Intrinsic data quality of corporate contributions to OpenStreetMap: Assessing completeness

Jugal Patel[1], Corey Dickinson[1], Raja Sengupta[1], Dipto Sarkar[2]

[1]Department of Geography, McGill University
jugal.patel@mail.mcgill.ca, corey.dickinson@mail.mcgill.ca, raja.sengupta@mcgill.ca
[2]Department of Geography and Environmental Studies, Carelton University, dipto.sarkar@carleton.ca

ABSTRACT

OpenStreetMap, once a prime example of volunteered geographic information, has evolved into a critical data infrastructure where several multi-national corporations contribute data through paid editing teams. Corporate editing teams often contribute large volumes of data in a short amount of time, and the quality of these contributed data has seldom been evaluated. We test differences in data quality between five regions with established corporate contribution histories, and one region with little to no corporate contributions. We find that the rate with which data quality has improved on consistent road segments from 2014 to 2021 varies between our selected corporate-contribution-dense and corporate-contribution-sparse regions. That is, corporate contributions have different impacts on data quality than other non-corporate contributions. While OpenStreetMap remains a vibrant peer-produced data store for volunteers, governments, and increasingly corporations to collaborate in production of accessible spatial data, new motivations to contribute necessitate a revaluation of produced data.

1. Introduction

Since 2016, several large corporations (including Apple, Facebook, Microsoft, Uber) have increasingly contributed data to OpenStreetMap (OSM), an oft-cited example of volunteered geographic information (VGI) (Anderson et al, 2019). Corporate contributors (CC) represent a distinct editor-type as users are compensated and thus their contributions are not volunteered. As corporations employ editing teams and state-of-the-art editing techniques aided by artificial intelligence, CC can edit substantial amounts of information quickly (Ibid; Anderson & Sarkar, 2020;2022). While corporate contributors are responsible for an increasing proportion of edits (or contributions) to OSM, there remain unanswered questions related to the quality of their contributions (Dickinson, 2021). In this study we provide a quantitative evaluation of the impact of corporate contributions to OSM data quality. Research corporate contributions to OSM is relatively novel. Anderson & Sarkar have identified the need for future research to explore mapping patterns within CC and the need to examine how CC relate to the identity of OSM as a VGI (Ibid).

While the most straight forward data evaluation methods involve comparing against extrinsic sources, such as ground reference information or authoritative data sources; lack of data availability, licensing terms, and costs render this untenable (Baron et al, 2014; Senaratne et al,

 Volume DOI: 10.17605/OSF.IO/N6DU5

2017, Antoniou & Skopliti, 215). Intrinsic quality (IQ) assessments offer a convenient alternative, allowing for evaluation of network quality without an external source. Research on the applicability of intrinsic data quality assessments has traditionally focused on contributor types, semantic accuracy, and contributor history (Madubedube et al, 2021). Sehra, Singh and Rai present an overview of the various approaches to assessing OSM data quality (Sehra et al, 2017). Building upon previous attempts to standardize these systems, we maintain standards for assessing intrinsic data quality, using completeness as a metric of IQ. In our case, completeness is operationalized as a measurement of how close the network may be to real-world fidelity – in which case, the network would be fully complete.

By utilizing IQ analysis techniques, we hope to evaluate the impact of corporate editors on OSM and answer the research question "Do areas with high rates of corporate editing within OSM have different data quality patterns than areas without CC?"

2. Methods & Data

Our approach quantifies intrinsic data quality across five regions that have established histories of corporate contributions: Dallas, Egypt, Jamaica, Thailand, and Singapore. We compare changes to data quality in these regions to Denmark, a region with few CC and little to no corporate contributions. Despite this, Denmark possesses a well-developed OSM presence due to a robust local mapping community (Coast, 2015). We emphasize two commonly accepted priorities in assessing OSM data: usability and attribute completeness; and extend them to corporate contributions. We explore whether the variation with which intrinsic quality of OSM data changes is a response to corporate contributions.

