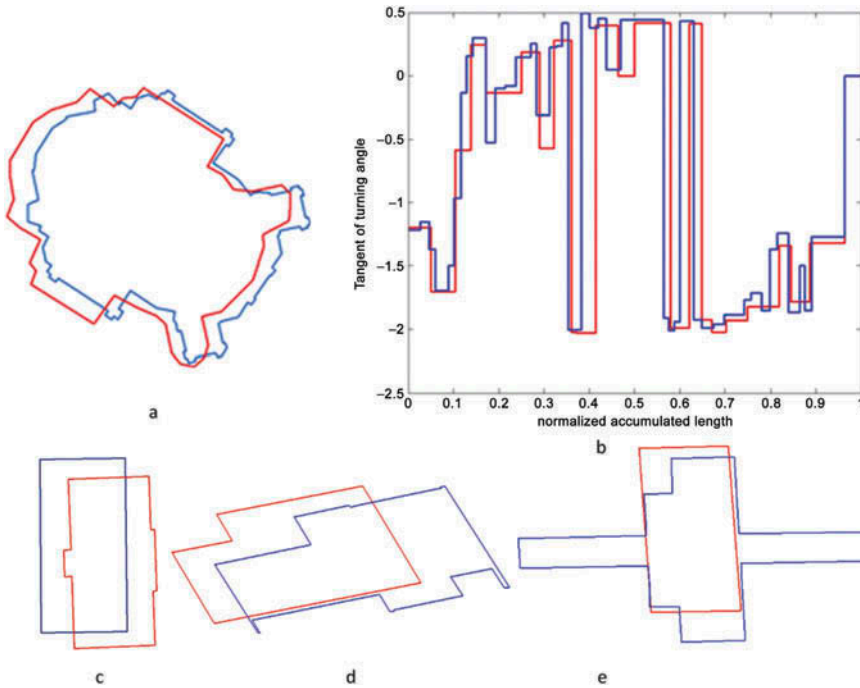
As shown in Table 5, the average offset of OSM building footprints to ATKIS building footprints is 4.13 m with the standard deviation of 1.71 m. The largest offset is nearly 15 m, while the smallest offset is less than a centimeter. The distribution of the offsets is close to a normal distribution with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

Note that the precision of building footprints data in ATKIS is  $\pm 0.5$  m, while the precision of Bing maps imagery in Munich is estimated as 3–4 m 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 a 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, and 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



**Figure 7.** Polygon similarity calculated from their turning function distance.

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 different. In order to make a comparison globally, the similarities have to be normalized by setting the rectangularity of a polygon (polygon A in Equations (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.

$$\text{Rectangularity} = \frac{\text{Area (polygon)}}{\text{Area (MBR)}} = S_n(A, \text{MBR}) \quad (3)$$

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

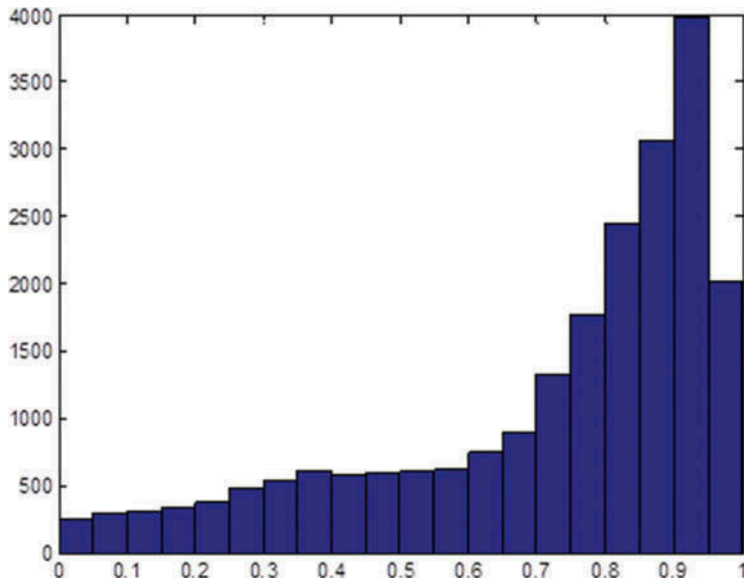
$$\frac{1 - S_n(A, B)}{d(A, B)} = \frac{1 - S_n(A, \text{MBR})}{d(A, \text{MBR})} \rightarrow S_n(A, B) = 1 - d(A, B) \frac{1 - S_n(A, \text{MBR})}{d(A, \text{MBR})} \quad (4)$$

