Crowdsourced Geospatial Data Quality: Challenges and Future Directions

Anahid Basiri[a]*, Muki Haklay[b] Giles Foody[c], Peter Mooney[d]

[a]Centre for Advanced Spatial Analysis, University College London, London, The United Kingdom, a.basiri@ucl.ac.uk;

[b]Department of Geography, University College London, London, The United Kingdom; m.haklay@ucl.ac.uk; ;

[c]School of Geography, The University of Nottingham, Nottingham, The United Kingdom; giles.foody@nottingham.ac.uk;

[d]Department of Computer Science, Maynooth University, Maynooth, Ireland; peter.mooney@nuim.ie;

Crowdsourced Geospatial Data Quality: Challenges and Future Directions

Introduction

A decade ago, Volunteered Geographical Information (VGI) was identified as a new source of information that would blur the traditional boundary between producers and the consumers of data (Goodchild, 2007). This form of information has been recognised by multiple names, including crowdsourced geospatial data (Heipke 2010) and user-generated geographic content (Fast and Rinner 2014), to name but a few. Many applications and services benefit from user-generated content contributed by a wide range of users through crowdsourcing projects. VGI made it possible for a much wider group of contributors to create and share geographical information. Despite the success and popularity of many VGI projects, such as OpenStreetMap (OSM), researchers continue questioning the reliability and fitness for use of crowdsourced data (Senaratne et al. 2017, Basiri et al. 2016a, Arsanjani et al. 2015, Foody et al., 2015, Koukoletsos et al. 2012, Haklay et al., 2010, Salk et al. 2015).

The belief that VGI is contributed by the public, some contributors with little experience and expertise of geospatial data, might have contributed to the perception of the unreliability of this data source. Such issues have impeded the adoption of crowdsourced geospatial data in several projects. While one can argue the importance of the individuals and their levels of expertise based on the concepts of "the wisdom of the crowd" and the collective decision, some questioned the representativeness, i.e. the structure of the crowd and "power of the elites", in many crowdsourcing projects (Leszczynski and Elwood, 2015, Ballatore and De Sabata, 2018).

This special issue of the International Journal of Geographical Information Science looks at the challenges and future directions of crowdsourced geospatial data with particular attention to the issues stem from data quality and biases of VGI. This editorial highlights how these issues are discussed and addressed by the articles of this special issue and how the papers highlight emerging technologies, concepts, platforms, debates, and methodologies and techniques within VGI and suggest future research directions. This special issue gathered papers on the topics of crowdsourced geospatial data quality (Ballatore and Arsanjani, 2018), thematic uncertainty and consistency across data sources (Hervey and Kuhn, 2018), trust issues within VGI (Severinsen et al., 2019), and contributors behaviour and interactions (Truong et al., 2018).

Crowdsourced Data Quality Challenges

VGI Data Quality Issues

Crowdsourced geographic data quality has been the main core part of many research and studies; Fonte et al. (2015), Senaratne et al. (2017), Fonte et al. (2017), Basiri et al. (2016a), Antoniou and Skopeliti (2015), Goodchild and Li (2012) and several other studies reviewed VGI's quality assessment and assurance methods. There are several ways to classify the quality assessment methods but the following categories are commonly mentioned in literature with different titles; (a) comparing data against "authoritative" spatial data (Dorn et al., 2015; Koukoletsos et al., 2012) (b) user's and/or machine learnt rules and patterns for checking the entries (Basiri et al., 2016a, Ali et al., 2014; Jilani et al., 2013; Neis et al., 2012; Basiri et al., 2016b; Leibovici et al., 2017) (c) gatekeeping and weighting users' entries (e.g. with respect to the their experiences, expertise, proximity, number of their entries, history and changesets) (Ciampaglia et al., 2018; McGreavy et al., 2017). Having a better understanding of the quality of VGI may help the adoption of crowdsourced geospatial data in some projects as the perception of unreliability may impede the adoption. To address the issues on trust, transparency and reliability Truong et al. (2018) looked at the contributors behaviour and their interactions. They qualified the behaviour of contributors to OpenStreetMap (OSM) through a multigraph approach to reproduce contributor's interactions in a more comprehensive way. Ballatore and Arsanjani (2018) looked at the origin and development of Wikimapia and discussed some aspect of Wikimapia, including the project's intellectual property and strategies for quality management. Hervey and Kuhn (2018) explored uncertainty with locational data obtained from social networks. They presented a taxonomy of things that can be located from social network posts and a means to describe them to users. Severinsen et al. (2019) present a formulaic model to addresses VGI quality issues by quantifying trust in VGI. Their 'VGTrust' model assessed information about a data author, and the spatial and temporal trust associated with the data they create in order to produce an overall VGTrust rating metric.

