Using geometric and semantic attributes for semi-automated tag identification in OpenStreetMap data

Müslüm Hacar[*1]

[1]Department of Geomatic Engineering, Yildiz Technical University, Istanbul, Turkey

March 24, 2021

Summary

OpenStreetMap is one of the successful volunteered geographical information projects. Participants contribute to this crowdsourced project by adding geometric and semantic data. However, both missing geometric and semantic data still cause completeness problems. In this paper, a semi-automated approach is suggested to identify the values of *leisure* tag of polygon features. The approach uses geometric (rectangularity, density, area, and distances to bus stop and shop) and semantic (amenity) data and estimates the key values using random forest classifier. In short, the results show that tag identification was conducted in three districts of Ankara with f-scores 78%, 86%, and 87%.

KEYWORDS: VGI, OpenStreetMap, tag identification, crowdsourcing, completeness

1. Introduction

Research on volunteered geographical information (VGI) is a challenging work assessing data quality. There are two type of studies evaluating the quality of VGI data: (1) using reference data, and (2) without using it. Using reference data has an enormous advantage determining the geometric or semantic accuracy by comparing two sources (Haklay, 2010; Girres and Touya, 2010; Mondzech and Sester, 2011; Zhang and Malczewski, 2018). However, it cannot give any information on volunteers' preferences for attributes, drawing trends and the diversity of attributes (Mooney and Corcoran, 2012a). Without reference data, intrinsic quality assessment can be conducted by geometric and semantic measures. Mooney, Corcoran, and Winstanley (2010) assessed OpenStreetMap (OSM) data quality by examining the polygon formations of hydrography and forested areas. They remarked that using satellite images as overlapping map it was easier to draw hydrographic features and boundaries than to draw the boundaries of the forest area. Mooney and Corcoran (2012a, 2012b) assessed updated data more than 15 times by country. They remarked that most of OSM data changed less than three and the number of contributors had no strong relationships with the number of tags. Jilani, Corcoran, and Bertolotto (2014) developed a machine learning prediction model to estimate the *highway* = * classes in OSM roads. As a result, more than 50% of *residential, pedestrian, primary, motorway, primarylink* and *motorwaylink* were predicted correctly, and less than 40% of the *cycleway, bridleway, path, secondary* and *secondarylink* were correct.

Tag names that are preferred and accepted by users are listed with their definitions in OSM: Map Features (2021). It is expected that OSM contributors respect the list during the tag selection for a geographic object to ensure the compatibility with other users' choices. Davidovic, Mooney, and Stoimenov (2016) evaluated how OSM volunteers respected OSM Wiki web page in 30 different urban areas. The same types of objects were generated with different tags in different areas. Basiri, Amirian, and Mooney (2016) proposed an approach to generating new objects or editing existing data from raw trajectory data by using data mining technics. Hacar (2020) assessed the tag adding trends of the OSM contributors. He remarked that while OSM is a good option in terms of tag diversity, the data completeness still needs to be increased.

[*] mhacar@yildiz.edu.tr

Each tag may have its own significance during spatial analysis. However, the most important issue during an analysis is the incompleteness of a target tag. Therefore, the attributes of OSM need to be enriched. One of the important tags is *leisure* tag since it applies to all kind of facilities where people can spend their spare time (Funke and Storandt, 2017). This study suggests a semi-automated approach to identify *leisure* tag values of polygon features. The proposed approach uses geometric measures and semantic data to estimate the key classes using random forest classifier. The following section presents OSM data, the study area, and the proposed approach with brief description of the measures. In the third section, an experiment was conducted with already known leisure values to test the predicted values. The results are evaluated by attribute (feature) importance. Then, last section concludes with the discussions about tag identification and future perspective.

2. Material and Method

2.1. Study area and the data

This study was conducted with OSM data in three districts of Ankara, the capital city of Turkey. All of the districts are located in urban area (**Figure 1** and **Table 1**). Both polygon and point data of OSM was used. While point data contains the point of interest (POI): *highway=bus_stop*, *shop=**, and *amenity=**, polygon data has all features represented by area such as building, pool, sea, boundary, and so on. Target tag in this study is *leisure=**. The most common values in *leisure* are *pitch, swimming_pool, park, garden*, and so on (Taginfo, 2021). To test the sufficiency of the proposed approach, the leisure tagged polygons were selected as training and test data.

