Assessment and Visualization of OSM Building Footprint Quality

Fabian Müller, Ionuț Iosifescu, Lorenz Hurni

Institute of Cartography and Geoinformation, ETH Zürich

Abstract. OpenStreetMap (OSM) is a crowd sourcing project, mostly known for its open data on street topology and land use. However, since 2007, OSM contributors have started to also map building footprints. In several cities or even countries, these footprint datasets are now deemed to be near completion and increasingly used as an alternative to proprietary data. In this paper we are investigating the completeness and other quality characteristics of the OSM building footprints for an entire country, namely Switzerland, by comparing them with official data sources.

The overall quality of OSM data has been the subject of some research especially regarding the street network. However, an automated process to deliver comprehensive quality metrics has rarely been discussed for building footprints and even more scarcely implemented for general use. Moreover, in order to use OSM data and thus profit from their permissive licensing in both science and commerce, a detailed analysis in such a scale that it is applicable for investigation areas of any scope, be it county or even country level, is irremissible, even though, due to the constantly changing nature of OSM data, such analyses do not hold their validity for very long.

Using reference data, the OSM building footprint data may thus be evaluated in terms of data quality and suitability. To achieve this, a comparison algorithm based on the building footprints' respective centroid distance and a shape signature function has been developed and implemented in Java.

Keywords: OpenStreetMap, swisstopo, building footprints, comparison

1. Introduction

While OpenStreetMap (OSM), a crowd sourcing project, mostly known for its open data on street topology and land use, becomes increasingly popular and is thus also used as a provider for building footprint data under a

liberal licence, completeness and quality of its data vary wildly. Comprehensive quality metrics enable users to decide whether OSM data are fit for their purposes or whether proprietary data must be purchased. While similar studies have already been undertaken for road networks in Germany (Zielstra & Zipf 2010), natural features (Mooney et al. 2010) and general data in France (Girres & Touya 2010), this work aims to enable any interested party to easily recreate comprehensive quality metrics for any region of their choosing by releasing both the program and the result dataset for Switzerland to the general public, allowing up-to-date information to be created and taken into account at decision-making processes and in research.

Completeness may easily be determined in a crude fashion via a buildings per square kilometer count. Quality, however, is best determined via the similarity of the geometries of two building footprints suspected to represent the same building. Various methods to determine quality have been assessed; however, the most reliable metrics have been generated using the turning algorithm described by Arkin et al. (1991), as suggested by Fan et al. (2014) and subsequently used for building footprints in Munich with satisfactory results (Fan et al. 2014a).

By searching for matching buildings between the comparison and the reference dataset and determining the turning function for all found matches, a dataset containing both all matched buildings with their turning value θ and all unmatched buildings is returned.

Next to the OSM data as comparison data, the topographic landscape model (TLM) data from swisstopo, a set of vector features with an accuracy of between 0.2 and 1.5 metres for building footprints, was used as reference data. Both comparison and reference data were then adjusted and parsed into a PostGIS database. Subsequently, a Java program searched for matches and calculated their turning value θ, writing the results into the database.

The Result Dataset has been visualised in QGIS with an innovative approach via hexagonal grids incorporating three levels of detail to display areas of higher population density with better accuracy whilst still retrieving interpretable results for areas of lower population density. Both the developed program and the visualisation methodology may be easily adapted for countries for which a reference dataset of similar quality is available, or to compare different temporal versions of building footprint data. In the spirit of openness promoted by the ICA-OSGeo Open Geospatial laboratory at ETH Zurich, the program will be released to the general public under an open source licence.

2. Materials and Methods

2.1. Creating the Match Dataset

In a preliminary step, the OSM building footprint data covering the extent of Switzerland were downloaded from the project on September 16th 2014 and subsequently adjusted as to be usable for the shape comparison algorithm. In detail, any invalid geometries were fixed, all geometries merged, and eventually multipolygons turned into single polygons; leaving just the building outlines and ignoring features like inner courts. The TLM data were already fit for use. Both datasets were then imported into a PostGIS database and their centroids calculated.

