ISPRS Int. J. Geo-Inf. **2014**, *3*, 750-763; doi:10.3390/ijgi3020750

OPEN ACCESS

ISPRS International Journal of

Geo-Information

ISSN 2220-9964

www.mdpi.com/journal/ijgi/

Article

A Quality Study of the OpenStreetMap Dataset for Tehran

Mohammad Forghani [1] and Mahmoud Reza Delavar [2,*]

[1] Department of Surveying and Geomatic Engineering, College of Engineering, University of Tehran, Tehran 14174, Iran; E-Mail: mo.forghani@ut.ac.ir

[2] Center of Excellence in Geomatic Engineering in Disaster Management, Department of Surveying and Geomatic Engineering, College of Engineering, University of Tehran, Tehran 14174, Iran

***** Author to whom correspondence should be addressed; E-Mail: mdelavar@ut.ac.ir; Tel.: +98-21-611-142-57; Fax: +98-21-880-088-37.

Received: 8 November 2013; in revised form: 22 April 2014 / Accepted: 8 May 2014 / Published: 22 May 2014

Abstract: There has been enormous progress in geospatial data acquisition in the last decade. Centralized data collection, mainly by land surveying offices and local government agencies, has changed dramatically to voluntary data provision by citizens. Among a broad list of initiatives dealing with user generated geospatial information, OpenStreetMap (OSM) is one of the most famous crowd-sourced products. It is believed that the quality of collected information remains a valid concern. Therefore, qualitative assessment of OSM data as the most significant instance of volunteered geospatial information (VGI) is a considerable issue in the geospatial information community. One aspect of VGI quality assessment pertains to its comparison with institutionally referenced geospatial databases. This paper proposes a new quality metric for assessment of VGI accuracy and as well as for quality analysis of OSM dataset by evaluating its consistency with that of a reference map produced by Municipality of Tehran, Iran. A gridded map is employed and heuristic metrics such as Minimum Bounding Geometry area and directional distribution (Standard Deviational Ellipse), evaluated for both VGI and referenced data, are separately compared in each grid. Finally, in order to have a specific output as an integrated quality metric for VGI, its consistency with ground-truth data is evaluated using fuzzy logic. The results of this research verify that the quality of OSM maps in the study area is fairly good, although the spatial distribution of uncertainty in VGI varies throughout the dataset.

Keywords: VGI; spatial data quality; fuzzy; open street map dataset

1. Introduction

Only a few special companies and national agencies in each country have collected geospatial data in the past. However, recent years have witnessed an influx of websites and web-based tool sets due to the emergence of Web 2.0 that has offered the possibility of crowd-sourced geospatial information acquisition and has brought about rapidly growing availability and accessibility of data in the geo-domain [1].

Web 2.0, geo-referencing, geotags, GPS, graphics and broadband communication are among the technologies that identified by Goodchild [2] as technologies which have made this activity possible. Goodchild also coined the term "Volunteered Geographical Information" (VGI) to describe this special case of user-generated content.

The term "VGI" mainly indicates the deliberate act of providing geospatial information, often by untrained users. Moreover, some other terms such as "Neogeography" [3], or "wikification of GIS" [4] are utilized to describe this act; however, all these terms imply the same concept of turning passive consumers into active producers of geospatial information.

In this context, OpenStreetMap website (OSM) is considered as one of the well-known VGI projects. The history of this project dates back to about ten years ago, and like other similar and famous websites such as Wikipedia, its gathered information can be described as user-generated content (UGC) [5]. In other words, OSM is an online map of the world which is open source and editable. The map is mainly created by volunteer users around the globe through collection and contribution of geospatial data either through users' own GPS data tracks, digitally tracing aerial images, or data acquisition from miscellaneous free sources. All OpenStreetMap data can be downloaded free of charge and in vector format, which has led to their widespread use.

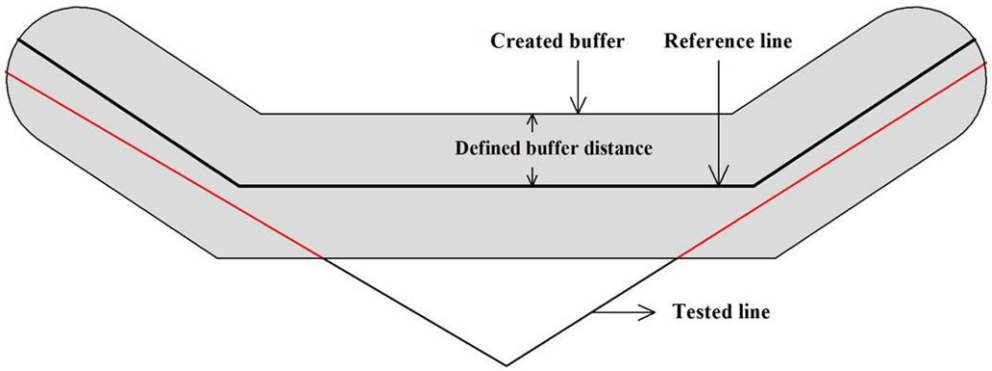
OSM data is considered beneficial for three main reasons [6]: firstly, being free, secondly being up-to-date, and thirdly providing global coverage, even for the less developed parts of the world. Despite these benefits, there exist some general concerns about OSM and VGI data. Data quality is one of the main problems in this context which can also be attributed to other online UGS-based portals [5]. Data quality can be defined as fitness and appropriateness for use, or how suitable some data is for meeting specific needs or fulfilling certain requirements for resolution of a certain problem.

