VALIDATION ANALYSIS OF OPENSTREETMAP DATA IN SOME AREAS OF CHINA

Pingping ZHOU[a] ,Wei HUANG[b] ,Jie JANG[b]

[a]China University of Mining &Technology, Beijing,China-zhpingping89@163.com
[b]National Geomatics Center of China,China-(huangwei,jjie)@nsdi.gov.cn

KEY WORDS: OpenStreetMap, data quality, evaluation, quality element, weighted coefficient, validation analysis

ABSTRACT:

The rapid development of computer technologies has given rise to the increase of open source web-based map services such as OpenStreetMap, a global vector data created by volunteers for free use. There is a concern about the quality and usability of the OpenStreetMap data because the volunteers that contribute the data generally lack the sufficient cartographic training. This paper focuses on the data quality analysis method for OpenStreetMap. A model for usability evaluation has been proposed. A benchmark between OpenStreetMap data and the1:10 000 topographic data in some areas of China has been done to verify the proposed model, and the method proves to be effective.

1. INTRODUCTION

With the advent of Web 2.0, geographic information service pattern has undergone tremendous change. People gradually become geographic information providers through uploading their data, which was termed Volunteered Geographic Information (VGI) by Goodchild[1]. OpenStreetMap (OSM) is one valuable application of VGI.

OSM was initiated by Stephen Coast in July 2004 at the University College London. Since its establishment, OSM is expanding scale, the number of registered users from hundreds in the middle of 2004 increase to more than five hundred thousand in November 2011.As an online map collaborative plan, provided voluntarily by individuals involving the capture, processing and dissemination of geographic information, the project aims to create and distribute vector data for the world because most maps thought of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways. There are three sources for OSM to obtain vector data, including hand-held GPS receiver trace data from users, donations from institution and organization, vectorization of images such as Landsat, Yahoo, Imagery, etc.

Spatial data is the base of geographic information and its quality is directly related to the accuracy of spatial analysis and operation. Though OSM project has many advantages, there are concerns about how the OSM quality is and what aspects of application it can meet for the volunteers that contribute lack professional knowledge and sufficient cartographic training. This paper focuses on the OSM data quality analysis method, and proposes an evaluation model.

Many scholars have performed a series of researches on OSM data quality [2-4]. Initially, Mordechai Haklay focused on the positional accuracy and length completeness of England OSM data through comparison with the Ordinance Survey's Meridian 2 dataset. The methodology used to evaluate the positional accuracy was based on Goodchild and Hunter (1997) and Hunter (1999). The comparison was carried out by using buffers 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. The completeness used the formula calculated as: Σ(OSM roads length) - Σ(Meridian roads length).The analysis shows that OSM information can be fairly

accurate: on average within about 6 meters of the position recorded by the Ordnance Survey, and with approximately 80% overlap of motorway objects between the two datasets. Subsequently, Aamer Ather, Ourania Kounad et al. constantly enriched the research content and carried out the study from the road location accuracy, data completeness and attribute accuracy. At the same time, the study areas have gradually extended from England to other areas. Recently, Blazej Cipeluch et al. compared the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps from the road coverage and POI numbers. In relation of road coverage, they covered the XML format data to the Google maps, Bing maps and OpenStreetMap with Openlayer tool to make a comparison. For lack of vector data as reference data, POI queries for Bing and Google maps relied on manual search from the web-page queries, which resulted that the operation process wasn't well controlled and POI was easy to miss. Overall, there were three problems on the analysis of OpenStreetMap. Firstly, the operation efficiency was low for the difficulty to obtain high accuracy vector data as reference data .Secondly, the method to access data completeness and attribute accuracy needed to improve, for example, the formula to analysis length completeness couldn't completely reflect the data quality when some important roads missed .Thirdly, researches stated above analyzed the data quality from the quality element, which easily showed how each quality element was but had difficulty in describing what fields the data was fit for .

