Assessing the Quality of OpenStreetMap Contributors together with their Contributions

Jamal Jokar Arsanjani
University of Heidelberg
Geoinformatic Department
Berlinerstrasse 48, 69120
Heidelberg, Germany
Jokar.arsanjani@geog.uni-heidelberg.de

Christopher Barron
University of Heidelberg
Geoinformatic Department
Berlinerstrasse 48, 69120
Heidelberg, Germany
christopher.barron@gmx.de

Mohammed Bakillah
University of Heidelberg
Geoinformatic Department
Berlinerstrasse 48, 69120
Heidelberg, Germany
mohamed.bakillah@geog.uni-heidelberg.de

Marco Helbich
University of Heidelberg
Geoinformatic Department
Berlinerstrasse 48, 69120
Heidelberg, Germany
marco.helbich@geog.uni-heidelberg.de

Abstract

In this paper, the volunteers' contributions to the OpenStreetMap (OSM) project is evaluated based on comparative investigations with administrative data of Germany provided by the Federal Agency for Cartography and Geodesy. Several data quality aspects, including for instance positional accuracy, completeness, and semantic accuracy, are analyzed and compared considering their contributors. Accordingly, several categories of OSM contributors are characterized based on the quantity and quality of their shared data. As such "beginners", "regular mappers", "intermediate mappers", "experts", and "professional mappers" are identified. The categorization of contributors proves the 90-9-1 rule applies in this study as well. A small number of contributors are professional and share the most information accurately, conversely, a large number of contributors contribute to OSM with the least amount of contribution and the minimal quality.

Keywords: Data quality, OpenStreetMap, contributors, volunteered geographic information

1 Introduction

Recently, geography has faced a new era in collecting spatial objects, numeric, and descriptive information on geographic objects, which has led it into a new concept so-called Neogeography [13]. This is the consequence of advancement of mobile technologies and broadband communication, among others, in the frame of Web 2.0. This has been caused due to a continuous influx of geoinformation from the Internet, mainly gathered through collaborative mapping projects. One of the most well-known and popular examples of it is the OpenStreetMap (OSM) project, which aims at providing a free editable map and the underlying data of the world dedicated by its inhabitants.

As the OSM project is totally open and no observation on the contributions is applied, the reliability and credibility of the shared data has to be considered before using them for any purposes. Incorrect, incomplete, and in general inaccurate data itself cause biased outputs and might result in trustworthy outcomes in geospatial applications. Moreover, this "error" can further be propagated in the modeling practices and decision making process. Thus, considering the quality of the data is of utmost importance. So far, a number of investigations on evaluating the quality of OSM data have been carried out [9]. Among the OSM feature sets, street network has been more paid attention to, and among data quality criteria, only completeness and positional accuracy have attracted much attention rather than the remaining items [9, 10].

Therefore, the main objective of this paper is to address the semantic accuracy of the contributions in addition to the completeness and the positional accuracy. Consequently, the contributors are ranked based on their quality and quantity of their contributions. This is in particular of importance because identifying the accuracy of the OSM data ensure developers how trustworthy this data source for the application is.

1.1 Geodata Quality and Uncertainty

Two major concepts, i.e. accuracy and precision, are often used for addressing the quality of data, which are semantically different and misused in the geospatial literature [14]. Precisely, *accuracy* is the degree to which data on a map or in a geodatabase matches with the reference data or values, while *precision* addresses the level of measurement and correctness of information in a geodatabase [6]. But generally speaking, the concept of “data quality” is sometimes incorrectly interpreted as data precision, uncertainty, or error. Geodata with high locational precision are often called high-quality data. However, the sense of quality is far beyond the concept of locational precision [14].

In general, the quality of geodata should be internally and externally recognized. One the one hand, internal quality concerns the data production standards and specifications, which is based on the errors existence in the data. Several standard organizations (e.g., ISO, ICA, FGDC, and CEN) define internal quality based on five aspects (also called the “famous five”): 1) attribute accuracy, 2) positional accuracy, 3) temporal accuracy, 4) logical consistency, and 5) completeness [5]. These data properties are introduced to the users through metadata files attached to datasets by producers [2]. On the other hand, external quality considers whether a dataset is suitable enough for a particular purpose and addresses the concept of “fitness of use” (FoU: [2, 8, 11]). In other words, users expect a dataset to meet/exceed some expectations, therefore, each quality measurement aspect does not concern them very much [3].

