sciendo

Polish Cartographical Review
Vol. 52, 2020, no. 3, pp. 94–107
DOI: 10.2478/pcr-2020-0010
Received: 26.11.2020
Accepted: 4.12.2020

MELIH BASARANER
Yildiz Technical University
Department of Geomatic Engineering
Division of Cartography
Istanbul, Turkey
orcid.org/0000-0002-4619-7801
mbasaran@yildiz.edu.tr

Geometric and semantic quality assessments of building features in OpenStreetMap for some areas of Istanbul

Abstract. Nowadays volunteered geographic information (VGI) and collaborative mapping projects such as OpenStreetMap (OSM) have gained popularity as they not only offer free data but also allow crowdsourced contributions. Spatial data entry in this manner creates quality concerns for further use of the VGI data. In this regard, this article focuses on the assessments of geometric and semantic quality of the OSM building features (BFs) against a large-scale topographic (TOPO) data belonging to some areas of Istanbul. The comparison is carried out based on the one-to-one matched BFs according to a geometric matching ratio. In geometric terms, various parameters of position (i.e. X, Y), size (i.e. area, perimeter and granularity), shape (i.e. convexity, circularity, elongation, equivalent rectangular index, rectangularity and roughness index), and orientation (i.e. orientation angle) elements are computed and compared. In semantic terms, BF type coherences are evaluated. According to the findings of geometric quality, the average positional difference was less than three meters. In addition, the perimeter values tended to decrease while area and granularity values tended to increase in OSM data against TOPO data. Those showed that the level of the detail of the OSM BFs was lower than TOPO BFs in general. This was also confirmed by the decreasing tendency of shape complexity according to the parameters of shape element. Orientation angle differences was often low except for some special cases. It was found that the scale of the OSM dataset, even though not homogenous, approximately corresponded to the lower limit of medium scale maps (i.e. 1:10,000) or a slightly smaller scale. According to the findings of semantic quality, in case of the presence of specific type definition, the coherence was rather high between OSM and TOPO BFs while the most OSM BFs did not have a specific type attribute. This study showed that the matching process needed some improvements while the followed approach was largely successful in the evaluation of the matched buildings from geometric and semantic aspects.

Keywords: OpenStreetMap, building features, geometric data quality, semantic data quality, topographic data

1. Introduction

Along with the developments in information and communication technologies such as Web 2.0, new forms of spatial data collection, map production and map use have emerged. This new paradigm is called with various terms but all with same or similar meaning, e.g. volunteered geographic information (VGI), crowdsourced mapping, collaborative mapping, spatial citizenship and neocartography. Today, citizens have also become a kind of spatial data producers alongside governmental and private mapping organizations since they have access to online mapping tools, very high resolution remotely sensed data and location-aware mobile devices. In this context, OpenStreetMap (OSM) has gained considerable popularity and provided new opportunities for many stakeholders to access and utilize spatial data. Fast update cycle and free access policy have made OSM an alternative data source for those organizations. On the other hand, official spatial data collection and map production are carried out by adhering geographic and cartographic data specifications, therefore subject to quality control pro-

cedures. As a wide variety of individuals with different background (i.e. different levels of expertise or experience in geodesy, cartography or geography) contribute to the population of its content, OSM data exhibits heterogeneity from both geometric and semantic aspects. For this reason, OSM data should be investigated in terms of spatial data quality before being exploited (C.C. Fonte 2017, D. Sui et al. 2013, L. See et al. 2017, W. Cartwright 2012).

Spatial data quality includes the following quantitative elements (W. Kresse and K. Fadaie 2004): completeness, logical consistency, positional accuracy, temporal accuracy, semantic (thematic) accuracy. Among them, positional accuracy refers to accuracy of the position of features while semantic (thematic) accuracy refers to accuracy of quantitative attributes and the correctness of qualitative attributes, as well as the classification of features and their relationships. On the other hand, geometric accuracy is closely related to positional accuracy but usually requires more detailed description. If the feature has a non-point geometry (i.e. a line or a polygon), not only its position (i.e. centroid) but also the other descriptive elements of its geometric structure (e.g. size and shape) are handled. Quality measures and quality indicators can be utilized for VGI quality assessment. Adhering mainly to ISO principles and guidelines (currently ISO 19157:2013 Geographic Information – Data Quality - iso.org/standard/32575.html), quality measures refer to the elements that can be used to detect discrepancies between contributed spatial data and the ground truth mainly by comparing them with authoritative data. When authoritative data were no longer available for comparisons and established measurements were no longer sufficient to evaluate VGI quality, more internal approaches are used to evaluate VGI quality, known as quality indicators such as various participation biases, contributors' expertise level and background (H. Senaratne et al. 2017). Briefly, three principal approaches are usually followed for OSM data quality investigation (A. Basiri et al. 2016): (a) comparing data against authoritative spatial data, (b) user's and/or machine learnt rules and patterns for checking the entries, (c) gatekeeping and weighting users' entries.

