UCL DEPARTMENT OF GEOGRAPHY

YEAR 2020-21

MODULE CODE:	**GEOG0163**
MODULE NAME:	**Data, Politics and Society**
COURSE PAPER TITLE:	**Representation Bias Exists in Volunteered Geographic Information- Necessary to Mitigate**
WORD COUNT:	**2775**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☒

GEOG0163 Data, Politics and Society

Representation Bias Exists in Volunteered Geographic Information- Necessary to Mitigate

How to mitigate representation bias in Volunteered Geographic Information platform - take OpenStreetMap as case study

Table of Contents

Blurb

Achieving high data coverage, accuracy and credibility requires understandability of the representation bias of Volunteered Geographic Information (VGI) and methodology of diluting the deviation through data production and consumption processes. In that case, democracy of VGI could be guaranteed, while the inequality of interests for different stakeholders could be avoid.

Introduction

In the context of global informatization, the public population has expressed an explosive interest in creating, assembling and disseminating geographic information through the internet. Volunteered Geographic Information (VGI) appeared as a crowdsourced geospatial approach to allow the public to create geographic information, supplementing formal qualification of official agencies (Goodchild, 2007). In previous research, VGI has been applied to predict forthcoming events or correcting decisions at high resolution with low cost, including land use estimation and traffic system assessment (Jiang et al., 2015 and Yang, 2020). However, according to Kitchin (2014), data science is not neutral, instead, it is a social construct involving assumptions, traditions, political and ethical biases of different individuals together. Therefore, the representativeness of ‘the crowd’ in the VGI approach could significantly affect the dataset’s fairness and further affect the experimental results of scientific research.

Considering this weakness, this article will deeply discuss the origin and corresponding impacts of representation bias within the VGI platform - OpenStreetMap. Among the VGI platforms, OpenStreetMap (OSM) is the most popular one, which can provide updated and freely editable geographic information data (Minaei, 2020). Therefore, the discussion and research on OSM aim to improve its utilization efficiency and also provide some reference for other VGI platforms in reducing the representation bias. After the investigation of representation bias

characteristics, suggestions to mitigate the bias through data production and data consumption processes will be proposed. From the data production perspective, improving engagement motivation by training local people and increasing the understandability of the project could possibly avoid the representation bias by extending data coverage. This method could be integrated with the technological contributions from global community, which could further mitigate the representation bias by indirectly improving data accuracy and collection convenience. In addition, introducing standardized data and provenance editing policies could probably improve the representative bias under the premise of ensuring citizens' privacy. From the data consumption perspective, triangulating VGI data with external datasets and internal datasets all could assess and mitigate representation bias, although the crosschecking also limited by language restrictions. Another approach is combining geographical thinking with machine learning method, applying importance weighting model to mitigate representation bias but take corresponding risks brought be modelling bias.

Origin and impacts of representation bias

Biased contributors

Previous research has suggested that demographic biases in participation could increase the limitations, leading to inequal geospatial knowledge and imperfect decision-making process. These biases affected by the imbalanced percentage between contributors and contributions, as well as contributors' characteristics, including age, educational background and gender (Gardner et al., 2020). Haklay (2016) had recognized the participation inequality in online systems for user-generated content. In OpenStreetMap, the proportion of registered users without any contribution could reach 70 percent, while the majority of remaining participants choose to infrequently create geographic information, accounting for 29.9 percent and last 0.01 percent population provide most of the information (Budhathoki, 2010).

High contributors with certain backgrounds and interests will impact recorded data types and topics. For example, lack of interest in accessible facilities for disabled persons might probably result in the incomplete datasets and indirectly leads to the low attention of corresponding policies.