VGI Biases

While quality issues of crowd-sourced data have been studied widely, the identification and estimation of biases in crowd-sourced projects have not received the same attention. This is due mainly to the lack of availability of the (geo-) demographic data of the contributors, which is either unrecorded (e.g. OpenStreetMap) or inaccessible due to commercial interest (e.g. in the now defunct Google MapMaker). Therefore understanding the impacts of demographic biases on crowdsourced maps is challenged by a lack of data on these data (Mullen et al., 2015; Basiri et al., 2018; Gardner and Mooney; 2018, Haklay, 2016; Gardner et al., 2018). Millar et al (2018) looked at the biases and studies the lack of citizen science monitoring programs. The study focuses on natural and demographic biases related to the location, accessibility, size and general attractiveness of lakes in Ontario.

Any VGI project is biased in one or more ways (Basiri et al., 2018). At the first glance, it seems that all the data contributed through VGI projects are “voluntary response samples”, which are always biased as they only include people who have chosen to volunteer (DeMaio, 1980). Whereas a random sample would need to include people whether or not they choose to volunteer (Goyder, 1986). Thus inferences from a voluntary response sample are not as credible as conclusions are based on a random sample of the entire population. While crowdsourcing projects are technically open to the whole population, and of course, anyone should be able to contribute, recent studies (Mullen et al., 2015; Gardner et. al. 2018; Zhu et al., 2017, Yang et al., 2016) have shown that even the most popular crowdsourced projects, such as OSM, are biased by the contribution patterns of its contributors, i.e. that a small percentage of the community contribute the greatest proportion of activity (the ‘long tail effect’ or 90-9-1 rule (Haklay, 2016)). Ballatore and Arsanjani (2018) studied the popularity of the project using behavioural data from Google Trends and compared the geography of Interest in Wikimapia with OpenStreetMap, from a temporal and spatial perspective. And found while OpenStreetMap attracts more interest in high-income countries, Wikimapia emerges as relatively more popular in low- and middle-income countries, countering the received notion of VGI as a Global North phenomenon. Therefore, we might questions the use of the terms “crowd” and “public” used in many crowdsourcing and public participatory projects by virtue of this skewed pattern of participation. This excludes the projects which may require a relatively higher experience level, access to some resources, or may limit participation to a specific geography or particular time interval due to the nature of the project (Morschheuser et al., 2018).

In addition to voluntary response bias, the volunteers, as individuals, can have different aspects and levels of quality of judgement and decision making (Hammond, 2000). Their decisions, opinions, and preferences could be significantly represented and/or influence their contribution (e.g. data). Although there are some arguments based on the concepts of "the wisdom of the crowd" trying to undermine or counter- balance the impacts of the individuals' biases on the collective decision, there are two challenges to this notion: Firstly, representativeness, i.e. the structure of the crowd and "power of the elites", in many crowdsourcing projects have been questioned (Comber et al., 2016), (See et al., 2013). For example, both Elwood (2010) and Leszczynski and Elwood (2015) have problematized participation biases in VGI on the grounds of a failure of crowdsourced mapping projects to represent the interests of the wider public, specifically those of women. Similarly, Ballatore and De Sabata (2018) have explored the extent to which VGI are representative of the wider population of the geospatial units they represent.

The issues of representation could, therefore, be an issue in terms of biases, however, some believe that the super active contributors are experts and so it is better to leave some decisions in their hands. While Giles (2005) and Rajagopalan et al. (2011) showed that collective decision-making can be more accurate than experts' comments, accuracy does not necessarily show all the aspects of quality and might not be even loosely correlated with potential bias. In terms of biases Greenstein et al. (2017) found the knowledge produced by the crowd are not necessarily less biased than the knowledge produced by experts. Ciampaglia et al. (2018) confirmed this by using Wikipedia contents, however, they found both biases and data quality could be moderated if substantial revisions and supervisions (of the gatekeepers) were implemented.

The second challenge to the notion of the "wisdom of the crowd", is the process of many VGI projects which is not based on a collective decision but instead on crowd “participation”. The difference is relatively implicit but highly important; the participants do not vote for/against every single decision or entry. The collection of individual decisions does not necessarily mean the collective decision making. Therefore the wisdom of the crowd may not be relevant to such projects as the individual bias can remain at micro-level. As the crowd makes decisions individually in a participatory project, the results of an individual’s contributions could be biased. Therefore for these projects the case of “given enough eyeballs, all bugs are shallow” (Raymond 1998) is no longer valid as there is not enough revision/votes for each piece of information contributed by volunteers. Ciampaglia et al. (2018) found that crowd-sourced content can also produce a large sample with a great variety of biased opinions.

