# QUALITY ANALYSIS OF OPEN STREET MAP DATA

WANG Ming [a, *], LI Qingquan [b], HU Qingwu [a], ZHOU Meng [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University,
129 Luoyu Road, Wuhan 430079, China - (m_wang, huqw, zhoumeng)@whu.edu.cn
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sengsing,
Wuhan University, 129 Luoyu Road, Wuhan 430079, China - qqli@whu.edu.cn

**KEY WORDS:** Crowd sourcing geographic data, OSM, road network, quality elements, quality assessment

**ABSTRACT:**

Crowd sourcing geographic data is an opensource geographic data which is contributed by lots of non-professionals and provided to the public. The typical crowd sourcing geographic data contains GPS track data like OpenStreetMap, collaborative map data like Wikimapia, social websites like Twitter and Facebook, POI signed by Jiepang user and so on. These data will provide canonical geographic information for pubic after treatment. As compared with conventional geographic data collection and update method, the crowd sourcing geographic data from the non-professional has characteristics or advantages of large data volume, high currency, abundance information and low cost and becomes a research hotspot of international geographic information science in the recent years. Large volume crowd sourcing geographic data with high currency provides a new solution for geospatial database updating while it need to solve the quality problem of crowd sourcing geographic data obtained from the non-professionals. In this paper, a quality analysis model for OpenStreetMap crowd sourcing geographic data is proposed. Firstly, a quality analysis framework is designed based on data characteristic analysis of OSM data. Secondly, a quality assessment model for OSM data by three different quality elements: completeness, thematic accuracy and positional accuracy is presented. Finally, take the OSM data of Wuhan for instance, the paper analyses and assesses the quality of OSM data with 2011 version of navigation map for reference. The result shows that the high-level roads and urban traffic network of OSM data has a high positional accuracy and completeness so that these OSM data can be used for updating of urban road network database.

## 1. INTRODUCTION

Crowd sourcing geographic data is an opensource geographic data which is contributed by lots of non-professionals and provided to the public (Giles, 2006; Heipke, 2010; Howe, 2008). As compared with conventional geographic data collection and update method, the crowd sourcing geographic data from the non-professional has characteristics or advantages of large data volume, high currency, abundance information and low cost and becomes a research hotspot of international geographic information science in the recent years (Goodchild, 2008; Goodchild, 2009; Heipke, 2010).

While discussing the processing and application method, the primary problem is to analyse the quality of crowd sourcing geographic data (Goodchild, 2007). As there have been lots of problems such as information redundancy, devoid of information in corwd sourcing geographic data, it is necessary to built the quality analysis model, assessment method or system of corwd sourcing geographic data before applying them. At present many geographic information experts abroad has done some research on OSM data quality and built preliminary OSM quality quality assessment model (Ather, 2009; Jan,etc., 2011). To solve the quality problem of crowd sourcing geographic data obtained from the non-professionals, a quality analysis model for OpenStreetMap crowd sourcing geographic data is proposed in this paper. Firstly, a quality analysis framework is designed based on data characteristic analysis of OSM data. Secondly, a quality assessment model for OSM data by three different quality elements: completeness, thematic accuracy and positional accuracy is presented. Finally, take the OSM data of Wuhan for instance, the paper analyses and assesses the quality of OSM data with 2011 version of navigation map for reference.

## 2. QUALITY ANALYSIS OF OSM

There are mainly three factors to influence the quality of OSM. Firstly, data collection and mapping are completed by the non-professionals who are lack enough geographic knowledge and effective training, there exists some artificial error. Secondly, the collected data may be from different data sourcing with different precisions. Thirdly, the data collected by different GPS from different volunteers may have different precisions. From the above, it is useless to assess the quality of OSM with conventional methods which is valid in normal map quality assessment. The simple and valid method is to compare OSM with reference data to analyse and assess the quality of OSM based on quality assessment model built with proper quality elements.

### 2.1 Quality Elements of OSM Data

On the basis of uniformed spatial reference, matched object classification, and consistent attribute description, the quality analysis of OSM data should be carried out with suitable quality elements to establish quality assessment model, by which the quality parameters are calculated. And then through a

* Wang Ming, Ph.D candidate, majors in data analysis and mining of crowed sourcing check-in data.

comprehensive analysis of these parameters, the completeness, accuracy, and availability of the data are evaluated.