Similarly, as mentioned above, the characteristics of contributors are also major driver of representation bias. Among them, gender is the most comprehensive but easily overlooked one, since gender could be seen as non-binary but a collection of feminine and masculine characteristics (Gardner et al., 2020). According to Gilbert et al. (2008), different socio-economic status, lifestyles and interests to computer science and technology make women recognize OSM editing as a time-consuming activity. For example, household duties would limit women's online time and activities (ibid). Since men produced 95 to 98 percent contributions for OSM, failure to represent the public interests of the "crowd", it has been critically claimed to be an unsuccessful practice for democratizing geographic knowledge (Gardner et al., 2020). Under this context, online interaction, concerned facilities and policies naturally skewed to men, which further affect women's participation enthusiasm and hot topics, forming a vicious circle.

Biased targets skewed towards powerful stakeholders

Apart from the highly heterogeneous in terms of data contribution, the heterogeneous in coverage should also be considered. It is essential for governments to encourage OSM development, as the voluntary action can reduce the financial burden and supplement authoritative data to some extent (Minaei, 2020). However, because of the economic backwardness, resource deficiency and less-educated condition in remote regions, inequalities of representativeness appear, in terms of coverage, quality and currency. In other words, the openness of VGI is a relative concept, which remains largely the privilege of people who have access to the internet, possess suitable technology (such as smartphones) and master the related capability to record geographic information (Gardner et al., 2020 and

Goodchild, 2007). Besides, the build-in language bias of most VGI platforms is also a problem, only supporting Roman alphabet and English (ibid.). In that case, public opinion orientation becomes increasingly biased.

In addition, considering the unequal rights controlled by involved stakeholders, VGI-related policing targets also skewed towards urban citizens. According to Huck et al. (2021), contributions could successfully respond to acute emergencies (such as earthquakes or floods) in urban areas, however, leaving under-mapped rural regions suffer from chronic humanitarian crises (such as poverty or conflict). For example, Haidi, an instance of the poorest countries in the world, had little demand for geographic mapping, leading to inadequate representativeness in standard platforms (Huck et al., 2021). However, information sometimes is the least available in most needed areas, as VGI platforms, such as OSM, often become the only available information source for politically unstable areas (Goodchild, 2007). Therefore, it is important to investigate the relevant motivation and barriers for individuals' engagement with VGI schemes, in order to mitigate the digital divide in time if possible.

Biased thematic attributes

The preference for different topics might lead to the phenomenon of insufficient representativeness, while the situation can be improved with appropriate usage and guidance. Generally, the less popular thematic attributes would receive little information considering the representativeness of VGI. As suggested by previous studies of Zhang & Zhu (2018), most public would focus on the geometric perspectives of VGI data and present interest in the specific thematic attributes, such as land cover. However, from another perspective, the diverse OSM completeness can reflect the importance hierarchy of specific information in human life. For example, Minaei (2020) mentioned the diversity of OSM road networks in Iran reflects corresponding concerns, worthiness and requirements on different roads, which might be information worth investigating. In addition, men and women even have different preference on mapping, which could be concerned and properly

utilized under reasonable context. According to Gardner et al. (2020), men concern more about the accuracy of the cartographic representation, while women concern more about the initial visibility of new data. In that case, guiding women to engage in and contribute more to humanitarian mapping is reasonable, as there is an emphasis on addition rather than modification.

Representation bias mitigation in data production process

Training locally & acting globally

Conducting project under the specific local context and integrating with available resources is a potential method. According to previous research, transforming complex real-world properties into geographic information on the map often loss local knowledge and context, therefore, reduce the accuracy (Longley et al., 2015). In the same vein, “Modifiable Areal Unit Problem”, caused by making inferences about the individual characteristics from the related group (such as administrative unit) (Wong, 2009), also applicable to non-shape information. In other words, the variability and inconsistency of the spatial analysis are influenced by the modify areal unit boundaries. In that case, the predictive models could be successfully trained with only local observations in the divided small spatial units and applied for prediction and decision-making processes of these same units. Although there would still be spatial bias in sub-areas, the representation bias could be eliminated to a great extent. In addition, it is important to train local people if necessary. In particular, experts and professors should promote better commination methods, concerning about not only local people’s educational level but also their technological literacy level. Besides, the designed interfaces of related applications for projects should meet the traditions, requirements and preferences of local participants, striving to expand data coverage and improve representativeness.