whereby  $d(A, \text{MBR})$  is the dissimilarity of MBR to footprint A calculated by Equation (1), and  $d(A, B)$  is the dissimilarity of footprint B to footprint A calculated by Equation (1).

The principle of the normalization process in Equation (4) can be explained as follows: the ratio of the normalized dissimilarity to the dissimilarity calculated using Equation (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



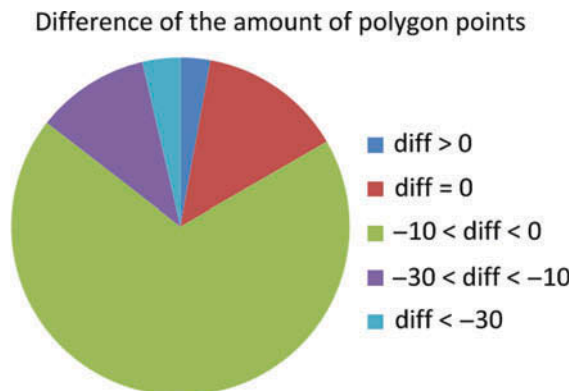


**Figure 8.** Distribution of shape similarities of corresponding building footprints.

Figure 7a) is  $d(A, MBR) = 1.47$ . The rectangularity of the ATKIS footprint is 0.72, which is treated as the normalized similarity of the MBR to the ATKIS footprint  $S_n(A, MBR) = 0.72$ . The dissimilarity of the OSM footprint (red polygon in Figure 7b) is calculated using Equation (1),  $d(A, B) = 1.18$ . Then the normalized similarity of the two footprints in Figure 7a can be obtained as  $1 - 1.18 \times (1 - 0.72)/1.47 = 0.78$ . Furthermore, Figure 7 shows three more pairs of building footprints, their normalized similarities are 0.92 (Figure 7c), 0.51 (Figure 7d), and 0.11 (Figure 7e), respectively.

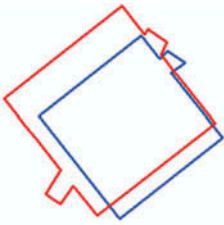

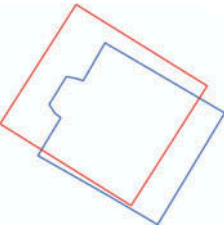

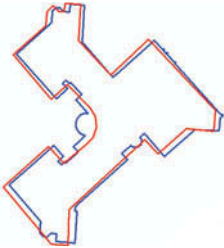

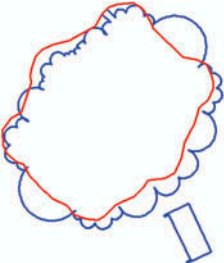

Figure 8 shows the distribution of similarities among corresponding building footprints in OSM and ATKIS. Obviously, there is a concentration peak between 0.7 and 1. It means that most of the building footprints in OSM have a high similarity (more than 70% similar) to their corresponding ones in ATKIS.

In order to find the reason for the dissimilarity of the corresponding building footprints in the two data sets, the numbers of points which form building footprints are analyzed. The chart in Figure 9 denotes that most of the building footprints in OSM



**Figure 9.** Chart diagram of the differences in terms of number of points.

**Table 6.** Examples of building footprints in OSM and ATKIS.

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

contain up to 10 points less than their corresponding ones in the ATKIS. In other words, OSM building footprints are a slightly simplified version of ATKIS building footprints.

