Good Practice Guidelines for Assessing VGI Data Quality

[1,2]Cidália C. Fonte, [3]Lucy Bastin, [4]Linda See, [5]Giles Foody, [6]Jacinto Estima

[1]Department of Mathematics,
University of Coimbra,
Coimbra, Portugal
cfonte@mat.uc.pt

[2]Institute for Systems Engineering
and Computers at Coimbra
(INSEC Coimbra)
Coimbra, Portugal

[3]School of Engineering and
Applied Science,
Aston University,
Birmingham, UK
l.bastin@aston.ac.uk

[4]International Institute for Applied
Systems Analysis (IIASA)
Laxenburg Austria
see@iiasa.ac.at

[5]School of Geography
University of Nottingham
Nottingham, UK
giles.foody@nottingham.ac.uk

[6]Information Management
School (IMS)
Universidade Nova de Lisboa
Lisbon, Portugal
D2011086@novaims.unl.pt

Abstract

Assessing the data quality of Volunteered Geographic Information (VGI) is important for determining the fitness-for-use of VGI for different applications. This paper provides guidelines for good practice that may assist in quality assessment, in particular recommendations on additional data that could be used and procedures that could be implemented that will facilitate the assessment of VGI quality. Finally, the role of protocols is discussed in terms of how they might improve data quality assessment.

Keywords: Volunteered Geographic Information, quality, positional accuracy, thematic quality, credibility

1 Introduction

With open access to high resolution satellite imagery, the proliferation of mobile devices and the interactivity of Web 2.0, mapping has been transformed. The idea of citizens as sensors [7] has become a reality, and georeferenced data of many different forms are now being collected and digitized. Referred to by some as Volunteered Geographic Information (VGI) [7], this relatively new source of data is growing steadily. OpenStreetMap (OSM) is the most successful example of a VGI initiative that has mapped parts of the world in very rich detail [11] while georeferenced photograph collections such as Flickr and Panoramio are exploding in volume.

There is considerable potential for using VGI for different applications, e.g. to complement authoritative data collected by mapping agencies or to provide new sources of information for validating land cover and land use maps [4, 6], yet there are barriers to adoption, in particular questions around the quality of the data and the credibility of the volunteers. The assessment of VGI quality is therefore fundamental for determining its fitness-for-use in specific applications. Several aspects of data quality can be assessed, such as positional quality (precision and accuracy), thematic quality (level of detail and accuracy), credibility of the data and of the volunteer, completeness, currency and logical consistency.

In this paper we present a set of good practice guidelines that may be useful for assessing VGI quality with a focus on positional and thematic accuracy, as well as on volunteer credibility. Finally, we discuss the role that data collection protocols could bring to the quality assessment of VGI in the future.

2 Criteria for the assessment of VGI quality

There are a number of criteria which can be used to assess the quality of spatial data [8], which can also be applied to VGI. The first and most frequently used criterion to examine VGI quality is positional accuracy; see e.g. various studies on positional accuracy of OSM [1, 9, 10, 14]. Positional accuracy or quality of VGI is usually associated with data georeferenced as points, lines or areas, such as road junctions or buildings. Portable data collection technologies are now capable of delivering a spatial precision exceeding ±10m [2]. When combined with the increasing availability of Web-based maps and very high resolution satellite imagery for digitizing, it is not surprising that the positional accuracy of VGI has increased, and is now appropriate for a wide range of applications. In fact, VGI is often acquired at a level of precision far finer than that needed by national mapping agencies. The positional accuracy of points representing geotagged photographs may also be considered and analysed, particularly since the location recorded by the device that took the photograph and the subject depicted in the photograph may be offset by a certain distance.

Another criterion for the evaluation of VGI is thematic quality. This assesses the accuracy of classes or thematic tags associated with specific locations or objects placed in geographical space, such as classes assigned to pixels in a land cover map, a tag assigned to a linear entity or a polygon, as for example a highway, river, building or green area. Some research has examined how well volunteers can classify satellite imagery from the Geo-Wiki application [3, 15] but more research is needed in this area.

