Assessing the accuracy of 'crowdsourced' data and its integration with official spatial data sets

Maythm Al-Bakri and **David Fairbairn**
School of Civil Engineering and Geosciences
Newcastle University
Newcastle upon Tyne, NE1 7RU, England, UK
{m.m.s.al-bakri, dave.fairbairn}@newcastle.ac.uk

***Abstract*— A key challenge in supporting effective spatial data integration is the assessment of data quality from different sources. This paper presents a methodology for assessing positional and shape quality for formal data, such as Ordnance Survey (OS), and crowdsourced data, such as OpenStreetMap (OSM) information, with the intention of assessing possible integration. It is based on the measurement of discrepancies among the data sets: results show that the accuracy of OS data is very close to a reference data set, but the positional and shape accuracy of OSM data does not match the reference or the OS data sets.**

Keywords: data integration; data quality assessment; crowdsourced data; positional accuracy; semantic accuracy

I. INTRODUCTION

With the rapid development of geospatial data collection technologies, and the growth of the World Wide Web (WWW) for different applications, a large amount of geospatial data can be disseminated and shared via the Internet. Today there is a wide variety of geospatial data sources available on the Internet such as the Google Map service, the OpenStreetMap (OSM) project, Yahoo imagery and others. OSM is a project designed to create and provide free spatial data based on volunteered efforts, personal computers and the Internet. Similarly, various government agencies provide a wide variety of topographic, thematic and cadastral mapping online. The Ordnance Survey (OS) OpenSpace service, for example, allows the embedding of formal OS digital maps in web applications in scales ranging from the whole of Great Britain down to street level. The US Census Bureau is a further example of a formal (governmental) spatial data agency, through which road vector data covering all of the United States is available. Thus, broadly speaking geospatial data available on the web can be categorised by the community that collects and collates it: data collected by state-sponsored or commercial organisations can be considered as Formal Data (FD), but we must also recognise alternative public-sourced data collection as 'crowdsourced' or User Generated Content (UGC) (not limited to spatial data).

These spatial data sources are managed by different communities and their data is also different in respect to quality, coverage, and purpose. Geospatial data from different sources often has variable accuracy levels due to differing data collection methods, and potentially, therefore data accuracy may not meet the user requirements in varying organizations. Many efforts have been made to integrate multiple geospatial datasets to improve overall accuracies (Omran and van Etten, 2007): such data integration can produce more accurate results and more reliable information than that obtained from a single source. However, the integration process remains one of the main challenges facing spatial data users. The wide variety of geographic information, notably its diversity in collection time, purpose, data quality and many other characteristics, leads to difficulty in integration.

A study has been carried out to assess the quality of OSM information and OS data by comparing them to reference field survey (FS) data sets. Standard high quality field survey was used to create a definitive reference data set. This accurate data set is produced using the highest precision survey instruments, the data from which can be used as the benchmark for FD and UGC data comparisons. The OS data used for comparison is taken from the MasterMap topographic layer, while the OSM information has been downloaded from CloudMade and Java OpenStreetMap (JOSM) editor.

This paper will focus on the geometric accuracy assessment primarily. This includes the evaluation of positional accuracy and shape fidelity. The resultant measures of accuracy will assist in an assessment of the possible role of 'crowdsourced' data in the flowline of official spatial data handling. Section II relates to metric testing including identifying and testing individual points, presenting and applying the buffer overlap method for comparing line symbols, and presenting moments and deriving moments for area symbols. Section III discusses initial work in applying semantic accuracy models to assist in semantic accuracy testing. Section IV gives the summary and conclusion.

II. MEASURES AND ANALYSIS

Different measures and analyses are required for the testing of differing data types. Section A presents the method of positional accuracy assessment. Section B explains in detail the technique of linear accuracy assessment. Section C illustrates the measures used for detecting area shape difference.

A. *Evaluation of Positional Accuracy*

According to Devillers and Jeansoulin (2006), "positional accuracy may be defined as the degree to which the digital representation of a real-world entity agrees with its true position on the earth's surface" (p.147).

