Flächennutzungsmonitoring VIII
Flächensparen – Ökosystemleistungen – Handlungsstrategien

IÖR Schriften Band 69 · 2016

ISBN: 978-3-944101-69-9

The intrinsic quality assessment of building footprints data on OpenStreetMap in Baden-Württemberg

Hongchao Fan, Anran Yang, Alexander Zipf

Fan, H.; Yang, A.; Zipf, A. (2016): The intrinsic quality assessment of building footprints data on OpenStreetMap in Baden-Württemberg. In: Meinel, G.; Förtsch, D.; Schwarz, S.; Krüger, T. (Hrsg.): Flächennutzungsmonitoring VIII. Flächensparen – Ökosystemleistungen – Handlungsstrategien. Berlin: Rhombos, IÖR Schriften 69, S. 253-260.

The intrinsic quality assessment of building footprints data on OpenStreetMap in Baden-Württemberg

Hongchao Fan, Anran Yang, Alexander Zipf

Zusammenfassung

In dieser Arbeit stellen wir ein Konzept zur Bewertung von der Qualität von Gebäudegrundrissen aus OpenStreetMap (OSM) ohne Verwendung von Referenzdaten vor. Insbesondere wird der Verlauf der Bearbeitung von Stützpunkten und Attributen der Objekte untersucht. Sieben Indikatoren Bewertung der intrinsischen Datenqualität wurden definiert. Für die vorliegende Studie ist am Beispiel von Baden-Württemberg eine PostgreSQL-Datenbank erstellt worden, um ein räumlich-zeitliches Datenmodell zu implementieren, welches sowohl einzelne Objekte als auch Bearbeitungsereignisse (Events) verfolgen kann. Vorläufige Ergebnisse zeigen eine relativ hohe Qualität der OSM-Gebäudedaten, wobei eine Steigerung der Qualität hinsichtlich Semantik, Geometrie und Positionsgenauigkeit als Beitrag der freiwilligen OSM-Bearbeiter zu beobachten ist.

Summary

In this work, we propose a framework to assess the quality of OpenStreetMap (OSM) building footprints data without using any reference data. More specifically, the OSM history data will be examined regarding the development of attributes, geometries and positions of building footprints. In total seven quality indicators are defined for the intrinsic quality assessment. For our case study in the federal state of Baden-Württemberg (BW), Germany, a PostgreSQL database is established based on a spatiotemporal data model which can track both individual objects and editing events on OSM. The preliminary experiments show that the quality of building footprints in BW is relatively high. And the quality in terms of semantics, geometries and positions are getting increasingly high over the time thanks to the considerable contribution of OSM volunteers.

1 Introduction

OpenStreetMap (OSM) is considered one of the most successful and popular volunteered geographic information (VGI) projects, and it has attracted significant and sustained interest in academia, industry, and governmental agencies. Currently, there are almost three million registered members (OSM, 2016), which has led OSM to grow rapidly. With the rapid development of OSM in recent years, especially, 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 latest statistics (the values are derived from our internal OSM database, which is updated daily), the number of buildings in OSM is above 200 million, thereof 18.4 million building footprints in Germany. The research of Fan et al. (2014) demonstrated that the data regarding building footprints on OSM has a high degree of 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 are highly similar to objects originating from authoritative datasets (ATKIS Basic DLM). Moreover, there is more and more information about building heights and roof structures, which is required for the 3D reconstruction. However, due to the crowd-sourcing approach of OSM, the data quality is often doubted.

Kunze et al. (2013) 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 (Kunze et al. 2013; Hecht et al. 2013). As criterion of quality assessment, the work mainly analysed the area difference of a group of buildings within hexagon/square instead of individual correspondence. Fan et al. (2014) addressed the OSM building completeness in Munich, Germany. The authors compared OSM building footprints data with ATKIS data in terms of completeness, semantic accuracy, position accuracy, and shape accuracy. Klonner et al. (2015) also addressed the building completeness and conducted a data quality analysis of building footprints in Bregenz, Austria. Most recently, authors employ authoritative datasets to compare the quality of OSM in terms of the completeness and thematic accuracy (Törnros et al. 2015; Dorn et al. 2015).