While professional mapping agencies employ applied standards, specifications and advanced technical methods to make sure their information production is of optimum credibility and authority, geospatial information offered by VGI projects such as OSM have a distributed nature of data gathering and loose co-ordination in terms of standards. As a result, user-generated geospatial information is of questionable credibility and there is a need to assess the actual appropriateness and suitability of OSM data.

A large number of research works conducted on VGI quality assessment focus on quantitative analyses carried out through comparison of VGI with referenced maps. These studies have been carried out using OSM data in different countries such as the UK [7], Ireland [8], Germany [9] and France [10].

One of the preliminary comparison analyses on OSM was carried out by Haklay [7] whose research involved an initial assessment of positional accuracy and completeness for Great Britain data through the comparison of motorway segments between OSM and Ordnance Survey (OS) Meridian 2 datasets. For assessment of positional accuracy, he followed the simplest version of Increasing Buffer Method (IBM) [11] to compare the linear objects, using a predefined buffer value to calculate the corresponding overlap percentage (Figure 1); and to assess the data completeness, he calculated the total length per sequence km for the two datasets and compared them.

Figure 1. Increasing Buffer Method (IBM): determining the level of accuracy based on proportion of the tested line within the created buffer around the reference line [11].



A visual comparison between OSM, Google Maps and Bing Maps was conducted in Ireland to study completeness, currency and accuracy of data [8]. Zielstra and Zipf [9] evaluated OSM data in Germany by comparing it with TeleAtlas dataset. They split the data into tiles of one Square Kilometer, calculated the total length of the datasets for each tile and evaluated the completeness of data based on the calculated values. VGI data quality was studied in France by extending Haklay's work, providing the assessment of a larger set of spatial data quality elements (*i.e.*, geometrics, attribute, semantics and temporal accuracy, logical consistency, completeness, lineage and usage), and using different methods of quality control. For instance, Hausdorff and average distance methods were used for positional accuracy assessment [10].

To make a contribution to his discussion, this paper attempts to evaluate the quality of OSM data. OSM is also compared with the reference map produced by Municipality of Tehran, Iran, based on some innovative metrics.

The advantage of employing the new metrics is achieving a more comprehensive view of the quality of geospatial data under study and identifying any uncovered incompatibilities in the data. This can help with identifying geographic regions with less vagueness and more accuracy for use in various GIS applications.

With regard to the structure of the paper, Section 2 outlines the experimental analysis of OSM data and provides a discussion of spatial data quality metrics on the VGI domain. Section 3 concludes the study by providing a discussion of the results.