The paper assesses the quality of OSM with three quality elements: data completeness, attribute accuracy and position accuracy. Data completeness is consist of length completeness and name completeness, attribute accuracy includes two aspects: name accuracy and tyep accuracy, position accuracy is used to assess the geometric accuracy of OSM roads.

### 2.2 Calculation Model of Quality Elements

**2.2.1 Data Completeness**: Data completeness includes length completeness and name completeness. Length completeness reflects the geometric quality and data coverage, name completeness means the completeness of name attribute.

Length completeness is an assessment of how many roads are expected to be found in the database but are missing as well as an assessment of excess data that should not be included. It can be calculated as the percentage of the length of the tested dataset $L_{OSM}$ to the length of the reference dataset $L_R$. Shown as formula (1):

$$Q_L = L_{OSM} / L_R \quad (1)$$

Road name is an important attribute information of road, the name completeness of OSM road reflects the attribute quality and usibility of OSM. Name completeness includes name attribute completeness and name length completeness. Name attribute completeness $Q_{S_N}$ can be calculated as the percentage of the amount of named raod in tested dataset $S_{OSM}^N$ to the amount of name road in the reference dataset $S_R^N$. Shown as formula (2):

$$Q_{S_N} = S_{OSM}^N / S_R^N \quad (2)$$

Name length completeness is calculated with the metric of road length. Shown as formula (3):

$$Q_{S_L} = S_{OSM}^L / S_R^L \quad (3)$$

Where $S_{OSM}^L$ is the length of named road in tested dataset, $S_R^L$ is the length of named road in reference dataset.

**2.2.2 Attribute Accuracy**: Attribute accuracy is consist of name accuracy and type accuracy. It reflects the accuracy of OSM road attributes.

Name accuracy can be calculate as the percentage of the length of tested dataset in which it's name is the same as reference dataset to the total length of tested dataset . Shown as formula (4):

$$Q_{L_M} = L_{OSM}^M / L_{OSM} \quad (4)$$

Tye accuracy is used to assess the accuracy of road type of tested dataset. It can be calculated as the percentage of the length of tested dataset in which it's type is the same as reference dataset to the total length of tested dataset . Shown as formula (5):

$$Q_{L_T} = L_{OSM}^T / L_{OSM} \quad (5)$$

Different from the name accuracy experiment, as the road type classification is different between these different datasets, it is necessary to built the corresponding relationship of road type between the tested dataset and reference dataset.

**2.2.3 Position Accuracy**: Position accuracy is the most important element to assess the geometric accuracy and usibility of crowd sourcing geographic data. The position accuracy calculation was based on a buffer technique developed by Goodchild and Hunter in 1997. In this technique, a buffer of width x is created for the reference road so as to calculate the proportion of the tested road that lies within the buffer. The width x indicates the half width of the real road. The formula is shown as follow:

$$Q_{L_P} = L_{OSM}^P / L_{OSM} \quad (6)$$

Where $L_{OSM}^P$ is the length of tested dataset lies within the buffer, $L_{OSM}$ is the total length of tested dataset.

## 3. ANALYSIS AND DISCUSSION OF EXPERIMENT

### 3.1 Experiment Data

The chosen area of study in this paper is urban area of Wuhan in China. The area of Wuhan chosen for this experiment is about 948.46 km2. The region which includes 9 municipality named Hongshan, Wuchang, Qingshan, Jiang'an, Jianghan, Qiaokou, Dongxihu and Hanyang covers the whole city centre of Wuhan. 1471 roads are dispersed throughout the region and their total length is 3466 km. The OSM data derives from the OpenStreetMap website and the geographic datum used in OpenStreetMap is the WGS-84/long datum while the reference data is the 2011 version of navigation map with position accuracy of 4 meters from NavInfo.