In addition, despite of the data collection process, allowing people in physically distant places to make suitable technological contributions and increasing accessibility of duplicated geographic information through multiple platforms could

possibly indirectly improve the credibility and representativeness of the datasets. Haitian earthquake case could be an example to overcome representation bias and alleviate the disaster by collecting information locally and making action globally. According to Zook (2010), although large amount of population could not help collecting local data in person, they indeed provide simple and useful data collection methods and geospatial mapping tools with solid baseline datasets for building upon, such as OpenStreetMap road data and other essential operational information, to promote global cooperation. In addition, by aggregating multi-source information, corresponding rescue teams could make correct response to the disaster (ibid.).

Provenance of VGI

VGI is originally driven by altruism inherent in voluntary individual effort and believe the goodness of contributors, therefore, the platform did not intervene the existence of asserted geographic information without citation, reference or authority (Goodchild, 2007). However, with the notice of potential profit-driven editing, it is increasingly important for scientific research to pay attention to the source of information. Provenance was defined as “information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness” (W3C, 2013), linking with series of metadata elements and sub-metadata elements. As early as the mid-1990s, the Content Standards for Digital Geospatial Metadata has been published by the U.S Federal Geographic Data Committee, seen as the description of geographic datasets (Goodchild, 2007). In that case, processing and investigating metadata were recognized as an effective and necessary step for searching, evaluating and utilizing geographic information. For example, if data consumers could recognize the background of the data producers on OSM, they could judge whether the data is suitable and credible enough to be analyzed.

Several details should be taken into account, in order to ensure the effectiveness of this method. There is no suggestion for individuals’ reputation for quality. Learning from the experience of profit-driven spatial crowdsourcing platforms (such as Uber)

mentioned by Tong et al. (2019), the relevant requirements, ratings and algorithms could be referenced and improved, mitigating surveillance pressure but becoming means to ensure the data quality and source. Another thing affecting people's cooperation on provenance provision is privacy concern. In that case, the terms and conditions of the VGI platforms should be provided properly, avoiding misunderstanding and unawareness (Jones et al., 2019). In addition, data anonymization and pseudonymization could be applied for preventing the possibility of compromising on privacy (ibid.). Comparing with the traditional mapping agencies, most VGI platforms are lack of corresponding standards and specifications to govern the geographic information production. Therefore, it is time to introduce appropriate policies, standardizing data editing and provenance recording.

Representation bias mitigation in data consumption process

Triangulate VGI data

Data accuracy and consistency of OSM is relatively difficult to be guaranteed, since the original intention is respecting grassroots activities and encouraging powerless speak. In response to this problem, triangulate OSM data with other conventional datasets (such as censuses) could be a potential solution. For instance, Dorn et al. (2015) had practiced the method of measuring thematic accuracy and completeness of specific spatial data by comparing the OSM data with an authoritative German reference dataset. Furthermore, the flood risk management project in Brazil applied VGI data to improve the coverage of monitored areas and promote the decision-making process, using this low-cost means to collect updated information for regions without appropriate monitoring stations (Horita et al., 2015).

However, Zhang & Zhu (2018) claimed that the method was limited by the availability of updated reference datasets to compare against. In contrast, comparison with comments within OSM platform seems to be a supplement. As shown by Seto, Kanasugi, & Nishimura (2020), OSM provides an opportunity for contributors to provide additional information and resolve errors using OSM Notes (a supplementary

tool to improve the quality of OSM following same format, see Table 1). Similarly, Zook (2010) also argued that crosschecking and peer-produced mapping approaches could express the superiority over traditional means. Generally, the correction action occur among anonymous contributors could probably promote the interactions between users to practice “open” quality management of “open” data, although there might be multilingual issue, various data structure and different geographical distribution of OSM Notes (ibid.).