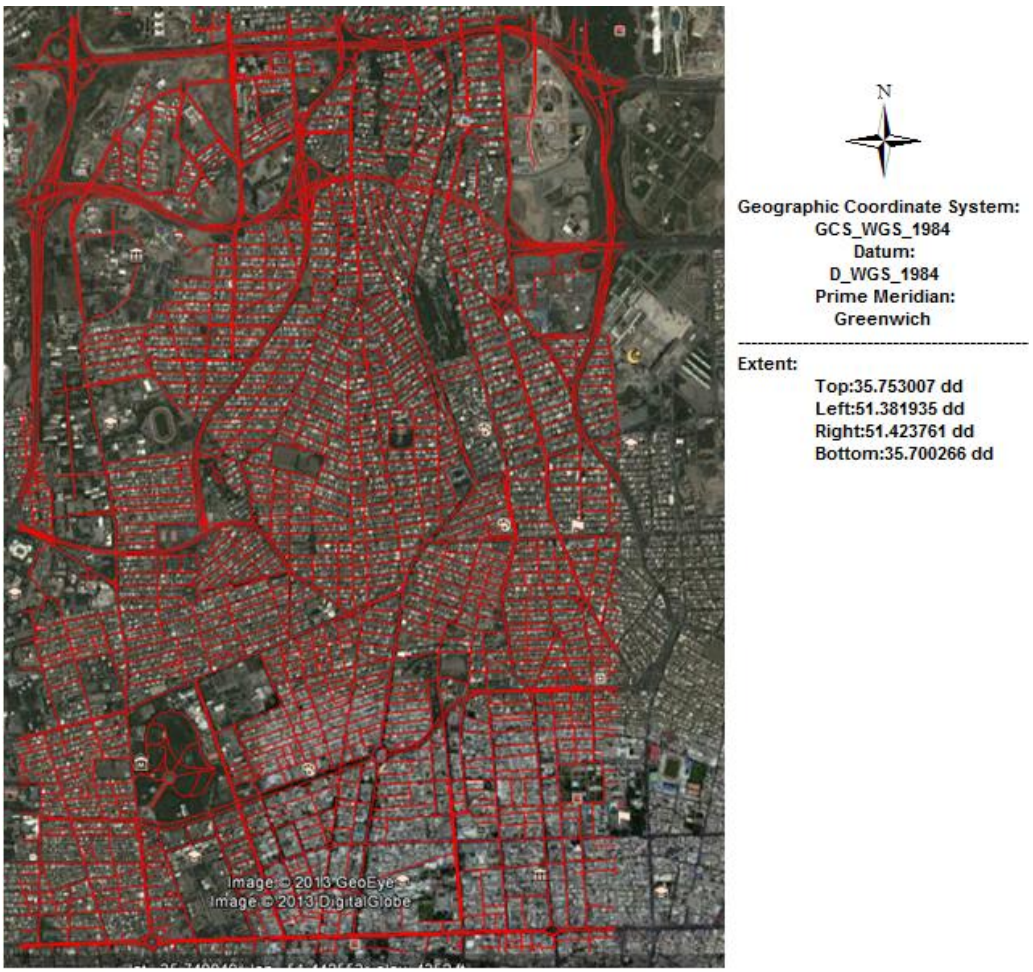
2. Methodology

This section introduces the proposed method for quality analysis of OSM data for Tehran. The zone under study, with an approximate area of 20 km^2, is one of the central urban zones in Tehran, the capital of Iran.

2.1. Available Datasets

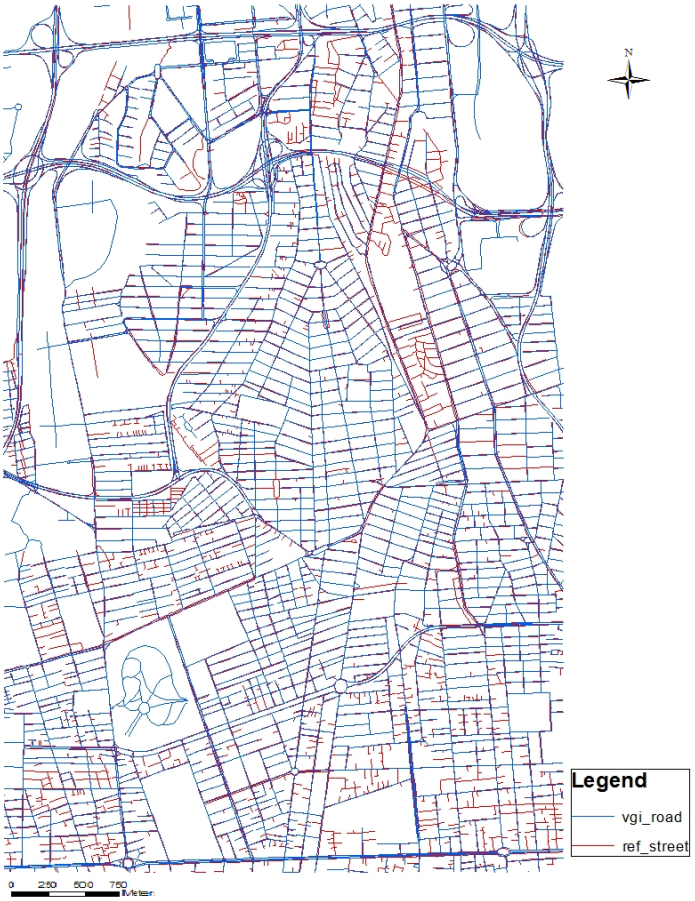
Geo-data can be extracted from the OpenStreetMap through two main methods. The first way is defining a particular area and downloading the contained XML information from OpenStreetMap. The second way is using websites such as "Geofabrik" or "Cloudmade" which offer free downloadable OpenStreetMap data in different formats such as XML and Shapefile. The latter was used in the current study. OSM data for Tehran was downloaded in shapefile format from Cloudmade and then the desired urban zone was clipped (Figure 2).

Figure 2. OpenStreetMap data of the study area overlaid on Google Earth.



On the other hand, in order to determine VGI quality in quantitative terms, a comparison with official data of a higher quality (data produced by accepted quality standards) is required. In this study, the reference map at a scale of 1:2000 that was produced by the Municipality of Tehran, in Shapefile format, has been used as ground-truth data (Figure 3).

Figure 3. Overlaying the OpenStreetMap (OSM) map with the reference map.



2.2. Assessment Method

The methodology of this study consists of three steps:

1. Split of datasets into tiles using grids. This step enables the comparison of the two sets in each cell grid;
2. Evaluation of quality metrics on a tile-by-tile basis for both VGI and reference maps;
3. Combination of the evaluated metrics through a fuzzy set approach and calculation of a quantitative term representing the quality of OSM data in each cell.

