2014 11th International Conference on Information Technology: New Generations

A Systematic Study of OpenStreetMap Data Quality Assessment

Sukhjit Singh Sehra
Dept of Computer Science& Engineering,
GNDEC,
Ludhiana, Punjab, India
Email: sukhjitsehra@gmail.com

Jaiteg Singh
Department of Computer Application,
CIET,
Punjab, India
Email: jaitegkhaira@yahoo.co.in

Hardeep Singh Rai
Dept of Civil Engineering,
GNDEC,
Ludhiana, Punjab, India
Email: hardeep.rai@gmail.com

***Abstract*—The evolution of web has changed the way of interaction with the user. Web 2.0 encouraged more contribution from the user of varying level of mapping experience and is called Crowd Sourcing. OpenStreetMap is also the outcome of Crowd Sourcing. It is collecting huge data with help of general public, researchers have started analysing the data rather than collecting it. The aim of this study is to review the research work for assessment of OpenStreetMap Data. It is concluded that the most of research work on assessment of OpenStreetmap data has been done for countries like Germany, UK & USA. But the authenticity and accuracy of reference data still unanswered. Another issue that is concluded by this review, in context to Indian subcontinent, is the requirement of through analysis of OpenStreetMap data.**

I. Introduction

A map is a "snapshot" of an spatial information, a single picture of a constantly changing database of geographic information [1]. From cave paintings to ancient maps of Babylon, Greece, and Asia, through the age of exploration, and on into the 21st century, map has been an integral part of mankind for long time back. Traditional maps are less convenient and non-interactive and therefore less efficient than digital maps, one of the main reasons traditional paper maps are being superseded by digital maps, is that a paper map cannot be updated. On average, 5% of roads are altered in some way every year [2]. So with a paper map which is only 2 years old, there are close to a 1 in 10 chance of being wrongly directed [2]. People continued to use and create map and the art of creating map is called cartography or mapmaking.

Earlier cartography was a specialised job, but the evolution of web made it possible that non-commercialised people can also contribute. Web 2.0 [3] encouraged greater collaboration among internet users and other users, content providers, and enterprises. This movement provided revolutionary new methods of sharing and computing data by crowdsourcing movement similar to wikipedia [4]. In regard to the geographical data the crowd-sourced movement is known as volunteered geographic information(VGI), also name it as Neography in regard to web 2.0 [5], so it is a special case of this web phenomenon and has been applied in many popular websites such as: Wikimapia, OpenStreetMap(OSM), GoogleMap, Flickr [6].

There are companies, who provide map data, which is not collected by crowdsourcing.The two major data providers are NavTeq and TeleAtlas. However, these data are costly, quickly outdated and restricted to specific areas covered by the data acquiring companies. Large companies have invested large sums of money to purchase smaller companies to acquire their data e.g. in 2007, Nokia acquired NavTeq, in 2006, Microsoft acquired the Imagery and Remote Sensing Company Vexcel [7].

A. OpenStreetMap

The focus of this paper is on a collaborative project, which used crowdsourced approach is called OpenStreetMap, created in year 2004, to provide free editable map of the world [8]. Two major driving forces behind the establishment and growth of OpenStreetMap have been restrictions on use or availability of map information across much of the world and the advent of inexpensive portable satellite navigation devices. OpenStreetMap is based on the concept of crowdsourcing, also called wikification of GIS [9], which encourages the volunteers worldwide to contribute through the collection of geographic data. The data of OpenStreetMap is useful because Firstly, the data is completely free with an open content licence. Secondly, it is current as it constantly being updated by the subscribed users who can also add points of interest important to them. Finally, OpenStreetMap has the potential to establish volunteers from all over world including less developed regions, where obtaining data can be difficult for most commercial mapping companies [10].

To compete with OpenStreetMap, Google introduced a tool called Map Maker in 2008, that enabled users to contribute data themselves. This tool was only available for areas with no or little commercial data coverage, e.g. India, Pakistan, Iceland and within a short time, large areas were mapped in this crowd-sourcing manner [7].

Also the data of OpenStreetMap has matured enough. In the last year alone, some of the biggest commercial industries have switched from Google Maps, to OpenStreetMap to power their map apps or websites, because google began charging for heavy use of its data. The growing list of names now includes Foursquare, Wikipedia and other startup websites such as MapBox, Skobbler and CartoDB, and government agencies also selected to use OpenStreetMap data as base map [13].