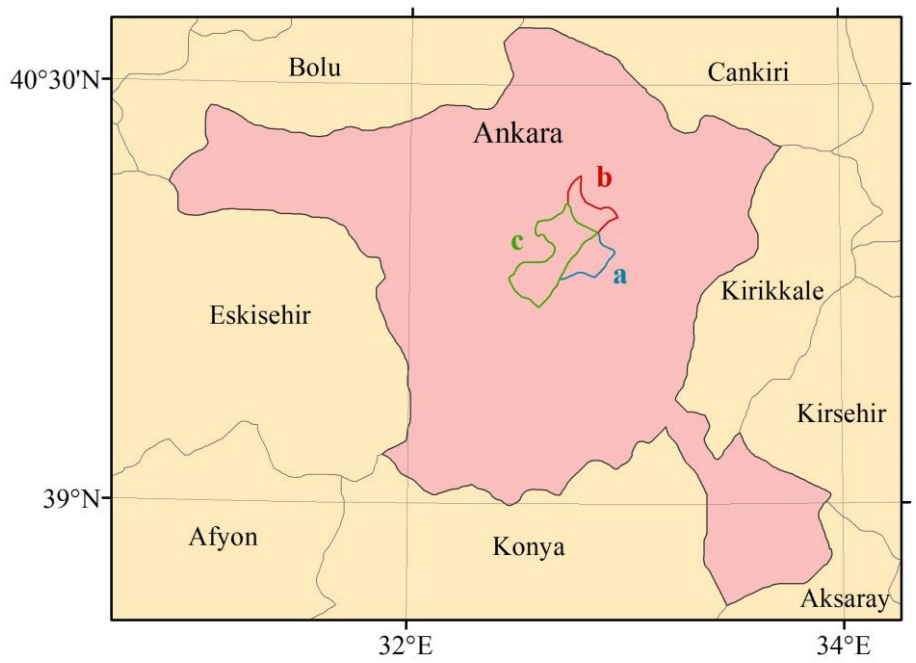

Figure 1 Ankara city: (a) Cankaya, (b) Kecioren, and (c) Yenimahalle districts.

Table 1 Statistics of the study areas.

District	Number of polygons	Number of polygons with *leisure=** tags
Cankaya	43654	1200
Kecioren	19824	488
Yenimahalle	40566	1308

2.2. The proposed approach

The approach consists of two stages: (1) identifying the possible polygon objects manually and (2) identifying the tag value automatically. First stage is managed by OSM contributors determining only suitable polygon objects that may have *leisure* tag. A contributor needs only OSM geometric data. In this stage, he/she decides which polygon geometry can be a leisure centre. After this manual identification, second stage is initiated with transfering semantic information and the computation of geometric measures. The closest POIs with *amenity* classes are assigned to the polygon features as semantic classes. The approach uses five geometric measures as rectangularity, density, area, and distances to bus stop and shop. **Figure 2** shows the workflow of the generation of the attributes. The simple rectangularity estimation R is calculated as the ratio of the area of a polygon against the area of its minimum bounding rectangle (Rosin, 1999). While distance to bus stop is computed between the centroid of the polygon feature and its closest POI with *highway=bus_stop* tag, distance to shop is computed between the centroid of the polygon feature and its closest POI with *shop=** tag. Density value is computed using the building centroids as density points in ArcMap 10 (Silverman, 1986; ESRI, 2020).