### 3.2 Experimental results and analysis

The experimental results of OSM accuracy assessment are shown in table 1.

| | | | |
|---|---|---|---|
| Length Completeness $Q_L$ | $L_{OSM}$ | $L_R$ | 38.0 % |
| | 3465977.447m | 9129950.159m | |
| Name Attribute Completeness $Q_{S_N}$ | $S_{OSM}^N$ | $S_R^N$ | 36.0 % |
| | 1471 | 4092 | |
| Name Length Completeness $Q_{S_L}$ | $S_{OSM}^L$ | $S_R^L$ | 26.1 % |
| | 2380133.9m | 9129950.159m | |
| Name Accuracy $Q_{L_M}$ | $L_{OSM}^M$ | $L_{OSM}$ | 51.4 % |
| | 1780632.921m | 3465977.447m | |
| Type Accuracy $Q_{L_T}$ | $L_{OSM}^T$ | $L_{OSM}$ | 32.2 % |
| | 1116337.777m | 3465977.447m | |
| Position Accuracy $Q_{L_P}$ | $L_{OSM}^P$ | $L_{OSM}$ | 51.5 % |
| | 1785989.309m | 3465977.447m | |

Table.1 Results of OSM Accuracy Evaluation in Wuhan

As table 1 shown, in the urban area of Wuhan, the length completeness of OSM is 38.0%, the name completeness based on name account is 36.0%, the name completeness based on length is 26.1%, the name accuracy is 51.4%, the tyepe accuracy is 32.2%, the position accuracy is 51.5%.

**3.2.1 Experiment and Analysis of Length Completeness**: As shown in Table 1, the length completeness of OSM is 38.0%. In general, the length completeness of OSM road is relatively low. The comparative experiment results have shown that the region with high road length completeness are almost in the districts with high urbanization level and dense population while the data collected in some districts which is far from the city centre are relatively insufficient so that the length completeness of OSM road in these region is extremely low.

For understanding the relationship between the length completeness of OSM road and inter-regional differences, the experiment performs statistical analysis of length completeness of different districts in Wuhan, as shown in Figure 1. From the figure, it is clearly that the district with highest road length completeness is Wuchang, in which the length completeness is 42.8%; followed by Hanyang and Hongshan, in which the length completeness is over 35.0%. But in Qingshan, Dongxihu and Qiaokou, the length completeness is much lower. Know then, the length completeness of OSM is largely constrained by urbanization level and population density.

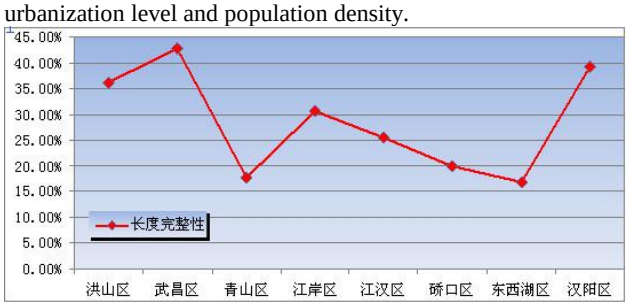


Figure.1 Length Completeness of District in Wuhan

In addition, the experiment performs statistical analysis of length completeness of OSM roads in different types, as shown in figure 2. It can be seen that the length completeness of high level road is much better, such as highway (64.2%), urban highway (90.4%), national road (72.4%) and so on, while the length completeness of low level road is less than 15.0%. As rural roads has the largest proportion of the tested dataset about 37.8%, it at its best embodies the length completeness of level roads and its length completeness is 40.4%. The results show that few pedestrian will participate in crowd sourcing data collection so that the collected data is not enough in these pedestrian path.

**3.2.2 Name Completeness**: As shown in Figure1, the name completeness based on name account of OSM in Wuhan is 36.0%. The reason of low name completeness is that the length completeness of OSM is not very high (38.0%) so that there are fewer named road in the tested OSM dataset. While the name completeness is calculated as the percentage of the amount of named raod in tested OSM dataset to the total amount of OSM roads, the name completeness of OSM can be upgraded to 73.3% (as shown in Table 2). It indicates that the name completeness of finished OSM road is very high. Furthermore, the unnamed roads in the test OSM dataset are mainly located in suburban district. Besides the limited data volume produced from users in these areas, It is also caused by lack of administrative standard management of rural roads compared to downtown area.

| | | | |
|---|---|---|---|
| NameAttribute Completeness $Q_{S_N}$ | $S_{OSM}^{N}$ | $S_{OSM}$ | 73.3 % |
| | 1471 | 2006 | |
| Name Length Completeness $Q_{S_L}$ | $S_{OSM}^{L}$ | $L_{OSM}$ | 68.7 % |
| | 2380133.9 m | 3465977.447 m | |

Table.2 Name Completeness with Improved Method

In addition, the name completeness based on length is 68.7% as shown in table 2. In contrast of the two methods' results, it is obvious that the name completeness of long distance road is lower than that short distance road.

