Human Computation 1 (2014) 569–592
<Copyright and creative commons notice>
ISSN: 2330-8001, DOI: <Insert DOI here>

Assessing the Topological Consistency of Crowdsourced OpenStreetMap Data

SUKHJIT SINGH SEHRA, ASSISTANT PROFESSOR, GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA, INDIA.

JAITEG SINGH , ASSOCIATE PROFESSOR, CHITKARA INSTITUTE OF ENGINEERING & TECHNOLOGY, PUNJAB, INDIA.

HARDEEP SINGH RAI, PROFESSOR, GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA, INDIA.

ABSTRACT

OpenStreetMap is world leader in collecting map data contributed by users, called crowdsourcing. But we have little knowledge about the people who collect it, their skills, knowledge or patterns of data collection. Also OpenStreetMap has loose coordination and no top-down quality assurance processes. This makes map data more vulnerable to errors and incomplete. To make the map data navigable, it must not have errors. The current proposal has been conducted to identify errors OpenStreetMap data. Small area of Punjab has been taken as test data for finding inconsistencies. It has been concluded that data contains lots of such errors and is not mature enough.

1. INTRODUCTION

Map has been an integral part of mankind for long time back. From cave paintings to into the 21st century, people have created and used maps as the essential tools to help them define, explain, and navigate their ways through the world. Just a few years ago, mapping was primarily used in the car for navigation. But now it enables everything from lucrative location-based services to game-changing autonomous driving (Newcomb, 2014).

The modern cartography has been strongly integrated with Geographical Information System. Earlier cartography was a specialised job, but the evolution of web made it possible that non-commercialised people can also contribute. The Web started with web 1.0 or “read only” web, but Tim (O‘Reilly, 2005) discussed the concept of Web 2.0 or “read-write” web. The client side

technologies such as Ajax and javascript framework are used for Web 2.0 development (O'Reilly, 2005). Web 2.0 encourages greater collaboration among internet users and other users, content providers, and enterprises (Hudson-Smith et al., 2009). This movement provided revolutionary new methods of sharing and computing data (Hudson-Smith et al., 2009; Turner, 2006; Goodchild, 2009; Haklay et al., 2008) by crowdsourcing movement similar to Wikipedia (Howe, 2006). In regard to the geographical data the crowd-sourced movement is known as VGI (volunteered geographic information), also name it as Neography in regard to web 2.0 (Goodchild, 2009; Haklay et al., 2008) others call it collaborative mapping (Fischer, 2008), 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 (Kounadi, 2009).

On the other side, with the advent of inexpensive portable satellite navigation devices as well as GPS enabled smartphones made easy for users to contribute to crowd sourced OpenStreetMap. The accuracy of smartphones based on iOS, Android, Windows Mobile, BlackBerry OS and Symbian mobile operating systems has been already tested and approved (Golicher, 2013). As of December 2012, the worldwide smartphone market had Android as its top operating system, counting on 68.3% of market share, followed by iOS with 18.8% and Blackberry with 4.7%. Forecasts show that by 2016 Windows Phone will overtake the third place, with 11.4% market share, while Android and iOS will remain in its previous positions, with 63.8% and 19.1% respectively. All these mobile phone are not only used for calling but comes with tons of applications. These applications are used for Navigation and many Location Based Service applications which enable Geotagging. According to one survey, Indians become the better users of Smartphone than the American users as per a survey report published by Google along with IPSOS and MMA (Mobile Marketing Association). The survey declares that, for many Indians Smart phones are not just a medium to call or message but a smart device to meet their multiple needs such navigation, social networking, share market etc.

As geography is big business, the map data providers such as Google, NavTeq, Mapmyindia, TomTom and TeleAtlas are in race to acquire the world map data. Becuase, they see that non-geospatial applications are integrated with spatial data. But their is huge cost invovled in map production, e.g. Navteq is spending approx $400m amount in creation and distribution of maps and has nearly 3400 work force. All this cost would generated from the users only. In addition big 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 (Schmitz et al., 2008) and in 2014 acquire waze for $1.5bn, becuase it was producing huge user generated content.