2. RESEARCH METHOD

2.1 OSM Quality Evaluation Model

It has always been a concerned problem for data producers and users about how to accurately assess the spatial data quality. In order to better describe it, quality elements are used to express spatial data quality .In recent years, scholars has made deep studies on what elements should be used, but have not yet formed uniformed quality elements, and following elements are usually used:

Linage – about the history of the dataset such as how it was collected and evolved.

Position accuracy – the accuracy of features or geographic objects in either two or three dimensions.

Attribute accuracy –the degree attribute field values adherence to the true data.

Logical consistency – the degree of adherence to logical rules of data structure, attribution and relationships.

Completeness –presence and absence of objects in a dataset

Temporal quality – this is a measure of the validity of changes in the database in relation to real-world changes and also the rate of updates.

Considered the reality and operability, we conduct the study from completeness, position accuracy, attribute accuracy and logical consistency which is regarded as first level element. Then we subdivide these four elements into more detailed quality element as second level element, completeness divided into length completeness and name completeness and attribute accuracy divided into type accuracy and name accuracy. Logical consistency is assessed through whether it satisfies the topology rules.

It uses weight coefficient to express the contribution each element of spatial data quality on the results of comprehensive evaluation, and the weight coefficient reflects the relative importance of each element participating in the evaluation. Four methods are commonly used to decide the weight coefficient, including subjective experience judgment, expert investigation or consultation, vote from the judgment panel and analytic hierarchy process (AHP). After reading related documents [5-6] and analyzing the importance of the element, we make the distribution among these elements including first level and second level element.

Table 1 quality element and weight coefficient

First level element	Weight coefficient	Second level element	Weight coefficient
Position accuracy	0.3	Position accuracy	1
Attribute accuracy	0.3	Name accuracy	0.6
		Type accuracy	0.4
Completeness	0.3	Length completeness	0.6
		Name completeness	0.4
Logical consistency	0.1	Logical consistency	1

Evaluation model connects the quality elements by adopting weighted coefficient. After getting each element's score, we summary the score according to the weight coefficient.

$$Q=\sum_{i=1}^{4} P_i Q_i \qquad (1)$$

Where P_i= weight coefficient of some quality element

Q_i =score of some quality element

The data is regarded as unqualified if one element's score is less than 60.

Data quality rating is classified according to current standards.

Table 2 quality score and level

score	90-100	75-89	60-74	<60
level	excellent	good	qualified	unqualified

2.2 How to calculate each quality element

1. Attribute Accuracy

Attribute accuracy is considered from two aspects, including type accuracy and name accuracy.

Road type is divided based on its position, roles and service function in the road system, which reflects road importance to some extent. Chinese roads are divided into national highway, provincial road, county road, township road and dedicated road according to administrative level and in the first four kinds, road encoding is G,S,X and Y respectively. In the road type accuracy calculation, scaling factor is subdivided into 0.6 and 0.4，G,S,X and Y roads weighted 0.6 and other roads weighted 0.4. G, S, X and Y road type accuracy is calculated as follows:

Q_{111}=100*(the length of roads with right type G, S, X and Y)/ (the total length of roads with right type G, S, X and Y and those should be modified to G, S, X and Y)

The calculation method of other roads type accuracy is similar to the calculation method above:

Q_{112} =100*(the length of roads with right other type)/ (the total length of roads with right other type and those should be modified to others)

Then type accuracy is got through the following method:

$$Q_{11}=0.6*Q_{111}+0.4*Q_{112}$$

In terms of road name annotation for the same road, there are four cases: both have name for reference data and experiment data; reference data has name but experiment data doesn't which is considered in the name completeness; experiment data has name but experiment data doesn't; both have no name.

For the name accuracy analysis, wrong name roads refer to that both have name but don't match. The following is the formula of name accuracy:

Q_{12}=100-100*(the length of road with wrong name in OSM)/ (the length of road with name in OSM)

Then attribute accuracy is calculated:

$$Q_1=0.4*Q_{11}+0.6*Q_{12}$$

2. Data Completeness

This paper focuses on length completeness and name completeness. The length completeness refers to whether there are missing roads compared with reference data. The name completeness means the reference data has name but the experiment data doesn't have for the same road.