1.2 OSM contributors

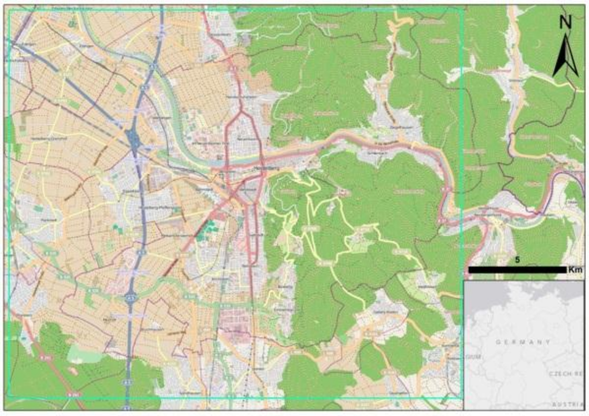
Individuals are contributing to OSM either constructively or damagingly. According to early efforts on identifying contributors to OSM (e.g., [1]), five major groups of users are characterized: neophytes, interested amateurs, expert amateurs, expert professionals, and expert authorities. However, depending on the type of VGI service and kind of contributions, the contributors vary. Therefore, a more representative approach for identifying users is needed. Neis & Zipf [12] categorized the OSM mappers into several categorizes based on the quantity of contributions as hit-and-run mappers, newbies, casual mappers, heavy mappers, heavy mappers 2.0, addicted mappers, crazy mappers, and bots. However, identifying the contributors based on not only the quantity but the quality of their contributions is still lacking. This paper aims at fulfilling this issue by cross comparing contributors' data with their quality.

2 Materials

2.1 Study site

As an experimental study site, the city of Heidelberg, located in the German state of Baden-Württemberg, is chosen. The selected spatial extent of interest is shown in figure 1, which covers the administrative boundary of Heidelberg. This sample area of Germany is selected, because it has received a lot of contributions and also contains heterogeneous landscapes, comprising urban and rural areas.

Figure 1: Spatial extent of the Heidelberg



2.2 Data Pre-processing

Two main datasets are used in this research, a) OSM data and b) official data provided by the Federal Agency for Cartography and Geodesy (BKG), which is prepared by a national authority in Germany responsible for providing geodata. The OSM data was downloaded through the Geofabrik website on November 12th, 2012 and the BKG data represents the latest version of street network (2011). For both datasets, every street segment was considered and imported to the geodatabase for the main analysis.

3 Methods

In total, 346 contributors contributed 15,350 road segments, which are in total 2,624 km road networks within the study site. Once the datasets were prepared, the data quality analysis was applied in order to initially evaluate aspects of data quality criteria and subsequently rank OSM contributors according to the quantity and quality of their contributions. As such, positional accuracy, temporal accuracy, logical accuracy, semantic accuracy for each individual contributor is computed and discussed. Each data quality aspect is discussed as follows.

Positional accuracy: This quality criterion was carried out through a buffer analysis as recommended by Goodchild and Hunter [4]. Under the premise of projecting the true fact of road networks, for each road type of the BKG data e.g., primary roads, secondary roads, motorways, residential dissimilar buffer distances of 3, 5, 10, and 15 meters for each side were applied to create the polygonal coverage of road network. To be noted that no road segment was deleted from this analysis as contributors may not recognize the type of roads properly or the road types mentioned in the official datasets have changed by the time of contributions to the OSM. The amount of overlap between OSM data and BKG polygonal coverage is interpreted as a function of positional accuracy, i.e., how good the OSM contributions can be matched with official data at certain thresholds.

Temporal accuracy: Whereas the position and attributes of geographic objects are not covered in most of datasets, this accuracy criterion cannot be further investigated. However, as OSM is receiving contributions over time and records them temporally in the OSM full history file, in future this criterion can be fulfilled by people once the data is relatively complete and contributors begin to modify the attributes and position of objects over time. This can be achieved even better than administrative data, because the collaborative approach seems to be wiser due to high cost of data collection.