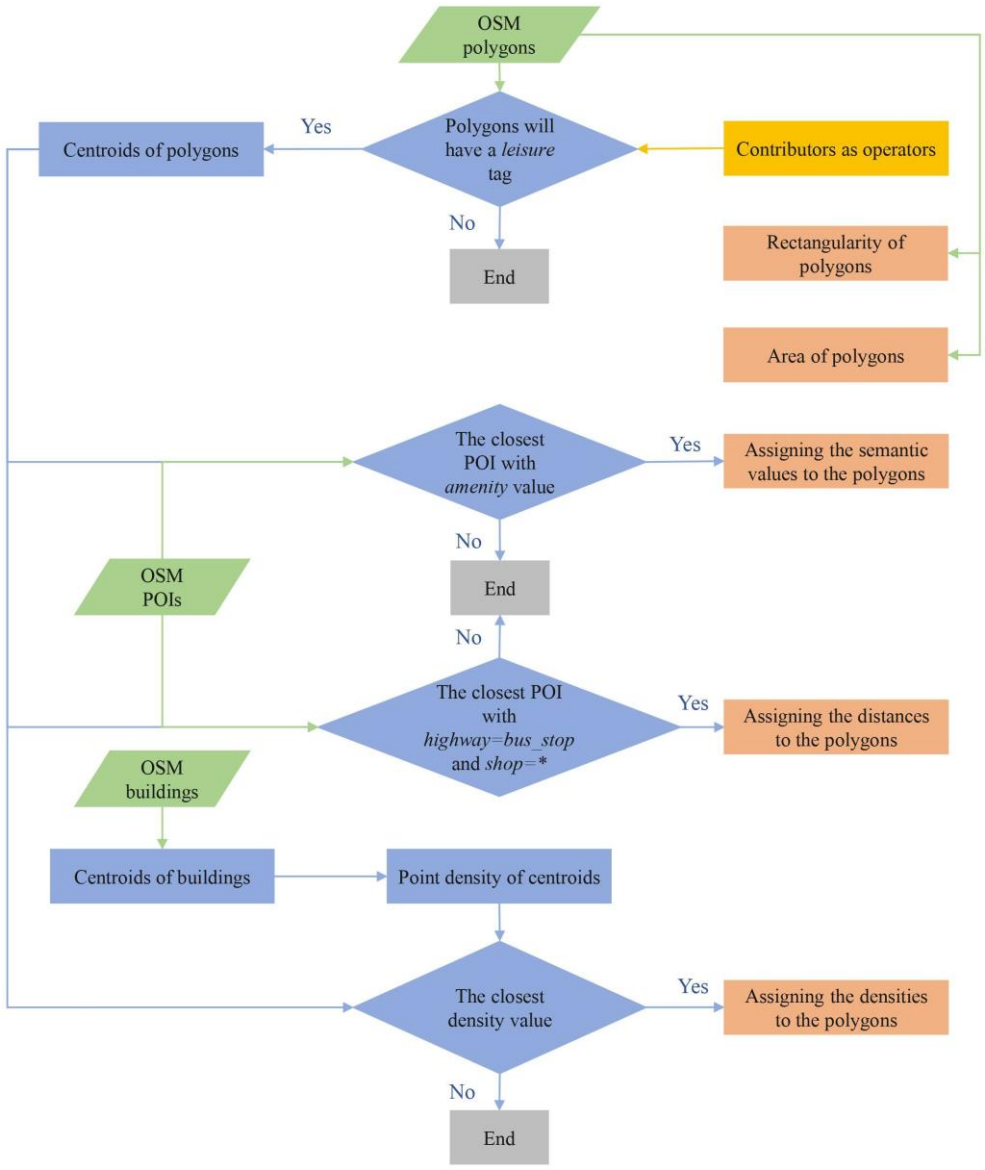

Figure 2 The workflow of the attribute generation.

Second stage ends with predicting the value of *leisure* tag by using the geometric and semantic attributes as independent variables (rectangularity, density, area, the distances to bus stop and shop are the continuous variables, and *amenity* tag value is nominal variable) in random forest classifier.