2.3. Quality Metrics

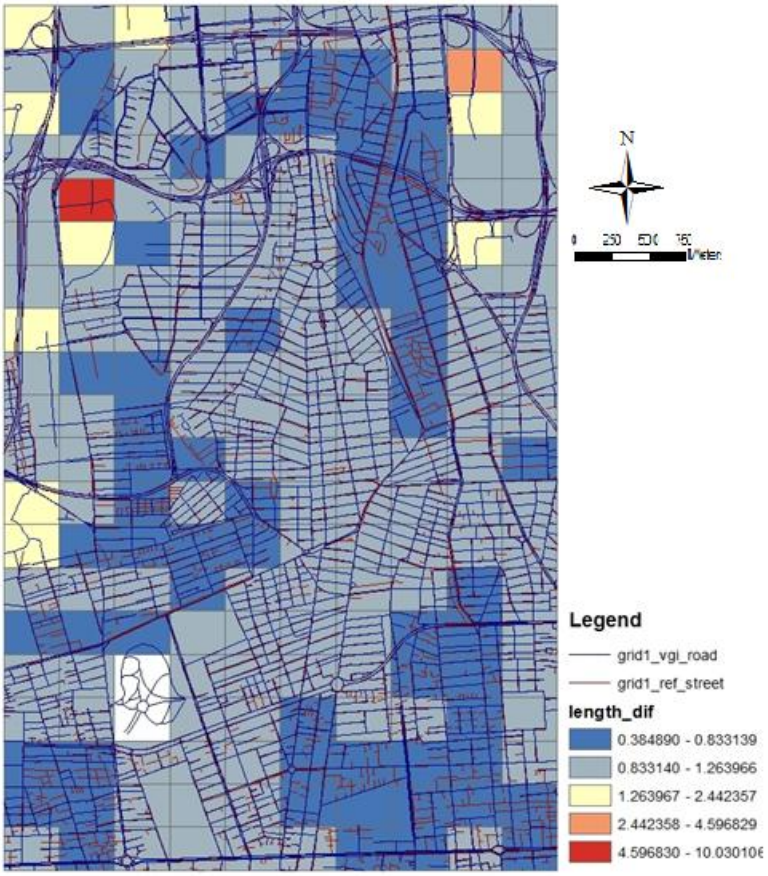
In this study, the following four measures were utilized for comparison of the two datasets.

2.3.1. Road Length

This metric is generally used to investigate the completeness of OpenStreetMap data in comparison with the reference map. The completeness of a road network can be determined by calculating the total length of the roads in one of the OSM datasets within a tile area and then comparing it with that of the reference map within the same area using Equation (1) (Figure 4). A difference in the overall length indicates inconsistency between the two datasets.

OSM road network completeness = |1 − (Σ(OSM roads length)/Σ(reference roads length))| (1)

Figure 4. The calculated difference between total length of roads contained in each cell in OSM and the reference map.



2.3.2. Minimum Bounding Geometry

The smallest convex polygons enclosing road features in each grid of both datasets were obtained and then their areas were calculated and compared (Figure 5). It is clear that more diversity between the calculated areas of Minimum Bounding Geometries indicates more inconsistency between the two datasets.

2.3.3. Directional Distribution (Standard Deviational Ellipse)

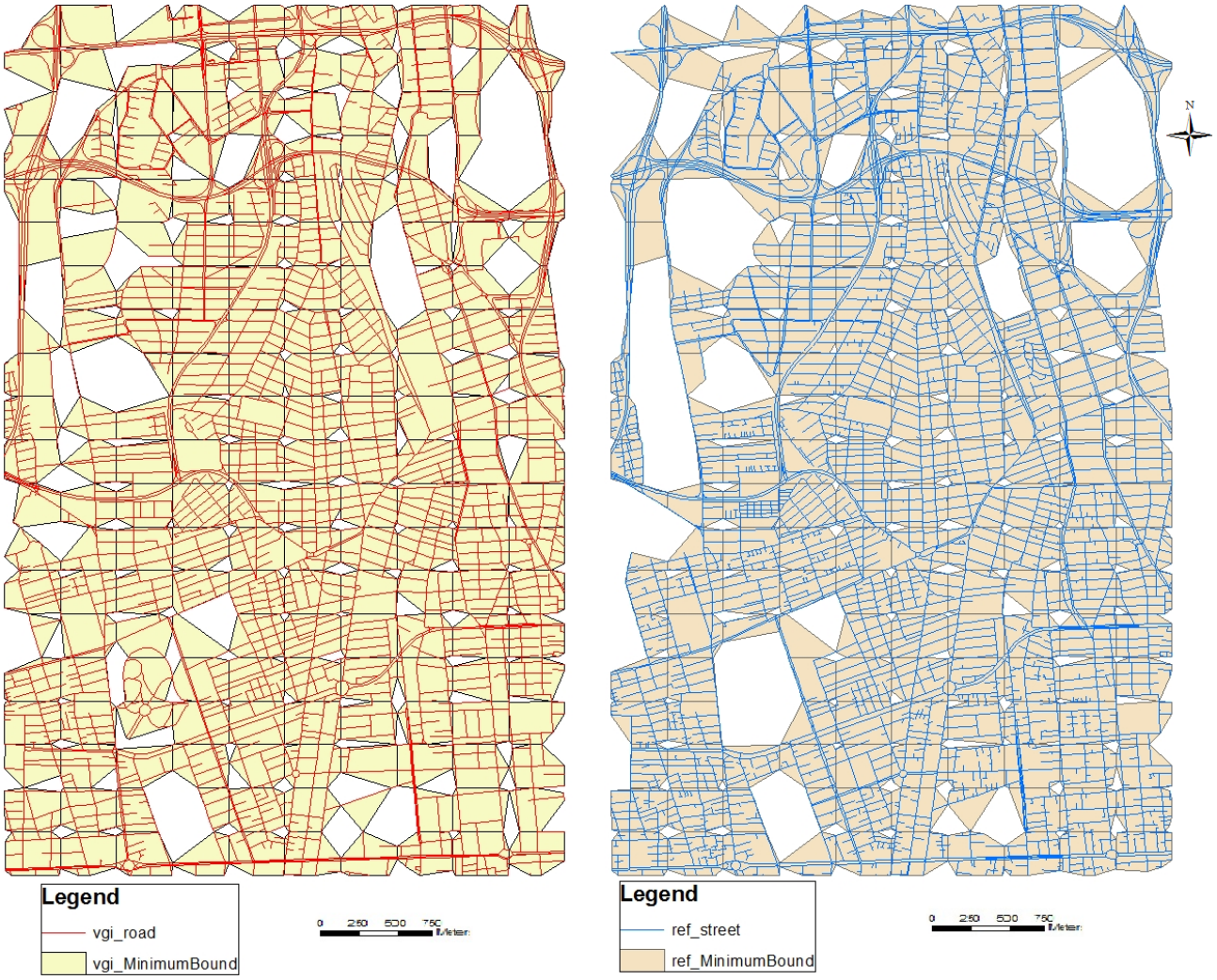
Directional Distribution metric summarizes the spatial characteristics of geospatial features such as central tendency, dispersion, and directional trends. Therefore, if Standard Deviational Ellipses for each grid of the two datasets are obtained and their directions are compared, the compatibility of the two datasets can be assessed (Figure 6).