In Table 6, four typical examples of the difference in OSM (red lines) and ATKIS (blue lines) are demonstrated. Only in very few cases, footprints in OSM are a little bit

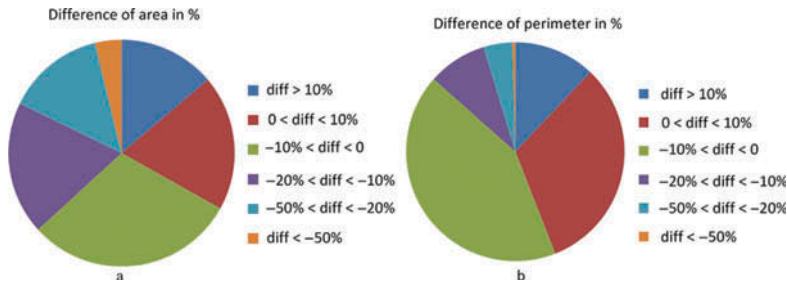


Figure 10. Chart diagrams of the differences in terms of area and perimeter.

more complicated than those in the ATKIS data. For instance, in Table 6a, the fire escape was digitalized as a part of the footprint in OSM, while it was neglected in the ATKIS data, because the footprint in OSM is digitalized according to the image of the roof from a bird's eye view, and hence the fire escape cannot be differentiated from the main part of building. In most cases (Table 6b–d), footprints in OSM are simplified. The more complicated a footprint is in reality, the larger the difference between OSM and ATKIS. There are three major reasons. Firstly, it is difficult to follow the architectural details according to roofs from a bird's-eye 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 a high similarity, in terms of shape, to the one in the reality.

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

## 6. Conclusion and future works

This article presents an approach to assess the quality of OSM building footprints data. A case study in Munich (Germany) was conducted. The results show that OSM building footprints data has a high completeness in terms of area covered. Almost all of the constructed area in the city is mapped as buildings in OSM. However, OSM building footprints data still lacks 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 a similar contrast to the ground and roads in their surroundings. The third type of building is referred to as villas in forest areas. The occlusion by vegetation makes it difficult to identify on Bing Image. On the other hand, there are new findings on OSM. More than 1200 newly constructed buildings are found in OSM which are not recorded in the 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 m offset on average to their corresponding ones in ATKIS. The footprints in OSM are highly similar to those in ATKIS in terms of shape. Most of the OSM building footprints are almost identical to those in ATKIS. However, there is a slight difference. Many buildings in the OSM footprints consist of lesser polygon points than the ATKIS footprints. Some architectural details are missing if buildings are complicated in structure. Furthermore, attributive information is not very rich.

The main reason for the afore mentioned 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 a result of the distortion of buildings due to the oblique aspect of the sensor used, while the missing geometrical detail is caused by the limited resolution of Bing map images. The semantic accuracy of OSM in dense urban areas is rather low because many buildings in densely constructed areas 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 the power of VGI: a huge number of volunteers for contribution and high frequency of updating the data.

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 regions (i.e. Baden-Wuerttemberg) containing large, medium, and small cities, as well as rural areas, will be evaluated against the authority data.

## Funding

This work is supported by NSFC (National Natural Science Foundation of China) projects No: 41101443 and 41201425, and the Klaus Tschira Foundation (KTS) in Germany.

## Note

1. The field of building name in ATKIS is 100% filled. However, only about 5.24% of the buildings in Munich have individual names, while most of them have a value of 'nameless' for the attribute field.

## References

- Arkin, E.M., *et al.*, 1991. An efficiently computable metric for comparing polygonal shapes. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13 (3), 129–137.