This study follows the first approach and makes an assessment of geometric and semantic quality of OSM BFs against a large-scale topographic (TOPO) data. Although there are various methods for the quality evaluations from various aspects for different cities or regions of the world, the OSM building features (BFs) of Istanbul have not been investigated comprehensively from both geometric and semantic aspects. For this purpose, this article presents various parameters to evaluate the geometric quality as well as uses an original approach for semantic quality evaluation based on the matched BFs.

2. Related work

There are many studies dealing with OSM data quality (A. Basiri et al. 2016, H. Senaratne et al. 2017, A. Basiri et al. 2019). Pertaining to the BFs, several studies are available from different perspectives. In this scope, R. Hecht et al. (2013) propose object-oriented methods to examine the completeness of OSM building footprint data based on official data from national mapping and cadastral agencies. An analysis conducted in Germany in November 2011 showed 25% completeness in the states of North Rhine-Westphalia and 15% in Saxony and continued to increase in the following year. H. Fan et al. (2014) evaluate the quality of building footprints in OSM data against ATKIS for Munich with respect to completeness, semantic accuracy, positional accuracy and shape accuracy. Their findings demonstrate that the OSM building footprint data in Munich has a high completeness and semantic accuracy as well as positional accuracy of about four meters. J. Nowak Da Costa (2016) proposes a new index, called the matching feature area-based completeness, to evaluate the completeness of OSM BFs against an official dataset of the Polish Mapping Agency and also presents a simple method to update the official register. Y. Xu et al. (2017) propose an autoencoder neural network trained with the samples obtained by matching the building footprints in OSM and official data for Toronto, in which several measures, including data completeness, positional accuracy, shape accuracy, semantic accuracy and orientation consistency are employed as inputs. M.A. Brovelli and G.A. Zamboni (2018) perform a comparative evaluation of the spatial accuracy of BFs compiled from Topographic Database and OSM belonging to

the Lombardy region. The study utilizes an automated search algorithm of homologous pairs between two different maps. Their findings show that the quality of the OSM BFs is comparable to that of the regional technical authoritative map at the scale of 1:5000. I. Maidaneh Abdi et al. (2020) present a framework for extracting the relative spatial accuracy of OSM building data using machine learning methods. Following a multi-criteria data matching, the process attempts to establish a statistical relationship between the external quality of the OSM data (i.e. obtained in comparison to the reference spatial data) and the measures of the internal quality of the OSM data (i.e. the OSM features themselves) to estimate external quality when the reference data is not available. K.T. Jacobs and S.W. Mitchell (2020) explore OSM data quality in Ottawa-Gatineau, focusing on historical map features and contributor data to understand how users contribute to the database and their ability to do this correctly. Unsupervised machine learning analysis reveals the cluster of "OSM validators/experts" and then it is used for data quality evaluation.

3. Methodology

Geometric quality assessment is performed by comparing various geometric parameters computed for TOPO and OSM BFs while the semantic quality assessment is based on the comparison of type (function) attribute of both datasets. Both assessments are carried out with the matched BFs.

3.1. Geometric quality assessment

Geometric quality assessment is performed based on four elements: position, size, orientation and shape (fig. 1). For this purpose, the following parameters are employed for those elements.

Position element is defined by a coordinate pair and for polygonal features, their centroids are used for this purpose and computed with equations 1 and 2.

$$X = \frac{1}{6A}\sum_{i=1}^{n}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (1)$$

$$Y = \frac{1}{6A}\sum_{i=1}^{n}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (2)$$

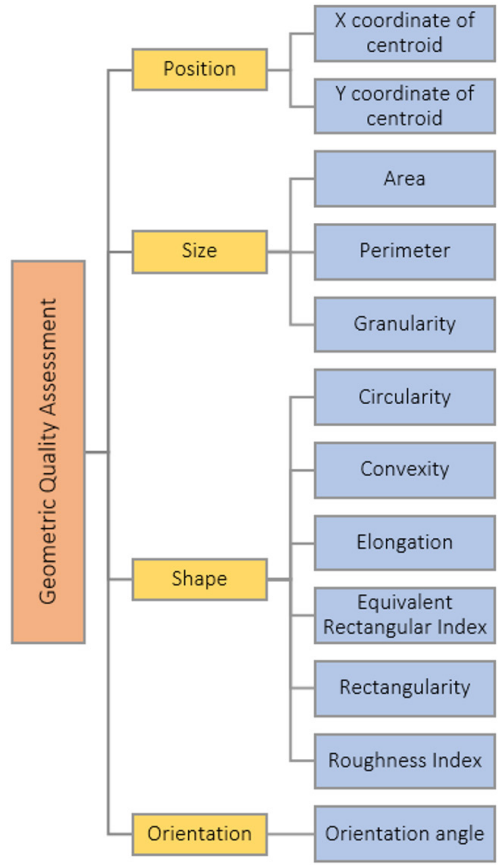

Fig. 1. The elements and the parameters of the geometric quality assessment

where X, Y are the centroid coordinates of a polygon, x_i, y_i are i-th vertex coordinates of a polygon, A is the area of a polygon and n is the number of the vertices of a polygon.