It usually compares both data's length at length completeness analysis, which can reflect experiment data's detailed degree, but can't ensure the major roads exist when experiment data is more detailed than reference data for on that condition, experiment data length must be larger than reference data length. For redundant path in the experimental data, it doesn't participate in the score calculation for reference data is of high precision and can satisfy some certain applications. Hence, the author makes some changes on the previous calculation method. The previous length completeness is calculated as the percentage of the length of the experiment dataset to the length of the reference dataset. The current calculation formula is as follows:

Q_{21}=100-100*(the length of OSM missing road in reference data) / (total length of reference data)

For name completeness, we also just consider the condition names are missing compared with reference data and it is calculated as follows:

Q_{22}=100-100*(the length of OSM road missing the name) / (total length of OSM data)

Then Q_2 can be got through:

$$Q_2=0.6*Q_{21}+0.4*Q_{22}$$

3. Position Accuracy

As to position accuracy, Tveite [7] defines two aspects of linear accuracy: a) positional point accuracy: positional accuracy can easily be given for well-defined points on the line (e.g. the end-points). For the rest of the line, it is difficult to say anything about positional accuracy and to quantify it, b) shape fidelity: to be able to say something about the accuracy of a line, it is useful to talk about its shape fidelity as compared to another line. The shape fidelity should indicate to what extent the curvature of two lines are similar.

Positional point method can't be used for the endpoints don't match between OSM data and experiment data. In 1997, Goodchild and Hunter proposed Positional accuracy of

digitized linear features [8]. Its idea is to regard the reference data as true data for reference data is with higher accuracy, and normally the deviation of experiment data with reference data should be within a range. As figure 1 shows, a buffer of width x(x equals to half of road width) is created for the reference feature so as to calculate the proportion of the experiment feature that lies within the buffer. The method has many advantages: ①it is relatively insensitive to extreme outliers; ②needn't match between the datasets.③is easy to operate for it bases on a simple overlay process that could be done in most vector GIS programs.

In this calculation, scaling factor is subdivided into 0.55 and 0.45; G, S, X and Y roads weighted 0.55 and other roads weighted 0.4. G, S, X and Y roads and other roads position accuracy are calculated as follows:

Q_{31}=100*(G, S, X and Y road length that fall into the buffer) / (G, S, X and Y road length)

Q_{32}=100*(other road length that fall into the buffer) / (other road length)

Q_3=0.55* Q_{31}+0.45* Q_{32}

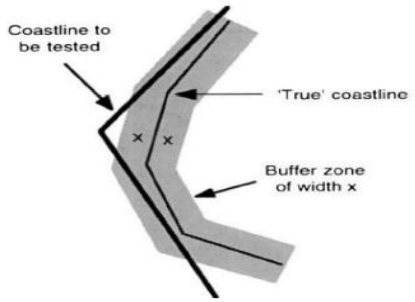

Figure 1 Goodchild & Hunter buffer comparison technique (source: Goodchild & Hunter 1997)

4. Logical Consistency

Logical consistency describes data structure (including conceptual, logical or physical structure), element property and the degree of the relationship between them in conformance with rules. It refers to the reliability of the data on the relationship, including data structures, data content, spatial attributes and thematic attributes, especially internal consistency on the topographical properties. Data logical consistency is validated by establishing topographical rules that include the road can't overlap as well as self-overlap. The formula to get this element's score is as follows:

Q_4=100-100*(the length of OSM road that don't obey these rules)/ (the total length of OSM roads)

3. METHOD VALIDATION

1:10 000 national basic data is chosen as reference data to validate the method of OSM vector road data quality，and three cities Handan, Lanzhou, Nantong are chosen as study areas.