In year 2004, to provide free editable map of the world, the a collaborative project which used crowdsourced approach is called OpenStreetMap. 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. OpenStreetMap has been used during earthquake situation in Haiti named as Haiti Crisis Map. 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 (Ather, 2009)

But their are map quality issues related to OpenStreetMap, which are given as: -

– No moderation and a minimal imposed data model on data being uploaded.
– Each individual has his own intents, purposes, motivations, knowledge, skill level and tools.
– Digitisation errors.
– Inaccurate mapping devices.
– Uneven geographical spread of contributor efforts.
– No proper interpretation of aerial photographs.

Because of above reasons OpenStreetMap data of India has been tested. In addition to it, another reason is that the most of the Indian map data i.e. basic roads network and city name, is donated by Automotive Navigation Data (AND) (OpenStreetMap.org, 2013). It is possible that the errors may have come from AND.

In this paper, the focus is on assessment of OpenStreetMap data for topological inconsistencies in selected city of Punjab, India. These inconsistencies must be removed to ensure map data integrity. The structure of paper is organised in different sections, the next section discuss about OpenStreetMap. Section 3 addresses topological inconsistencies and section 4 elaborate the methods used for the finding the errors, section 5 describe the results and in last conclusion.

1.1. Crowdsourced OpenStreetMap Data

OpenStreetMap project started in 2004 by Steve Coast with an idea to provide free editable map data of world. 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, which encourages the volunteers worldwide to contribute through the collection of geographic data.

OpenStreetMap has three main components are: -

– Node.
– Way.
– Relation.

Relation data structure, consists other to components i.e. node and ways to represent relationship between components. On the 4th July, 2007, Automotive Navigation Data donated the entire road networks for India to OpenStreetMap.

1.1.1. Why OpenStreetMap

When proprietary map providers spending billion dollars in creation and distribution of maps. The Crowdsourced maps such as OpenStreetMap produces huge spatial data, with less effort and minimum cost (Goodchild, 2007). OpenStreetMap produces plentiful labelled data at no cost. The researcher would work on the devising method to use the data rather than collecting the data. As many prior studies in algorithmic learning considered smaller training sizes due to traditional costs of labelling. Now there may be greater benefits to such learning techniques. Also Crowdsourcing

Table 1. ***Comparison of OpenStreetMap & Proprietary Map Providers***

OpenStreetMap	Proprietary Map Provider
Rendered maps or raw data	Rendered maps
No Developer Key	Developer Key
Multiple APIs	Single API
Several providers	Single provider

offers coverage of many places around the world where there is no or very small commercial coverage by commercial vendors. e.g. even the slum areas are also mapped.

In addition to this, when other map providers are trying to monetise the map data availability. The OpenStreetMap respects communities, their work and privacy. The ownership of the content will always be with the user. It uses Creative Commons Attribution-ShareAlike 2.0 license, which allows the user to download data for the applications and analysis. Table 1 discusses the few more benefits of using OpenStreetMap data as compared to Proprietary map data providers.

1.1.2. OpenStreetMap Indian Users Statistics

OpenStreetMap has strong community of around 1.9 million contributors as shown in fig 1. For India, the number of registered users are 2877. The total no of 7,561,749 nodes, 433,747 ways and 6,094 relations exist in the map data. There two type of edits, as shown in the fig 2, v1 means objects created but are never modified (version 1) and last edit are the objects which modified at least one more time after creation.

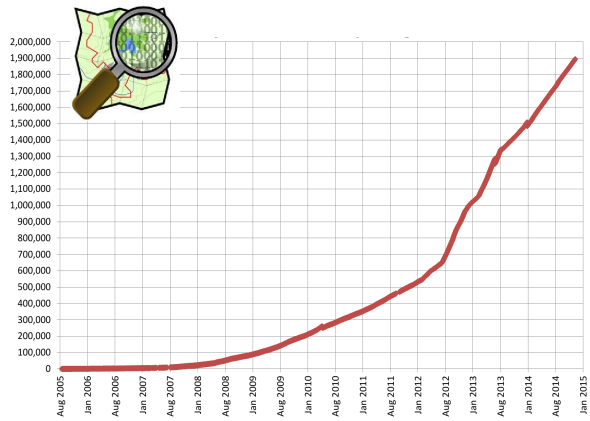
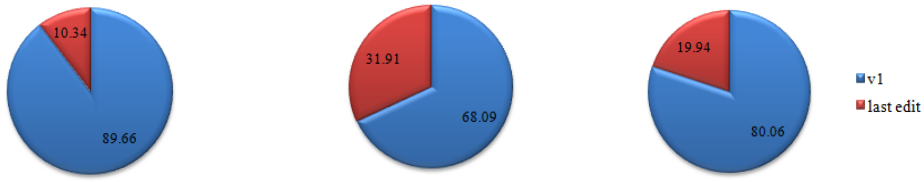