Table 1: OSM Notes properties in OSM Notes (Seto, Kanasugi, & Nishimura, 2020)

Property	Description	Description Example
lon	Longitude	0.1000000
lat	Latitude	51.0000000
id	OSM Notes ID	16,659
date_created	First created note timestamp	2019-06-15 08:26:04 UTC
status	OSM Notes status	Open
date	Comment timestamp	2019-06-15 08:26:04 UTC
uid	OSM account id number	1234
user	OSM account name	username
action	OSM Notes status	open
text	OSM Notes contents	ThisIsANote

Importance weighting model

Another approach is to apply machine learning method to weight datasets by importance weighting function. According to Zhang & Zhu (2018), computing the loss in learning classifiers could properly mitigate the representation bias. In the next year, Zhang & Zhu (2019) conducted experiments to verify the feasibility and rationality of their methodology. As shown in Figure 1 & 2, it is possible to assess and modify the sample representativeness by analyzing the related environmental covariates using proper machine learning models. Through the experiment, the researchers tried to adjust the representativeness of the expected sample and simulate that of the target phenomenon by changing the corresponding parameters (ibid.). In other words, observations in under-represented places could probably receive higher weights, in order to mitigate representation bias of the datasets. After applying the approach to habitat suitability study of the red-tailed hawk, the result suggested that weighted datasets could better reflect the species occurrence

locations (Zhang & Zhu, 2019), therefore, this methodology was proven to be effective for improving prediction accuracy.

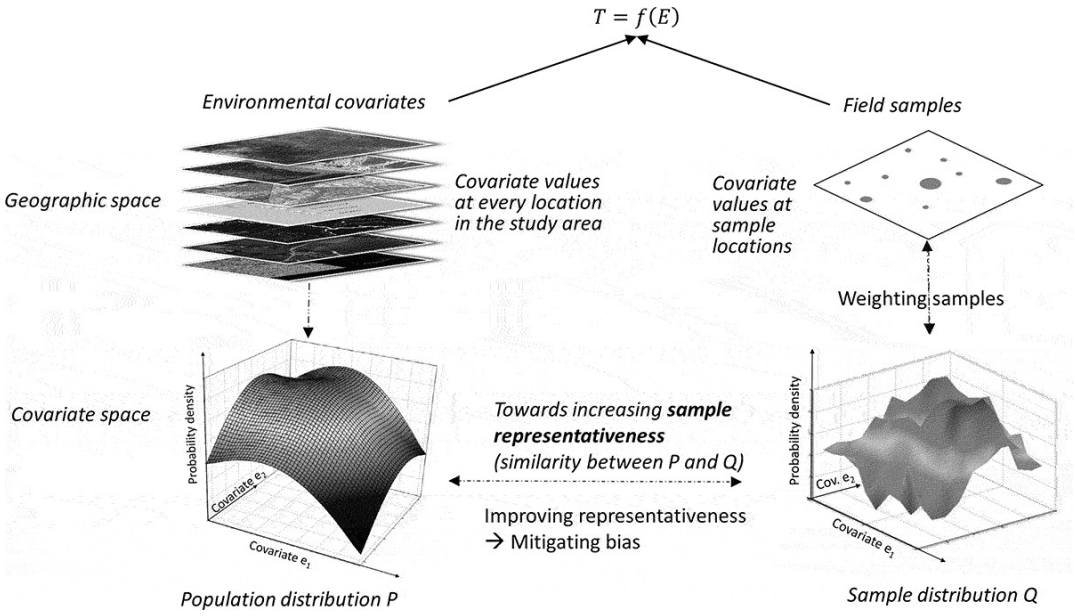

Figure 1: The basic principle of importance weighting model for representation bias mitigation (Zhang & Zhu, 2019)