3.1 Calculation the Attribute Accuracy

As mentioned in section 2，road type error falls into two kinds: Road type that should be coded has empty code or has wrong road type; or road type that should be null has code. Table 3 shows type accuracy, the data distribution of type accuracy sees the appendix 1.

Table 3 type accuracy

City / Length	Handan		Lanzhou		Nantong	
	G,S type	Other type	G,S type	Other type	G,S type	Other type
Wrong type Road	11067.39	0	9728.73	17507.22	45830.06	0
Corresponding total road	16905.65	41941.56	70629.08	367463.62	45830.06	300493.78
Error radio	65.47%	0	13.77%	4.76%	100%	0
Score	60.72		89.83		40	

Table 3 shows the type accuracy gap among three cities is very large, and the road that should be coded always misses the type. Name accuracy analysis is performed by visual contrast to find out the road with wrong name for the number of road with name is small (the data distribution of name accuracy sees the appendix 2).

Table 4 name accuracy

City / Length	Handan	Lanzhou	Nantong
Wrong name road	0	10473.86	0
Road with name	5838.27	58926.63	2286.94
Error percentage	0	17.77%	0
Score	100	87.05	100

In Handan and Nantong, the road with name is only one, which leads to the score is full or zero. In Lanzhou, the errors are mainly wrong description to orientation, for example Binhe Mid Road leveled into Binhe East Road.

Table 5 attribute accuracy

City	Handan	Lanzhou	Nantong
Score	84.29	88.16	76

Table 5 shows the attribute accuracy scores.

3.2 Calculation the Completeness

ArcGIS is applied to select the missing road of OSM from reference data (the data distribution of length completeness sees the appendix 3, the data distribution of name completeness sees the appendix 4).

Table 6 length completeness

City / Length	Handan	Lanzhou	Nantong
Experiment data	58847.21	265792.95	346323.83
Reference data	93180.39	1021518.4	969285.10
Missing road	36557.16	693975.6	691996.23
Missing radio	39.23%	67.94%	71.39%
Score	60.77	32.06	28.61

Length completeness of OSM is much lower than that of reference data. In the three cities, Handan roads are most detailed and evenly distributed; Lanzhou Roads are mainly located in the central areas, where data details are nearly same to that of reference data, but the roads are missing severely in other areas, only several roads distributed; Nantong roads have even distribution and poor details.

Table 7 name completeness

City / Length	Handan	Lanzhou	Nantong
Reference data with name	50144.88	289267.22	302914.98
Experiment data with name	5838.27	5892.632	2286.94
OSM road	58847.21	265792.95	346323.83
Missing name road	39628.12	204537.54	286303.19
Missing percentage	67.34	76.95	82.67
Score	32.66	23.05	17.33

Table 8 Completeness score

City	Handan	Lanzhou	Nantong
Score	49.53	28.46	24.10

3.3 Calculation the Position Accuracy

Reference data contains part road edges, which can be used to generate road surface. As to the road without edges, centerlines are applied to generate buffer, whose distance is respectively 11.25 meters and 7.5 meters in G, S roads and other roads (the data distribution of position accuracy sees the appendix 5).

Table 9 position accuracy

	Handan		Lanzhou		Nantong	
	G,S type	Other type	G,S type	Other type	G,S type	Other type
Road fall into buffet	16247.45	23820.09	59484.774758	234220.439386	43828.77	224287.24
Corresponding total road	16297.81	27059.29	68463.346213	274485.718232	57270.62	284899.21
radio	99.69	88.03	86.88	85.33	76.53	78.73
Score	94.44		86.18		77.52	

As the table 9 shows, the position accuracy in three cities is very high, for instance, of which Handan's national highway and provincial road reaches to 99.69.

3.4 Calculation the Logical Consistency

Topographical rules are built to validate the logical consistency, including must not overlap and must not self-overlap, and errors are found out through error inspector of Topology Tools in ArcGIS. The results show logical errors of Handan and Nantong don't exist. Lanzhou data has one self-overlap error and six overlap errors, and the lengths are 301.87meters and 11600.34meters respectively. so the final scores are 100，97.28 and 100.