A specifically written Java program utilising the open source Java GIS toolkit GeoTools connected to the database. For each polygon representing a building footprint in the comparison dataset, the nearest polygon in the reference dataset was determined – provided such a polygon existed within a certain distance of their respective centroids. This maximum distance may be set in the program's parameters; for the purposes of this paper, 20 meters were used. Any centroid distance larger than 20 meters thus disqualified a building in one dataset from being a building's counterpart in the other dataset.

If a match was found, the polygons' turning function Θ was calculated. The turning function is a translation-, scale- and rotation-independent representation of a polygon based on the relative length of each perimeter section and its angle relative to the following perimeter section. The integral of two turning functions may then be used as a polygon distance function and thus as a similarity metric, providing their turning value θ and yielding intuitive results in most cases and allowing building footprints beneath a certain threshold to be defined as identical.

The mean centroid of both geometries, the IDs of both buildings and their θ were then written into a new result dataset within the PostGIS database. Buildings from either initial dataset without match were also inserted into the Result Dataset, bearing null values instead of a matched building's ID and a calculated θ. The result dataset thus included both matched and non-matched buildings as a georeferenced point cloud.

2.2. Visualising the Result Dataset

The Result Dataset was both too extensive and to erratic to be meaningfully visualised in a concise manner as points. Thus we approached the data using various binning methods, in which the points are summarised in different ways: in a hexagonal grid structure (hexagonal binning), in the communal boundaries of Switzerland, and in a combination of both, a

newly invented technique for which we propose the name “layered hexagonal binning”.

For hexagonal binning, hexagonal grid structures of various scales were created using the “Create grid layer” method of QGIS’ MMQGIS plug-in. In detail the grid structures were created with hexagon perimeters of 1 km, 2 km, 4 km, 8 km and 16 km. The communal boundaries of Switzerland were taken from swisstopo’s free dataset “swissBOUNDARIES 3D”.

After importing the hexagonal grids as well as the boundaries into the PostGIS database, simple point-in-polygon queries created a new table for each grid and the boundaries, allowing the data to be displayed as choropleth maps.

For layered hexagonal binning, the communal boundaries were merged into three classes depending on their building density. These classes were then used to display only hexagonal bins of a certain radius – the higher the building density, the smaller the displayed radius, and thus: the higher the level of detail.

The dataset was also visualised in form of an automatically generated atlas using QGIS’ atlas functionality. The result was a set of 2435 maps, one for each commune of Switzerland, displaying both the centroids of TLM and OSM buildings without matches and the centroids of matched buildings as to prove insight into which streets, quarters or areas were already present in OSM and how completely they were mapped. The point signatures showing the latter were scaled according to their θ, allowing for a swift judgement regarding the quality of buildings mapped in OSM in comparison to their TLM counterparts.

3. Results

An overview of the result dataset in 1-km-bins *(Figure 2)* shows that at least some matches were found almost everywhere where data of both initial datasets existed, except for most of the cantons of Ticino, Obwalden and Fribourg. In some places, especially in the cantons of Basel-Stadt and Basel-Landschaft, all TLM buildings were matched by OSM buildings, while the opposite was not true: In these areas the OSM dataset is – quantitatively – better than its TLM counterpart.

There also are a few areas were both OSM and TLM buildings are found, but none of them were deemed the same building by the algorithm. These areas thus serve as an indicator for the quality of the matching process – assuming that the TLM dataset is almost complete and OSM building footprints are not created where no buildings are found in reality, such

areas should not be found with a perfect matching algorithm. However, these assumptions are most certainly only partially correct at best – the TLM dataset, for example, may not include illicit constructions and, due to its publishing interval, lag behind, and the OSM dataset may not necessarily be up to date regarding demolitions and new constructions, especially where data are achieved via satellite image tracing as opposed to fieldwork.