2.3.4. Median Center

The last quality metric employed in this study was median center. The median center of a cell implicates the location that minimizes overall Euclidean distance to the features contained in that cell.

Thus, the distance of median centers of the two datasets in each grid cell can be considered as a metric for their comparison (Figure 7).

Figure 5. (**a**) Minimum bounding geometry for features of each tile for OSM. (**b**) Minimum bounding geometry for features of each tile for the reference map.



(**a**) (**b**)

2.4. Fuzzy Model

Spatial data quality consists of three parts including the definition of elements of spatial data quality, establishment of metrics for measuring these elements and finally the communication of data quality [12]. Although there are different elements of spatial data that are quality defined, such as completeness, positional accuracy, temporal accuracy and lineage, since VGI is a new trend in Geospatial Information Science, there is a confusion as to how and which element of spatial data quality can apply to VGI. On the other hand, the defined metrics do not measure only and exactly one quality element, but rather, each of them covers some integrated aspects of spatial data quality.

Figure 6. (**a**) Standard Deviational Ellipse for features of each cell for OSM. (**b**) Standard Deviational Ellipse for features of each cell for reference map.

(**a**) (**b**)

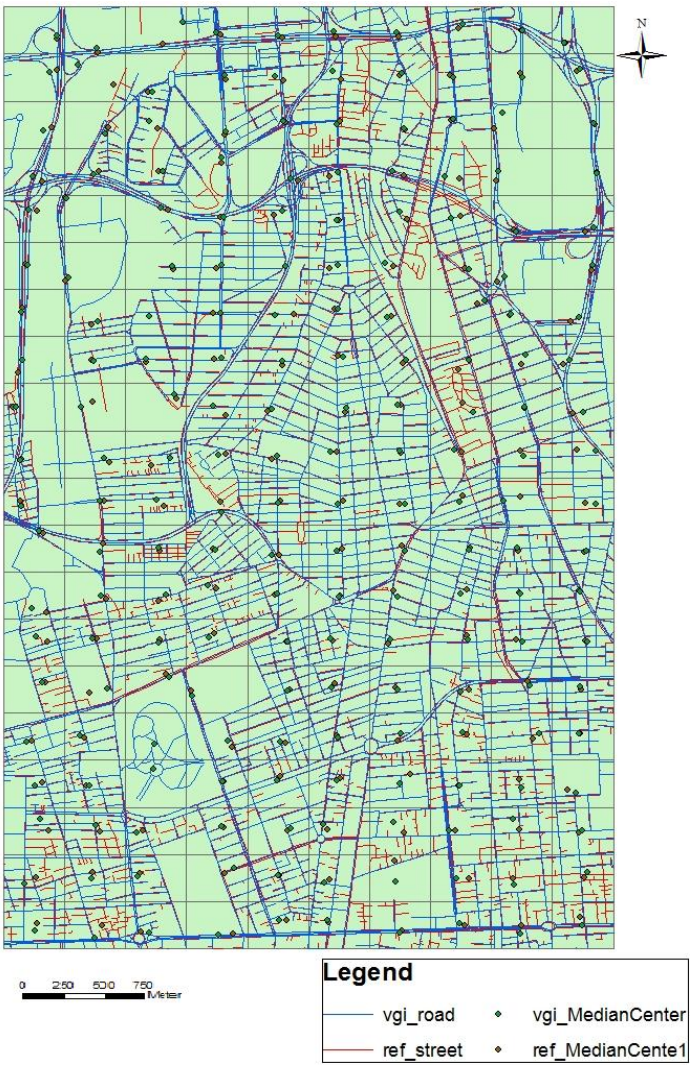
As a result, the measured metrics need to be integrated to have a specified value, indicating the quality of OSM map in comparison with that of the reference map. For this reason, a flexible method is required to handle vagueness or uncertainty surrounding the combination of quality metrics [13]. One of the most well-known methods in this case is fuzzy logic that allows flexible weighted combinations.