Finally all quality elements are summarized as table 10：

Table 10 Final results

Cities	Handan	Lanzhou	Nantong
attribute accuracy	84.29	88.16	76
completeness	49.53	28.46	24.10
position accuracy	94.44	86.18	77.52
logical consistency	100	97.28	100
final score	78.48	70.57	63.29
quality level	unqualified	unqualified	unqualified

The quality level of three cities is evaluated according to the quality rating standard, and the results including the assessment result of each quality element basically conform to experts' evaluation results.

4. CONCLUSIONS

The conclusions can be drawn from above example results:

(1) The quality evaluation model proposed can better reflect the quality of spatial data and be applied to carry on comprehensive assessment of each element. The setting that qualified data must ensure each element score is more than 60 in the calculation can effectively avoid the occurrence of the extreme situation.

(2) The quality evaluation model is easy to operate.

(3) As the scores show, the data of three cities is very poor in data completeness including length completeness and name completeness; and has some advantages in attribute accuracy, position accuracy and logical consistency.

1. Type accuracy

Figure 2 Handan Type Accuracy

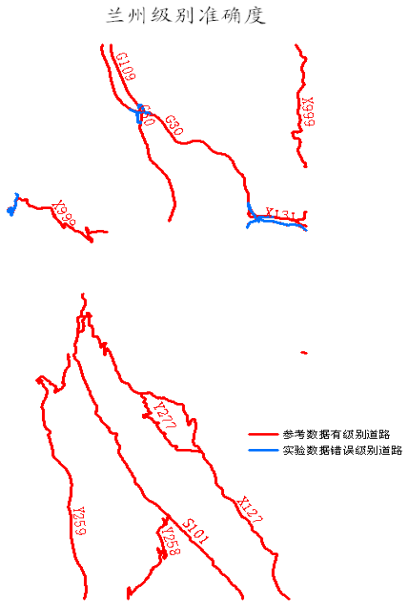

Figure 3 Lanzhou Type Accuracy

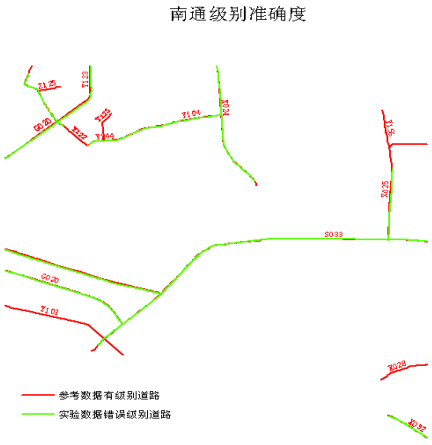

Figure 4 Nantong Type Accuracy

Reference

[1] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

[2] Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design*, *37*(4), 682.

[3] Ather, A. (2009). A quality analysis of openstreetmap data. *ME Thesis, University College London, London, UK*.

[4] Kounadi, O. (2009). Assessing the quality of OpenStreetMap data. *Msc geographical information science, University College of London Department of Civil, Environmental And Geomatic Engineering*.

[5] Du Daosheng, & Wang Zhanhong. (2000). The spatial data quality model research. *Chinese journal of image and graphics*: A series, (7), 559-562.

[6] Ceng Yanwei, & Gong Jian. (2004). The method to evaluate and control spatial data. *Journal of wuhan university, information sciences, 29* (8), 686-690

[7] Goodchild, M. F., & Hunter, G. J. (1997). *A simple positional accuracy measure for linear features*. International Journal of Geographical Information Science, 11(3), 299-306.

[8] Tveite, H. (1999). *An accuracy assessment method for geographical line data sets based on buffering*. International journal of geographical information science, 13(1), 27-47.