Logical consistency: It addresses the trustworthiness of the topological and logical relationships encoded between the dataset segments. There is no indicator to measure it quantitatively; however, visually this element is of a major concern for collaboratively collected data like OSM. Nonetheless, depending on the type of data, the degree of goodness varies, i.e., point data such as POIs don’t face this problem as much as polygonal entities do. Polyline data such as roads contain such a problem on a variety of cases. Such topological inconsistency can be seen at the road junctions, beginning and ending of the road segments. Nonetheless, these issues can be resolved by applying some automatic functions for correcting topological errors as well as connecting the lines endings and beginnings. Nevertheless, some tools like the OSM inspector have been recommended by the OSM developers for checking and resolving this problem. As contributors are not trained to map, always OSM data preparation procedure must be considered. More importantly, as no scale for OSM data has been yet defined, logical existence of some road segments is questioned.

Completeness: This aspect is considered as a ratio of the length of the shared road segments by each volunteer to the total shared road segments. As completeness is meant to measure the degree of data absence in a dataset to a reference dataset, a completeness index is proposed in order to measure the degree of data completeness and its relation to the contributors, i.e., how much of the whole dataset is contributed by each individual user. A completeness index is defined, which describes how much of the whole data is shared by each volunteer. This allows us to investigate how complete the volunteers contribute to map a particular area.

Semantic accuracy: This item reports how good the data represents the geographical entities in terms of their types and their attributes and semantics. For the case of OSM, two major items are subjected to measure: a) type of roads and b) attributes of roads. The whole semantic accuracy is measured via checking whether road type and road name of each segment are given or not. If so, the road segment has the maximum semantic accuracy. If less than it is achieved, then it can be proportionally measured.

4 Results

Once each quality aspect is measured quantitatively, the indices should be combined in order to measure the overall quality. A cross table is designed to compare the three measured indicators namely positional accuracy, completeness, and semantic accuracy. A set of conditional rules are defined to classify contributors such as "beginner", "regular", "intermediate", "expert", and "professional mappers" based on the quantity and quality of their contributions. Table 1 represents the range of quality aspects.

Table 1: The domain of each accuracy evaluation item

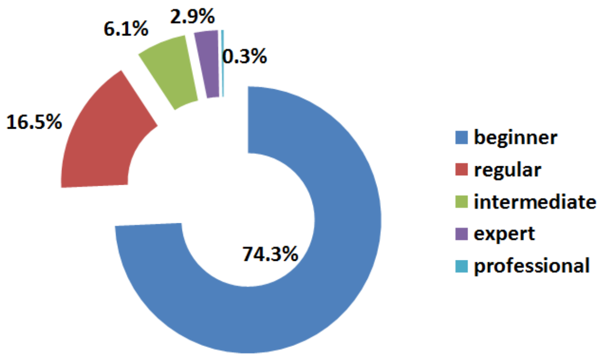
Quality item	Measured values domain
Positional accuracy (PA)	0-100%
Completeness Index (CI)	0-21.55%
Semantic accuracy (SA)	50-100%

Overall, the contributors have mapped the area at a range of 0,01 to 21,5%, where the positional accuracy of their shared data is measured between 0-100%. Moreover, the semantic accuracy of their data is measured between 50 to 100% as their data contain at least the type of roads, which means their contribution contain 50% semantic attributes. The indices are measured per contributor, i.e., the values are an average value of the whole line segments that a particular user has contributed. Each index is categorized into 5 classes i.e., PA1,PA2,…,PA5,CI1,…,CI5,…. Subsequently, each category of contributors is labeled by a set of indices.

"Beginner mappers" are those, who have the minimal expertise about mapping so that the positional accuracy of their contributions as well as their semantic accuracy of objects. These people do not contribute actively. Therefore, their contributions lack of fulfilling the completeness. "Regular mappers" share their data and only take care of importing data and digitizing objects with less attention on the positional and semantic accuracy of objects. "Intermediate mappers" map geographical objects whenever they have time and mind about the positional accuracy and less attention on the semantic accuracy. "Expert mappers" try to map geographical features based on their knowledge of the area and pay much attention on the positional accuracy as well as adding semantic information of the objects. Finally, "professional mappers" share the best of their knowledge of terrestrial objects to a large amount of data and even gather some relevant information to attach to the respective objects to present the most trustworthy contributions to OSM. Certainly, these people digitize objects at a professional level that is being done in commercial agencies.