Figure 1. ***The statistics of registered users (OpenStreetMap.org, 2014b)***



(a) Nodes (b) Ways (c) Relations

Figure 2. OpenStreetMap Data of India (OpenStreetMap.org, 2014a)

1.1.3. Map Data Contribution Methods

The most common approach is recording map data using a Global Positioning System (GPS) receiver and edit the recorded information using one of the various freely available editors such as JOSM or Merkaator (Sehra et al., 2013). The users provide additional information about the collected data by adding attributes and store the final results in the OpenStreetMap database. Users do not require any specialised GPS receiver for mapping, as it has been tested that smartphones (Golicher, 2013) can be considered as device for mapping. The accuracy of GPS has been checked and found satisfactory. Microsoft Bing supports the project by providing various aerial images as background layer. The Problem with this layer is it can be outdated and no uniform coverage is provided. Other web based tools are Potlatch2 and iD editor as shown in fig 3, but only registered users can upload the changes to OpenStreetMap.

1.1.4. OpenStreetMap Data Repositories

OpenStreetMap license terms allow the user to download and distribute the data. This data can be downloaded from the following repositories under Creative Commons Attribution-ShareAlike 2.0 license terms:-

1. Overpass API :- Using this API, user can download a bounding box from a mirror of the OpenStreetMap database.
2. Planet OSM :- It provides regularly-updated copies of the complete OpenStreetMap database.
3. Geofabrik Downloads :- It provide regularly-updated extracts of continents, countries, and selected cities
4. Metro Extracts :- It provides extracts for major world cities and their surrounding areas can be downloaded.

The Indian map data for the proposed work is downloaded from geofabrik repository in shapefile (.shp) format. Further the test data of Ludhiana, Punjab is extracted after applying the clipping function available in Quantum Geographic Information System (QGIS) (Quantum, 2011). For clipping the test data, City, State administrative boundaries (polygons) are used as per Indian census 2011 data.