3. Evaluation of the experimental results

An experiment was conducted with the data in section 2.1 to test the second part (automated) of the proposed approach. The aim of the experiment is to learn about the measure performance and how accurate the tag values are predicted. Five geometric and one semantic attributes are used as features for tag value identification in random forest. The classification was carried out with Sci-Kit Learn python package (Pedregosa et al., 2011). After the prediction model trained with randomly selected data (80% of datasets), it estimated the *leisure* tag values in the rest of the datasets. **Table 2** shows the brief results of the prediction model. Also, the importance values of the features were computed using Gini impurity to show the significance order of the attributes in the model (**Figure 3**) (Breiman et al., 1984). Results show that the model estimated the *leisure* tags semi-automatically with f-score at least 78%. The most significant attribute for the estimation is clearly seen as rectangularity and area of source polygon. This means that the polygons with *leisure* tags have recognizable shape patterns.

Table 2 Statistical results of the study areas.

District	Average precision	Average recall	Average f-score
Cankaya	77	79	78
Kecioren	87	88	86
Yenimahalle	87	89	87

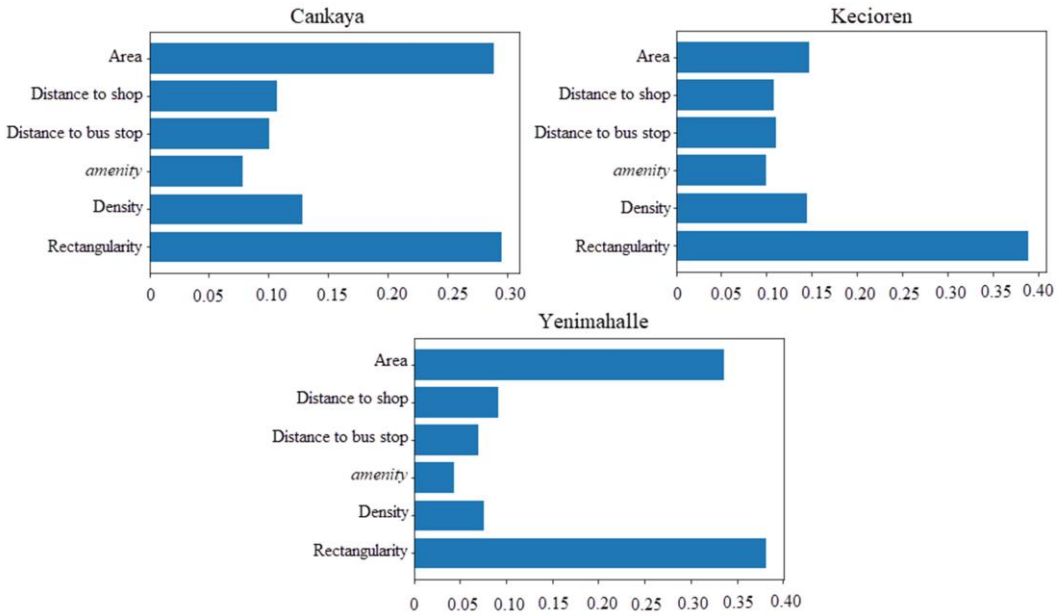

Figure 3 Feature importance for tag identification in random forest classifier.

Table 3 gives the statistics of the highest numbers of leisure tag values both predicted correctly and incorrectly. While most of the *pitch* and *park* values were predicted correctly, almost half of the *swimming_pool* values were predicted incorrectly in each district. Specifically, *playground* values were also incorrectly predicted in Cankaya. The polygons with *swimming_pool* value have high rectangularity rate. This means that the shapes of most of the swimming pools are almost rectangle. However, the other measures of these polygons do not have a regular distribution as the measure of rectangularity has. For instance, there is no explicit association between mean area sizes of correctly and incorrectly predicted polygons in accordance with the districts (**Table 3**). This kind of irregular

distribution in the attributes may result in incorrect predictions.

Table 3 Statistics of correctly and incorrectly predicted polygons.