The average θ of all 1,266,121 matched buildings is 1.2504, with individual θ's ranging from 2.17E-8 to 9.44. Of the matched buildings 4,942 have a θ worse than 5, while 192,461 matched buildings have a θ better than 0.1. Examples of such θ are shown in *Figure 1*.

In the Swiss communes, the average θ ranges from 0.37 to 2.51 *(Table 1)*, counting only communes with ten or more matched buildings. On average, the average θ is 1.25 – quite close to that of all matched buildings individually –, and the average θ of the median is 1.22.

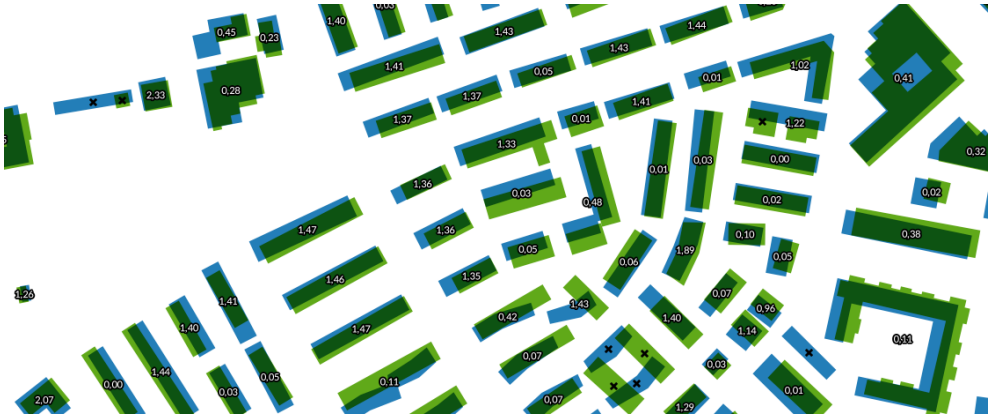


Figure 1. *Exemplary data of Zurich – blue marks TLM data, green marks OSM data. The numbers denote the turning value θ.*

	Average θ	Matched buildings	Total buildings
16 Flerden	0.37	14	196
17 Marchissy	0.37	11	147
…	…	…	…
586 Muttenz	*(median)* 1.22	457	3250
…	…	…	…
2263 Oberrohrdorf	2.44	756	864
2266 Rivaz	2.51	41	69
Average	1.25	556	996