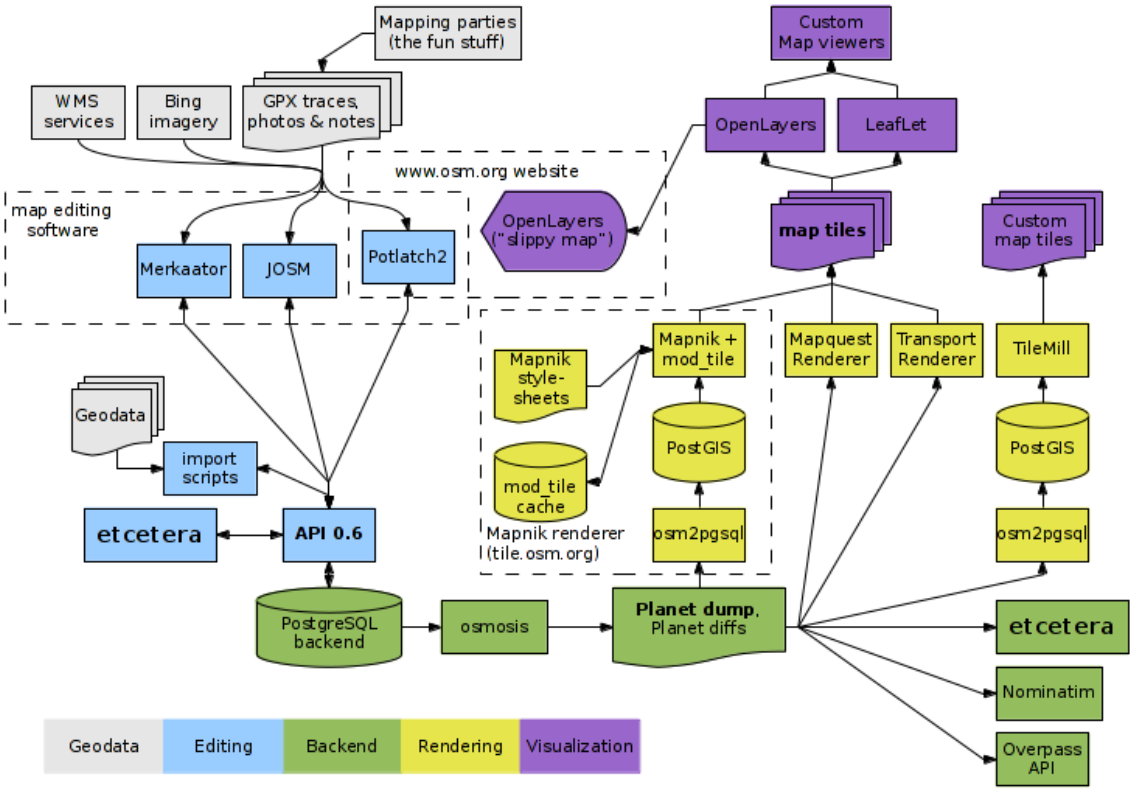


Figure 3. ***Methods of Uploading and Retrieval of OpenStreetMap Data (OpenStreetMap, 2014)***

2. ELEMENTS OF OPENSTREETMAP DATA ASSESSMENT

There are specific elements defined and used by the researchers for the assessment of OpenStreetMap data (Girres and Touya, 2010; Statistics Canada, 2013), which given as follows : -

1 Lineage - It describes the history of the spatial data, including descriptions of the source material from which the data were derived, and the methods of derivation. It also contains the dates of the source material, and all transformations involved in producing the final digital files or map products.
2 Positional accuracy - It is the spatial and geometric accuracy of the data. Positional accuracy of map data is represented as discrepancy between mapped point and reference point. Various methods are used for assessing the positional accuracy. Metrics used for spatial accuracy are root mean square error, standard deviation or confidence interval.
3 Attribute accuracy - It refers to the accuracy of the quantitative and qualitative information attached to each feature (such as population for a population centre, street name, census subdivision name and code).

4 Topological consistency - It describes the trustworthiness of the topological and logical relationships between the dataset segments (Joksic and Bajat, 2004). These relations typically involve spatial data inconsistencies such as incorrect line intersections, polygons not properly closed, duplicate lines or boundaries, or gaps in lines. It deals with the structural integrity of a given data set based on formal framework for modelling of spatial data and relationships among objects. These types of errors must be corrected to avoid incomplete features and to ensure the data integrity.
5 Completeness - It describes the completeness of objects and their attributes as compared to reference map data. It illustrates the commission of an entities in spatial database related to their number in real world. It is an aspect of fitness-for-use. Fitness-for-use term is referred to decision making for accessing whether database meets the requirements of a particular application.

All these elements depend upon the user mapping experience & tools used along with other support system also need reference data set for assessment as done by most of the researchers. But this paper focuses only on the topological inconsistencies, which can be identified within the map dataset.

3. TOPOLOGICAL CONSISTENCY

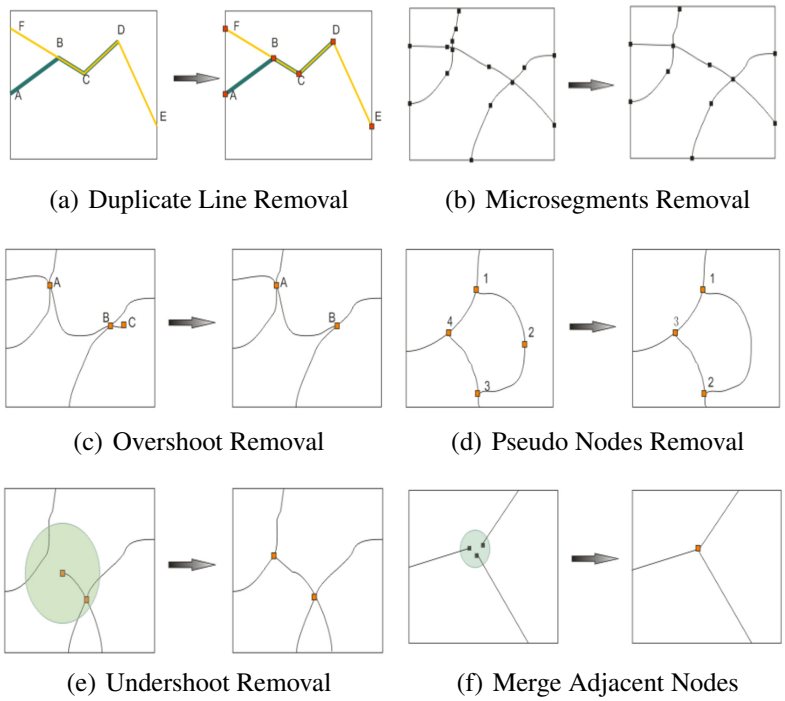
Topological consistency describes the trustworthiness of the topological and logical relationships between the dataset segments (Joksic and Bajat, 2004). These relations typically involve spatial data inconsistencies such as incorrect line intersections, polygons not properly closed, duplicate lines or boundaries, or gaps in lines. It deals with the structural integrity of a given data set based on formal framework for modelling of spatial data and relationships among objects. These types of errors must be corrected to avoid incomplete features and to ensure the data integrity.