Both the positional and thematic quality of spatial data are traditional criteria for quality assessment. However, the third aspect considered here is credibility, which is particularly relevant to VGI, both in terms of the credibility of the data and the credibility of the person who provided the VGI. The credibility of a person (e.g. a volunteer) is the degree to which the information provided by that user can be trusted. Credibility of the data is used to denote the quality of the information contained in that observation.

Other criteria for the assessment of spatial data quality, but not dealt with in this paper, include: completeness, i.e. the degree to which an area is covered by one or more features in space or time; currency, which refers to how up-to-date the data are; and logical consistency, or the assessment of the data with reference to other data, either from the same source or from independent (and sometimes authoritative) data sources. These aspects will be dealt with in more detail in a future publication.

3 Good practice guidelines for quality assessment

In many cases, the assessment of VGI quality requires additional data such as metadata, local knowledge or comparison with other sources of GI, both volunteered and authoritative. In this section we make some initial steps towards the definition of good practice guidelines for what additional data should be recorded or used, which can add value to the assessment of VGI quality, or what additional procedures could be implemented. These guidelines specifically address positional and thematic quality, and volunteer credibility, and result from an analysis of the several sources of error and uncertainty, methodologies used to assess quality and the data that may be collected in VGI initiatives.

3.1. Guidelines for assessing positional quality

3.1.1 Collect information from multiple contributors

If several contributors provide information about the same feature or phenomenon, an assessment of positional consistency can be made. Additionally, this information may be used to conflate data, providing the location based on data contributed by all volunteers. For example, it has been shown that the positional accuracy of roads in OSM improved with an increase in the number of contributors, illustrating that Linus' Law applies to this source of VGI [10].

3.1.2 Store historical information

VGI is dynamic since information is continuously being added or updated. We would therefore recommend that records of historical data provided by the volunteers should be retained. This would enable the identification of change and assess the stability of positional information over time.

3.1.3 Store data about the methodology used to determine the position

The position of features or phenomena may be determined using several procedures, which have different degrees of reliability, precision and accuracy. These are described below with guidelines for good practice.

Positioning over a geo-referenced image

The location of a phenomenon may be assessed by positioning it on top of a satellite or aerial image. It would then be useful to record:

a) the nature of the image with particular regard to issues such as its spatial resolution and spectral composition, which will provide information about the difficulty the volunteer had in identifying the features, and the accuracy of the position provided;
b) The date when the image was collected, including the year, month, day and even time of day. This may give information about the season, which may influence, for example, the condition of the vegetation, the phenology, the degree of human occupancy in touristic regions, the amount of traffic depending on, e.g., whether it was rush hour or not, the amount of light, and the direction of shade in the image, etc.;
c) Whether specific instructions were given to the volunteers about where to locate the features. Ideally, information regarding where to place the photographs should be provided, including the location from which the photograph was taken and its orientation. Other information regarding points of interest should also be specified, e.g. whether the points correspond to the building centroid or to the entrance of the building that gives access to the point of interest.

Positioning over a map

If the phenomena are positioned using the information provided by a base map, it is useful to record:

a) The type of map used, such as a map made by volunteers, a thematic map, or a topographic map created by a national mapping agency;
b) The map scale and/or the minimum mapping unit, both of which provide a measure of the maximum precision attainable;
c) The date of the map, which is important for determining the currency of the data;
d) The level of generalization of the features and classes present in the map.

These data may provide information about the reliability of the base map used.