Preprocessed historical archive data are hosted on Amazon Web Services in an S3 bucket as well-known textual representations of geometry. We query pre-processed historical OSM data stored here using AWS-Athena. We then cross-reference this global historical archive of user edits with the list of known corporate contributors (Anderson et al, 2019). The authors maintain a list of known corporate-editing associated accounts, allowing us to connect individual edits with corporate organizers using editor usernames. This list is derived from collating lists of editors maintained and shared by each corporation in order to be compliant with the Organized Editing Guidelines of OSM (Chapman, 2018).

We limit the scope of our work to years 2014 through to 2021, and to land-based transportation networks. Within the OSM data structure, land-based transportation networks include roadways, pedestrian paths, cycling paths, and other similar types of non-motorway land-based transport network data (OSM Foundation, 2021). We merge these data, using R 4.04 to enable geospatial analysis, carried out using QGIS 3.16 (R, 2022; QGIS, 2022). Specifically, we tie known edit locations using usernames associated with known organized editing teams (Anderson et Al, 2019). We then intersect six polygons (that of Thailand, Jamaica, Egypt, Singapore, Dallas, and Denmark) with merged datasets to show where (within these six polygons) and when (by year) edits occurred by which corporate contributor. Figure 1 shows a brief overview of the six regions and their corporate contribution history.