Fuzzy logic is a convenient way to map an input space to an output space. Although fuzzy overlay and weighted overlay look similar to each other, the two are built on different foundations. Fuzzy overlay is based on set theory, while weighted overlay is based on linear combinations. So in the combine step, fuzzy logic explores the interaction of the possibility of the phenomenon belonging to multiple sets and the interaction of the inaccuracies in the membership of the sets, as opposed to weighted overlay which is based on a relative preference scale. Therefore, Fuzzy logic has proved to be of great appeal to researchers as it is appropriate for handling vagueness in geospatial domain, specifically to analyze the relationships and interaction between all the sets for the multiple criteria in the overlay model.

Regarding the idea of fuzzy logic [14], the evaluated metrics would be characterized into classes. However, due to imprecision of thought, vagueness and ambivalence, the boundaries between classes

are not always exactly sharp and it is not clear-cut whether something belongs to a class or not. Both of these sources of inaccuracies can cause imprecision in assigning cells to specific classes. In human language, these imprecisions are qualified through modifiers, such as very, slightly and moderately and fuzzy logic performs process more like natural human thinking.

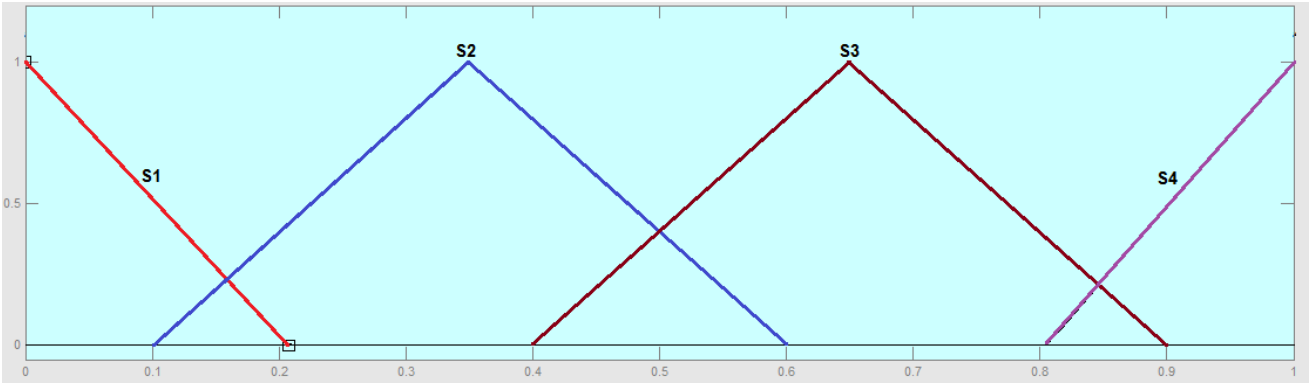
Figure 7. Median centers of OSM and reference dataset features in each grid cell.



In Fuzzy logic, the classes are defined as sets. In the classical set theory, an object is a member of a set if it has a membership value of 1, otherwise the membership value of 0 is assigned. By contrast, in fuzzy set theory, membership value can take on any value between 0 and 1 reflecting the degree of membership certainty. For example, for road length as one of the input criteria, each tile value will be transformed or assigned a value between 0 and 1 on the possibility of that Road length value being a member of High-inconsistency class (set). The value 1 indicates full certainty that the value is in the set, and 0 indicates with full certainty that it is not in the set. All other values are some level of possibility, with the higher values indicating more likelihood of membership. The process of transforming the original input values to the 0 to 1 scale of possibility of membership is called the fuzzification process. In this study, one output parameter, indicating the uncertainty of OSM and four already defined input parameters were considered and four classes were used for each of these

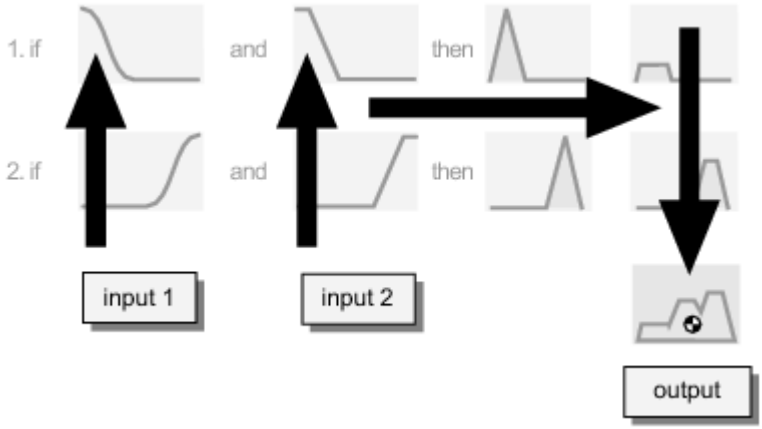
parameters to make them categorical variables. Then, a triangular membership function that had been obtained by visual analysis was assigned to each of them. A case in point is the output parameter. For this variable, four implications, including strong-consistency (S1), adequate-consistency (S2), moderate-consistency (S3) and week-consistency (S4), were defined (Figure 8). The meaning of strong-consistency (S1), for instance, is that there is a high consistency between OSM and the reference map.