Even some governmental agencies have donated map data for the enrichment of OpenStreetMap data. In 2007 US census Bureau's imported their Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) map data to OpenStreetMap [14]. Many small and big organisations

978-1-4799-3187-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ITNG.2014.115

donated their map data to OpenStreetMap, one such organization is Automotive Navigation Data(AND), it donated the entire street map of the Netherlands as well as road networks for China and India [15]. When commercialised sector has been motivated towards OpenStreetMap and also the demand of Location Based Service applications is growing, all this ultimately requires the emphasised analysis of geographical information provided by OpenStreetMap.

B. User Data Contribution to OpenstreetMap

The contribution of new data to the project can be accomplished in different ways. The most classical, yet still most common, approach is to record data using a GPS receiver and edit the collected information using one of the various freely available editors such as JOSM, web based Polatch 2 and iD editor released in May 2013. The user provides additional information about the collected data by adding attributes and stores the final results in the OSM database. In addition to this method advent of inexpensive portable satellite navigation devices as well as GPS enabled smartphones made easy for users to contribute data. The accuracy of smartphones based on iOS, Android, Windows Mobile, BlackBerry OS and Symbian mobile operating systems has been already tested and approved [11]. In addition if the accuracy of any smartphone is not as compared to the Professional GPS receiver, still the mapper can map to good accuracy as Microsoft Bing support the project [12] by providing various aerial images as background layer, which allows the OSM members to digitize data such as streets from the images very easily and correctly. The only problem using satellite imagery is that it can be outdated. Anonymous changes to the database are no longer supported; however, any Internet user who registers for the project can add information to the map and change existing data.

C. Map Data Sources

Two main datasets are used in all research papers, a) OpenStreetMap data may be downloaded through the Geofabrik or Cloudmade website in different file formats and b) reference data provided by the any governmental or mapping Agency for Cartography. For both datasets, every street segment was considered and imported to the geodatabase for the main analysis.

The paper has been segmented into different sections, second section discusses about spatial quality assessment work done in chronological order on OpenstreetMap data with drawbacks. The third section discusses about the elements used for evaluation of spatial data quality the last part is conclusions drawn on the basis of review and future directions.

II. Quality Assessment Work

OpenStreetMap is growing and it is expected that in 2013, number of contributor would cross 1 million mark [8]. These contributors produce huge spatial labeled data, with a variety of ways and varying levels of effort. So when labeled data is easy to come by, the focus of the researcher would be on working with the labeled data rather than collecting it [16].

Numerous scientists have contributed their research work on assessment of quality of the OSM in recent years and still open to further research. Li [17] discussed that volunteered geographic information (VGI) is a phenomenon of recent years, offering an alternative mechanism for the acquisition and compilation of geographic information. He discussed the issues involved in the determination of quality for geospatial data, and traced the history of research on VGI quality. But the preliminary research work on assessment of OSM was initiated in 2008 with the debate on need of accuracy and correctness of compiled information within the world of Web 2.0. Researcher began comparing OSM data with ground data available with different governmental agencies. The approaches used for comparing the map data are dicussed by [6], [10], [18] on the basis of completeness, positional accuracy, temporal accuracy, and thematic accuracy. First such assessment of OSM data was conducted for Great Britain, and compared ordnance survey (OS) geodata with OSM data [18]. Haklay [19] in his research work buffered British Ordnance Survey data to determine which percentage of the OSM roads were covered. Also, he created a raster, summed up the lengths of the roads in each cell and compared them. Ather [10] extended this work to the OS Master Map for selected parts of London. He additionally compared completeness of road names. A commonly applied technique for matching different road networks is graph matching [20]. Suitability of OSM data for navigation is investigated in [21]. The analysis of Germany started with comparison of commercial mulinet proprietary map data from TomTom [21], [22] compared with street map data from different proprietary geodata providers. Both research works concluded the nearly similar results that OpenStreetMap data shows a high degree of detail in urban areas but in rural area poor degree of attribute accuracy. The main difference between the approaches is that [21] discussed the geographical discrepancies within Germany, while others simply concluded the completeness of OSM in comparison to other map datasets.

In 2011, Ciepluch [30] discussed that not everyone contributes data of the same quality. The reason for that is a lack of practice and knowledge which can be improved by practice and experience in map making. The research discussed semi automatic approach for the quality assessment. Ludwig [22] described a methodology to compare OSM street data with Navteq for all populated roads in Germany. The methodology was based on a matching between the street objects of OSM and Navteq adopted from [23] which allows for object-wise comparison of geometries and thematic attributes. They also compared and matched the OSM and Navteq data models. They split the OSM objects into segments by intersecting them with buffers around the Navteq objects. To establish correspondence between each Navteq object and its OSM segments they compared Navteq objects and their OSM candidates only by geometry length, category and name (attributes they consider reliable). Finally, they calculated relative quality measures: relative object completeness (percentage of Navteq objects that has a match), relative attribute completeness, difference in speed limits and positional differences (using 5 m, 10 m and 30 m buffers). They found that “oneway” is more often missing from OSM compared to Navteq in uninhabited areas (48.8%) than in inhabited ones (28.1%) and “speed limit” is missing for 80.7 % of objects in inhabited areas and for 92.6% of objects in uninhabited areas. They concluded that the relative completeness of attributes seems to be the higher the more relevant the attribute is for non-motorized usage. Another