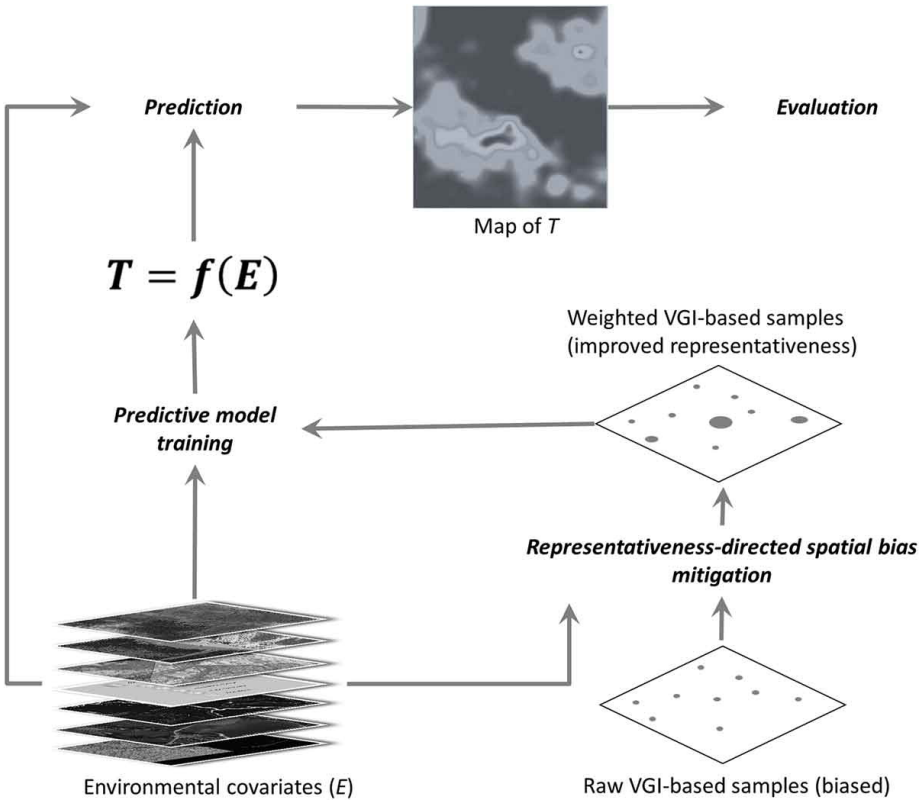

Figure 2: Methodology overview (Zhang & Zhu, 2019)

However, this sufficiently large datasets are required to train the model, in order to estimate the optimal parameters for the function and ensure the prediction accuracy. In addition, it is difficult to achieve in high-dimensional cases. The reason for that could be explained by machine learning theories. Witten et al. (2013) had mentioned that conventional method for dimension reduction would greatly reduce the interpretability of the model if choosing to remain its accuracy. Therefore, the non-interpretable methods could probably not draw practical and usable conclusions to facilitate predictions and decisions. When considering the artificial intelligent technology, the complexity of the models would also bring “black box” issues, no one knows the specific operation of algorithms, therefore, it is difficult to find the cause of the error or crash and then upgrade the model (Rai, 2020). Furthermore, researchers also questioned who should be responsible for the potential discrimination and unfair representative data caused by algorithm bias (Vaughan, 2019). Overall, it might not be a suitable choice to trade the increased modelling bias for the representation bias that might be mitigated.

Conclusion

In conclusion, the information era has brought opportunities for generating geographic information using VGI. As demonstrated by the article, representation bias could appear because of biased contributors, biased targets skewed towards powerful stakeholders and biased thematic attributes and lead to inequality and inaccuracy of research results. The origin and impacts of representation bias could be seen as necessary background information for spatial data scientists and guide the bias mitigation methodology. To mitigate the representation bias, approaches from data production and data consumption perspectives were all mentioned. Considering data production process, the integration of local and global resources and the comparison between data and its provenance could improve data coverage, accuracy and credibility. Considering data consumption process, the combination between external and internal crosschecking and the collaboration of geography and

machine learning fields could probably mitigate the representation bias, although there might be modelling defects. Overall, it seems reasonable to consider representative bias from a comprehensive perspective, and even cooperate with other regions or fields (see Figure 3). This essay has provided an overview of the possible approaches for future research, but it is still necessary to verify the effectiveness and possibility of the combination of these methods, as well as whether the proposed disadvantages and risks can be avoided. Therefore, more research and experiments should be conducted for promoting conclusive solutions.