 Volume DOI: 10.17605/OSF.IO/N6DU5

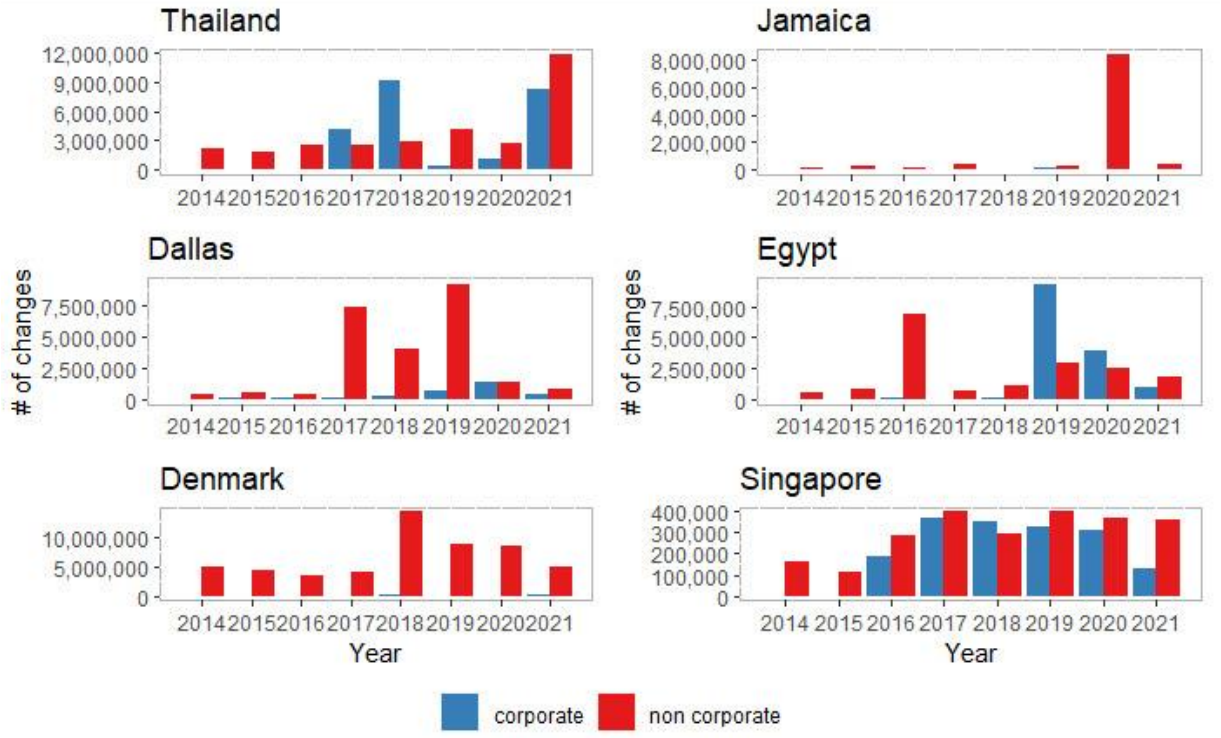

Figure 1: Above is an overview of six selected regions of interest and their corporate-contribution history. Note Jamaica, a region with a history of CC experienced large non-corporate editing event in 2020.

Our regions are based on their geographic distribution and their varied levels and types of corporate contribution. Additionally, these regions all share a history of being relatively data sparse before corporate contributions became prominent - these regions each now have at least one active CC. These criteria led previous research to highlight these regions as being interesting cases for study of CC to OSM (Anderson & Sarkar, 2020). Research focused on these areas have also shown that there is co-editing interactions between the corporate and non-corporate editors implying that data here is co-produced (Dickinson, Patel, & Sarkar, 2022).

Completeness as an intrinsic data quality indicator can be derived as both a measure of the entire network's completeness (NC) and of the completeness of specific attributes of individual segments within that network (AC). Our analysis traces the extent to which the network has been mapped over time for each region of interest. Results confirm a common assumption: total length of networks increase over time.

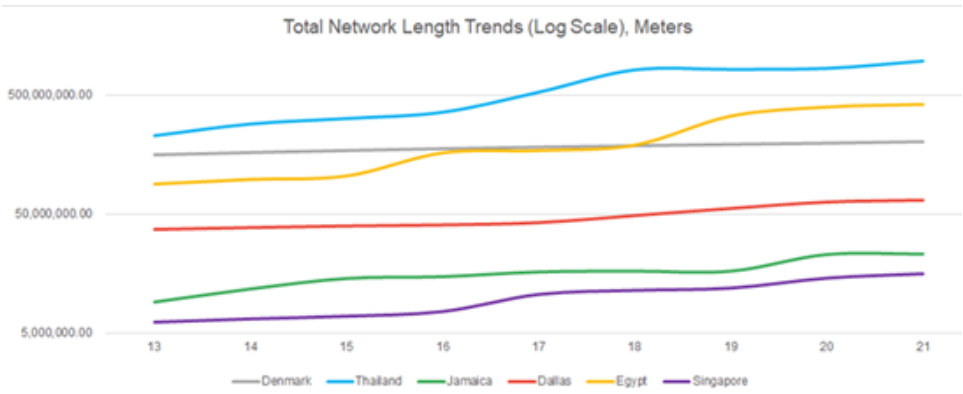

Figure 2: Above shows log-scaled network length over time for each region of interest. Each region increases its presence on OpenStreetMap, with our corporate-contribution control region, Denmark, maintaining a consistent total mapped network length

We examined Network Completeness (NC: proportion of network which has been mapped) by exploring how the road network length changes over time. Greater changes in the total network length result in a greater NC.

$$NC = \Delta \frac{(Total\ Network\ Length)}{Year}$$

Attribute completeness (AC: percentage of features within a network that have additional tags) was incorporated as a feature for considering usability. Additional tags, such as speed limits and turn restrictions are useful for routing. Here, we explore AC in the context of roads with non-null entries within their "name" attribute.

$$AC = \Delta \frac{Total\ Network\ Length \cap (name\ \&\ maxspeed\ attributes \neq null)}{Year}$$

Road segments without names or maximum speed attributes are not suitable for GPS navigation, inappropriate for real-world routing, and have a lower quality level than roads with more complete attribute information. This holds true even for road segments not intended for usage by cars (such as pedestrian paths, railways, or bikeways), as names and attribute information are necessary for all types of land-based transport. The motivation to increase the length of the network often outpaces the rate of map attribute addition. Relatedly, "map seeding", previously explored by Nagaraj (2017), in which the laying out of an inaccurate skeleton network by early contributors allows for future refinement (Nagaraj, 2017). For metrics, we consider only those segments – denoted by their OSM_ID – that were consistent across 2014-2021. This is key as this allows us to understand change in completeness over time.