researcher [24] statistically analysed the routing process using OpenStreetMap road data of the inner city of Hamburg. They didnt compare different network datasets but examined the relation between completeness of one-way information and driving time estimations.

A similar approach was used in France to analyse OSM data [25]. The results of this research showed the advantage and flexibility, but also concluded the problem of the heterogeneity of the data specifically for France. This is because of different data sources that have been used in OSM and also the differences in the work by the project participants in France. In 2011, the first studies that analysed the quality of OSM outside of Europe were conducted [26]. In this research work the OSM project data had been compared with proprietary data from TomTom (TeleAtlas) and Navteq for the entire state of Florida (USA) and four specific cities within the USA. In comparison to the results for Germany or England, the discrepancies between the rural and urban areas in the USA showed an opposite tendency. In Florida, the rural data was, in parts, even more complete than that of the proprietary datasets in the relative comparison conducted. [27] compared the amount of pedestrian-related data between freely available sources, i.e., OSM and/or TIGER, and proprietary providers, i.e., Tele Atlas Multinet and/or Navteq Discover Cities. They concluded that integration of pedestrian-only segments can lead to a more realistic assessment of service areas when compared to using networks that contain only streets that are passable by cars and that the assessment of VGI data quality, especially OSM data, is an ongoing issue of high importance for successful geo-applications [5], [19].

Other analyses in 2012 [28] assessed the effect of network data integration from different sources on the length of computed shortest paths for pedestrians and concluded that data integration leads to an increased value for users of pedestrian routing applications but that combining OSM and other commercial datasets cannot be considered for implementation due to current licensing issues. Neis [29] assessed the completeness of the OSM street network via a relative comparison (street network length, no. streets without names, no. turn restrictions) between OSM and a commercial dataset provider (TomTom formerly known as Tele Atlas). They noted though that for comparison the TomTom dataset is suitable only for street network data for car-specific navigation. They also evaluated logical consistency using an internal test, whereby topological and thematic consistency is determined. Concerning turn restrictions, researcher discussed that although the number of turn restrictions available in the OSM dataset is continually increasing, it will probably take several more years before OSM achieves the same level as TomTom, based on the current status and development. Apart from England, no studies have been conducted to date over a period of several years and for an entire country [19].

The only research work on assessment of Indian Road Network [31] concluded that, the possible future direction may be to study the economic growth along the highways and analysis of the road network of metropolitan cities and rural regions, which were not included due to the unavailability of actual data such as OpenStreetMap data.

In 2013 many researchers have been aggressively working in the area of assessment of OpenStreetMap, but assessment work for OpenStreetMap data of India has not initiated yet.

III. Elements of Spatial Data Quality

It is clear by the investigation of the literature that OpenStreetMap data is rich and detailed, containing huge amounts of data which is relevant to people on the ground i.e. the people who collected it. The contributors produce huge spatial labelled data, with a variety of ways and varying levels of effort. So when labelled data is easy to come by, the focus of the researcher would be on working with the labelled data rather than collecting it [16] Researchers have been working on quality assessment of OpenStreetMap data and compared it with the map data of their own subcontinent captured by the governmental or private agencies. But for evaluation of OpenStreetMap data different quality parameters are required, Kresse [32] and Oort [34], discussed that aspects of quality are - lineage, positional accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy, temporal information. In addition [34] discussed are Usage/purpose/constraints, Variation in quality, Meta-quality, and Resolution. For OpenStreetMap another aspect that could be checked is verifiability of tags. Most of these are briefly discussed as under but Most of the researchers have worked on two aspects i.e. positional accuracy and completeness.

A. Positional Accuracy

Positional accuracy of OpenStreetMap Data represented as discrepancy between mapped Point and Reference Point. Various researchers have performed the comparison using buffer algorithm of [4] to determine the percentage of line from one dataset that is within a certain distance of the same feature in another dataset of higher accuracy. Another approach used is grid-based approach [2], [19], [21].