Assessing such positional accuracy can be eased by adopting established standards for spatial data accuracy, helping the measurement and reporting of accuracy across a whole dataset. Existing standards such as National Map Accuracy Standards (NMAS) generally focus on testing paper maps not digital data. Further, this standard measures the errors at the map scale instead of ground scale, which can be considered problematic when changing the mapping system into digital formats that can be output at varying scales (Congalton and Green, 2009). Alternative spatial data accuracy standards, such as the National Standard for Spatial Data Accuracy (NSSDA), have now been developed.

NSSDA specifies that the positional accuracy can be reported at ground scale rather than map scale, allowing it to be used with digital map data as well as hard copy maps. Furthermore, the NSSDA provides a formal approach to how the tested points should be identified, measured and distributed across the map. It suggests that twenty or more test points are required to effectively test data accuracy. These points must be well defined, easy to measure and found in both tested and reference data sets. The ideal distribution of tested points is even with at least 20 percent of the points in each quadrant when the data set covers a rectangular area. The intervals between points should be at least 10 percent of the diagonal distance of the total area of the data set (Paul, 2008). The NSSDA uses Root Mean Square Error (RMSE) to estimate the positional accuracy. Simply, it can be computed as follow:

$$RMSE=\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\delta x_i^2+\sum_{i=1}^{n}\delta y_i^2\right)} \qquad (1)$$

Where, n is the number of check points and δx_i^2 ,δy_i^2 are the direct linear distance mismatches for the *i*th checked point in x and y directions.

Areas chosen for testing covered two contrasting case studies, one of an urban area (Cramlington, Northumberland), and one of an open peri-village, rural landscape (Clara Vale, Northumberland) as shown in Fig. 1 (A & B respectively). The reason behind this choice is to identify and test data quality using both 'hard' (manmade) and 'soft' (natural) features. Each area has a self generated reference data set, the open-access data obtained through the OSM project, and OS MasterMap data. A total of 40 check points were selected and distributed based on the principles of the NSSDA.

The RMSE values comparing the reference data set with the samples from the tested data sets were 0.478m (OS) and 9.661m (OSM) in the urban area, and 1.843m (OS) and 11.032m (OSM) in the peri-village area.

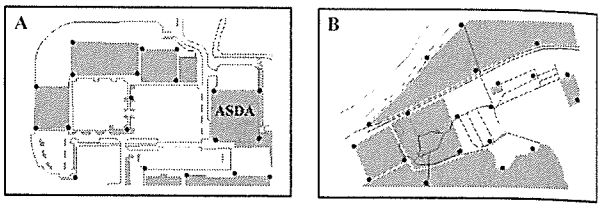

Figure 1. The distribution of points for positional accuracy assessment.

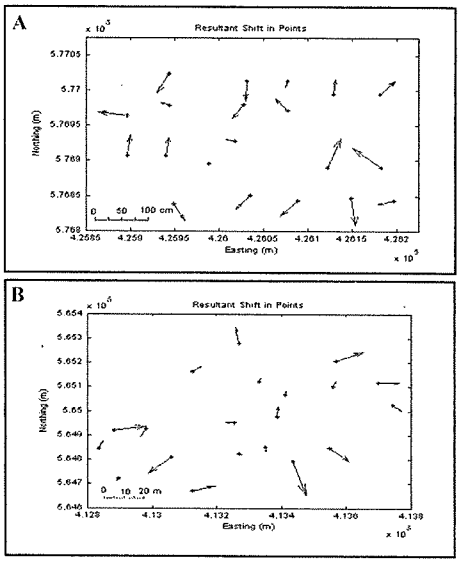

Figure 2. Identifying the magnitude and direction of differences between (A) FS-OS data and (B) FS-OSM data