**3.2.3 Name Accuracy**: As shown in Table 1, the name accuracy of OSM data in Wuhan is 51.4%. There are many factors that influence the name accuracy, such as unnamed road data in the reference dataset, existent of the OSM data unable to match to, affiliation issue of the transition sections connecting to several roads, overlapping of overpasses and roads beneath in two-dimensional datasets and so on.

To evaluate the influence of these factors like redundant OSM roads and unnamed reference roads, the name accuracy is calculated after eliminating these roads and the results are shown in Table 3.

| | |
|---|---|
| Name Accuracy without Redundant Data | 51.4% |
| Name Accuracy without Unnamed Date | 29.1% |
| Name Accuracy without Both Above | 58.1% |

Table.3 Name Accuracy without Redundant Data

As shown in Table 3, the redundant data records in OSM dataset have little influence in the name accuracy because the ratio of named and unnamed roads is similar in this part of dataset. While unnamed roads in reference dataset have much greater impact on the results since many records originally regarded as correctly named no longer participate in the test after the elimination, which take a relatively large proportion. After the elimination of both parts mentioned above, the name accuracy rises up significantly, the reason of which is that through this elimination, more redundant data are eliminated while correctly named data records remain intact, which directly leads to larger proportion of correctly named data records.

In addition, the difference between datasets in naming rules also gives greater impact on the name accuracy of OSM data. The results would be affected if the roads with incorrect names due to problems of naming conventions are extracted and corrected manually. In an additional test, the OSM data records of 240545.62 meters roads that can be corrected directly are extracted for name corrections. And the name accuracy after the correction is 58.3%. Although the results is of a little subjectivity since people were invoved in the process of name matching, they can still be able to reflect the influence of different naming rules to the name accuracy qualitatively.

**3.2.4 Type Accuracy**: As shown in Table 1, the typ accuracy of OSM data in Wuhan is 32.2%. The difference of typ accuracy in various districts is insignificant in Wuhan. The major factor affecting accuracy is that there are many long-distance trunk roads with incorrect typ attributes in OSM data. And the misclassification of OSM data is mainly caused by the inconformity between OSM and navigation map in classification criterion. The most significant difference is that there is no “national road” type in OSM classification criterion thus all OSM roads matched to the “national road” type in reference dataset are regarded as misclassified. At the same time, the length of OSM roads matched to the “national road” type in reference dataset is about 205040.300 meters, represents 5.9% of the whole dataset, and brings greater effect on the result of type accuracy test.

**3.2.5 Position Accuracy**: According to the position accuracy assessment model, the position accuracy of OSM data is 51.5%, as shown in Table 1. After eliminating redundant data records from OSM dataset, the position accuracy rises up to 60.4%, as shown in Table 4. The results show that the redundant data also bring greater effect on position accuracy of OSM.

| Position Accuracy $Q_{L_P}$ | $L^P_{OSM}$ | $L_{OSM}$ | 60.4 % |
|---|---|---|---|
| | 1785989.309 m | 2920386.236 m | |

Table.4 Position Accuracy without Redundant Data

The results of position accuracy of raods in different levels are shown as follows, in figure 3.

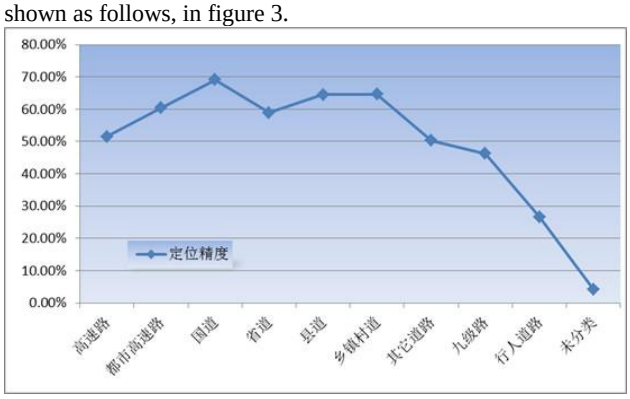


Figure.3 Position Accuracy of Different Rank Road

As shown in Figure 3, in general, the position accuracy of high-level roads is higher than low-level roads. And the reason why the position accuracy of pedestrian paths is particular low among all low-level roads is that there are very few data collecting and uploading consciously carried out by pedestrians, which causes insufficient of data volume. Besides that, the constant construction in urban area and hardness in defining the outlines of the sidewalks are also relevant.