Figure 8. Membership function plot of output variable.



The next step is using a fuzzy rule-based system to integrate the inputs and evaluate the output parameter. Mamdani's fuzzy inference method is the most commonly seen fuzzy methodology. Figure 9 shows the employed rule-based inference system diagram.

Figure 9. Fuzzy rule-based inference system diagram.

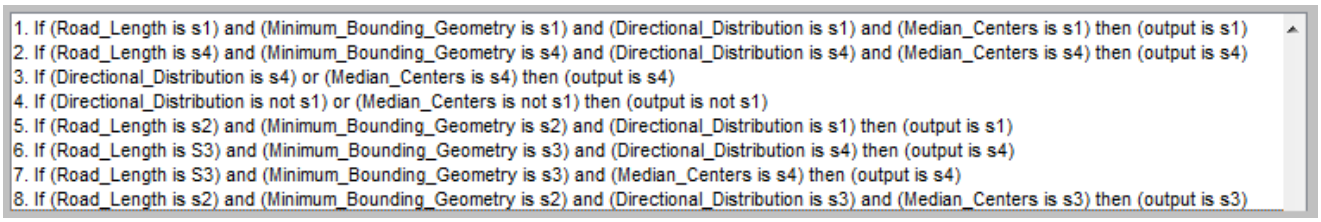


Based on the above diagram, a variety of operators such as the fuzzy AND and fuzzy OR can be employed to combine the input membership values based on some defined rules. In this study, to find appropriate rules, a visual comparison was done between OSM map, the reference map and each of the criteria evaluated maps. A number of employed rules are presented in Figure 10.

The result of applying fuzzy operators is to obtain one number that represents the result of the antecedent for the rule and after that, the consequent of rule is implied as a fuzzy set represented by a membership function reshaped by antecedent. Because the evaluation is based on the testing of all of the rules in system, the rules must be combined. Aggregation is the process by which the fuzzy sets

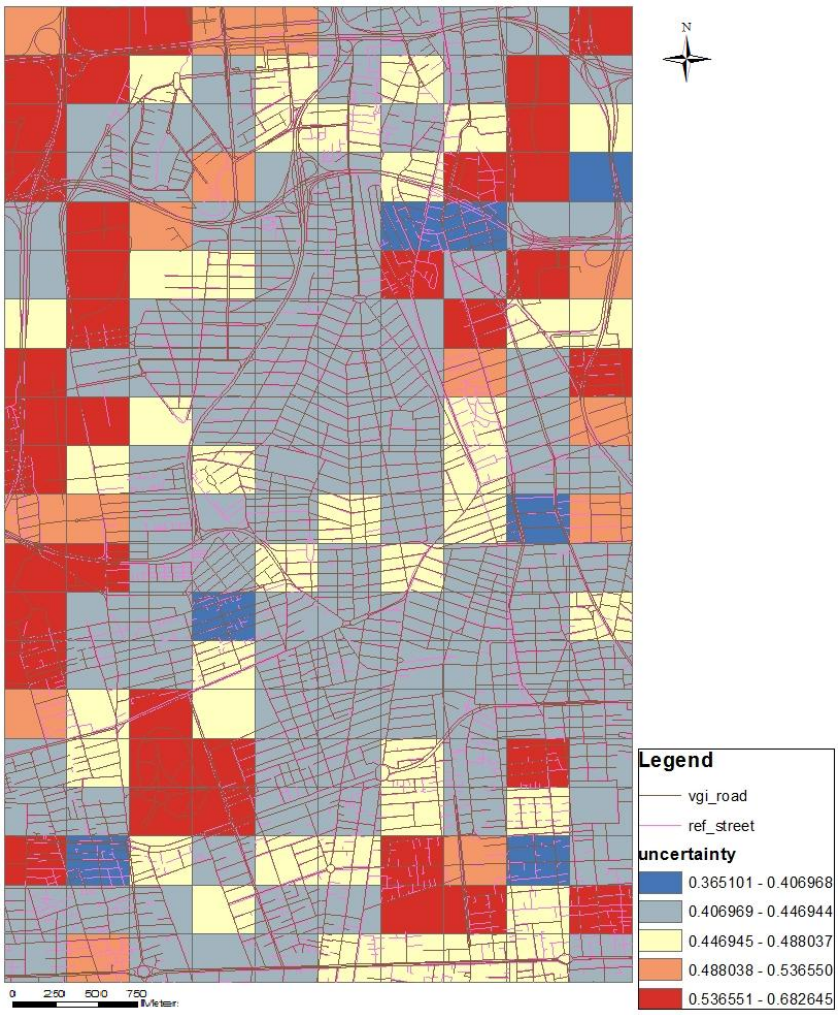
that represent the output of each rule are combined into a single fuzzy set. After the aggregation process, there is a fuzzy set for output variable that needs defuzzification. Perhaps the most popular defuzzification method is the centroid calculation, which returns the center of area under the curve of the membership function associated with the aggregate output fuzzy set.