The abovementioned methods of quality assessment rely on the access of reference datasets which are unfortunately, in many cases, not available due to contradictory licensing restrictions or high procurement costs. Therefore, intrinsic quality assessment has been introduced in the recent years. Many existing approaches examined OSM data by checking the change history of features. For example, Keßler and Groot (2013) evaluated feature-level attributes such as the number of versions, the stability against changes and the corrections and rollbacks of features so as to infer OSM features' quality. Barron et al. (2014) have developed a comprehensive analysis framework, called iOSManalyser, for investigating the intrinsic data quality of OSM based on its mapping history. In their work, a broad range of more than 25 different methods and indicators were presented to evaluate the quality of an OSM dataset.

This work is dedicated to the intrinsic quality assessment for building footprints data on OSM. First of all, a conceptual framework is developed regarding the intrinsic quality assessment for OSM building footprints data. At the same time, preliminary results of the intrinsic quality assessment in Baden-Württemberg, Germany, will be demonstrated. Furthermore, a conceptual framework for the effective analysis of OSM history data will be presented in order to carry out the intrinsic quality assessment for OSM data.

2 The conceptual framework of intrinsic quality assessment for OSM building footprints data

In total, seven indicators are defined for the intrinsic quality assessment of building footprints data on OSM:

In terms of completeness:

(i) the development of built-up area over time,

(ii) the development of building count over time,

(iii) the development of positional accuracy attributes over time,

(iv) the average vertex displacement of a building footprint when edited by OSM contributors,

and in terms of shape accuracy:

(v) the orthogonality of building footprints,

(vi) the parallelism of building footprint edges to the nearby line segments of roads, and

(vii) the fragmentariness of patterns formed by building footprints.

While the first three indicators and the fifth one are quite obvious and easy to be understood, the fourth, sixth and seventh indicator will be elaborated in the following, in order to give a better depiction of how they are calculated.

(1) Calculation of the average displacement of a vertex on a building footprint polygon after several editing processes on OSM

It is assumed that a vertex $P(x, y)$ of a building footprint polygon is (re-)edited many times $P_i(x_i, y_i)|(i = 1, \dots, n)$ by different contributors on OSM. Due to the non-rectangular shape on the vertex and some other reasons such as shadow or occlusion by vegetation, the contributors might have problems to see the exact vertex on the aerial or satellite images. For this reason, the exact position might be estimated and corrected differently by different OSM contributors. As the result, there are many records of positions of the referred vertex. In order to estimate its most likely location, a grid-based accumulation space is generated at first. Then the number of points falling within the accumulation cells is counted. The cell with the largest number is then the most likely location of the vertex. The average displacement to this point is then the positional accuracy.

(2) Calculation of the parallelism of building footprint edges

Normally, the edges of buildings immediately adjacent are parallel to the nearby street. In most cases, building footprints are decomposed into long line segments, while the street nearby consists also of line segments. In these cases, the parallelism is very easy to be calculated. However, there are some cases in which one of the edges of the building footprint is curved and just like the street nearby as demonstrated in figure 1. In this case, the edge of building footprint is denoted as L_A, while the street nearby is represented as L_B.

In the first step, L_A is converted into a sequence of points $(P_{a,1}, \cdots, P_{a,n})$ with small and equal intervals. For point $P_{a,i}$, its foot point $P_{\perp a,i}$, which is perpendicular to L_B is calculated. If $P_{\perp a,i}$ is located on L_B, the distance from the point to L_B, is calculated as the Euclidean distance from $P_{a,i}$ and $P_{\perp a,i}$, namely, $d_i(P_{a,i}, P_{\perp a,i})$. It is assumed that there are k points $(P_{a,i}, \cdots, P_{a,m})$ $(k = m - i, i \geq 1 \; and \; m \leq n)$ on L_A that have foot points $(P_{\perp a,i}, \cdots, P_{\perp a,m})$ perpendicular to L_B. The distance between L_A and L_B is then the RMS (root mean square) of $(d_i, \cdots, d_m)$, d_{rms}. The RMSE (root mean square error) d_{rmse} is then used to evaluate the parallelism of L_A and L_B. In this work, the distance between two points is set at 0.3 m, which is sufficiently small compared with a line segment of a road in the physical world, which can ensure that the RMSE can be used to evaluate parallelism.