Positioning using GNSS measurements

When measurements are made using receivers from the Global Navigation Satellite System (GNSS), the measurement procedure and the upload of the position may be done in two ways: the measurement is made automatically when the data are collected and uploaded, e.g. when taking a picture and uploading an EXIF file; or it may be collected separately and uploaded later. The second approach may be considered as less reliable, since there may be mistakes in the insertion of the positional information, or the volunteer may have moved between where they measured the phenomenon and where they determined the location. Additional information to assess quality, which would be useful if recorded, includes:

a) The type of GNSS receiver used;
b) The number of measurements used to determine the location;
c) The date and time of the measurement, which enables the determination of the Dilution of Precision associated with the measurements;
d) The number of satellites used for positioning. Since the date and time of the measurements enable the identification of the available satellites, some of them may be obstructed by features such as vegetation, buildings or even the terrain.

Conflating data provided by volunteers

When the position of the feature is obtained through the conflation of data provided by several volunteers (or from data involuntarily obtained, e.g. from mobile phones), it would be useful to know:

a) The amount of data for a given feature that has been used to obtain the indicated location;
b) The degree of variability of the data used to determine the most probable value;
c) The dates and times associated with the collection of data about a particular feature.

Alternatively, interested users can be provided with access to the raw data.

3.2 Guidelines for assessing thematic quality

The quality control of thematic data may be facilitated if some procedures are implemented during the data collection process, such as:

a) Collecting information from multiple contributors, which enables, as for the positional quality, checking the consistency of the results or assigning a label through data conflation, whenever divergent data are provided, using, for example, latent class analysis [5];
b) Asking volunteers for a confidence rating with the tags;
c) Keeping historical information for the same reasons as outlined above for positional quality;
d) Collecting additional information such as the amount of time taken to assign a label and/or if the volunteer used instructions or consulted training materials between assessing a point and providing a label. This may provide indirect information about the confidence of the volunteer in the assignment of the tag or label. Other metadata might be consulted, e.g. the prevailing atmospheric condition at the time of data collection, which may be relevant for the collection of some biological or environmental data.

Collecting additional data with georeferenced photographs would also be very useful, e.g. the orientation of the photograph, a description of whether the surrounding area is homogeneous or heterogeneous, the date when the photograph was taken, data about the exposure of the photograph and the type of camera used, etc. This type of additional metadata would provide valuable information to assist in applications such as land cover and land use mapping.

3.3 Guidelines for assessing volunteer credibility

The credibility of the volunteer may be used as an indicator of the reliability of the data provided. Volunteer credibility may be separated into volunteer expertise and volunteer trustworthiness. A volunteer’s expertise may be assessed using metadata about the volunteer, such as age, education, profession or interests. To assess the volunteer's trustworthiness, a range of approaches may be used including:

a) Use control information, such as test sites where information provided by experts or selected volunteers is available, which can be used to assess the contributions of each volunteer; see e.g. [15];
b) Use historical data provided by the volunteer, such as the number of times their contributions were corrected by other volunteers, selected volunteers or experts;
c) Use information about the where the volunteer is located. Here we assume that the closer a volunteer is to the location of the data that was uploaded by them, the more reliable the data will be. An example of the value of local knowledge is shown in [12].

3.4 Generic good practice guidelines

Some general practices may be implemented that can contribute to the production of more reliable information, such as:

a) Implement automatic means to check the data provided by the volunteers whenever possible. These approaches may use additional data or metadata and make an automatic check of whether the data provided are likely to be correct. To this aim, for example, the geographic approach proposed in [7] may be used, which consists of comparing several sources of data to identify if inconsistencies are present. For example, check if a traffic accident was posted on or near a road, or if a forest fire occurred in a region where there is actually a forest.
b) Enable volunteers to identify erroneous contributions (regarding positions or attributes). This may provide valuable information about the contributors themselves and also about difficulties in assigning classes or the credibility of locations of phenomena. Information that could be corrected by volunteers includes: incorrect georeferencing of photographs; identifying erroneous road types or building use; or a misinterpreted class;
c) Enable discussions among the volunteers whenever difficulties are found, such as the best class to assign to a particular location. This may enable the sharing of locally relevant information; improve the understanding of ontologies; self-correction; and quality control.