Topological errors, which occur during digitising and data exploration processes, are also known as semantic errors (Ubeda and Egenhofer, 1997). Topological errors exist due to violation of predefined topology rules. The most common topology errors in map data are shown in fig 4 .

– Duplicate Lines
– Overshoots
– Undershoots
– Micro Segments
– Pseudo Nodes
– Merge Adjacent Endpoints
– Self Intersection

4. METHODS OF TOPOLOGICAL ERROR DETECTION

If map is not topologically correct then that map may not be useful in terms of navigation. OpenStreetMap data is not navigable in its original form (Neis et al., 2011). To check test map data for topological errors, two map data file formats are utilised, one is XML and other Shapefile. The files are utilised as per requirement of the algorithm. The algorithm used for the detection of the topological errors is a plugin in OpenJump (Steiniger and Hunter, 2012). Another GIS software



(a) Duplicate Line Removal (b) Microsegments Removal

(c) Overshoot Removal (d) Pseudo Nodes Removal

(e) Undershoot Removal (f) Merge Adjacent Nodes

Figure 4. Various Topological Errors (Chang, 2008).

used in QGIS, which provides integration with many other packages e.g. POSTGIS, GRASS, GeoServer and MapServer.

4.1. Micro-segment Vertices Detection

It is the first step before detection and cleaning process starts, because this algorithm uses a distance threshold to decide if two features should be snapped together or not. The presence of micro-segments less than threshold are undesirable. The algorithm used to detect microsegments chooses a point to be removed from the geometry if deformation is minimal (Michaël Michaud, 2014) e.g. micro segments have two vertices as shown in fig 5. The problem with this algorithm is that it does not check if the vertex removed by the algorithm is also used by another geometry. This removal may break topology consistency. The fig 6 shows the used test data of Ludhiana city and detected micro-segments in it.

4.2. Network Topology Detection & Correction

This algorithm processes dataset representing a network and detect topological errors like node mismatches, undershoot and overshoots already discussed above as shown in fig 4. It identify

(a) Micro Segments (b) After Cleaning of Micro Segments

Figure 5. Cleaning Micro-segments

(a) Ludhiana City Data (b) Detected Mirco-segments

Figure 6. Micro-segment Detection

nodes in network based on distance threshold between a node of the layer to process and a feature of the reference layer (Michaël Michaud, 2014).

4.2.1. Detection of Nodes

All those nodes would be detected, where 1st feature is at less than 3 meters from a reference node, 2nd and 3rd features are at 1 meter from segment, at 2 meters from another one and at more than 3 meter from any vertex 4th feature is at more than 3 meter from the feature is at less than 3 meter from a reference.

4.2.2. Correction of Nodes

1st and last features will be properly snapped to reference, 2nd and 3rd features will be snapped, but a vertex needed to be inserted into the reference layer.

Another option available with this algorithm is to deal with degree 3+ nodes of the processed layer. It would be useful to decrease the tolerance parameter for these nodes, because if they are not snapped on the reference layer, there would be higher probability that it should not be snapped. The results obtained after processing the test data set are shown in fig 7.

Figure 7. Results from Network Cleaning Algorithm

4.3. Coverage Gaps Detection

This algorithm is used for detection of defects like overlaps or gaps as shown in fig 8. Nearly 95% or more can be detected with this algorithm (Michaël Michaud, 2014).