In this algorithm, preprocessing step followed [33] for each dataset, to separate road segments sharing the same street name are merged in order to provide a single line string for either street. The junctions within the datasets are then extracted by determining all point coordinates where exactly two distinct line strings cross each other. This approach admittedly ruled out junctions where three or more streets cross but has been preferred for the sake of clarity [33]. The concatenated names of the streets crossing each other served as an identifier for a given junction. These identifiers are then used to select and spatially compare corresponding junctions among the datasets. The deviation of the junction point coordinates from the corresponding points in the defined reference data set has been used as a measure of positional accuracy.

B. Completeness

Completeness is another very important attribute for analysis of Spatial data [22], [4], [6], [10], [18]. it describes the completeness of objects and their attributes. To prepare the dataset for completeness comparison, a grid at a resolution of 1km(normally) is created, then the comparison is performed to find out the difference between OpenStreetMap and proprietary data, to avoid the inclusion of coastline objects and small slivers of grid cells, all incomplete cells with an area less than a square kilometre were eliminated [21], [30]. Ciepluch [30] developed a PHP script to automate the process of grid

generation for an arbitrary geographical area.The development of the script allows to run the script as an optional component in a work-flow of PHP programs used for this research.

C. Temporal accuracy

It describes the date of data observation, type of update such as creation, modification, deletion, unchanges, and validity periods for spatial data records [6], [10], [18]. The quality of this element can be computed by the degree to which the information describes adequately spatial entities. In case of OpenStreetMap data, even the position and attributes of geographic objects are not covered in most of datasets, so not much work has been done on this element. In addition, the information is used by the researchers, who wanted to compare history of the node with current status of node [35]. Further this information could be used for declaring bad users (who intend to input wrong data to OpenStreetMap data)

D. Logical consistency

It describes the trustworthiness of the topological and logical relationships between the dataset segments [27], [28]. There is no indicator to measure it quantitatively; however, visually this element is of a major concern for collaboratively collected data like OpenStreetMap. This information focusses mainly on Polygon date such as roads contain such a problems. This is also called topological inconsistency and can be seen at the road junctions, beginning and ending of the road segments. OpenStreetMap data contains lots of such errors [29], before using map data for navigation purposes, it needs to preprocessed. There some tools such as JOSM, OpenJUmp and webbased OSM Inspector and keepright, which used to correct the topological inconsistency.

IV. Conclusion & Future Scope

This review paper concludes that OpenstreetMap is generating huge dataset with the help of non-commercialised users of varying level of mapping experience, due to this it contains some anomalies. So the assessment becomes vital to give maturity to OpenStreetMap data. But as OpenStreetMap is gaining popularity, the number of absolute and relative errors are falling. Findings of Neis [29] on turn restriction, useful for street navigation, that OpenStreetMap would take approximately five years to be at par with proprietary data set. But with the in depth study of the assessment work, it can be concluded that Openstreetmap is quite developed and mature as compared to geodata from commercial vendors. Many organisations such as Wikipedia & Foursquare have recognized Openstreetmap and using its data commercially. But till now not much has been done for contribution, this may be due to unawareness of OpenStreetMap. The future scope of this assessment are as follows:-

- The advent of crowdsourcing has created a variety of new opportunities for improving upon traditional methods of data collection and annotation. This in turn has created intriguing new opportunities for data-driven machine learning. So Machine learning has been identified as area which can be used for handling the OpenStreetMap data [16]. So by combining these quality control measures and machine learning approaches, a model would be devised that can check for the user anomalies in data by users.
- Finding and eliminating data discrepancy and thus increasing spatial accuracy and consistency of OpenStreetMap datasets.
- Statistically analyse routing and navigability of OpenStreetMap road network of India by comparing this data with governmental agency data.

References

[1] Merriam, D., Kansas 19th century geologic map, Kansas Academy of Science, Transactions 99 pp 95–114.

[2] TomTom, portable GPS car navigation systems. http://www.tomtom.com, Accessed on 12th Nov 2013.

[3] O' Reilly, T., What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, O'Reilly Media: Cambridge, MA, USA, 2008. http://oreilly.com/web2/archive/what-is-web-20.html, Accessed on 19th May 2013.

[4] Haklay, M., Alex S., and Chris P., Web mapping 2.0: The neogeography of the GeoWeb, in *Geography Compass 2.6*,2008, 2011-2039.

[5] Goodchild, M.F., Spatial Accuracy 2.0. in *Proceeding of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, China,* , 2008, 25–27.

[6] Kounadi, O., Assessing the quality of OpenStreetMap data *MSc. Dissertation, University College of London Department of Civil, Environmental And Geomatic Engineering*, August 2009.

[7] Schmidt, M., Weiser, P., Web Mapping Services: Development and Trends, in Springer Lecture Notes On Online Maps with APIs and WebServices, 2012, 13–21.

