Position Paper for the GIST Workshop on 'The Role of VGI in Advancing Science'

VGI data quality: a dynamic way of assessing OpenStreetMap's accuracy

Thomas Koukoletsos, PhD Student – UCL, thomas.koukoletsos.09@ucl.ac.uk

The rapid growth of crowdsourced or Volunteered Geographical Information (VGI) has led to the creation of a contemporary source of spatial data, which theoretically has two major advantages: it is free and potentially up to date. Practically, however, there are several implications. The data is not always or literally free, it is not up-do-date universally, and definitely it cannot be used by everyone or be provided from everyone due to skills and knowledge limitations.

However, one of the hard-to-tackle challenges of VGI is its quality because the credibility of the contributors, as well as the accuracy and the completeness of the provided data are unknown. This makes decision makers reluctant to use VGI, no matter how attractive source of Geographic Information it may be. Existing spatial data quality evaluation tools consider datasets of having a uniform completeness or positional, temporal and thematic accuracy, which make them unsuitable for the inherently heterogeneous VGI. New or revised quality evaluation methods that could take into consideration the spatial variations of VGI are necessary, driven by the question 'is this data source fit for purpose X in area Y?' rather than the more general 'is this data source fit for purpose?'. This paper presents part of a PhD research project aimed at proposing the development of an automatic suite of methods for quality evaluation of VGI.

As a case study, OpenStreetMap (OSM) is a VGI project of growing significance, importance and world coverage. Its significance urged researchers to comment on the necessity for quality assessment tools (Haklay & Weber, 2008; Haklay, 2008; Goodchild, 2008), and in terms of its spatial accuracy research already carried out provided results better than the anticipated in the selected areas that it was tested (Haklay, 2008). However, the previous researchers on spatial quality of OSM addressed specific areas of limited size and their results cannot be representative of the whole dataset due to VGI heterogeneity. In addition, the simplified version of Goodchild's and Hunter's (1997) methodology that they followed provided fixed results that cannot be applied for other purposes with stricter or looser accuracy requirements. Finally, lack of automation hinders repetition of the procedure over time, and if we consider the dynamic and potentially continuous VGI contribution, by the time that the analysis is completed, results may have already become obsolete.

The first tool of the proposed procedure will address positional accuracy and completeness. Three inputs are considered: OSM data, data of higher quality from Ordnance Survey (OS) and an appropriate tessellation according to which the data will be split and individually tested, using an iterative, dynamic and interactive method of comparison. The tessellation can refer either to areas of specific size, or to data contributed by specific users, tackling the user-credibility issue. Results are provided for every tessellation cell individually, giving a more detailed quality insight of VGI. The same procedure can be carried out in the future, when one or both of the datasets are updated.

The comparison of common point features between the two datasets is accomplished by comparing their coordinates. This method is straightforward, adopted in many similar cases and is even standardised (FGDC-STD 1998a, FGDC-STD 1998b). The challenge, however, lies in identifying the same semantic points in the two datasets in order to make the comparison possible. The problem will be addressed using a proximity checking algorithm.

Comparing linear features is more difficult; there is no standardised procedure, however researchers have

addressed the problem from different aspects. The methods selected here have been proposed by Goodchild and Hunter (1997), Hunter (1999) and Tweite and Langaas (1999). An increasing buffer is applied on the 'reference' linear dataset and the percentage of the 'tested' dataset that is contained in the buffer is calculated. The buffer size can be defined by the user according to the user's accuracy requirements. The comparison is carried out for each grid cell. Accordingly, the inverse procedure can be selected; for a desired matching percentage and a user-defined increment step of buffers, an iterative procedure begins and results of matching percentage for each buffer will help the user to decide if OSM data in a specific area is fit-for-purpose, or if compromises have to be made. The buffer size can be regarded as the spatial accuracy of OSM.

The words 'tested' and 'reference' regarding the datasets are deliberately not assigned to a specific dataset, because the comparison should be possible both ways: to test VGI against higher quality data, and vice versa. This is achievable by applying the buffer in both the datasets (Tweite and Langaas, 1999), which allows to define where OSM has better coverage than OS dataset. Results refer either to the calculation and analysis of the length of the tested line inside the buffer (Goodchild and Hunter, 1997; Hunter, 1999), or to the calculation and analysis of the area of the created polygons when buffer is applied in both the datasets (Tweite and Langaas, 1999). Visualisation of the results includes positional accuracy for each grid cell, data existing only in one of the two datasets, matching percentage achieved in each of the increasing buffer steps etc.

The proposed suite of methods can also be applied to other vector VGI projects. It can provide a continuous evaluation of VGI, it can answer a customisable for different areas and cases 'fit-for-purpose' question and it can easily be carried out over time simply by repeating the process with the updated data. Additionally, it can provide information on missing data from the datasets. Finally, by defining a localised positional accuracy of VGI, questions about the credibility of the user or the quality of the equipment used become of minor importance. Future work will include tools to address temporal and thematic accuracy.

References

FGDC, 1998a. FGDC-STD-007.1-1998 *Geospatial positioning accuracy standards. Part 1: Reporting methodology*. Virginia, USA: Federal Geographic Data Committee

FGDC, 1998b. FGDC-STD-007.3-1998 *Geospatial positioning accuracy standards. Part 3: National standard for spatial data accuracy.* Virginia, USA: Federal Geographic Data Committee

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 DOI: 10.1080/136588197242419

Goodchild, M.F., 2008. *Assertion and Authority: The Science of User-Generated Geographic Content*. Available at: http://www.geog.ucsb.edu/~good/papers/454.pdf

Haklay, M. and Weber, P., 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing,* 7(4):12-18. DOI: 10.1109/MPRV.2008.80

Haklay, M., 2008. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, In *Environment and Planning* (under review).

Hunter, G.J., 1999. New Tools for Handling Spatial Data Quality: Moving from Academic Concepts to Practical Reality. *URISA Journal,* 11(2):25-34.

Tveite, H. and Langaas, S., 1999. An accuracy assessment method for geographical line data sets based on buffering. *International Journal of Geographical Information Science*, 13(1):27-47. DOI: 10.1080/136588199241445