Future Directions

The papers within this special issue looked at some of the challenges and issues of crowdsourced geographic data, including data quality, biases, and trust issues. They also provided some solutions to either address or have a better understanding of the implications of these issues. It seems that research focus of research on VGI has been moving towards the structure of the ‘crowd’ and volunteers (geo-)demographic biases and the impact of having such biases in different VGI projects and research on how to promote diversity of contributors communities, addressing the issues of transparency and trust while protecting the privacy of the contributors, working on intellectual property of crowdsourced data and projects. It seems that future research look at VGI beyond just a way to create maps but as a complex but more democratic, reproducible and open but reliable system engaging society and promoting diversity, collaborations, and wider engagement.

References

Ali, A.L., and Schmid, F., (2014), Data quality assurance for volunteered geographic information. In International Conference on Geographic Information Science, pp. 126-141. Springer, Cham.

Antoniou, V., and Skopeliti, A. (2015). Measures and indicators of VGI quality: An overview. ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, 2, 345.

Arsanjani, J. J., Mooney, PA., Schauss, A., (2015), Quality Assessment of the Contributed Land Use Information from OpenStreetMap versus Authoritative Datasets. In OpenStreetMap in GIScience, Experiences, Research, and Applications (LNCS), edited by J. Jokar Arsanjani, A. Zipf, P. Mooney, M. Helbich, 37–58. Berlin Heidelberg: Springer.

Ballatore, A., and Arsanjani, J.J., (2018), Placing Wikimapia: an exploratory analysis. International Journal of Geographical Information Science: 1-18.

Ballatore, A., and De Sabbata, S. (2018), Charting the geographies of crowdsourced information in Greater London, Proceedings of the AGILE Conference, Lund, Sweden.

Basiri, A., Haklay, M., Gardner, Z. (2018). The Impact of Biases in the Crowdsourced Trajectories on the Output of Data Mining Processes. Association of Geographic Information Laboratories in Europe (AGILE).

Basiri, A., Jackson, M. Amirian, P., Pourabdollah, A., Sester, M., Winstanley, A., Moore, T., and Zhang, L., (2016a). Quality assessment of OpenStreetMap data using trajectory mining. International Journal of Geospatial information science 19, no. 1: 56-68.

Basiri, A., Amirian, P., Mooney, P. (2016b), Using crowdsourced trajectories for automated OSM data entry approach. Sensors, 16(9), 1510.

Ciampaglia, G., Nematzadeh, A., Menczer, F., Flammini, A., (2018), How algorithmic popularity bias hinders or promotes quality. Scientific reports8, no. 1: 15951

Comber, A., Mooney, P., Purves, R. S., Rocchini, D., & Walz, A. (2016), Crowdsourcing: It matters who the crowd are. The impacts of between group variations in recording land cover. PloS one, 11(7), e0158329.

Dorn, H., Törnros, T., Zipf, A., (2015), Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. ISPRS International Journal of Geo-Information 4, no. 3: 1657-1671.

DeMaio, T.J. , (1980), Refusals: Who, where and why. Public Opinion Quarterly 44, no. 2: 223-233.

Elwood S (2010). Geographic information science: Emerging research on the societal implications of the geographical web. Progress in Human Geography, 34(3), 349-357.

Fast, V. and Rinner, C., (2014), A systems perspective on volunteered geographic information. ISPRS International Journal of Geo-Information, 3(4), pp.1278-1292.

Fonte, C., Vyron, C., Bastin, L., Estima, J., Arsanjani, J.J., Laso Bayas, J.C., See, L., Vatseva, R., (2017), Assessing VGI data quality. Mapping and the citizen sensor: 137-163.

Fonte, C.C., Bastin, L., See, L., Foody, G. and Lupia, F., 2015. Usability of VGI for validation of land cover maps. International Journal of Geographical Information Science, 29(7), pp.1269-1291.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Comber, A. (2015). Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. The Cartographic Journal, 52(4), 336-344.

Gardner, Z., Mooney, P., Dowthwaite, L., Foody, G., (2018), Gender differences in OSM activity, editing and tagging. Proceedings of GISRUK 2018 Conference, Leicester, 17-20th April.

Gardner, Z. and Mooney, P., (2018). Investigating gender differences in OpenStreetMap activities in Malawi: a small case-study. Proceedings of AGILE Conference, Lund, Sweden, 12-15th June.