[8] http://en.wikipedia.org/wiki/OpenStreetMap Accessed on 19th May 2013.

[9] Sui, D.Z., The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS, *Computers Environment and Urban Systems*, 2008, 32, 1–5.

[10] Ather, A., A Quality Analysis of OpenStreetMap Data M.E. Thesis, University College London, London, UK, May 2009.

[11] Golicher, D., Accuracy of an android cell phone GPS in the UK. URl http://duncanjg.wordpress.com/2011/05/08/ accuracy-of-an-android-cell-phone-gps-in-the-uk/, Accessed on 19th May 2013.

[12] Bing; OSM Wiki. URL http://wiki.openstreetmap.org/wiki/Bing Accessed on 12th Nov 2013.

[13] http://idealab.talkingpointsmemo.com/2013/ 01/year-of-the-map.php?m=1, Accessed on 12th Nov 2013.

[14] http://wiki.openstreetmap.org/wiki/TIGER, Accessed on 12th Nov 2013.

[15] http://wiki.openstreetmap.org,Accessed on 12th Nov 2013.

[16] Lease, M., On Quality Control and Machine Learning in Crowdsourcing in *Proceeding of Human Computation, AAAI Workshop*, August 2011, 97–102.

[17] Li, L., Goodchild, M. F., Assuring the quality of volunteered geographic information in *Spatial Statistics*, Volume 1, May 2012, 110–120, ISSN 2211-6753, 10.1016/j.spasta.2012.03.002.

[18] Hakley, M., How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England in *Environment and Planning B: Planning and Design*, August 2008.

[19] Haklay, M., How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets in *Environment and Planning B: Planning and Design*, 2010, 37(4), 682–703.

[20] Zhang, M. and Meng, I., An iterative road-matching approach for the integration of postal data in *Computer, Environment and Urban Systems*, 2007, 31, 597–615.

[21] Zielstra, D., Zipf, A., A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany in *Proceedings of 13th AGILE International Conference on Geographic Information Science, Guimar Portugal*, May 2010, 10–14.

[22] Ludwig, I., Voss, A., Krause-Traudes, M., A Comparison of the Street Networks of Navteq and OSM in Germany in *Advancing Geoinformation Scienced of a Changing World Lecture Notes in Geoinformation and Cartography 2011, published by Springer*, 65–84.

[23] Walter, V., Fritsch, D., Matching spatial data sets: a statistical approach in *International Journal of Geographical Information Science*, 13.5 (1999), 445–473.

[24] Fessele, M., Poplin, A., Statistical analysis of routing processes using OpenStreetMap road data of Hamburg with different completeness of information about one-way streets in *Proceedings of GeoValue'2010*, 2010, 87–92.

[25] Girres, J.F., Touya, G., Quality assessment of the French OpenStreetMap dataset in *Transaction in GIS*, 2010, 14, 435–459.

[26] Zielstra, D., Hochmair, H.H., Digital street data: Free versus proprietary in *GIM International*, 2011, 25, 29–33.

[27] Zielstra, D., Hochmair, H.H., A comparative study of pedestrian accessibility to transit stations using free and proprietary network data in *Transportation Research Record: Journal of the Transportation Research Board*, 2011, 2217, 145–152.

[28] Zielstra, D., Hochmair, H.H., Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data in *Proceedings of Transportation Research Board - 91st Annual Meeting, Washington, DC, USA*, 22–26, January 2012.

[29] Neis, P., Zielstra, D., Zipf, A., The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011 in *Future Internet*, 2012, 4(1), 1-21, doi:10.3390/fi4010001.

[30] Ciepluch, B., Mooney, P., Jacob, R., Winstanley, A., Sketches of Generic Framework for Quality Assessmenft of Volunteered Geographical Data in *IEEE Geoscience and Remote Sensing Society (GRSS)*, 2011 , 1–5.

[31] Mukherjee, S., Statistical analysis of the road network of India in *Indian Academy of Sciences* , Vol. 79, No. 3, September 2012, 483–491

[32] Kresse, W and Fadaie, K., Standardization of Geographic Information. Springer, 2004.

[33] Helbich. M., Neis. P., Amelunxen. C., Zipf A.,Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata, Herbert Wichmann Verlag, Berlin, pp 24–33

[34] Oort, P., Automatically and Accurately matching objects in geospatial datasets PhD Thesis, Wageningen University, Wageningen, Germany,

[35] Neis P, Zipf A. Analyzing the Contributor Activity of a Volunteered Geographic Information Project The Case of OpenStreetMap. ISPRS International Journal of Geo-Information. 2012; 1(2):146-165.