3. Results

We used a nested analysis of variance; independent two-group Wilcoxon-Mann-Whitney u-tests; and Kruskal-Wallis rank sum tests to compare two independent groups: 1) road segments in Denmark; and 2) road segments in Thailand, Jamaica, Egypt, Dallas, and Singapore. Our tests are designed to test we are whether Denmark segment metrics are distinct from regions where CC have increasingly contributed to mapping efforts (Thailand, Jamaica, Singapore, Egypt, and Dallas). Jamaica experiences a surge in non-corporate contributions in 2020. We perform tests for each metric and region compared to Denmark. Our metrics examined were the

 Volume DOI: 10.17605/OSF.IO/N6DU5

year-to-year percent changes in each metric (Segment Length [NC], segment name [AC], segment speed [AC]) for each network. Nested analysis of variance was performed for each completeness metric comparing the means of each region and contribution group (CC or not). Tukey Honest Significant Differences (Tukey HSD) were collected for each analysis of variance, which confirmed intuition that each region would be distinct (yielding statistical significance when tested for difference in means). These results are summarized below in Table 1 with each column representing a completeness metric.

Table 1: Summary of tests for measuring change in three metrics related to intrinsic data quality

	Change in segment lengths (NC)	**Change in number of named attributes (AC)**	**Change in number of speed attributes (AC)**
Nested Analysis of Variance	F = 191806 p-value <2.2e-16	F = 2293310 p-value <2.2e-16	F = 10252710 p-value <2.2e-16
Wilcoxon-Mann-Whitney u-test	W = 1.2639e+13, p-value < 2.2e-16	W = 1.062e+13, p-value < 2.2e-16	W = 2.464e+10, p-value < 2.2e-16
Kruskal-Wallis rank-sum test	chi-squared = 656954, df = 1, p-value < 2.2e-16	chi-squared = 119206, df = 1, p-value < 2.2e-16	chi-squared = 119220, df = 1, p-value < 2.2e-16

We find that across all regions and metrics tests show that there is a statistical difference between corporate and non-corporate contributions. We interpret this to mean that for each metric it is likely that changes due to corporate contributions differ from non-corporate contributions in Denmark. This may reflect geographic variation between very different places, and therefore limits the extent to which these results can be generalized. While our statistical analysis reveal that these two groups differ statistically, the manifestation of that difference can be interpreted through examination of other trends in our data. Regarding completeness, our analysis shows that for both network completeness and attribute completeness there is a trend towards areas with higher rates of CC having more volatility in terms of improvement rates. This is reflected in figure 2, below.