It matched segments and these are considered matched, when: -

- Segments are not topologically equals (AB is not equal to CD and AB is not equal to DC)
- Minimum distance between segments is less than the user-defined tolerance,
- The angle between both segment is less than the user-defined tolerance
- The orthogonal projection of each segment on the matched one is non null

4.4. Bad features in OpenStreetMap

Bad features in OpenStreetMap data are detected using Geometry Metric validation algorithm in Openjump and QGIS. it is tested based on the following rules enforced on the data: -

- Basic topology
- Disallow repeated consecutive Points
- polygon orientation
- minimum segment length (.001 is taken here)
- minimum angle in degree(1.0 is taken)
- minimum polygon area (.01)
- Geometries are simple i.e. do not self intersect

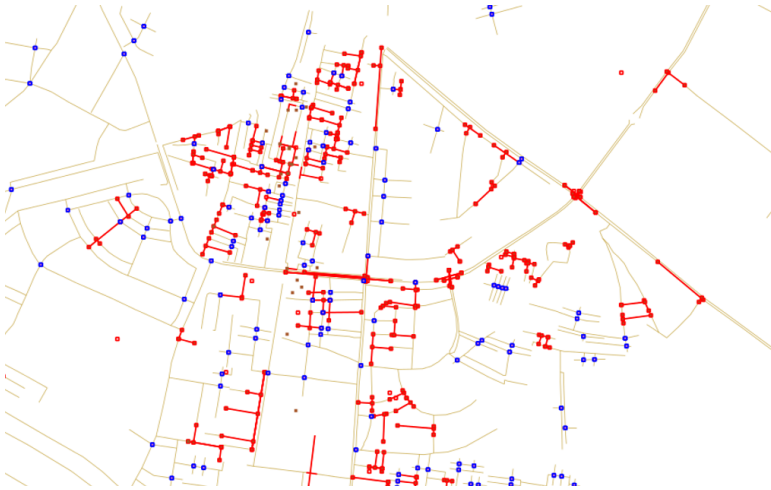


Figure 8. Outcome of Coverage Gaps Algorithm

5. RESULTS

Test data of Ludhiana city was extracted after the clipping the same from the OpenStreetMap data of India using QGIS clipping function. The total feature of this data set are 3210 map segments and 20550 nodes. When this data is processed using various algorithms. It has been found that 37 micro-segments are detected. Network cleaning algorithm found 309 mismatch features and 618 points, Coverage gap algorithm with distance tolerance 1.0 and angle tolerance 22.5 without fence, found gap segments of 12 features and 24 points, overlaps of 959 features and 8607 points. Further it has been found that the map data has bad features 32 and 05 points.

In QGIS topology checker, it identified total of 5497 errors in which 5034 are dangling errors, one duplicate node, 24 mulitpart-geometry and rest are pseudo nodes. The geometry errors found are 90.

6. CONCLUSION

OpenStreetMap is world leader in collecting map data con- tributed by users. But users are of different backgrounds and have varying level of mapping experience. Also it uses loose coordination and no top-down quality assurance processes. This makes map data more vulnerable to errors and incomplete. But OpenStreetMap, showing the increasing relevance of open and crowd-sourced data also for commercial purposes. There are about 60+ companies using its data commercially. So it require much emphasis on map qualtiy assessment.

As shown using statistics that the data contains most of nodes with version v1 only that means that mostly, have come from AND. As their version is v1, these values are never modified by the user. This shows the less activity in this area, and is more prone to errors. During this work, city data of Ludhiana, Punjab was taken and processed using different algorithms. It has been found that map

data contains lots of topological errors. So the map data for the city, is not mature enough to be used for navigational purposes. The future work will focus on removing topological errors from the OpenStreetMap data of Punjab.

7. REFERENCES

Ather, A. (2009). A Quality Analysis of OpenStreetMap. (2009). ftp://ftp.cits.rncan.gc.ca/pub/cartonat/Reference/VGI/Dissertation-OpenStreepMap-Quality-Aather-2009.pdf

Chang, K. T. (2008). *Introduction to GIS*. McGraw-Hill.

Fischer, F. (2008). Collaborative mapping–How wikinomics is manifest in the geo-information economy. *Geoinformtics* 11, 2 (2008), 28–31. http://www.gisaci.upol.cz/filesftp/Geoinformatics-02-2008.pdf

Girres, J and Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transaction in GIS* 14 (2010), 435–459. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2010.01203.x/abstract

Golicher, D. (2013). Accuracy of an Android cell phone GPS in the UK. (2013). http://duncanjg.wordpress.com/2011/05/08/accuracy-of-an-android-cell-phone-gps-in-the-uk/

Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* 3, 2 (2009), 82–96. http://www.tandfonline.com/doi/abs/10.1080/17489720902950374