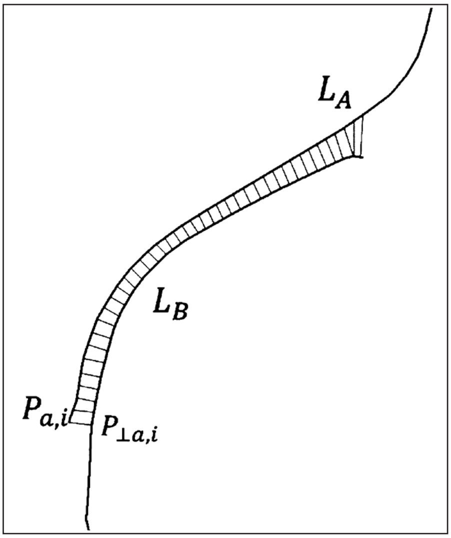

Fig. 1: The parallelism of edges of building footprints to road network (Source: own work)

(3) Calculation of the fragmentariness of patterns formed by building footprints

Buildings with similar shapes and sizes in the same area can form patterns if they are distributed regularly. This kind of knowledge can be used for intrinsic quality assessment of OSM building footprint. Firstly, the building footprints in a local area will be compared in terms of shapes and sizes, whereby a turning function is applied for the similarity. Secondly, the centroids of similar building footprints are used to estimate the distribution of the buildings by using regression or partition regression models. In the next step, it is checked whether there are buildings intersecting with the pattern but not considered as similar buildings in the pattern. This building should share the shape and size to the buildings in the pattern according to the abovementioned hypothesis. A pseudo building footprint with the same shape and size will be calculated at its position. And its orientation can be computed by the interpolation of the pattern regression. Finally, the positional accuracy and shape accuracy can be calculated by comparing the OSM building footprint at that location and the pseudo building footprint. It is also possible to obtain the displacement of the orientation.

3 The spatio-temporal data model for OSM history data

In order to conduct the intrinsic quality assessment of OSM data, the full history of OSM data has to be made available. The original format of the OSM history is very terse and capable to express rich information about the entities, their relationships, and their temporal changes. The geometry oriented models used by editing tools cannot provide an equivalent representation of the data. In this work, we developed a spatiotemporal data model not only for the purpose of assessing building footprints data, but also for the purpose of a comprehensive investigation of OSM features in terms of data quality and user behaviour.

We first discuss the pure temporal model of OSM history based on the concepts of entity and event. In the so-called Interval Based Model, an entity is homogeneous across the interval but an event is not, that is, an entity over and the entity over is the same entity, which is not the case for an event. For example, a building in 2010-2011 will be the same building as in 2010-2012, but the improvement of a building from 2010 to 2011 is only part of the improvement from 2010 to 2012. In the Instant Based Model, the entity and the event are not so distinguishable in theory, but usually we can recognize them in the context. For example, "nodes added" is clearly an event, while "road 111 at 2016-01-01 00:00" is an entity.

We propose a model including four types, which are entities over intervals, events over intervals, entities at instants, and events at instants. The objects over intervals have two essential properties "enter_time" and "exit_time", while objects over instants have one property named "timespot". Granted that there is a function $\mathrm{Eval}(\mathrm{e}, \mathrm{t})$ calculates the value of the entity e at time t. The semantic of Eval is trivial in the instant based view, but in the interval based view, Eval is inconclusive unless $\forall \mathrm{i}' \forall \mathrm{i}'' \sqsubset \mathrm{i}$, $\mathrm{Eval}(\mathrm{i}') = \mathrm{Eval}(\mathrm{i}'')$. For convenience, we further define the predicate *Version* in the interval view as:

$$Version(i) \equiv Conclusive\big(Eval(i)\big) \land \big(\forall i'(i' \prec i \lor i' \succ i) Eval(i) \neq Eval(i')\big)$$