Size element is defined by area, perimeter and granularity (minimum edge length) for polygonal features. Area (A), perimeter (P) and granularity (G) are computed with equations 3, 4 and 5, respectively.

$$A = \frac{1}{2}\sum_{i=1}^{n}(y_i x_{i+1} - x_i y_{i+1}) \quad (3)$$

$$P = \sum_{i=1}^{n}\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (4)$$

$$G = \min_{i=1\rightarrow n}\left(\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}\right) \quad (5)$$

Shape element is usually defined by various shape indices because it is often difficult to cha-

racterize a shape with a single index (M. Basaraner and S. Cetinkaya 2017). In this context, circularity, convexity, equivalent rectangular index, rectangularity and roughness index are employed for polygonal features. Equivalent rectangular index yields values in the range (0,1.128] while the others in the range (0,1]. Shape complexity increases as the value approaches zero. When they are computed, the outer boundaries of complex polygons (i.e. polygons with holes) are regarded. This is important for most shape parameters to yield meaningful values in terms of shape complexity.

Circularity (*CI*) measures the area deviation between a polygon and its equal-perimeter circle and computed based on the area and perimeter of a polygon (equation 6). It reveals how similar a polygon is to a circle in shape.

$$CI = \frac{4\pi \mathrm{A}}{P^2} \tag{6}$$

Convexity (*CNV*) measures the areal deviation between a polygon and its convex hull (CH) (equation 7). It reveals the degree to which a polygon is curved inward or outward.

$$CNV = \frac{\mathrm{A}}{A_{CH}} \tag{7}$$

Elongation (*E*) is the ratio of the short edge's length (L_s) to the long edge's length (L_l) of the minimum area bounding rectangle (MABR) (equation 8). It is not directly relating to shape complexity, but can help distinguish between compact and non-compact shapes.

$$E = \frac{L_s}{L_l} \tag{8}$$

Meanwhile, the MABR of a polygon is different from its minimum bounding rectangle (MBR). The latter corresponds to the horizontal rectangle formed by the extreme coordinates of a polygon. On the other hand, the former is obtained by means of the convex hull (CH) of a polygon. The CH is rotated iteratively in a way that one of its edges becomes horizontal each time and then the MBR is generated. Consequently, the MABR corresponds to the minimum-area MBR rotated in the reverse direction by the original angle of the CH's respective edge. In practice, the main difference is that the MABR is not affected from the orientation. This is important for shape analysis. For example, two rectangles of the same size but different orientations yield the same-size MABRs while their MBRs have different sizes (M. Basaraner and S. Cetinkaya 2017, Z. Li 2007).

Equivalent rectangular index (*ERI*) measures perimeter deviation between a polygon and its equal-area rectangle, derived by scaling the MABR, and improves the drawback of *REC* being too sensitive to the long and thin protrusions (equation 9).

$$ERI = \sqrt{\frac{\mathrm{A}}{A_{MABR}}} \times \frac{P_{MABR}}{\mathrm{P}} \tag{9}$$

Rectangularity (*REC*) measures the areal deviation between a polygon and its MABR (equation 10). It reveals the degree of resemblance of a polygon to a rectangle.

$$REC = \frac{\mathrm{A}}{A_{MABR}} \tag{10}$$

Roughness index (RI) is a measure of compactness but more sensitive to the intrusions and protrusions and less sensitive to the eccentricity of a polygon than CI, being the most typical compactness measure. It is computed based on average length of radial distances between the centroid and densified boundary points (μ_{rd}), area and perimeter of a polygon (equation 11). The number of points interpolated on the boundary for BFs is chosen as 300, as recommended in M. Basaraner and S. Cetinkaya (2017).

$$RI = \frac{1}{\pi(1+4\pi)} \times \frac{\mu_{rd}^2}{A + P^2} \tag{11}$$

Orientation element is defined for polygons by orientation angle (θ), i.e. the angle between the horizontal axis and the long edge of the MABR of a polygon and computed with equation 12 (fig. 2).

$$\theta = arctan \frac{\Delta Y_{MABR}}{\Delta X_{MABR}} \quad (0° \leq \theta < 90° \text{ if shape of the MABR is square, } 0° \leq \theta < 180° \text{ otherwise}) \tag{12}$$

After all of the parameter values are computed, their comparisons are made. Since the same

formula is used for all the parameters, it is given in a general form in equation 13.

$$\Delta PRM = PRM_{TOPO} - PRM_{OSM} \quad (13)$$

where ΔPRM is the value difference of the respective parameter, PRM_{TOPO} is the value of respective parameter obtained from TOPO and PRM_{OSM} is the value of respective parameter obtained from OSM.

