Increasing OSM Data Accessibility with the Analysis-Ready Daylight Distribution of OSM: Demonstration of Cloud-Based Assessments of Global Building Completeness

Jennings Anderson[1,*] and Timera W. Omidire[1]

[1] Meta Platforms, Inc., USA; jenningsa@fb.com, omidire@fb.com

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2022 Conference after peer-review.

Despite being one of the most open and freely available spatial datasets, OpenStreetMap (OSM) data accessibility remains a challenge. Data accessibility measures how easily end-users can access and use a given dataset for their needs [1]. With its unique data structure of nodes, ways, and relations connected via tags, OSM data is not immediately consumable by standard geospatial analysis tools. OSM Pre-analysis workflows require OSM data to be downloaded, parsed, and converted into more common formats, which means that novice end-users of OSM may lack the experience to readily access and use OSM data in decision-making.

Incorporating communities into spatial decision-making processes, such as mapping, is important because a). community members are experts on their communities and b). have a larger stake in final solutions which directly impacts their lived-experiences [2]. OSM empowers a variety of communities, including local governments [3], digital humanitarian groups [4], and even student groups [5], to help navigate and understand places of respective importance.

Research by Nirandjan et al. recently lowered barriers to using OSM data as a reference dataset of critical infrastructure [6]. After categorizing and quantifying particular types of OSM features, the authors released the data in formats more common in geospatial analysis, such as GeoTiffs [6]. This article's popularity (ranked 90th percentile on the publisher's website) demonstrates the importance of making OSM data—and datasets derived from OSM—more accessible by means of familiar data structures compatible with common tools. If OSM data were more accessible for analysis, would we see it used in more geospatial research and innovation at large [7]?

While many community-maintained tools exist to convert, extract, and download OSM data, each requires domain knowledge of the unique OSM data structure (nodes, ways, and relations). Furthermore, working at the country or planet-scale requires extensive computational resources. To further lower the barriers to entry for OSM data analysis and

Anderson, J., & Omidire, T.W. (2022). Increasing OpenStreetMap Data Accessibility with the Analysis-Ready Daylight Distribution of OpenStreetMap: A Demonstration of Cloud-Based Assessments of Global Building Completeness.
In: Minghini, M., Liu, P., Li, H., Grinberger, A.Y., & Juhász, L. (Eds.). Proceedings of the Academic Track at State of the Map 2022, Florence, Italy, 19-21 August 2022. Available at https://zenodo.org/communities/sotm-22
DOI: 10.5281/zenodo.7004501

extraction, we created the Analysis-Ready Daylight OpenStreetMap Distribution (ARD-OSM). Freely available to anyone, it is published on the registry of open data (RODA) on Amazon Web Services (AWS) [8]. Containing 1B OSM features, ARD-OSM is optimized for use with Amazon Athena, a serverless interactive query engine on AWS. Additionally, ARD-OSM has resolved the OSM data format into common geometries such as points, lines, and polygons. The dataset also includes pre-computed valuable attributes such as length, surface area, quadkeys, and geographic bounding boxes which are stored as additional metadata. To demonstrate the analytical capabilities of this dataset, next, we will perform a global OSM building density assessment.

Building density is a common metric in OSM quality research, often used to assess map coverage and completeness, such as studied by Yeboah et al. [9]. Measuring building density requires counting all of the buildings within a defined unit of spatial analysis. We use zoom-level 11 map tiles to create an analysis grid that encompasses the global built environment in fewer than 1M tiles. Then, we divide the building count by the area of each map tile to obtain the number of buildings mapped per square kilometer.

Since every feature in ARD-OSM includes the zoom-level 15 quadkey of the map tile in which it exists, we can use a more efficient SQL GROUP BY expression instead of a geospatial operator for aggregation. After setting up the *analysis_ready_daylight* table in Amazon Athena, this is the query used to count the number of buildings in each zoom-level 11 map tile:

```
SELECT      substr(quadkey, 1, 11) as z11_tile,
            count(id) as number_of_buildings
FROM  analysis_ready_daylight
WHERE tags['building'] IS NOT NULL AND release = 'v1.12'
GROUP BY substr(quadkey, 1, 11)
```

In May 2022, running in AWS region us-east-1, this query took 15 seconds and cost USD $0.10. In this way, a global accounting of all the buildings in OSM can be performed quickly, cost-effectively, and by anyone with a basic knowledge of common SQL. The results of this query show the density of mapped buildings in OSM to be highest in Europe with additional areas of high density where Humanitarian mapping campaigns have been active such as Nepal, South Eastern Asia, and isolated parts of Africa. This is consistent with the findings of Herfort et al. [10].