Accordingly, among 346 contributors, 257 contributors are "beginners", 57 "regular mappers", 21 "intermediate mappers", 10 "experts", 1 "professional mapper". Figure 2 charts the pattern of contributors in the study area.

Figure 2: Pattern of different OSM contributors according to the quality of their contributions



5 Conclusions and Future Work

The present paper considered a set of criteria for assessing the reliability of OSM data, using the case study of

Heidelberg, Germany. The corresponding BKG data was collected as reference data to evaluate the quality of OSM. The degree of completeness, positional accuracy, and semantic accuracy of the dataset were analyzed. However, other aspects of data quality were pointed out such as logical consistency as well as temporal accuracy although they could not practically be measured because they cannot be quantitatively measured. As this research ranked OSM contributors according to the quality and quantity of their contributions, a number of conditional rules were defined to categorize contributors into several categories such as "beginner", "regular", "intermediate", "expert", and "professional mappers". Whereas assessing the quality of VGI is an emerging concern and needs to be considered more comprehensively and quantitatively, such an investigation helps to assess the quality of the VGI data together with their contributors so that within data quality analysis the data giver is considered as well.

Of course, no data is complete and it represents the best representation of the real world entities. Hence, it is not always required to measure every single element of a dataset to see how close to reality it is, but the data could be evaluated from a different point of view i.e., for which purpose the data is going to be used. Therefore, the conception of fitness for use must be further considered and respective measurement of fitness for use indicators to be applied. The authors would like to recommend to research more deeply on the techniques and indicators on measuring the fitness for use of each dataset. The proposed approach for ranking OSM contributors in the present research seems to be a practical approach, however further analysis of contributors and their contributions as well as the defining more indicators for quantitative measurement is meant to be followed. Moreover, considering a larger study area so that more contributors are subject to categorization is recommended.

References

[1] David Coleman, Yola Georgiadou, Jeff Labonte. Volunteered Geographic Information : The Nature and Motivation of Produsers, *International Journal of Spatial Data Infrastructures Research*, vol. 4, pages 332–358, 2009.

[2] R. Devillers, Y. Bédard, R. Jeansoulin, and B. Moulin. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data, *International Journal of Geographical Information Sciences*, vol. 21, no. 3, pages 261–282.

[3] Rodolphe Devillers, Robert Jeansoulin. *Fundamentals of Spatial Data Quality*, ISTE, London, 2006.

[4] M. F. Goodchild and G. J. Hunter. A simple positional accuracy measure for linear features, *International Journal of Geographical Information Science*, vol. 11, no. 3, pages 299–306, 1997.

[5] Stephan C. Guptill, Joel Morrison. *Elements of Spatial Data Quality*, Elsevier Science, New York, 1995.

[6] M. Gervais, Y. Bedard, M. Levesque, E. Bernier, and R. Devillers, "Data Quality Issues and Geographic Knowledge Discovery," pp. 99–116.

[7] Muki Haklay, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets, *Environment and Planning B: Planning and Design*, 37(4), pages 682–703, 2010.

[8] Julian Hagenauer, Marco Helbich, Mining urban land use patterns from volunteered geographic information using genetic algorithms and artificial neural networks, *International Journal of Geographical Information Science*, 26(6): pages 963-982, 2012.

[9] Marco Helbich, Chritoph Amelunxen, Pascal Neis, & Alexander Zipf, Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In *Proceedings of GI_Forum 2012*. Salzburg, Austria, pages 24-33, 2012.

[10] T. Koukoletsos, M. Haklay & C. Ellul. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 2012.

[11] J. L. Morrison, Spatial data quality. In Guptill, S.C. and Morrison, J.L. (Eds.), *Elements of Spatial Data Quality*, Elsevier Science, New York, 1995.

[12] Pascal Neis, & Alexander Zipf. Analyzing the Contributor Activity of a Volunteered Geographic Information Project – The Case of OpenStreetMap, *ISPRS International Journal of Geo-information*, 1, pages 146-165, 2012.

[13] Andrew Turner. *Introduction to Neogeography*, Sebastopol, CA: O'Reilly Media, 2006.

[14] Pepijn Van Oort. Spatial Data Quality: From Description to Application (PhD Thesis), Wageningen University, the Netherlands, 2006.