Concerning orientation angle difference ($\Delta\theta$), following conditions are taken into account to obtain final value difference (equation 14). In this way, it is ensured that $\Delta\theta$ ranges in [-90°, 90°].

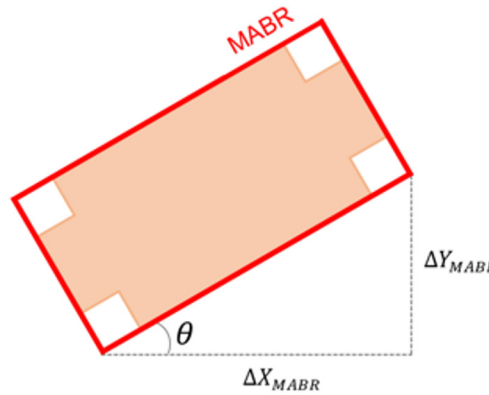

Fig. 2. MABR and orientation angle of a polygon

$$\Delta\theta = \begin{cases} Min(\theta_{TOPO} - \theta_{OSM},\ 180^o - (\theta_{TOPO} - \theta_{OSM})),\ if\ (\theta_{TOPO} - \theta_{OSM}) \geq 0^o \\ Max(\theta_{TOPO} - \theta_{OSM},\ -180^o - (\theta_{TOPO} - \theta_{OSM})),\ otherwise \end{cases} \quad (14)$$

Table 1. Corresponding classes between OSM and TOPO datasets in the semantic match table

OSM	TOPO		OSM	TOPO	
Type	Type 1	Type 2	Type	Type 1	Type 2
(non-specified)			monastery	religious	
apartments	dwelling		museum	official	
bank	commercial		office	commercial	
bar	commercial		place_of_worship	religious	
church	religious		police	official	
city_gate	historical	official	post_office	official	
clinic	official	commercial	public	official	
collapsed	ruin		public_bath	commercial	historical
college	school		public_building	official	
commercial	commercial		residential	dwelling	
courthouse	official		roof	porch	
dormitory	official	commercial	ruins	ruin	historical
ferry_terminal	official		school	school	
fountain	historical	official	shed	dwelling	
gym	sport facilities		shrine	religious	
hospital	official	commercial	stadium	commercial	
hotel	commercial		theme_park	commercial	official
house	dwelling		tomb	religious	
industrial	manufacturing	factory	townhall	official	
library	official		train_station	official	
marketplace	commercial	historical	university	school	

In addition, pertaining to the position element, the positional difference (ΔPST) is computed with X and Y coordinate differences (ΔX and ΔY) of corresponding TOPO and OSM BFs (equation 15).

$$\Delta PST = \sqrt{\Delta X^2 + \Delta Y^2} \qquad (15)$$

Accordingly, scatterplots of the parameters are obtained. Furthermore, the descriptive statistics (mean, median, standard deviation, minimum value and maximum value) about all of the parameters are calculated based on their absolute values.

3.2. Semantic quality assessment

From semantic aspect, first corresponding BF types (classes) are identified between OSM and TOPO data and a semantic match table is created for identical types of both datasets (tab. 1). This table also includes a row (record) for the non-specified OSM feature type. The classification (taxonomy) used in OSM is sometimes more specific than that used in TOPO. In addition, some of the types in OSM potentially correspond to more than one type in TOPO. Therefore, an additional TOPO type column is created to reduce the number of misclassifications that may result from this ambiguity. It is then analysed whether the BFs are assigned to a correct feature class through the "Type" attributes. Accordingly, the percentage of the correct type assignment (PCTA) is obtained.

3.3. Matching of the BFs

In order to compare the BFs from both datasets, they need to be matched. This is achieved with geometric matching. Depending on the factors such as level of detail (scale/resolu-

Fig. 3. Study area

tion), up-to-dateness, completeness and contributor interpretation, various cardinal relationships can emerge between the BFs of TOPO and OSM data. In this context, one-to-zero (1:0), one-to-one (1:1), one-to-many (1:n), many-to--many (n:m), zero-to-one (0:1) and many-to--one (n:1) relationships between the BFs can be confronted.

In this study, the geometric and semantic assessments are performed with the BFs that have 1-1 matching (m_{1-1}). They are considered to be of 1-1 matching if they have more than 70% overlap geometrically (equations 16 and 17). According to the equation 16, it is ensured that each pair of building polygons (i.e. PLG_{TOPO} and PLG_{OSM}) satisfies more than 70% overlap. In other words, if one of the building is larger than the other, the smaller one may meet the condition while it may not be met by the larger one. Therefore, the area of the larger one (i.e. maximum area) is used as the denominator to prevent this situation. The matching ratio (rt_m) is determined experimentally.

```
Select Columns:   *
from Tables:      TOPO_matched, OSM_matched
where Condition:  OSM_matched.obj intersects TOPO_matched.obj and (
                  CartesianArea ( overlap ( OSM_matched.obj ,
                  TOPO_matched.obj ) , "sq m" ) / Maximum (
                  CartesianArea ( OSM_matched.obj , "sq m" ) ,
                  CartesianArea ( TOPO_matched.obj , "sq m" ) ) ) > 0.7
                  and ( CartesianArea ( OSM_matched.obj , "sq m" ) > 25
                  and CartesianArea ( TOPO_matched.obj , "sq m" ) > 25 )
```