Fig. 2 illustrates the magnitude and direction of the errors in comparing FS and OS data in location (A) and comparing FS and OSM data in location (B). Similar observations are evident for other comparisons. It is clear that the RMSE value is very high for OSM compared to OS data. This is probably due to the fact that a number of different sources have been used to create the OSM maps including GPS tracks and Yahoo aerial imagery. Most of the OSM data measurements have been done using low accuracy GPS. Furthermore, some features are averaged by eye from many GPS positions and tracks recorded by several individuals over time. As James Derrick (2009) asserts, "Asda Cramlington is bounded by a public path to the East and North, and a road to the West and a precinct to the South. This means the footprint was inferred from neighbouring features but I can't easily distinguish where the North wall of the building is and the loading compound starts" (personal communication, April 2009). James Derrick is the person who uploaded the data of Cramlington shopping centre into the OSM project. The lower accuracy of the rural data set may be explained by the incompleteness of the data coverage resulting in fuzzy, interpolated or inferred boundaries such as the extent of woodland areas etc.

B. *A measure of Linear Geometric Accuracy*

Linear accuracy can be examined as an indication about the shape or curvature similarity between two lines, as well as positional accuracy of points along the line. When line features such as roads or railways are considered, comparison using point accuracy is insufficient to capture the geospatial complexity of linear features. While point measures may be

straightforward to understand and calculate, they do not capture all aspects of line accuracy.

In research on line accuracy measurements, the buffering method, and an assessment of buffer overlay has been a popular technique (Tveite, 1999; Goodchild and Hunter, 1997). The elements of this method are shown in Fig. 3: lines in each data set (e.g. X, Q) get a number of buffers of various sizes (e.g. XB, QB). It is an iterative approach because it will be impossible to estimate the appropriate buffer size in advance. The process of gradually increasing buffer size should be terminated when the results of measuring displacement or overlap seem to stabilize. By comparing the area of different zones, one is able to calculate overlap percentage or average displacement (DE) between two lines:

$$DE=\pi b_s \frac{Area(\overline{XB} \cap QB)}{Area(XB)} \quad (2)$$

where, b_s is the buffer width and Area ($\overline{XB} \cap QB$) is the area outside XB and inside QB.

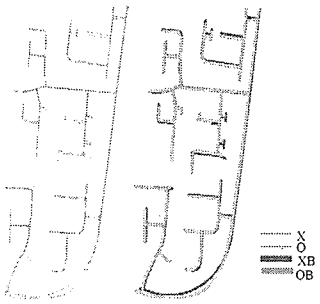

Figure 3. The buffer overlay method

Three data sets were used in the experiment (FS, OS and OSM) and the comparison was carried out for some main and minor roads in the Cramlington as shown in Fig. 3. Different samples were taken to evaluate the differences between the FS centreline and the location that is recorded in OS and OSM. The rest of the analysis was performed by creating a buffer around each data set, then evaluating the average displacement and overlap percentage.

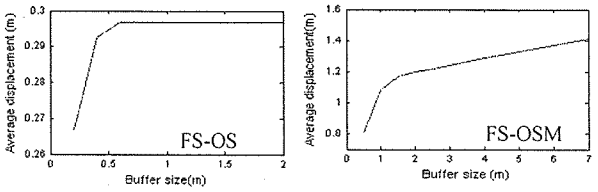

Figure 4. The overlap information.

Average displacement and overlap percentage for a number of buffer sizes is shown Figs. 4 and 5. The graphs must be expected to increase with increasing buffer size until they reach the optimum values of overlap percentage and average displacement of that line. After that the graphs will start to flatten out. The diagrams indicate a significant difference in the accuracy of OS and OSM with the OS data being closer to the reference data set: the average displacement for OS is about 0.3m, whereas it is about 1.5m for OSM. Similar observations can be made for overlap percentage between datasets.

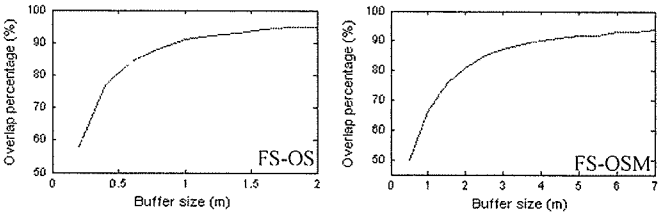

Figure 5. The overlap information.