Goodchild, M. F. (2007). in the World of Web 2.0. *International Journal* 2 (2007), 24?32. http://www.geoinformatics.cn/wp-content/uploads/citizensasvoluntarysensors.pdf

Haklay, M, Singleton, A, and Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass* 2, 6 (2008), 2011–2039. http://onlinelibrary.wiley.com/doi/10.1111/j.1749-8198.2008.00167.x/full

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4. http://sistemas-humano-computacionais.wikidot.com/local--files/capitulo:redes-sociais/Howe_The_Rise_of_Crowdsourcing.pdf

Hudson-Smith, A, Crooks, A, Gibin, M, Milton, R, and Batty, M. (2009). NeoGeography and Web 2.0: concepts, tools and applications. *Journal of Location Based Services* 3, 2 (2009), 118–145. http://www.tandfonline.com/doi/abs/10.1080/17489720902950366

Joksic, D and Bajat, B. (2004). Elements of spatial data quality as information technology support for sustainable development planning. 11 (2004), 77–83. http://www.doiserbia.nb.rs/Article.aspx?ID=1450-569X0411077J

Kounadi, O. (2009). *Assessing the quality of OpenStreetMap data*. MSc Dissertation. University College of London, Department of Civil, Environmental And Geomatic Engineering. ftp://ftp.cits.nrcan.gc.ca/pub/cartonat/Reference/VGI/Rania_OSM_dissertation.pdf

Michaël Michaud, . (2014). Personal Website : Topology Library. (2014). http://geo.michaelm.free.fr

Neis, P, Zielstra, D, and Zipf, A. (2011). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet* 4, 1 (2011), 1–21. http://www.mdpi.com/1999-5903/4/1/1

Newcomb, D. (2014). Telenav's switch to Open Source Mapping allows for more frequent software updates and traffic reports. (2014). http://www.pcmag.com/article2/0,2817,2458392,00.asp

OpenStreetMap, . (2014). Component Overview of OpenStreetMap. (2014). http://wiki.openstreetmap.org/wiki/Component_overview

OpenStreetMap.org, . (2013). Automotive Navigation Data. (2013). http://wiki.openstreetmap.org/wiki/AND_Data

OpenStreetMap.org, . (2014)a. OpenStreetMap ODbL acceptance and user ranks for the region of india. (2014). http://odbl.de/india.html

OpenStreetMap.org, . (2014)b. OpenStreetMap Statistics. (2014). http://www.openstreetmap.org/stats/data_stats.html

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. In *O' Reilly Media:*. Cambridge, USA. http://oreilly.com/web2/archive/what-is-web-20.html

Quantum, G. (2011). Development Team (QGIS). 2011. Quantum GIS Geographic Information System, Version 1.7. 0. Open Source Geospatial Foundation Project. Vancouver. *British Columbia, Canada. qgis. osgeo. org* (2011).

Schmitz, S, Zipf, A, and Neis, P. (2008). New Applications based on collaborative geodata?the case of Routing. In *XXVIII INCA international congress on collaborative mapping and space technology, Gandhinagar, India* (2008). http://koenigstuhl.geog.uni-heidelberg.de/publications/bonn/conference/cmap2008.cartography-bonn.subm.pdf

Sehra, S. S, Singh, J, and Rai, H. S. (2013). Assessment of OpenStreetMap Data - A Review. *International Journal of Computer Applications* 76, 16 (2013), 17–20. DOI:http://dx.doi.org/10.5120/13331-0888

Statistics Canada, . (2013). Spatial data quality elements. (2013). http://www.statcan.gc.ca/pub/92-195-x/2011001/other-autre/qua-eng.htm

Steiniger, S and Hunter, A. J. S. (2012). OpenJUMP HoRAE–A free GIS and toolbox for home-range analysis. *Wildlife Society Bulletin* 36, 3 (2012), 600–608. DOI:http://dx.doi.org/10.1002/wsb.168

Turner, A. (2006). Introduction to Neogeography. *O'Reilly Media, MA, USA* (2006). http://shop.oreilly.com/product/9780596529956.do

Ubeda, T and Egenhofer, M. (1997). Topological error correcting in GIS. In *Advances in Spatial Databases*, Michel Scholl and Agnès Voisard (Eds.). Lecture Notes in Computer Science, Vol. 1262. Springer Berlin Heidelberg, 281–297. http://dx.doi.org/10.1007/3-540-63238-7_35