Table 1. *An excerpt of the communes sorted by their average θ.*

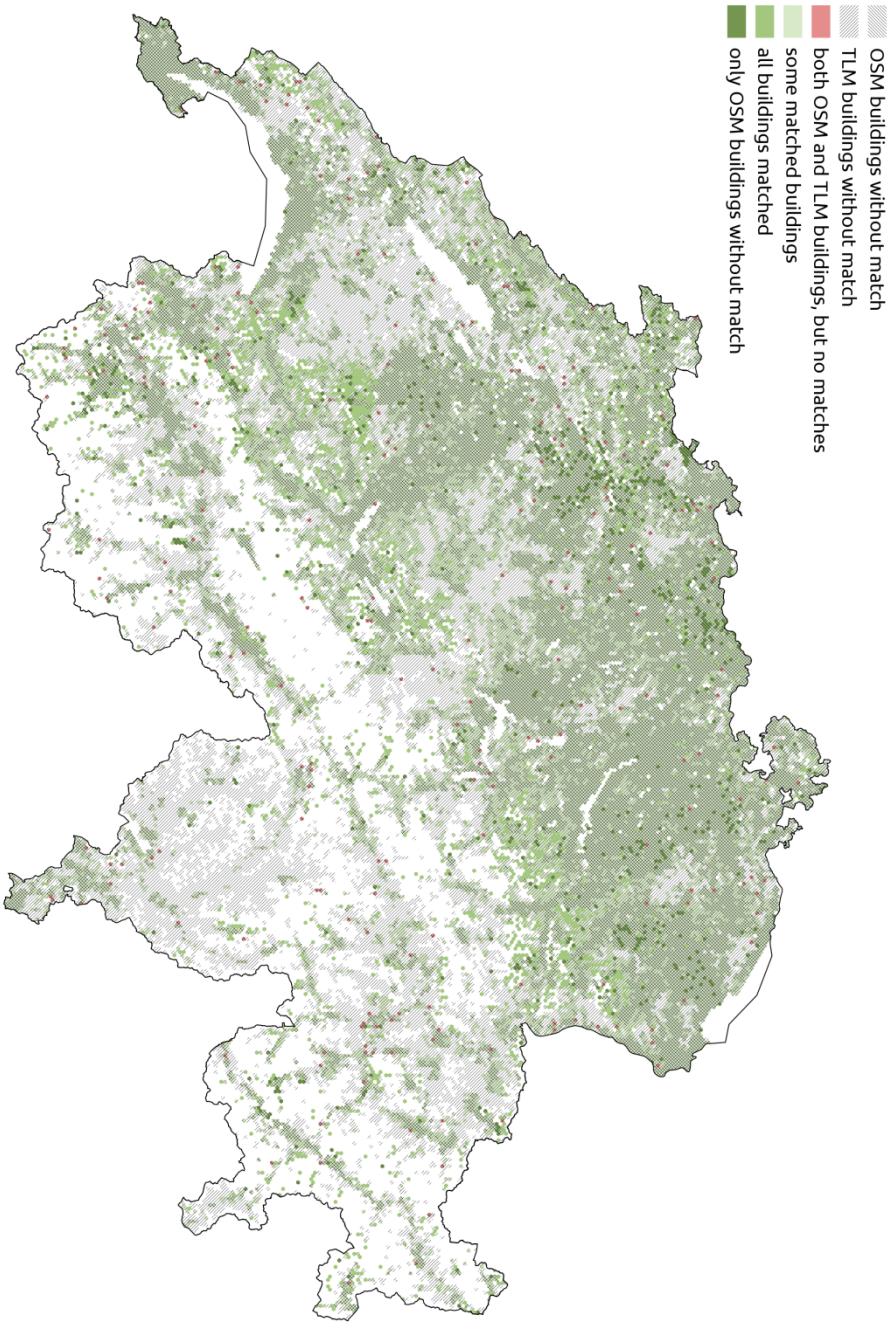


Figure 2. *An overview of the dataset in 1-km-bins.*

Three characteristic numbers may be derived from the dataset: The OSM/TLM building ratio *(Figure 3)* indicates the completeness of the OSM dataset, the matched/unmatched ratio *(Figure 4)* indicates the quantitative similarity of both datasets, i.e. whether they strive towards bijectivity in a given area, and the average θ *(Figure 5)* indicates the qualitative similarity of all matched buildings.