C. Area Shape Measure

To measure shape discrepancies between features, we use techniques based on invariant moments. Moments were first used for mechanics purposes other than shape description. Hu (1962) was the first to set out the mathematical foundation for two dimensional moments invariant and demonstrated their application to shape recognition. He proved that a proper combination of moments can provide translation, scale, and rotation invariant quantities. The geometric moments of order p+q with the basis set (x^p, y^q) can be defined as:

$$m_{pq}=\iint_N x^p y^p f(x,y)dxdy \quad (3)$$

For p,q = 0,1,2,.....

m_{pq} is the two dimensional moment of the function f(x, y).The order of the moment is (p+q) where p and q are both natural numbers.

The set of seven invariant moments have been proposed by Hu as in (4). This can be used for scale, position, and rotation invariant pattern identification.

$$\begin{aligned}
\Omega_1&=m_{20}+m_{02}\\
\Omega_2&=(m_{20}-m_{02})^2+4m_{11}^2\\
\Omega_3&=(m_{30}-3m_{12})^2+(3m_{21}-m_{03})^2\\
\Omega_4&=(m_{30}+m_{12})^2+(m_{21}+m_{03})^2\\
\Omega_5&=(m_{30}-3m_{12})(m_{30}+m_{12})[(m_{30}+m_{12})^2-3(m_{21}+m_{03})^2]\\
&\quad +(3m_{21}-m_{03})(m_{21}+m_{03})[3(m_{30}+m_{12})^2-(m_{21}+m_{03})^2]\\
\Omega_6&=(m_{20}-m_{02})[(m_{30}+m_{12})^2-(m_{21}+m_{03})^2]+4m_{11}\\
&\quad (m_{30}+m_{12})(m_{21}+m_{03})\\
\Omega_7&=(3m_{21}-m_{03})(m_{30}+m_{12})[(m_{30}+m_{12})^2-3(m_{21}+m_{03})^2]\\
&\quad -(m_{30}-3m_{12})(m_{21}+m_{03})[3(m_{30}+m_{12})^2-(m_{21}+m_{03})^2]
\end{aligned} \quad (4)$$

After Hu, several studies have explored further methods to compute moments invariant. In 1993, Chen published a paper introducing a convenient procedure to calculate the moments invariant along an object boundary. These moments are called improved moments invariant and are a reformation of Hu's moments. In this case, the one dimensional moment of order (p+q) over a general line is defined by (5):

$$m_{pq}=\int_C x^p y^p dl \quad (5)$$

The recognition of objects using moments of outlines are also possible and may lead to some simplification in computation when compared to the area moments.

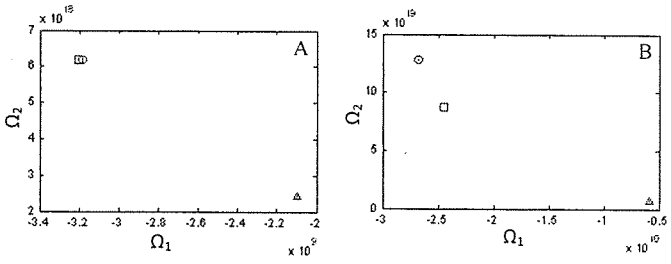

Figure 6. Plot of the first two improved moments invariant for three geospatial datasets: □ FS, ○ OS and △ OSM