In addition, experiments of position accuracy in different districts of Wuhan are carried out in this paper with results shown in Table 5. As shown in Table 5, the road network of suburban district like Dongxihu District and Qiaokou District, in which the position accuracy is relatively high, are mostly consisted of high-level roads and common roads, such as national roads, rural roads and so on. The proportion of sidewalks, Level-9 roads and unclassified roads is much lower in these districts than in others. Based on this phenomenon, a column is added to the table to demonstrate the proportion of high-level roads and low-level roads, from which the conclusion can be drawn that the position accuracy of roads is basically positive correlation to the gap between high-level roads and low-level roads in a certain district. This phenomenon shows that the reason of inconspicuous improvement of positional accuracy in downtown Wuhan, given the fact that the completeness is significantly better there, is that the low-level roads take large proportion in these areas and the construction activities happen frequently.

| District | High-level road | Low-level road | Gap | Position Accuracy |
|---|---|---|---|---|
| Qingshan | 30.8% | 40.8% | -10.0% | 42.1% |
| Hongshan | 46.3% | 17.3% | 29.0% | 57.6% |
| Hanyang | 59.3% | 17.6% | 41.7% | 57.8% |
| Jiang'an | 71.4% | 13.9% | 57.5% | 58.4% |
| Wuchang | 64.1% | 12.9% | 51.2% | 59.6% |
| Dongxihu | 50.4% | 1.3% | 49.1% | 68.8% |
| Qiaokou | 68.0% | 2.6% | 65.4% | 71.4% |
| Jianghan | 88.9% | 0.8% | 88.1% | 78.6% |

Table.5 OSM Position Accuracy of District in Wuhan

## 4. CONCLUSION

The emergence of crowd-sourcing geospatial data provides abundant data source for geospatial data update. Whereas the quality issue of crowd-sourcing data has a direct influence to the usage of these data, therefore it was extensively researched all over the world. The quality assessment of OSM data in China represented by Wuhan is implemented and the factors relevant to quality issue are analyzed in this paper. As the results show, the overall completeness and position accuracy of OSM data in Wuhan are both relatively poor. While as for various type of roads, urban highway and national roads have better completeness and position accuracy. Apart from that, OSM data provides more details than 2011 version of navigation data in part of Wuhan area. As a result, OSM data can be used as a new data source for the data acquisition and update of urban high-level traffic fundamental networks and downtown road networks in order to complement or partially replace traditional urban traffic network data acquisition and update approach.

## 5. REFERENCE

**References from Journals**:

Giles,J.,2006. Wikipedia rival calls in the experts. Nature, Vol. 443, Oct. 5, pp.493.

Goodchild, M.F., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal, 69, pp.211-221.

Goodchild, M.F., 2008. Commentary: whither VGI? GeoJournal, 72, pp. 239–244.

Goodchild, M.F., 2009. Geographic information systems and science: today and tomorrow. Procedia Earth and Planetary Science (special issue title: Proceedings of the International Conference on Mining Science and Technology), Vol. 1, Issue 1, pp.1037–1043.

Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing, 65, pp.550-557.

Howe, J. 2006. The rise of crowdsourcing. WIRED Magazine, Issue 14.06, June, http://www.wired.com/wired/archive/14.06/crowds.html, Retrieved on Feb. 20, 2011.

Jan De Leeuw, Mohammed Said, Lapezoh Ortegah, Sonal Nagda, Yola Georgiadou, Mark DeBlois. (2011). An Assessment of the Accuracy of Volunteered Road Map Production in Western Kenya. Remote Sensing, 3(2), pp.247-256.

**References from Other Literature**:

Ather,A.(2009). A Quality Analysis of OpenStreetMap Data. MEng Thesis, London, University College London.

## 6. ACKNOWLEDGEMENT

The authors would like to thank the Fundamental Research Funds for the Central Universities to support the project "Quality Analysis and Data Mining of Crowd Sourcing Check-in Data " (Grand No: 2012213020208).