Figure 10. A number of defined rules in inference system.



The system described above was used to integrate the inputs and evaluate the output parameter for each grid cell. Figure 11 presents an overview of the evaluated uncertainty of OSM map in comparison with the referenced map using fuzzy logic.

Figure 11. Evaluated uncertainty of the OSM dataset in comparison with that of the reference map on a tile-by-tile basis.



Employing the metrics such as length, smallest convex polygons and directional distribution (tile by tile) is actually easier than using the methods like IBM that require some preprocessing steps. Neither of the proposed measures alone can be of much use in the quality assessment of OSM data. For example, length measurement alone cannot say much about the quality. Two datasets with the same total length of roads may have different roads. However, each of these criteria can be an approximate indicator of the consistency between OSM data and the ground-truth map and the measurement resulted from integration of the theme can be more precise and useful in presenting the degree of consistency. This may be justified as their combination can reduce the uncertainty of each criterion and make an indicator which covers more quality aspects of OSM data. Thus, the fuzzy logic is employed to integrate them. Visually comparing Figure 3, which presents an overlay of the OSM map with a reference map, with Figure 11, which shows integrated quality metric in each tile, can illustrate the ability of the proposed method to evaluate the consistency of an OSM map with a reference map.

3. Conclusions

Although OSM data, as an instance of VGI data, has numerous advantages, there are some concerns about their usage. One such concern is the spatial data quality for whose assessment various methods have been proposed.

The current study, in addition to providing an overview of spatial data quality in VGI and OSM, attempted to assess the quality of free and voluntarily provided data by OpenStreetMap users. One of the central urban zones of Tehran was selected for the purpose of this analysis. The area under study consisted some places that had more complete data and some that were relatively empty.

The method applied in this study for quality assessment of OSM geospatial data was their comparison with the accurate existing data. In comparison with previous research studies, the present study was innovative in developing new quality metrics for OSM data assessment that can cover more qualitative aspects of VGI spatial data.

According to the obtained numerical results, the evaluated uncertainty for each of the grid cells of the study area (0.36–0.68) and defined linguistic terms in the fuzzification step, it can be stated that there are no areas with high-quality data but there are no areas with very low quality data either. However, the study area is generally composed of cells with medium quality data (about 80% of study area).

Although it was shown that VGI could reach a fairly good spatial data quality, it was confirmed that the actual problem was the way to deal with uncertainty of OSM because as presented in Figure 9, the quality of the OSM data in different tiles varies considerably. With regard to the results of the study, the main quality problem of OSM dataset is heterogeneity of the OpenStreetMap data in terms of their completeness in comparison with that of the reference map. Such heterogeneity leads to inconsistency between the OSM dataset and ground-truth data. While it is an accepted hypothesis that commercial datasets contain some randomly distributed errors, such random errors cannot be assumed as the cause of the evaluated level of uncertainty for OSM grid cells. This poses questions on reliability of VGI information, especially in precise GIS analyses. Therefore, it can be concluded that even though OSM data is very cost-efficient, its reliability and applicability as an alternative to commercial datasets depend largely on its actual application. This means that, regarding the required precision of spatial data and based on the area of interest, it can be decided whether the OSM data can be useful or not.

Author Contributions

Mahmoud Reza Delavar initiated the concept of the paper, critically revised the paper and took care of editorial and proof-reading issues. He is also mainly contributed in the abstract, introduction and conclusion of the paper. Mohammad Forghani contributed major parts of modeling, writing and structuring of the contents.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zielstra, D.; Zipf, A. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 10–14 May 2010; pp. 1–15.
2. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221.
3. Hudson-Smith, A.; Crooks, A.; Gibin, M.; Milton, R.; Batty, M. NeoGeography and Web 2.0: Concepts, tools and applications. *J. Locat. Based Serv.* **2009**, *3*, 118–145.
4. Sui, D.Z. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Comput. Environ. Urban Syst.* **2008**, *32*, 1–5.
5. Cooper, A.; Coetzee, S.; Kaczmarek, I.; Kourie, D.; Iwaniak, A.; Kubik, T. Challenges for Quality in Volunteered Geographical Information. In Proceedings of the AfricaGEO 2011 Conference, Cape Town, South Africa, 31 May–2 June 2011.
6. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120.
7. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Design* **2010**, *37*, 682–703.
8. Ciepluch, B.; Jacob, R.; Mooney, P.; Winstanley, A. Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science, Leicester, UK, 20–23 July 2010.
9. Neis, P.; Zielstra, D.; Zipf, A. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet* **2011**, *4*, 1–21.
10. Girres, J.F.; Touya, G. Quality assessment of the French OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459.
11. Goodchild, M.F.; Hunter, G.J. A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 299–306.
12. Servigne, S.; Lesage, N.; Libourel, T. Approaches to Uncertainty in Spatial Data. In *Fundamentals of Spatial Data Quality*; Devillers, R., Jeansoulin, R., Eds.; ISTE Ltd.: London, UK, 2006; pp. 179–210.

13. Bordogna, G.; Carrara, P.; Criscuolo, L.; Pepe, M.; Rampini, A. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Inf. Sci.* **2014**, *258*, 312–327.
14. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353.