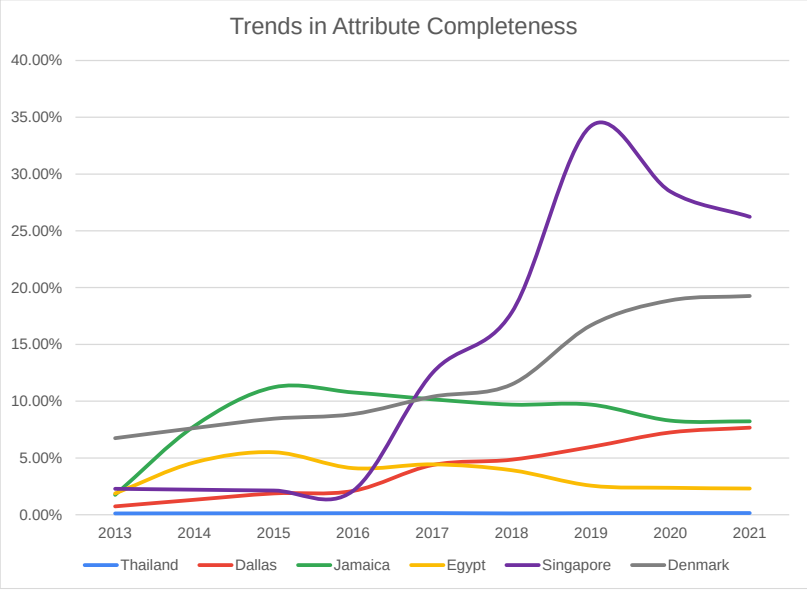

Figure 3: Trends in attribute completeness. Y axis represents percentage of network classified as complete in terms of the attributed we examined, i.e. the attributes “Max Speed” and “Name”

are present. Note Thailand's network is very poorly mapped in terms of speed limits but shows incremental improvement mirroring other trends, however it is not apparent at this scale.

4. Discussion & Conclusion:

Intrinsic completeness is measured by examining change over year, rather than against a true measure of completeness, and thus is a best guess at the map's fidelity to the real world. Our examination, as it is a preliminary study, is also limited to three forms of completeness, and does not address other geometry types within the map or regions where CC may be taking place.

While this article highlights the trends of data quality increase, it does not tease apart the quality assessment of contributions by corporate teams versus other mapping groups. As a crowdsourcing platform, data in OSM is co-produced by repeated editing of the objects by different members of the community. The appearance of CCs in OSM represents the latest evolution of the platform and represents another community of 'prod-users' in the OSM ecosystem. Consequently, there is significant interaction between CCs and non-CCs in data co-production in OSM, further reinforcing Linus's Law and the idea that OSM is a 'community of communities' (Solis, 2017). It is also worth noting that while CC have mostly been focused on the road network, their edits can have a map seeding effect. As a result of the reference laid down by the background road network, more features may get digitized in the overall area. Thus, to evaluate the true impact on the quality of data, an as-assessment across a wide variety of OSM features is required.

There is a trend towards increasing data quality in terms of gradual increase of network length, and completeness in terms of attributes needed for navigability. When compared to Denmark, our corporation-contribution-control group, the primary difference to note is not with regards to the quality of the data, but with respect to the rate at which the quality improves. Denmark is amongst the handful of nations which has traditionally had good data quality and active volunteer community (Mooney et al, 2010). While Denmark's levels of CC have recently begun to catch up with our other five regions, its historically low rates and its markedly different editing history serve to set it apart. That is, it serves as a good comparison set for this study.

Corporate contributors' goals are interconnected with navigation apps, delivery services, autonomous vehicle research, and other road-network based developments. This pattern is supported in our study, in that areas with high CC rates show very fast increases in the development of AC, which is crucial in allowing OSM data to be used for navigation

In this article, we focused on changes to key intrinsic data quality metrics for OSM. Temporal snapshots helped us quantify whether areas with CC differ from those without known corporate contributions. Based on our results, areas with high rates of CC are showing a statistically significant difference in rate of data quality change than areas with low rates. Nested analysis of variance and Tukey HSD summaries confirm that each region is in-fact statistically different from each other region when measuring completeness. In this article, we have focused only on the road network, however, CC is not limited only to the road network. For example, In Egypt, out of the 2 million edits done by Digital Egypt, 1.7 million involved addresses (Sarkar & Anderson, 2020). Thus, more quality evaluations across different map features are required before conclusively stating whether CC increases the overall data quality in OSM, though our research seems to indicate that it is likely to be doing so. Relatedly, the ways by which aggregation is applied in our work is limits analysis. That is, by taking country-level polygons and comparing these across one-another, we mix in polygon-level biases. Future work should aim to remove the need for country-level polygons as an important unit of analysis.

 Volume DOI: 10.17605/OSF.IO/N6DU5

References:

Anderson, J.; Sarkar, D. Curious cases of corporations in OSM. in proceedings of the academic track, state of the map 2020; Minghini, M., Juhász, L., Yeboah, G., Mooney, P., Grinberger, A.Y., Eds.; Zenodo: Online Conference, 2020.