District	The highest number of leisure tag value predicted		Mean area of polygons predicted (m^2)	
	Correctly	Incorrectly	Correctly	Incorrectly
Cankaya	*pitch*	*playground, swimming_pool*	6313.7	11615.9
Kecioren	*park*	*swimming_pool*	2273.8	1797.7
Yenimahalle	*pitch, park*	*swimming_pool*	4404.6	2072.9

4. Conclusion

This paper shows how significant the semantic and geometric attributes are. The study was conducted in three districts to prove the efficiency of the prediction model. All study area had similar results except insignificant differences. The results point out that each measure has different level of importance. While *amenity* tag had limited contribution in the model, shapes of the polygons are basically related to the correctness of the estimation. Also, the approach is unable to predict almost half of the *leisure* polygons with *swimming_pool* value because of the irregular distribution in the attributes. Therefore, for more accurate results, the approach requires new measures demonstrating the shape patterns of polygons like turning function (Arkin et al., 1991). Future research will focus on searching shape measures and other semantic tags in OSM, and making the first stage (manual) of the approach automated.

References

Arkin E M, Chew L P, Huttenlocher D P, Kedem K, and Mitchell, J S (1991). *An efficiently computable metric for comparing polygonal shapes*. Cornell University, Ithaca, New York.

Basiri A, Amirian P, and Mooney P (2016). Using crowdsourced trajectories for automated OSM data entry approach. *Sensors*, 16(9), 1510.

Breiman L, Friedman J, Stone C J, and Olshen R A (1984). *Classification and regression trees*. Chapman and Hall, CRC press, London.

Davidovic N, Mooney P, and Stoimenov L (2016). *An analysis of tagging practices and patterns in urban areas in OpenStreetMap.* AGILE 2016 Conference, Helsinki, Finland, 14–17 Jun 2016.

Funke S, and Storandt S (2017, May). *Automatic Tag Enrichment for Points-of-Interest in Open Street Map*. In International Symposium on Web and Wireless Geographical Information Systems, Shanghai, China, 8-9 May 2017. (Lecture Notes in Computer Science, vol 10181. Springer, Cham. https://doi.org/10.1007/978-3-319-55998-8_1)

Girres J F and Touya G (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, 14(4), 435-459.

Hacar M (2020). Analyzing the Contribution Trends of Volunteers by Comparing Tag Metadata of OpenStreetMap Residential Roads [In Turkish: OpenStreetMap Yerleşim-içi Yollarına Ait Etiket Bilgilerinin Karşılaştırılmasıyla Gönüllülerin Katkı Sağlama Eğilimlerinin İncelenmesi]. *Harita Dergisi*. 164, 77-87.

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.

Jilani M, Corcoran P, and Bertolotto M (2014, November). *Automated highway tag assessment of OpenStreetMap road networks*. 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 449-452), Dallas/Fort Worth, TX, US, 4-7 Nov 2014.

Mondzech J and Sester M (2011). Quality analysis of OpenStreetMap data based on application needs. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(2), 115-125.

Mooney P and Corcoran P (2012a). The annotation process in OpenStreetMap. *Transactions in GIS*, 16 (4), 561-579.

Mooney P and Corcoran P (2012b). Characteristics of heavily edited objects in OpenStreetMap. *Future Internet*, 4(1), 285-305.

Mooney P, Corcoran P, and Winstanley A C (2010, November). *Towards quality metrics for OpenStreetMap.* 18th SIGSPATIAL international conference on advances in geographic information systems (pp. 514-517), San Jose, CA, US, 2-5 Nov 2010.

OpenStreetMap Wiki (2021). Map Features. https://wiki.openstreetmap.org/wiki/Map_Features, accessed 20.01.2021.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, ... and Duchesnay É (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.

Rosin P L (1999). Measuring rectangularity. *Machine Vision and Applications*, 11(4), 191-196.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, CRC press, London.

Taginfo. (2021). OpenStreetMap Taginfo. https://taginfo.openstreetmap.org/, accessed 20.01.2021.

Zhang H and Malczewski J (2018). Accuracy Evaluation of the Canadian OpenStreetMap Road Networks. *International Journal of Geospatial and Environmental Research*, 5(2), 1-14.

Biography

Müslüm Hacar works as a researcher at Department of Geomatic Engineering, Yildiz Technical University. He did research in the main fields of cartography such as conflation (in doctoral thesis) and generalisation. He is currently working on quality assessment of VGI data and the behaviour of contributors.