Appendix

2. Name accuracy

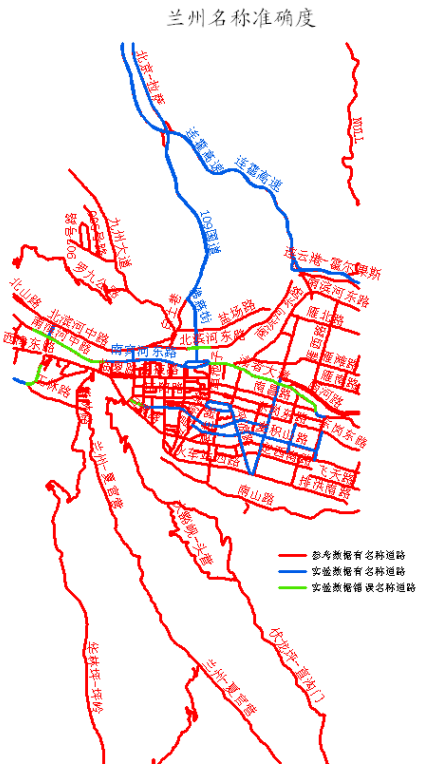

Figure 5 Lanzhou Name Accuracy

3. Length completeness

Figure 6 Handan Length Completeness

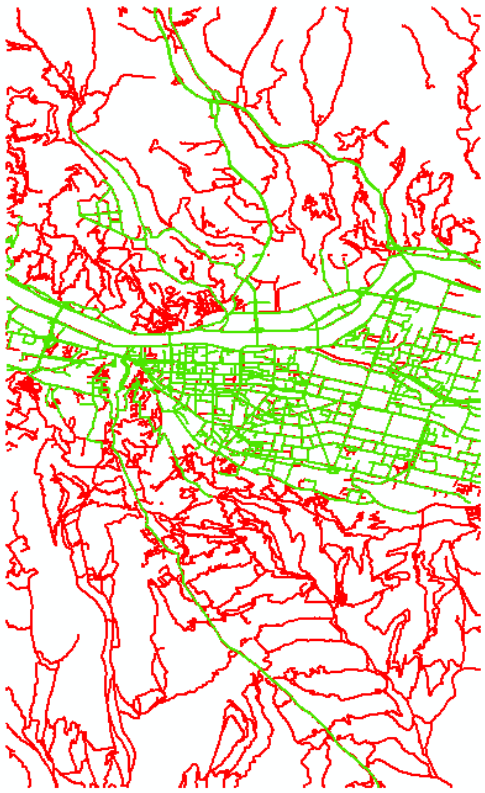

Figure 7 Lanzhou Length Completeness

Figure 8 Nantong Length Completeness

4. Name completeness

Figure 9 Handan Name Completeness

Figure 12 Handan Position Accuracy

Figure 10 Lanzhou Name Completeness

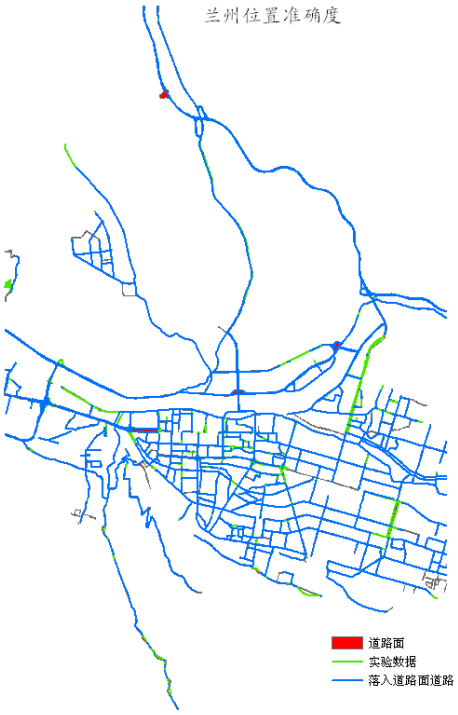

Figure 13 Lanzhou Position Accuracy

Figure 11 Nantong Name Completeness

5. Position accuracy

Figure 14 Nantong Position Accuracy