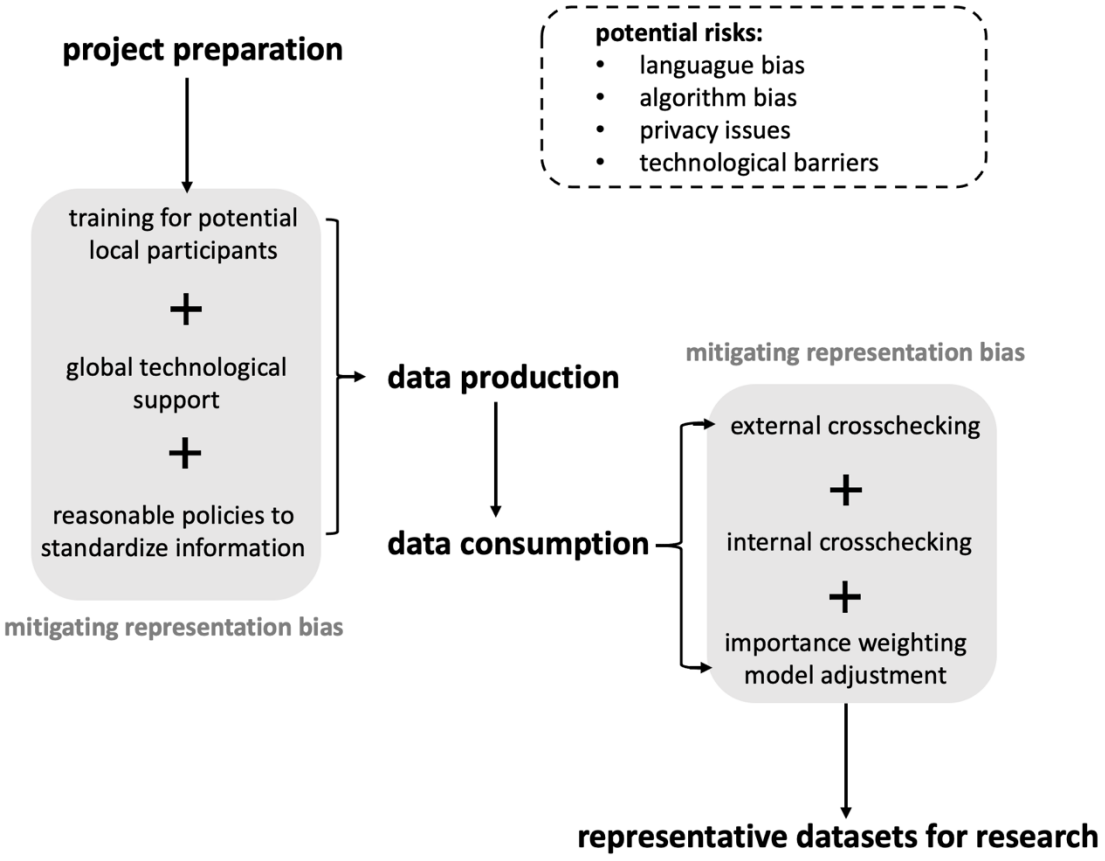

Figure 3: Methods review for mitigating representation bias in VGI
(Source: Author's own)

Reference list

Budhathoki, N. R. (2010). *Participants' motivations to contribute geographic information in an online community*. Ph.D Thesis. University of Illinois at Urbana-Champaign. Available at: https://search-proquest-com.libproxy.ucl.ac.uk/docview/863624855?pq-origsite=summon (Accessed: 23 December 2020).