Anderson, J.; Sarkar, D.; Palen, L. Corporate editors in the evolving landscape of OpenStreetMap. ISPRS Int. J. Geo-Inf. 2019, 8, 232, doi:10.3390/ijgi8050232.

Antoniou, V.; Skopeliti, A. Measures and indicators of vgi quality: an overview. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2015, II-3/W5, 345–351, doi:10.5194/isprsannals-II-3-W5-345-2015.

Barron, C.; Neis, P.; Zipf, A. A comprehensive framework for intrinsic OpenStreetMap quality analysis. Trans. GIS 2014, 18, 877–895, doi:10.1111/tgis.12073.

Chapman, K. Organised editing guidelines. OpenStreetMap Found. 2018, 1–4.

Cipeluch, B.; Jacob, R.; Winstanley, A.; Mooney, P. comparison of the accuracy of OpenStreetMap for Ireland with google maps and bing maps. In Proceedings of the Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences; 2010.

Coast, S. The Book of OSM; CreateSpace Independent Publishing Platform: California, United States of America, 2015; ISBN 1-5142-3274-X.

Dickinson, C. Inside the 'Wikipedia of Maps,' Tensions Grow Over Corporate Influence. Bloomberg.com 2021.

Dickinson, C., Patel, J., & Sarkar, D. (2022). Corporate editing and its impact on network navigability within OpenStreetMap. Editors, 49.

Foundation, O. 2021 Survey Results - OpenStreetMap Foundation Available online: https://wiki.osmfoundation.org/wiki/2021_Survey_Results (accessed on 15 July 2021).

Goodchild, M.F.; Li, L. assuring the quality of volunteered geographic information. Spat. Stat. 2012, 1, 110–120, doi:10.1016/j.spasta.2012.03.002.

Madubedube, A.; Coetzee, S.; Rautenbach, V. A contributor-focused intrinsic quality assessment of openstreetmap in mozambique using unsupervised machine learning. ISPRS Int. J. Geo-Inf. 2021, 10, 156, doi:10.3390/ijgi10030156.

Mooney, P.; Corcoran, P.; Winstanley, A.C. Towards quality metrics for OpenStreetMap. In Proceedings of the proceedings of the 18th sigspatial international conference on advances in geographic information systems - GIS '10; ACM Press: New York, New York, USA, 2010; p. 514.

Nagaraj, A. information seeding and knowledge production in online communities: evidence from OpenStreetMap. Ssrn 2017, doi:10.2139/ssrn.3044581.

QGIS Development Team QGIS Geographic Information System; 2021;

R Core Team R: A Language and Environment for Statistical Computing; Vienna, Austria, 2021;

Sarkar, D.; Anderson, J. Community interactions in OSM editing. Presented at the State of the Map 2021 (SotM 2021), online, 2021.

Sarkar, D.; Anderson, J.T. Corporate editors in OpenStreetMap: Investigating Co-editing Patterns. Trans. GIS 2022, n/a, doi:10.1111/tgis.12910.

Sehra, S.; Singh, J.; Rai, H. Assessing Open StreetMap data using intrinsic quality indicators: an extension to the QGIS processing toolbox. Future Internet 2017, 9, 15, doi:10.3390/fi9020015.

Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. (Muki) A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 2017, 31, 139–167, doi:10.1080/13658816.2016.1189556.

Solis, P. Building Mappers Not Just Maps: Challenges and Opportunities from YouthMappers on Scaling up the Crowd in Crowd-Sourced Open Mapping for Development. In Proceedings of the Annual Meeting of the Association of American Geographers; Boston, Massachusetts., 2017.

Wang, R.Y.; Strong, D.M. Beyond Accuracy: What data quality means to data consumers. J. Manag. Inf. Syst. 1996, 12, 5–33, doi:10.1080/07421222.1996.11518099.

 Volume DOI: 10.17605/OSF.IO/N6DU5