- 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. *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. *Transactions in GIS*, 14 (4), 435–459.
- Goetz, M., 2013. Towards generating highly detailed 3D CityGML models from OpenStreetMap. *International Journal of Geographical Information Science (IJGIS)*, 27 (5), 845–865.
- Goetz, M. and Zipf, A., 2012. Towards defining a framework for the automatic derivation of 3D CityGML models from volunteered geographic information. *International Journal of 3-D Information Modeling*, 1, 496–507.
- Gösseln, G. and Sester, M., 2003. Semantic and geometric integration of geoscientific data sets with ATKIS-applied to geo-objects from geology and soil science. In: *Proceedings of the ISPRS commission IV joint workshop 'challenges in geospatial analysis, Integration and Visualization II'*, 8–10 September, Stuttgart [on CDROM].
- Gröger, G., et al., 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 the United Nations regional cartographic conference for Asia and the Pacific*, 11–14 April, Kuala Lumpur.
- Guptill, S. and Morrison, J., eds., 1995. *Elements of spatial data quality*. Pergamon: Oxford.
- 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., et al., 2010. How many volunteers does it take to map an area well? The validity of Linus' Law to volunteered geographic information. *Cartographic Journal*, 47, 315–322.
- Helbich, M., Amelunxen, C., and Neis, P., 2012. Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. In: J. Strobl, et al., eds. *Angewandte Geoinformatik*. Berlin: Herbert Wichmann Verlag, 24–33.
- Huh, Y., et al., 2013. Line segment confidence region-based string matching method for map conflation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 78, 69–84.
- Kim, J.O., et al., 2010. A new method for matching objects in two different geospatial datasets based on the geographic context. *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. *Transactions in GIS*, 16 (4), 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* [online]. Student research project at the Technical University of Dresden. Available from: <http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-88141> [Accessed 8 November 2013].
- Li, L. and Goodchild, M.F., 2011. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2 (4), 309–328.
- Min, D., Zhilin, L., and Xiaoyong, C., 2007. Extended Hausdorff distance for spatial objects in GIS. *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*. New York: ACM, 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, and M. Sester, eds. *'IAPRS', Vol. 32/4, ISPRS commission IV symposium on GIS – between visions and applications*, 7–10 September, Stuttgart.
- Mustière, S. and Devogele, T., 2007. Matching networks with different levels of detail. *Geoinformatica*, 12 (4), 435–453.
- Neis, P., et al., 2010. Empirische Untersuchungen zur Datenqualität von OpenStreetMap – Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. In: *AGIT 2010. Symposium für Angewandte Geoinformatik*, Salzburg.

- Neis, P., Zielstra, D., and Zipf, A., 2012. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4 (1), 1–21.
- Olteanu, A. and Mustière, S., 2008. Data matching – a matter of belief. In: A. Ruas and C. Gold, eds. *Headway in spatial data mining, Lecture Notes in geoinformation and cartography*. Berlin: Springer, 501–519.
- OSM, 2013a. Stats – OpenStreetMap Wiki [online]. Available from: <http://wiki.openstreetmap.org/wiki/Statistics> [Accessed 5 August 2013].
- OSM, 2013b. Bing – OpenStreetMap Wiki [online]. Available from: <http://wiki.openstreetmap.org/wiki/Bing> [Accessed 10 April 2013].
- OSM, 2013c. Buildings – OpenStreetMap Wiki [online]. Available from: <http://wiki.openstreetmap.org/wiki/Buildings> [Accessed 10 April 2013].
- 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.
- Samal, A., Seth, S., and Cueto, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18 (5), 459–489.
- Uden, M. and Zipf, A., 2012. OpenBuildingModels – towards a platform for crowdsourcing virtual 3D cities. In: *7th 3D GeoInfo Conference*, 16–17 May, Quebec City, QC.
- Volz, S. 2006. An iterative approach for matching multiple representations of street data. In: M. Hampe, M. Sester, and L. Harrie, eds. *ISPRS workshop – multiple representation and interoperability of spatial data*, 22–24 February. Hannover: ISPRS, 101–110.
- Walter, V. and Fritsch, D., 1999. Matching spatial data sets: a statistics approach. *International Journal of Geographical Information Science*, 13 (5), 445–473.
- Xiong, D. and Sperling, J., 2004. Semiautomated matching for network database integration. *ISPRS Journal of Photogrammetry & Remote Sensing*, 59 (1-2), 35–46.
- Yang, B., Zhang, Y., and Luan, X., 2013. A probabilistic relaxation approach for matching road networks. *International Journal of Geographical Information Science*, 27 (2), 319–338.
- Zhang, M., 2009. *Methods and implementations of road-network matching*. Thesis (PhD). Institute for Photogrammetry and Cartography, Technical University of Munich, Munich.
- 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*, 10–14 May, Guimarães.