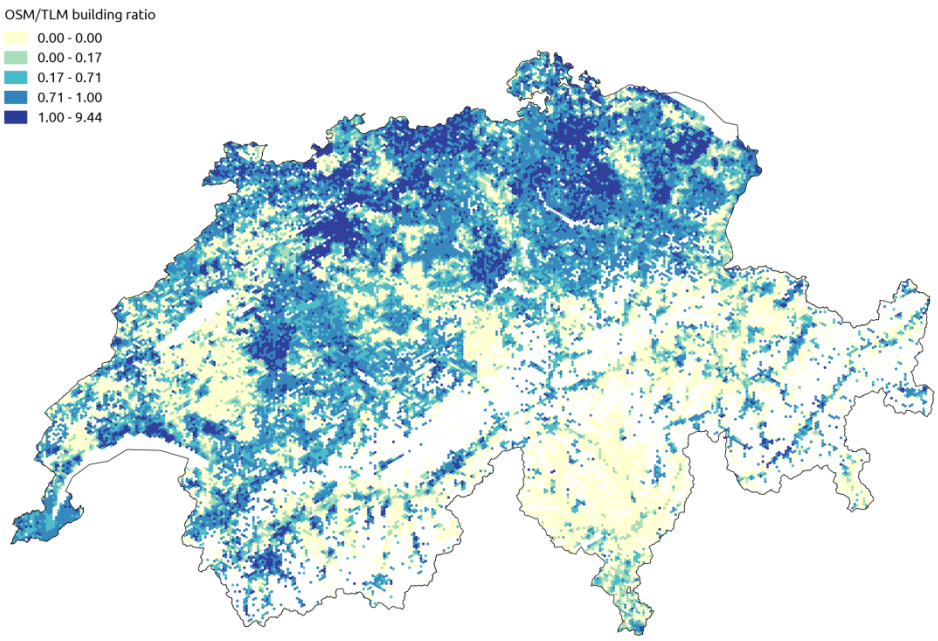


Figure 3. *OSM/TLM building ratio.*

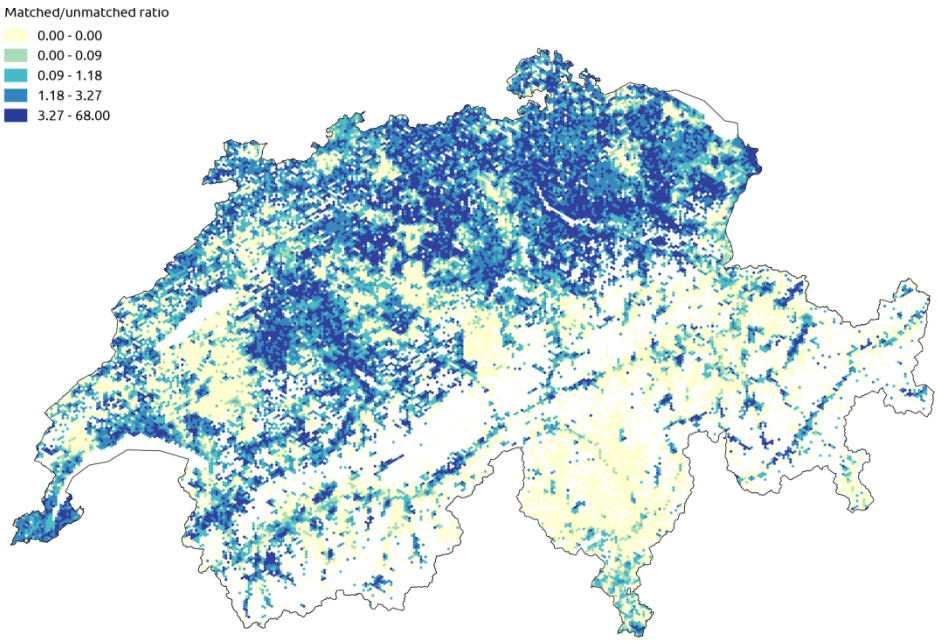


Figure 4. *Matched/unmatched ratio.*

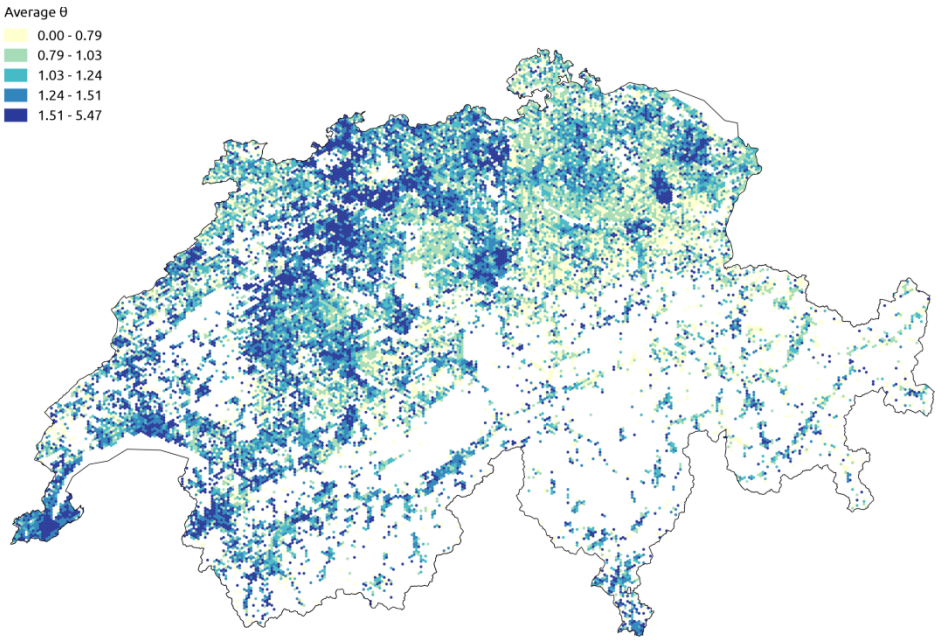


Figure 5. *Average θ.*

The layered hexagonal binning map *(Figure 6)* shows at a glance both