Giles, J. (2005), Internet Encyclopaedias Go Head to Head. Nature (438:7070), pp. 900–901.

Goodchild, M F., and Li, L.. (2012), Assuring the Quality of Volunteered Geographic Information. Spatial Statistics, 1: 110–120.

Goodchild, M. F. (2007), Citizens as Sensors: The World of Volunteered Geography. GeoJournal 69 (4): 211–221.10.1007/s10708-007-9111

Goyder, J. (1986), Surveys on surveys: Limitations and potentialities. Public Opinion Quarterly 50, no. 1 (1986): 27-41.

Greenstein, S., and Zhu, F., (2017), Do Experts or Crowd- Based Models Produce More Bias? Evidence from Encyclopædia Britannica and Wikipedia. Forthcoming, Management Information Systems Quarterly.

Haklay, M., (2016), Why is Participation Inequality Important? In Capineri & Huang (eds.) European handbook on crowdsourced geographic Information, Ubiquity Press, pp. 35-45.

Haklay, M., (2016), Why is participation inequality important?. Ubiquity Press.

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. (2010). 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.

Hammond, Kenneth R., (2000), Coherence and correspondence theories in judgment and decision making. Judgment and decision making: An interdisciplinary reader: 53-65

Heipke, C., (2010), Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing 65, no. 6: 550-557.

Jilani, M., Corcoran, P., Bertolotto, M. (2013). Automated quality improvement of road network in OpenStreetMap. In Agile Workshop (action and interaction in volunteered geographic information) (p. 19).

Koukoletsos, T., M. Haklay, and C. Ellul. (2012), Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. Transactions in GIS 16 (4): 477–498.10.1111/j.1467-9671.2012.01304.x

Leibovici, D., Rosser, J., Hodges, C., Evans, B., Jackson,M., Higgins, C. (2017), On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. ISPRS International Journal of Geo-Information 6, no. 3: 78.

Leszczynski A and Elwood S (2015). Feminist geographies of new special media. The Canadian Geographer, 59(1), 12-28.

McGreavy, B., Newman, G., Chandler, M., Clyde, M., Haklay, M., Ballard, H.L., Gray, S., Scarpino, R., Hauptfeld, R., Gallo, G.A., (2017), The Power of Place in Citizen Science. Maine Policy Review 26.2: 94 -95,

Morschheuser, B., Hamari, J., Maedche, A., (2018), Cooperation or competition–When do people contribute more? A field experiment on gamification of crowdsourcing. International Journal of Human-Computer Studies (2018).

Mullen, W.F., Jackson, S.P., Croitoru, A., Crooks, A., Stefanidis, A. and Agouris, P., 2015. Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. GeoJournal, 80(4), pp.587-605.

Neis, P., Goetz, M. and Zipf, A., 2012. Towards automatic vandalism detection in OpenStreetMap. ISPRS International Journal of Geo-Information, 1(3), pp.315-332.

Rajagopalan, M. S., Khanna, V. K., Leiter, Y., Stott, M., Showalter, T. N., Dicker, A. P., Lawrence, Y. R., (2011), Patient-oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database," Journal of Oncology Practice (7:5), pp. 319–323.

Raymond, E. (1998), The Cathedral and the Bazaar. First Monday, http://tinyurl.com/bqfy3s, accessed May 2018

Ricketts, J.A. (1990), Powers-of-ten information biases. MIS Quarterly: 63-77

Salk, C. F., T. Sturn, L. See, S. Fritz, and C. Perger. 2015. Assessing Quality of Volunteer Crowdsourcing Contributions: Lessons from the Cropland Capture Game. International Journal of Digital Earth 2015: 1–17

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C. and Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science, 31(1), pp.139-167.

See, L., Comber, A., Salk, C., Fritz, S., Van Der Velde, M., Perger, C., ... & Obersteiner, M. (2013). Comparing the quality of crowdsourced data contributed by expert and non-experts. PloS one, 8(7), e69958.

Zhu Di, Zhou Huang, Li Shi, Lun Wu & Yu Liu (2017) Inferring spatial interaction patterns from sequential snapshots of spatial distributions, International Journal of Geographical Information Science, 32:4, 783-805, DOI: 10.1080/13658816.2017.1413192

Yang, A., Fan, H., Jing, N., Sun, Y., Zipf, A., (2016), Temporal analysis on contribution inequality in OpenStreetMap: A comparative study for four countries. ISPRS International Journal of Geo-Information 5, no. 1: 5.