The definition of *Version* clearly suggests that it depends on how *Eval* is understood, which brings the confusion about versions of ways and relations. If *Eval* is defined as the tags and node/member entities referred to, then the result are exactly the same versions marked in the original data. However, if ways and relations are regarded as geometry with tags much more versions result. Let it be supposed that a "version" of a highway spanning from January 1st to February 3rd, of which one of its inner nodes N changes its position on January 29th. The "version" is in fact not a version since $Eval(Jan.1, Jan.29) \neq Eval(Jan.29, Feb.3)$. Instead, $Version$ in such case can be defined as:

$$Version_x(i) \equiv i \neq \emptyset \wedge i \in \{Version_x \cap (\cap_{1\rightarrow n}^{l} i_l) | Version_{xl}(i_l)\}$$

The five fundamental types in OSM history are defined as shown in figure 2. We preserve most of the original structures with some differences to make the whole model more consistent and convenient.

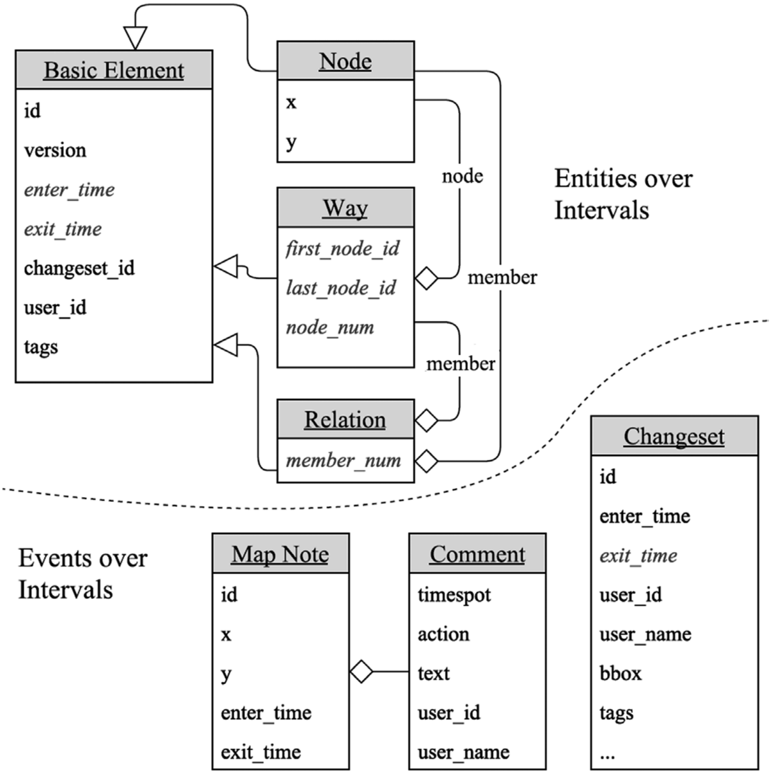

Fig. 2: Five major types corresponding to the original data (Source: own work)

4 Preliminary results in Baden-Württemberg

For the current stage, a PostgreSQL database has been established by using the spatiotemporal model introduced in section 3. We use the building footprints data in Baden-Württemberg (BW) for the test of the proposed intrinsic quality assessment.

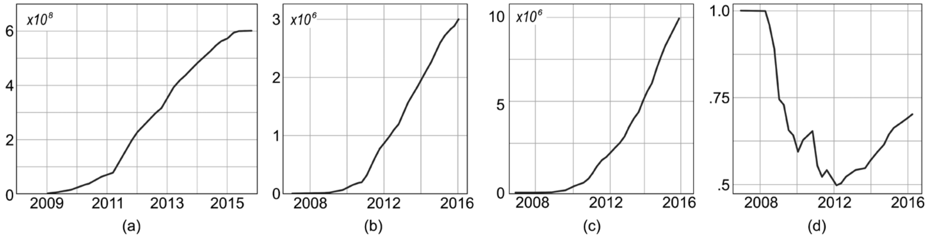

Fig. 3: The development of parameters of building footprints in BW (Source: own work)

As shown in figure 3a, the total area of building footprints in BW trend convergently, while the number (figure 3b) of building footprints still seems to rise with a steep slope. Figure 3c depicts that the number of buildings tagged with attributes is still increasing rapidly. And figure 3d represents the development of the percentage of buildings with rectangle shape. The valley in 2012 is followed by an increase of the percentage of rectangle-shaped building footprints. This reflects the reality that buildings were mapped in blocks as rectangles at the early stage of OSM development. Then OSM contributors started mapping buildings with complex shapes. From 2012 onwards, the polygons as group-buildings have been refined, so that individual buildings are digitized. This development means that both the semantic and the geometric accuracy of building data on OSM have been improved in the recent years.