the quality and the quantity of OSM's footprint dataset – the smaller the bin resolution, the higher the building density, and the lighter the colour, the better the shape similarity between OSM and TLM data.

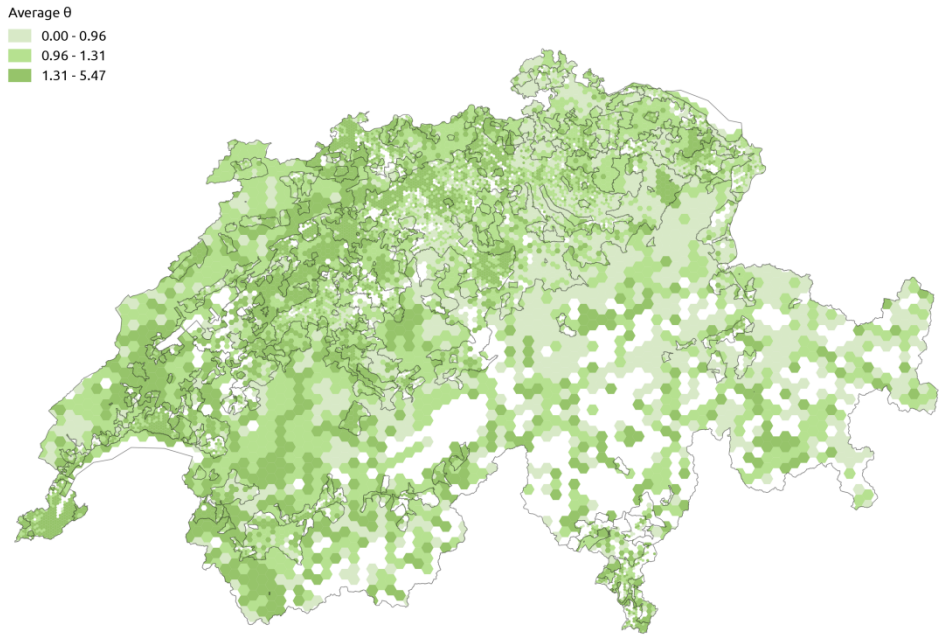
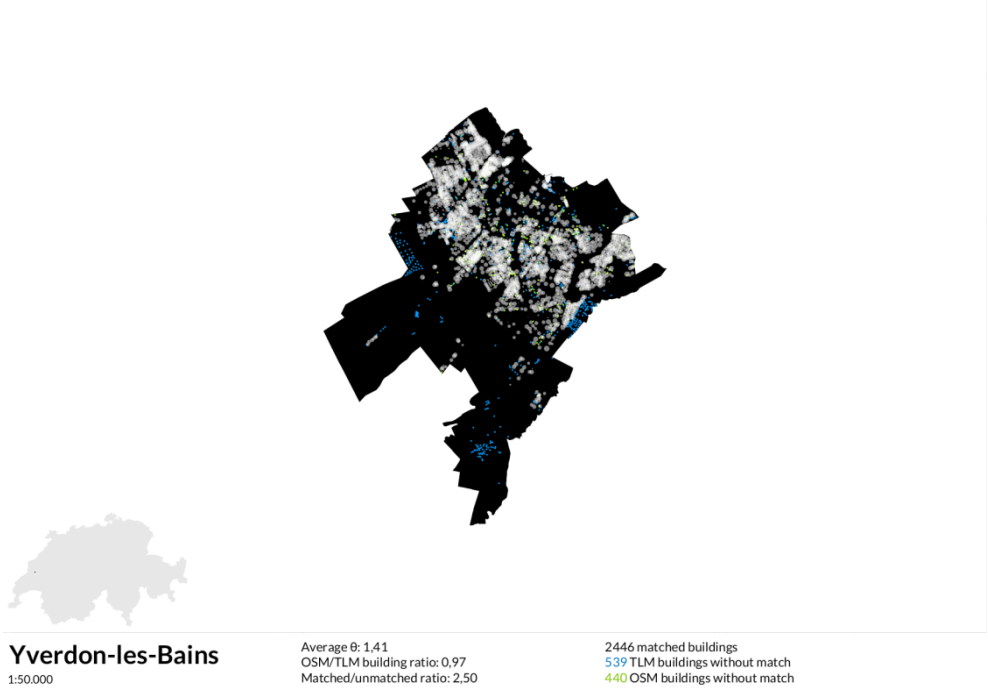