How should these densities be interpreted? Do denser regions have higher completeness in which most or all buildings are mapped? Building density is an intrinsic data quality measure, to further contextualize these findings, we need to perform an extrinsic assessment by comparing our results against an external dataset. A recent study confirmed the viability of referencing population data for building density assessment [11]; and Orden et al. demonstrate a three-step methodology using Meta's High Resolution Settlement Layer (HRSL) first requiring both vectorization and spatial aggregation to assess building completeness with respect to population in both the Philippines and Madagascar [12].

Because the HRSL is also published via RODA [13], it can be easily *joined* to our results. Once HRSL data is incorporated to obtain a measure of *buildings mapped per square kilometer per person*, we find that parts of Europe remain in the top tiers of density with the most buildings mapped per person. Nepal and many parts of South Eastern Asia, however,

are no longer in the same top tier of map coverage. Figure 1 depicts these differences. While there are many mapped buildings, the higher populations of these regions reveals that there are still many areas where the buildings have yet to be mapped. This yields a generally lower level of completeness than initially identified, which remains consistent with the findings of Herfort et al. [10]. Additionally, parts of the United States and New Zealand actually appear more complete with areas of lower density coinciding with regions of lower population, yielding a higher measure of map completeness than before.

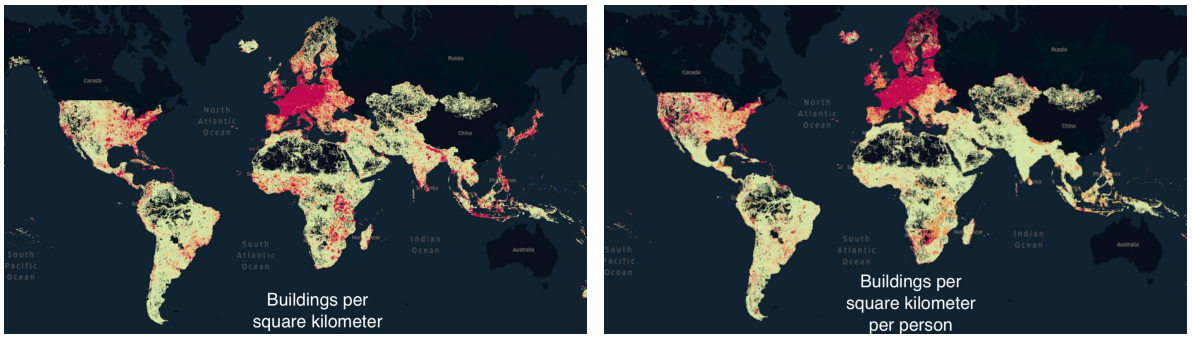

Figure 1. Subtle differences in building density in OSM when incorporating population density. Areas in red have highest density while areas in yellow have lower densities.

This case study cheaply and easily reproduced popular methods for both intrinsic and extrinsic data quality assessments of OSM building coverage without needing to download nor pre-process any OSM data. The analysis was done completely in the cloud on AWS using free and open data in RODA. Additional metadata in ARD-OSM enabled the query to run efficiently and cost-effectively. The same methodology can be applied to investigations of any other object type in OSM from hospitals to ice cream shops. We also recognize that ARD-OSM does not solve the needs of researchers looking to work with OSM history data. Other tools such as Ohsome, which utilizes the OpenStreetMap History Database are better suited for those types of historical analyses [14]. Researchers competent in the Java programming language can leverage the Ohsome API to interrogate the complete history of OSM features. Such investigations go beyond the current scope of ARD-OSM, which is optimized for efficient planet-scale exploration of the latest version of the map. Future innovations to the dataset and process could include pre-computing the complete history of each feature, but for the time-being, we recommend using Ohsome for such investigations.

Additionally, while ARD-OSM is currently available on AWS services, the raw data is a series of Parquet files intended to be used with PrestoDB, therefore, users are not locked into AWS. We see the creation of ARD-OSM as an example of what is possible when OSM researchers utilize distributed cloud-based database technologies and hope that future researchers can expand on these findings.