Fig. 4. SQL query for the data matching

$$rt_m = \frac{Area\,(PLG_{TOPO} \cap PLG_{OSM})}{Max\,(Area(PLG_{TOPO}, PLG_{OSM})} \quad (16)$$

Fig. 5. One-to-one matched BFs

$$m_{1-1} = \left\{ \begin{matrix} True, \, if \, rt_m > 70\% \\ False, \, otherwise \end{matrix} \right\} \quad (17)$$

4. Experimental study

In this section, the study area, data, software and some GIS data processing and analysis details are explained.

The study area, where old settlements are concentrated, mainly covers Fatih district and some parts of Beyoglu, Zeytinburnu and Eyupsultan districts in the European side of Istanbul (fig. 3). The BFs compiled from 1:1,000 scale topographic map (TOPO) data and OSM data were used belonging to the study area. TOPO data includes the most detailed and accurate spatial data used in topographic maps in Turkey. In Istanbul, the Metropolitan Municipality is responsible from their production and maintenance. OSM data was downloaded from Geofabrik web site (download.geofabrik.de/europe/turkey.html). The number of BFs in the TOPO data set was 109 351. The OSM data originally covered largest area but the number of OSM BFs was less than that of TOPO when the study area was regarded. Meanwhile, the OSM data was more up-to-date than TOPO data because the former was downloaded in 2020 and possibly produced in the last few years while the latter was the version produced around six years ago. However, this is a central and largely historical area where one can expect little change.

MapInfo Pro GIS software was used in the experimental study. The parameters were automatically computed with an add-on written in MapBasic. At the beginning of the experimental study, OSM data was converted from Geographic (WGS84) to the Gauss-Krüger Central Meridian 30° (ITRF96) projected coordinate system used by TOPO data to be able to integrate them and make computations. After the parameters were computed, the data matching was performed with a SQL query (fig. 4) and 1–1 matched BFs were obtained (fig. 5). During the matching, the BFs whose area was less than 25 sq m were also excluded from the experiment because they may high possibly represent some insignificant structures without specific type definition. Meanwhile, the query was executed twice by changing the order of the table because the new table was formed by the geometries of the table specified first in the from clause of SQL while the attributes came from both tables. Consequently, the number

Table 2. Semantic types and numbers of BFs in the OSM match table

OSM					
Type	**Number**	**Type**	**Number**	**Type**	**Number**
(non-specified)	4246	gym	1	public_bath	1
apartments	43	hospital	7	public_building	16
bank	1	hotel	16	residential	7
bar	1	house	2	roof	10
church	18	industrial	27	ruins	1
city_gate	1	library	4	school	18
clinic	1	marketplace	2	shed	1
collapsed	1	monastery	1	shrine	3
college	1	museum	7	stadium	2
commercial	8	office	1	theme_park	1
courthouse	1	place_of_worship	156	tomb	1
dormitory	1	police	2	townhall	1
ferry_terminal	2	post_office	1	train_station	2
fountain	1	public	6	university	18

Table 3. Semantic types and numbers of BFs in the TOPO match table

TOPO					
Type	**Number**	**Type**	**Number**	**Type**	**Number**
commercial	185	official	265	ruin	5
dwelling	3514	parking garage	2	school	152
factory	62	porch	58	sports facilities	11
gas station	9	power plant	1	transformer	3
historical	9	pump building	2	under construction	19
manufacturing	111	religious facility	233		

of BFs per the dataset was 4641 in the OSM match and the TOPO match tables (tables 2 and 3). Then, the OSM match table was joined with the semantic match table by OSM BF type. Clearly speaking, the corresponding BF types were assigned to OSM data from TOPO data according to the table 1. In this way, the OSM's corresponding TOPO BF types were compared to the original TOPO BF type coming from the matching through SQL. Thus, it became possible to calculate the PCTA for OSM BFs.

5. Results and discussion

Concerning geometric quality assessment, several descriptive statistics of the parameters categorized by the elements were given in table 4. Besides, parameter-specific scatterplots were shown in figs. 6, 7, 8 and 9.