Figure 6. *The layered hexagonal binning map derived from the result dataset.*

The atlas provides more detailed queries for a specific commune, displaying individual buildings as point signatures, as seen in *Figure 7.* Large-scale mapping projects may thus rely on these maps to deduce whether OSM building footprint data are comprehensive in their area. Furthermore, OSM contributors may use these maps to deduce where mapping efforts should be undertaken.

Figure 7. *Example of an average atlas page: Yverdon-les-Bains.*



4. Conclusion and Further Work

The greatest part of Switzerland still has some data gaps to be filled in the OSM dataset. However, in some areas, especially urban ones, OSM building footprints are already perfectly usable and sometimes even more comprehensive than their TLM counterparts. The quality of the OSM building footprints is generally good and comparable to that of TLM.

However, the visualised data are only a snapshot of OSM's continuing change. Further work might therefore compare OSM datasets from different points in time both with each other and with a reference dataset to reach conclusions regarding the speed and accuracy with which OSM building footprint data are changing and to what extent the focus of OSM contributors lies on mapping new areas and to what extent it lies on correcting already mapped data. Once multiple such analyses covering a period of one or two years exist, prognoses could be made as to when OSM building footprint data might match proprietary data in terms of quality and quantity. Furthermore, the dataset could be visualised in a manner optimised for the selected scale in an interactive atlas, allowing both a swift overview of the data and in-depth analyses of possible relationships.

References

Arkin E M, Chew L P, Huttenlocher D P, Kedem K, Mitchell J S B (1991) An Efficiently Computable Metric for Comparing Polygonal Shapes. IEEE Transactions on Pattern Analysis And Machine Intelligence, 13:3; pp: 209-216. doi:10.1109/34.75509

Fan H, Zipf A, Fu, Q (2014) Estimation of Building Types on OpenStreetMap Based on Urban Morphology Analysis. From: Huerta J et al. (eds., 2014) Connecting a Digital Europe Through Location and Place, Lecture Notes in Geoinformation and Cartography. Springer, Cham. doi:10.1007/978-3-319-03611-3_2

Fan H, Zipf A, Fu Q, Neis P (2014a) Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science 28:4. doi:10.1080/13658816.2013.867495

Girres J, Touya G (2010) Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS, 2010, 14:4; pp. 435–459. doi:10.1111/j.1467-9671.2010.01203.x

Mooney P, Corcoran P, Winstanley A C (2010) Towards quality metrics for OpenStreetMap. GIS ‘10 Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems; pp. 514-517. doi:10.1145/1869790.1869875

Zielstra D, Zipf A (2010) A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. 13th AGILE International Conference on Geographic Information Science 2010; Guimarães, Portugal.