4 The Role of Protocols

The use of strict protocols for data collection in citizen science projects is quite normal, particularly in the areas of biodiversity and conservation [13]. VGI, on the other hand, has mainly evolved as a bottom up initiative. A prime example is OSM, which is now a self-organizing community, and where scientific use of the data has not been the primary driver for data collection. OSM has guidance on how features can be tagged but it does not impose any minimum data collection protocols on volunteers. Although the community has freedom in what they choose to map, it also means that the data are more difficult to assess in terms of quality. We face similar challenges from georeferenced photographs from sites such as Flickr and Panoramio although some sites such as the Degree Confluence Project, Geograph and the Oklahoma Field Photo Library do require users to conform to some minimum set of protocols. The establishment of protocols may, on the one hand, provide valuable information which may make the data useful for additional applications. However, if protocols are imposed that are too demanding, they might demotivate the volunteers to contribute. Therefore, a balance needs to be identified so that protocols are not seen as restrictive but rather a way to help users in providing higher quality data. Moreover, protocols can be specified from a minimum set for a particular application to varying levels of good to a best set of protocols, where the latter two contain more detailed information.

5 Conclusions

Quality assessment of VGI remains one of the most important issues for determining the fitness-of-use of VGI for different applications. This paper presented some good practice guidelines that may provide valuable information to assess the positional and thematic quality as well as the volunteer credibility of VGI, enabling its potential utilization for a wider range of applications; this work is only a starting point and will continue to be developed further in the future. However, the implementation of some of these guidelines might require the definition of protocols for data collection, which has advantages and disadvantages, and therefore needs to be defined with caution.

Acknowledgements

The authors would like to acknowledge the contribution of COST Action TD1202. This work was also partly supported by a STSM Grant from COST Action TD1202. http://www.citizensensor-cost.eu.

References

[1] Canavosio-Zuzelski, R. et al.: A photogrammetric approach for assessing positional accuracy of OpenStreetMap© roads. *ISPRS International Journal of Geo-Information*, 2(2):276–301, 2013.

[2] Coleman, D.: Volunteered geographic information in spatial data infrastructure: An early look at opportunities and constraints. In: Rajabifard, A. et al. (eds.) *Spatially Enabling Society: Research, Emerging Trends and Critical Assessment*. pp. 1–18 Leuven University Press, Leuven, Belgium, 2010.

[3] Comber, A. et al.: Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, 23:37–48, 2013.

[4] Estima, J., Painho, M.: Flickr geotagged and publicly available photos: Preliminary study of its adequacy for helping quality control of Corine land cover. In: Murgante, B. et al. (eds.) *Computational Science and Its Applications – ICCSA 2013*. pp. 205–220 Springer Berlin Heidelberg, 2013.

[5] Foody, G.M. et al.: Assessing the accuracy of Volunteered Geographic Information arising from multiple contributors to an internet based collaborative project: Accuracy of VGI. *Transactions in GIS*, 17(6): 847–860, 2013.

[6] Foody, G.M., Boyd, D.S.: Using volunteered data in land cover map validation: Mapping West African forests. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(3):1305–1312, 2013.

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

[8] Guptill, S.C., Morrison, J.L.: Elements of Spatial Data Quality. Elsevier Science Limited, 1995.

[9] Haklay, M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37:682–703, 2010.

[10] Haklay, M. et al.: How many volunteers does it take to map an area well? The validity of Linus' Law to volunteered geographic information. *The Cartographic Journal*, 47(4):315–322, 2010.

[11] Jokar Arsanjani, J. et al.: The emergence and evolution of OpenStreetMap: a cellular automata approach. *International Journal of Digital Earth*, 8(1):74–88, 2015.

[12] De Leeuw, J. et al.: An assessment of the accuracy of volunteered road map production in Western Kenya. *Remote Sensing*, 3(12):247–256, 2011.

[13] Munson, M.A. et al.: A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution*, 1(3):263–273, 2010.

[14] Neis, P. et al.: The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(4):1–21, 2011.

[15] See, L. et al.: Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE*. 8(7): e69958, 2013.