As regards the position element, the differences in X direction were usually greater than in Y direction. It is difficult to interpret this but it seems to be a kind of systematic error that

Table 4. Difference statistics of the parameter values

Element	Parameter	Average	Median	Std. dev.	Min.	Max.
POSITION	ΔX (m)	2.06	1.98	1.31	0.00	9.78
	ΔY (m)	0.90	0.75	0.76	0.00	14.17
	ΔPST (m)	2.41	2.30	1.25	0.02	14.23
SIZE	ΔA (m^2)	28.95	13.86	75.71	0.00	3094.51
	ΔP (m)	3.86	2.41	6.36	0.00	251.88
	ΔG (m)	5.85	4.10	6.138	0.000	82.449
SHAPE	ΔCI	0.038	0.022	0.046	0.000	0.485
	ΔCNV	0.020	0.010	0.028	0.000	0.285
	ΔE	0.064	0.045	0.064	0.000	0.531
	ΔERI	0.024	0.011	0.035	0.000	0.370
	ΔREC	0.051	0.036	0.052	0.000	0.421
	ΔRI	0.035	0.016	0.050	0.000	0.514
ORIENTATION	Δθ (°)	8.7491	1.5086	22.9049	0.0002	89.9972

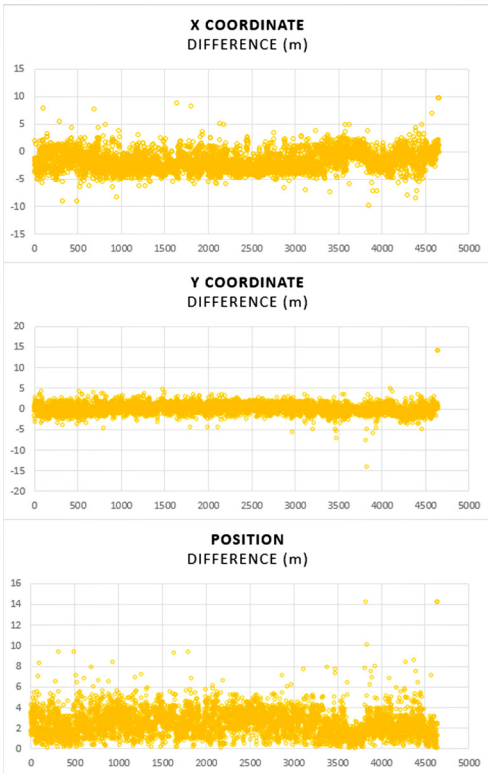

Fig. 6. Scatterplots of the position element

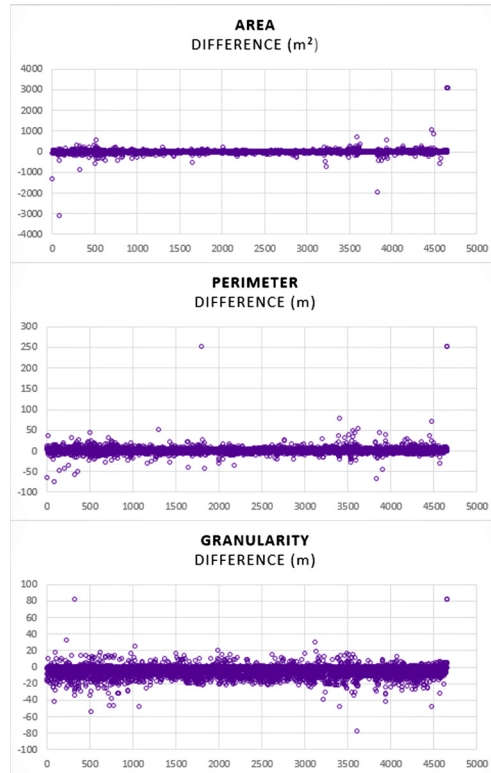

Fig. 7. Scatterplots of the size element

Table 5. The percentage of correct type assignment (PCTA) per OSM BF type

Type	PCTA	Type	PCTA	Type	PCTA
apartments	97.7	hospital	71.4	public_building	18.8
bank	0	hotel	18.8	residential	57.1
bar	0	house	50	roof	60
church	55.6	industrial	59.3	ruins	100
city_gate	100	library	50	school	83.3
clinic	0	marketplace	100	shed	0
collapsed	0	monastery	100	shrine	100
college	100	museum	28.6	stadium	50
commercial	25	office	100	theme_park	0
courthouse	100	place_of_worship	92.3	tomb	0
dormitory	100	police	100	townhall	100
ferry_terminal	50	post_office	100	train_station	100
fountain	100	public	50	university	61.1
gym	0	public_bath	100		

Fig. 8. Scatterplots of the shape element

might be induced from the base data. Average positional difference was 2.41 m (median 2.30 m) between OSM and TOPO data. Accordingly, in terms of positional and graphic accuracies required for map production (P. Kohlstock 2014), it can be stated that the scale of the OSM data approximately corresponds to 1:10,000 or a slightly smaller scale. However, the feature-specific quality differences should not be disregarded.

As regards the size element, for OSM data, the perimeters tended to decrease while the areas and granularity tended to increase when compared to TOPO data. These differences are likely due to the fact that the BFs are often interpreted at a lower level of detail and additional parts of the BFs such as porches are combined with the main BFs for the former. Those indicate that OSM data, as expected, is of a lower level of detail in general.