Dorn, H. et al. (2015). 'Quality evaluation of VGI using authoritative data - a comparison with land use data in Southern Germany', *ISPRS International Journal of Geo-Information*, 4 (3), pp. 1657-1671.

Gardner, Z. et al. (2020). 'Quantifying gendered participation in OpenStreetMap: responding to theories of female (under) representation in crowdsourced mapping', *GeoJournal*, 85 (6), pp. 1-18.

Gilbert, M. R. et al. (2008). 'Theorizing the digital divide: Information and communication technology use frameworks among poor women using a telemedicine system', *Geoforum*, 39 (2), pp. 912-925.

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

Haklay, M. E. et al. (2016). 'Why is participation inequality important?', in Capineri, C. et al. (ed.) *European handbook of crowdsourced geographic information*. London: Ubiquity Press, pp. 35-44.

Horita, F. E. A. et al. (2015). 'Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with Wireless Sensor Networks', *Computers & Geosciences*, 80, pp. 84-94.

Huck, J. J. et al. (2021). 'Centaur VGI: a hybrid human-machine approach to address global inequalities in map coverage', *Annals of the American Association of Geographers*, 111 (1), pp. 231-251.

Jiang, S. et al. (2015). ‘Mining point-of-interest data from social networks for urban land use classification and disaggregation’, *Computers, Environment and Urban Systems*, 53, pp. 36-46.

Jones, K. H. et al. (2019). ‘Toward an ethically founded framework for the use of mobile phone call detail records in health research. *JMIR MHealth and UHealth*, 7 (3), p. e11969.

Kitchin, R. (2014). ‘Big Data, new epistemologies and paradigm shifts’, *Big Data & Society*, 1 (1), pp. 1-12.

Longley, P. A. et al. (2015). ‘Representing geography’, in *Geographic information science & systems.* 4th edn. John Wiley & Sons, pp. 55-76.

Minaei, M. (2020). ‘Evolution, density and completeness of OpenStreetMap road networks in developing countries: The case of Iran’, *Applied geography (Sevenoaks),* 119, p.102246.

Rai, A. (2020). ‘Explainable AI: from black box to glass box’, *Journal of the Academy of Marketing Science*, 48 (1), pp.137-141.

Seto, T., Kanasugi, H. & Nishimura, Y. (2020). ‘Quality verification of volunteered geographic information using OSM Notes data in a global context’, *ISPRS International Journal of Geo-Information*, 9 (6), p. 372.

Tong, Y. et al. (2019). ‘Spatial crowdsourcing: a survey’, *The VLDB journal*, 29 (1), pp. 217-250.

Vaughan, A. (2019). ‘UK knew its photo checker could fail with dark skin’, *New scientist (1971),* 244 (3252), p.12.

Witten, D. et al. (2013). *An introduction to statistical learning : with applications in R / Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.* London: Springer.

Wong, D. W. (2009). ‘Modifiable areal unit problem’, *International Encyclopedia of Human Geography*, 7, pp. 169-174.

W3C (2013). *PROV-Overview*. Available at: https://www.w3.org/TR/prov-overview/ (Accessed: 23 December 2020).

Yang, L. et al. (2020). ‘Road extraction based on level set approach from very high-resolution images with volunteered geographic information’, *IEEE Access*, 8, pp. 178587-178599.

Zhang, G. & Zhu, A. (2018). ‘The representativeness and spatial bias of volunteered geographic information: a review’, *Annals of Gis: Geographic Information Sciences*, 24 (3), pp. 151-162.

Zhang, G. & Zhu, A. (2019). ‘A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena’, *International journal of geographical information science: IJGIS*, 33 (9), pp. 1873-1893.

Zook, M. (2010). ‘Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake’, *World Medical and Health Policy*, 2 (2), pp. 7-33.