For each sample data set, several polygons from the maps have been extracted. The E, N coordinates of the polygon shapes are stored and then processed by calculating moments on the outline of each polygon. In this research, the moments were extracted by applying (4) and (5) and then used for shape similarity measurements between different data sets. This has been carried out by calculating the differences between moments. For example, if (K) and (L) are two features from two data sets representing the same object, they produce a set of moment values represented by (k) and (l) and the distance between them can be given as j = k-l. The descriptor values (j) that are obtained from testing the data sets can be used to distinguish between object shapes; this value will be zero if (k) and (l) are identical. On the other hand, if they are different then the magnitudes of the coefficients will give a reasonable measure of difference enabling discrimination between shapes. Fig. 6 shows a plot of the first two improved moments (Ω_1 and Ω_2) for both hard and soft features in Cramlington (A) and Clara Vale (B). The results indicate that there are significant similarities between FS data and OS data sets in Cramlington. In contrast, the study found that most of the sampled OSM data does not match the reference or the OS data sets. Another important finding was that there is a significant separation between most of the FS, OS and OSM representations of the Clara Vale data sets. However, it is clear that the distance between FS and OS data sets is less than that between OSM and the other data sets.

III. SEMANTIC ACCURACY TEST

Accuracy measurement also covers the semantic data associated with the measured coordinate data set. Testing using standard confusion matrices was attempted, but problems resulted due to inconsistent classification schemes, among the OS and the OSM categories. Thus, a further study of the ontologies of each data set was undertaken, and a measure of 'closeness' of attribute information was attempted. The study is in the process of undertaking semantic distance tests in order to additionally assess the possibilities of data integration from a range of 'crowdsourced' and official sources. Semantic distance measures semantic similarity of two categories by counting the number of links that connect two concepts within the hierarchical relations. This is a structural approach, in which the more similar categories have lower numbers of links between them. An additional feature based model will also be applied for measuring semantic similarity from different classification systems. This approach will be based on two relations, namely intersection and difference. The semantic similarity can be measured by applying those two functions to features and their weighted depth in the hierarchy of classes of features (Feng and Flewelling, 2004).

IV. CONCLUSION

This paper presents a methodology to assess shape and positional quality among data sets for integration purposes. Different methods have been used for obtaining useful descriptions of geometric accuracy assessment (positional accuracy and shape fidelity). The NSSDA has been chosen to select and analyse tested points, then calculate the RMSE values, yielding comparative measures of positional accuracy. A second procedure, for line feature comparison, has compared a tested source to a reference source by establishing buffers around the lines in both data sets that can be analysed using simple statistics which consider the relative uniformity of such buffer zones (union, intersection etc.). Further shape metrics, including moments invariant, have also been applied to evaluate polygon accuracy. The results of this analysis show that the accuracy of OS data is very close to the reference FS data set. On the other hand, the analysis found that the positional and shape accuracy of OSM data does not match the reference or the OS data sets. Future steps for this study will more closely examine semantic similarity by encoding classifications as XML schema, and extend the methodology to other parameters, such as temporal accuracy. Completeness and lineage will also be incorporated into the integration model. Additional data sources beyond field surveyed information, such as images (e.g. Flickr), textual descriptions and user updating (e.g. TomTom MapShare) can also contribute to 'crowdsourced' data sets and their accuracy needs to be examined.

REFERENCES

Chen, C.-C. (1993). Improved moment invariants for shape discrimination. *Pattern Recognition.* 26 (5), 683-686.

Congalton, R. G. and Green, K. (2009). *Assessing the accuracy of remotely sensed data: priciples and practices.*USA: Taylor & Francis Group.

Devillers, R. and Jeansoulin, R. (2006). *Fundamentals of spatial data quality.* Great Britain: Antony Rowe Ltd, Chippenham, Wiltshire.

Feng, C. C. and Flewelling, D. M. (2004). Assessment of semantic similarity between land use/land cover classification systems. *Computers, Environment and Urban Systems.* 28 (3), 229-246.

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

Hu, M.-K. (1962). Visual pattern recognition by moment invariants, *IRE Transactions on Information.* 8,179-187.

Omran, E. E. and van Etten, J. (2007). Spatial-data sharing: applying social-network analysis to study individual and collective behaviour. *International Journal of Geographical Information Science.* 21(6), 699 - 714.

Paul, A. Z. (2008). Positional accuracy of spatial data: non-normal distributions and a critique of the national standard for spatial data accuracy. *Transactions in GIS.* 12(1), 103-130.

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.