Using the method presented in section 3, the parallelism of edges of building footprints to the nearby road line segments are calculated. Figure 4 shows that most of the buildings have an edge which is more than 90 % parallel to the immediately adjacent street/ road. This is almost consistent with reality. In other words, the positional accuracy of building footprints in BW seems to be very high.

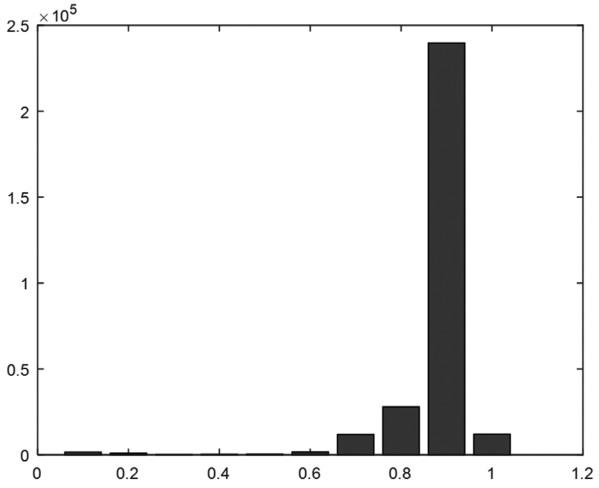

Fig. 4: The parallelisms of building footprints edges to their nearby roads (Source: own work)

5 Conclusion

In this paper, a framework is presented for the intrinsic quality assessment of OSM building footprints data. In total, seven quality indicators are suggested for the quality measurement without any reference data. The main idea is to observe the historic development of OSM data on the one hand. On the other hand, we are utilizing the knowledge of urban area to define indicators of intrinsic quality, because urban environment is man-made and almost everything there is based on certain rules. For instance, buildings are constructed with rectangles. Then the rectangularity of building footprints can be used as a kind of indicator for the quality of building footprint. Preliminary statistics of data on building footprints have been done by using the OSM history data in Baden-Württemberg, Germany. The experimental results show that data on building footprints in BW are mapped with relatively high accuracy and quality in terms of semantics, geometries and positions is still improving over time.

6 Literatur

Barron, C.; Neis, P.; Zipf, A. (2014): A comprehensive framework for intrinsic OpenStreetMap quality analysis. Transactions in GIS, 18(6), 877-895.

Dorn, H.; Törnros, T.; Zipf, A. (2015): Quality Evaluation of VGI Using Authoritative Data-A Comparison with Land Use Data in Southern Germany. ISPRS International Journal of Geo-Information, 4(3), 1 657-1 671.

Fan, H.; Zipf, A.; Fu, Q.; Neis, P. (2014): Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science, 28(4), 700-719.

Hecht, R.; Kunze, C.; Hahmann, S. (2013): Measuring completeness of building footprints in OpenStreetMap over space and time. ISPRS International Journal of Geo-Information, 2(4), 1 066-1 091.

Keßler, C.; de Groot, R. T. A. (2013): Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. In: Geographic information science at the heart of Europe, 21-37, Springer International Publishing.

Klonner, C.; Barron, C.; Neis, P.; Höfle, B. (2015): Updating digital elevation models via change detection and fusion of human and remote sensor data in urban environments. International Journal of Digital Earth, 8(2), 153-171.

Kunze, C.; Hecht, R.; Hahmann, S. (2013): Assessing the completeness of building footprints in OpenStreetMap: an example from Germany. In: 26th International Cartographic conference, 25-30).

OSM – OpenStreetMap (2016): http://wiki.openstreetmap.org/wiki/Stats, Accessed on 2016/08/15.

Törnros, T.; Dorn, H.; Hahmann, S.; Zipf, A. (2015): Uncertainties Of Completeness Measures In Openstreetmap – A Case Study For Buildings In A Medium-Sized German City. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 1, 353-357.