As regards the shape element, all shape parameters except for elongation yielded higher

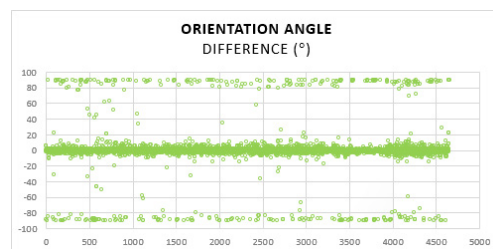

Fig. 9. Scatterplot of the orientation element

values for OSM data in general. This means that the shapes tend to be less complex in OSM data than in TOPO data. This is another evidence of the previous finding about the level of detail. On the other hand, the elongation is not directly related to shape complexity but its tendency of increase is likely to be related to the fact valid for the area. In this case, the length

difference of the main axes (non-compactness) of a BF may tend to increase.

As regards the orientation element, orientation angle differences were usually small but there were also some high differences. In practice, the angle between two rectangles can be maximum 90°; however, it is expected to be quite small for the BFs representing the same building from two different datasets. It was found that this was due to the two BFs of rectangular (square-like) shape but slightly elongated in approximately perpendicular directions. Therefore, the very high values should be accepted as outliers.

Some of the BF pairs that generated high geometric differences were shown in fig. 10.

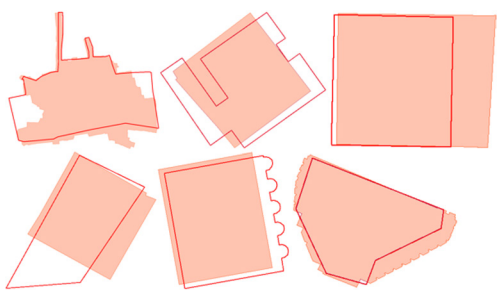

Fig. 10. Some examples of the TOPO (salmon) and OSM (red) BF pairs with high differences from geometric aspect

Concerning semantic quality assessment, 91.5% (4246 out of 4641) of the OSM BFs were not assigned a specific type. Table 5 shows the PCTA for OSM BFs. It was calculated 73.7% (291 out of 395) for the whole OSM data. Meanwhile, 80% of the non-specified OSM BF types were belong to “Dwelling” feature type in TOPO data. In general, OSM dataset has more specific types than TOPO dataset. The semantic mismatches may be caused by the presence of multiple functions for some BFs and the differences between the data up-to-dateness apart from possible misinterpretations of the contributors.

The results and previous experience show that the geometric matching process is challenging. If a lower matching ratio is chosen, the number of incorrect matches can increase and the values of the geometric quality parameters can differ more (O.E. Erden and M. Basaraner 2019). On the other hand, in the opposite case, the number of the matches decrease while the average parameter values tend to become closer. In this context, some thresholds may be developed to eliminate the mismatches if the matching ratio is set smaller. Another critical problem is the cardinal relationships between the BFs. Since the TOPO dataset is of a higher level of detail in general, there are numerous n:m (n > m) and n:1 relationships between TOPO and OSM BFs among others. In this case, the possibility of the matching will dramatically decrease because the matching ratio cannot be met.

This study presented quite a comprehensive comparison of OSM and TOPO BFs in terms of geometric and semantic quality in the case of Istanbul and employed various parameters for this purpose. In addition, the original interpretations of the findings were made. Regarding the limitations, the TOPO data was not so new and some differences, particularly from semantic aspect, might have arisen in relation to this factor. In addition, the matching process needs some improvements. In this respect, as mentioned above, alternative approaches can be adopted. From semantic aspect, BFs with multiple functions can be defined with different types in the respective data sets. Hence, point of interest (POI) data may be used for identifying multiple types. In addition, city information system data can be used instead of TOPO data since they are more detailed semantically.

6. Conclusions

This article presented geometric and semantic quality assessments of BFs in OSM data against TOPO data for some areas of Istanbul. In terms of geometric quality, position, size, shape and orientation elements, involving various parameters, were investigated for both data. X and Y differences were computed for the position element. Area, perimeter and granularity were computed for the size element. Convexity, circularity, elongation, equivalent rectangular index, rectangularity and roughness index were computed for the shape element. Orientation angle was computed for the orientation element. In terms of semantic quality, a semantic match table was prepared that included the identical BF types of both data. This

was followed by the geometric matching phase using a matching ratio based on areal overlap to obtain one-to-one matching BFs of TOPO and OSM data. The assessments were performed on these matched BFs. From geometric aspect, the differences of the parameter values were yielded for comparison, including positional shift. It was demonstrated that OSM BFs had usually a lower level of detail compared to TOPO BFs in the study area. Besides, it was found that the scale of the OSM data approximately corresponded to 1:10,000 or a slightly smaller scale even though it was not completely homogenous throughout the dataset. From semantic aspect, the most of the BFs were not assigned a specific type attribute in OSM data. Those were largely belong to the dwellings according to TOPO data. On the other hand, the percentage of correct type assignment was rather high in general among the specifically defined OSM BFs. As regards the matching, the ratio was set a bit high to eliminate the outliers. This led to a lower number of but better BF matches and thus slightly better statistics compared to the previous experience. Future work may focus on the improvement of the matching process. Another problem is how to deal with one-to-many or many-to-many relationships from the quality assessment perspective. Those need further investigation.