ARD-OSM contains 1B features—nearly every OpenStreetMap object—in common geospatial feature types such as points, lines, and polygons. To additionally aid researchers, features are enriched with additional metadata describing their location and physical attributes such as length or surface area. From a data accessibility perspective, we anticipate ARD-OSM in future research and innovation, curated by a wide range of end-users,

to readily integrate OSM data for decision-making processes which bring communities closer together.

References

[1] Open Data Charter (2015). Principles. Open Data Charter. Retrieved from https://opendatacharter.net/principles
[2] Keenan, P.B., & Jankowski, P. (2018). Spatial decision support systems: Three decades on. *Decision Support Systems, 116*(2019), 64–76.
[3] Johnson, P.A. (2017). Models of direct editing of government spatial data: challenges and constraints to the acceptance of contributed data. *Cartography and Geographic Information Science, 44*,(2), 128–138.
[4] Palen, L., Soden, R., Anderson, T.J., & Barrenechea, M. (2015). Success & Scale in a Data-Producing Organization: The Socio-Technical Evolution of OpenStreetMap in Response to Humanitarian Events. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, Seoul, Republic of Korea, 18-23 April 2015, 4113–4122.
[5] Maidin, M.A.A., Ahmad, F., Abidin, N.I., Suhaili, J., Awang, M., Rahman, M.A.A., Musa, M.K., Hamidon, N., Yusop, F.M., Syazwan, M.M.S., Harun, H., Hamid, N.H.A., & Kamil, N.A. (2021). Design campus map using OpenStreetMap digital software. In: Zaini, M.A.A., Jusoh, M., & Othman, N. (Eds.) *Proceedings of the 3rd International Conference on Separation Technology*, Springer, Singapore, 113–129.
[6] Nirandjan, S., Koks, E.E., Ward, P.J., & Aerts, J.C.J.H. (2022). A spatially-explicit harmonized global dataset of critical infrastructure. *Scientific Data 9*(1), 11–13.
[7] National Academy of Sciences (US), National Academy of Engineering (US) and Institute of Medicine (US) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age (2009) *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, DC, National Academies Press.
[8] Daylight Map Distribution of OpenStreetMap. Retrieved from https://registry.opendata.aws/daylight-osm
[9] Yeboah, G., Porto de Albuquerque, J., Troilo, R., Tregonning, G., Perera, S., Ahmed, S.A.K.S., Ajisola, M., Alam, O., Aujla, N., Azam, S.I., Azeem, K., Bakibinga, P., Chen, Y.-F., Choudhury, N.N., Diggle, P.J., Fayehun, O., Gill, P., Griffiths, F., Harris, B., Iqbal, R., Kabaria, C., Ziraba, A.K., Khan, A.Z., Kibe, P., Kisia, L., Kyobutungi, C., Lilford, R.J., Madan, J.J., Mbaya, N., Mberu, B., Mohamed, S.F., Muir, H., Nazish, A., Njeri, A., Odubanjo, O., Omigbodun, A., Osuh, M.E., Owoaje, E., Oyebode, O., Pitidis, V., Rahman, O., Rizvi, N., Sartori, J., Smith, S., Taiwo, O.J., Ulbrich, P., Uthman, O.A., Watson, S.I., Wilson, R., & Yusuf, R. (2021). Analysis of OpenStreetMap Data Quality at Different Stages of a Participatory Mapping Process: Evidence from Slums in Africa and Asia. *ISPRS International Journal of Geo-Information, 10*(4), 265.
[10] Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific Reports, 11*, 3037.
[11] Zhang, Z., Zhou, Q., Brovelli, M.A., & Wanjing, Li (2022). Assessing OSM building completeness using population data. *International Journal of Geographical Information Science*, *36*(7), 1443–146.
[12] Orden, A., Flores, R.A., Faustino, P., & Samson, M.S. (2020). Measuring OpenStreetMap building footprint completeness using human settlement layers. In: Minghini, M., Coetzee, S., Juhász, L., Yeboah, G., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2020 Online Conference*, 4-5 July 2020.
[13] High Resolution Population Density Maps + Demographic Estimates by CIESIN and Meta. Retrieved from https://registry.opendata.aws/dataforgood-fb-hrsl.
[14] Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.B., & Zipf, A. (2019). OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards*, *4*(1), 1–12.