Measuring OpenStreetMap building footprint completeness using human settlement layers

Ardie Orden[1,*], Ren Avell Flores[1], Pia Faustino[1] and Mark Steve Samson[1]

[1] Thinking Machines Data Science, Taguig, Philippines; ardie@thinkingmachin.es, avell@thinkingmachin.es, pia@thinkingmachin.es, marksteve@thinkingmachin.es

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2020 Online Conference after peer-review.

Non-government organizations and local government units use geographic data from OpenStreetMap (OSM) to target humanitarian aid and public services. As more people start to depend on OSM, it is important to study data completeness in order to identify unmapped regions so that OSM volunteers can focus their attention on these areas. In this study, we propose a method to measure the data completeness of OSM building footprints using human settlements data.

Specifically, we use Facebook's High Resolution Settlement Layer (HRSL), a dataset of built-up areas derived from satellite images, as a proxy for ground truth building footprints. We then measure data completeness by getting the "percentage completeness" of pixels which is computed using the total percentage of pixels within the intersection of the human settlement layer and the OSM building footprints. The method can be broken down into three steps: (1) convert the human settlement layer into a vector; (2) perform a spatial join to find the intersection between the vectorized human settlement layer and the OSM building footprints; and (3) calculate the data completeness based on areas from the vectorized human settlement layer that intersect with the building footprints.

Chepeish and Polchlopek [1] conducted a similar study measuring data completeness in OSM building footprints. We differentiate our work from Chepeish and Polchlopek [1] in three ways. First, for the human settlement layers, they used WorldPop which has a spatial resolution of 100 meters [2] while we used the High Resolution Settlement Layer (HRSL) from Facebook which has a spatial resolution of 30 meters [3]. Second, for the data processing, they rasterized the building footprints while we vectorized the human settlement layer. Third, for calculating the data completeness, they used a combination of geographic information system (GIS) and machine learning (ML) while we solely used GIS.

Using building footprints from January 2020 and human settlement layers dated June 2019 and October 2018, the percentage completeness is 32.75% and 10.89% for the Philippines and Madagascar, respectively. We found that in the Philippines, most of the unmapped buildings are in rural areas. When the pixels are aggregated to the municipality-level and plotted as a scatter plot of the urban percentage completeness vs. the rural percentage completeness, the municipalities appear to group together into two

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, July 4-5 2020. Available at https://zenodo.org/communities/sotm-2020
DOI: 10.5281/zenodo.3923033
© 2020 by the authors. Available under the terms of the Creative Commons Attribution (CC BY 4.0) license.

categories: sparsely mapped and thoroughly mapped. A possible explanation is that there are not enough OSM volunteers to map all municipalities and the OSM community focuses on thoroughly mapping high population municipalities rather than moderately mapping all municipalities. Interestingly, poverty incidence data from the Philippine Statistics Authority is not correlated with data completeness. Complete or incomplete OSM data in an area is not an indicator of wealth or poverty.

As this work has garnered interest from humanitarian organizations such as the Humanitarian OpenStreetMap Team (HOT), to whom regularly updated information on OSM data completeness is extremely valuable, we looked into ways to automate workflows in QGIS by using the built-in workflow builder tool (i.e. Processing Modeller) and by using the QGIS API. However, as we consider scalability and reproducibility important for this line of work, we ultimately deemed QGIS to be unfit for our use case. QGIS is not scalable because the data processing is not easily parallelizable and it is also not easily reproducible by developers who do not have training with GIS software. Thus, we decided to migrate our workflow to GeoPandas and rasterio and open-source our code [4]. Our workflow improved because (1) we were able to speed up the process by migrating to the cloud and increasing the computing resources; and (2) we were able to improve the reproducibility by allowing us to communicate our work more effectively to people who are not familiar with GIS.

For future research, we recommend exploring other human settlement datasets. The Global Human Settlement Layer (GHSL), for example, has a spatial resolution of 30 meters [5] which is comparable to WorldPop and HRSL. We also encourage further data analysis on the percentage completeness in order to get insights on how to improve the process of contributing to OSM.

References

[1] Chepeish, E., & Polchlopek, J. (2018). Estimate OSM building coverage completeness by comparing vs WorldPop raster. Retrieved from https://github.com/azavea/hot-osm-population.

[2] Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific data*, *4*(1), 1-4.

[3] Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E. B., & Dang, H. H. (2017). Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839*.

[4] Orden, A., & Flores, R. A. (2020). Supplementary code for "Measuring OpenStreetMap building footprint completeness using human settlement layers. Retrieved from https://github.com/thinkingmachines/osm-completeness.

[5] Florczyk, A. J., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., Melchiorri, M., Pesaresi, M., Politis, P., Schiavina, M., Sabo, F., & Zanchetta, L. (2019). GHSL Data Package 2019. JRC Technical Report, EUR 29788 EN, Publications Office of the European Union, Luxembourg.