References

Basaraner M., Cetinkaya S., 2017, *Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS*. "Intern. Journal of Geogr. Inform. Science" Vol. 31, No. 10, pp. 1952−1977.

Basiri A., Haklay M., Foody G., Mooney P., 2019, *Crowdsourced geospatial data quality: challenges and future directions*. "Intern. Journal of Geogr. Inform. Science" Vol. 33, No. 8, pp. 1588−1593.

Basiri A., Jackson M. Amirian P. Pourabdollah A., Sester M., Winstanley A., Moore T., Zhang L., 2016, *Quality assessment of OpenStreetMap data using trajectory mining*. "Geo-spatial Inform. Science" Vol. 19, No. 1, pp. 56−68.

Brovelli M.A., Zamboni G.A, 2018, *New method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints*. "ISPRS Intern. Journal of Geoinformation" Vol. 7, No. 8, p. 289.

Cartwright W., 2012, *Neocartography: opportunities, issues and prospects*. "South African Journal of Geomatics" Vol. 1, No.1, pp. 14−31.

Erden O.E., Basaraner M., 2019, *Geometric quality analysis of building footprints from OpenStreetMap data in comparison to topographic map data*. In: International Symposium on Advanced Engineering Technologies (ISADET), 2−4 May 2019, Kahramanmaras, Turkey.

Fan H., Zipf A., Fu Q., Neis P., 2014, *Quality assessment for building footprints data on OpenStreetMap*. "Intern. Journal of Geogr. Inform. Science" Vol. 28, No. 4, pp. 700−719.

Fonte C.C., Antoniou V., Bastin L., Estima J., Arsanjani J.J., Bayas J.-C.L., See L., Vatseva R., 2017, *Assessing VGI data quality*. In: G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C.C. Fonte, V. Antoniou (eds.), *Mapping and the citizen sensor*, pp. 137−163. London: Ubiquity Press.

Hecht R., Kunze C., Hahmann S., 2013, *Measuring completeness of building footprints in OpenStreetMap over space and time*. "ISPRS Intern. Journal of Geoinformation" Vol. 2, No. 4, pp. 1066−1091.

Jacobs K.T., Mitchell S.W., 2020, *OpenStreetMap quality assessment using unsupervised machine learning methods*. "Trans GIS" Vol. 24, No. 5, pp. 1280−1298.

Kohlstock P., 2014, *Kartographie − eine Einführung*, 3. Auflage. Paderborn: Schöningh (UTB).

Kresse W., Fadaie K., 2004, *ISO standards for geographic information*. Berlin: Springer.

Li Z., 2007, *Algorithmic foundation of multi-scale spatial representation*. Boca Raton: CRC Press.

Maidaneh Abdi I., Le Guilcher A., Olteanu-Raimond A.-M., 2020, *A regression model of spatial accuracy prediction for OpenStreetMap buildings*. "ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Inform. Sciences" Vol. V-4-2020, pp. 39−47.

Nowak Da Costa J., 2016, *Novel tool for examination of data completeness based on a comparative study of VGI data and official building datasets*. "Geodetski Vestnik" Vol. 60, No. 3, pp. 495−508.

See L., Estima J., Pődör A., Arsanjani J.J., Bayas J.-C.L.,Vatseva R., 2017, *Sources of VGI for mapping*. In: G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond., C.C. Fonte, V. Antoniou (eds.), *Mapping and the citizen sensor*, pp. 13−35. London: Ubiquity Press.

Senaratne H., Mobasheri A., Ali A.L., Capineri C., Haklay M., 2017, *A review of volunteered geographic information quality assessment methods*. "Intern. Journal of Geogr. Inform. Science" Vol. 31, No. 1, pp. 139−167.

Sui D., Goodchild M., Elwood S., 2013, *Volunteered geographic information, the exaflood, and the growing digital divide*. In: D. Sui, M. Goodchild, S.

Elwood (eds.), *Crowdsourcing geographic knowledge − volunteered geographic information (VGI) in theory and practice*, pp. 1−12. New York: Springer.

Xu Y., Chen Z., Xie Z., Wu L., 2017, *Quality assessment of building footprint data using a deep autoencoder network*. "Intern. Journal of Geogr. Inform. Science" Vol. 31, No. 10, pp. 1929−1951.

download.geofabrik.de/europe/turkey.html – Geofabrik downloads – Europe – Turkey (access 05.01.2020)

iso.org/standard/32575.html – ISO 19157:2013 Geographic information – Data quality (access 03.12.2020)