Proceedings of OSM Science 2023

Editors:

Marco Minghini
Hao Li
A. Yair Grinberger
Pengyuan Liu
Godwin Yeboah
Levente Juhász
Serena Coetzee
Peter Mooney
Alessandro Sarretta
Jennings Anderson

DOI: 10.5281/zenodo.10443403

Contents

OpenStreetMap as an emerging scientific field: Reflections from OSM Science 2023

A. Yair Grinberger[1*], Hao Li[2], Pengyuan Liu[3], Godwin Yeboah[4], Levente Juhasz[5], Serena Coetzee[6], Peter Mooney[7], Alessandro Sarretta[8], Jennings Anderson[9], and Marco Minghini[10,†]

[1] Department of Geography, The Hebrew University of Jerusalem, Israel; yair.grinberger@mail.huji.ac.il
[2] Department of Aerospace and Geodesy, Technical University of Munich, Ottobrunn, Germany; hao_bgd.li@tum.de
[3] School of Geographical Sciences, Nanjing University of Information Science and Technology, China; neil.pengyuanliu.0705@gmail.com
[4] Research Computing, Research Technology Platforms, University of Warwick, Coventry, United Kingdom; g.yeboah@warwick.ac.uk
[5] GIS Center, Florida International University, Miami, FL, United States; ljuhasz@fiu.edu
[6] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa; serena.coetzee@up.ac.za
[7] Department of Computer Science, Maynooth University, Maynooth, Ireland; peter.mooney@mu.ie
[8] Research Institute for Geo-Hydrological Protection, National Research Council, Padova, Italy; alessandro.sarretta@cnr.it
[9] Meta Platforms Inc., Salem, OR, United States; jenningsa@fb.com
[10] European Commission, Joint Research Centre (JRC), Ispra, Italy; marco.minghini@ec.europa.eu

* Author to whom correspondence should be addressed.
† The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

OpenStreetMap (OSM) started as a project in 2004 aiming at creating a digital and open map of the world via collaborative mapping, emerging over time to become a community (or a collection of communities) [1] or an ecosystem [2] around the project itself. This ecosystem encompasses local and global communities of data and software developers creating a large number of tools and services, e.g. for spatial data infrastructures [3], disaster response [4], and routing [5]. Additionally, a new scientific field focusing on OSM is emerging with academic researchers investigating the different scientific aspects of the living OSM community [6–8]. The Academic Track at the annual State of the Map (SotM) conference, with five editions from 2018 to 2022 has become a knowledge hub for gathering and sharing recent progress in OSM-related research and scientific findings directly with the broad OSM community. Moreover, the 2019 and 2020 editions of this Track have led to the first special issue of scientific articles dedicated to OSM [9], which first formalized and used the term “OSM Science” outside informal conversations in SotM and the OSM-science mailing list [10] (with one exception being Haklay’s reference to ‘OSM studies’ [11]). In its sixth edition, the Academic Track starts to use the new name of “OSM Science” referring to

Grinberger, A.Y., Li, H., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., Anderson, J., & Minghini, M. (2023). OpenStreetMap as an Emerging Scientific Discipline: Reflections from the OSM Science 2023
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443298

the emerging scientific discipline characterized by its unique focus on the OSM project, data, contributors, community, and applications. The proceedings of the OSM Science at the SotM Europe 2023 conference, taking place in Antwerp, Belgium on November 10-12, 2023 [12], include 17 short papers corresponding to 8 talks and 9 lightning talks presented at the conference. In this Editorial, we survey these papers by grouping them into diverse but also interrelated research topics following an interdisciplinary perspective [13].

The study of data quality in OSM has a profound record, since the "fitness-for-use" of OSM data may vary in different application scenarios and spatial-temporal contexts. Several abstracts included in these proceedings concern the topic of OSM data quality assessment. The first two investigate the completeness of OSM buildings and the attribute completeness of OSM roads, respectively. Oostwegel et al. [14] present a dynamic assessment of the completeness of OSM buildings on a global scale, which can be updated at the moment that a building is added, modified, or removed. Similarly, but for a city scale, Andorful et al. [15] explore the association of OSM road classes and Points-of-Interest (POI) categories in their proximity. The key findings are envisioned to guide mappers by suggesting road classes in scenarios where POIs are mapped before the roads themselves to ensure a simultaneous quality check of both road classes and POIs. Beyond buildings and roads in OSM, O'Brien [16] seeks to enhance the quality and completeness of OSM land use information with open datasets, such as governmental retail areas and geolocated energy performance data, in Great Britain. Inspired by the Nolli map, Scheck et al. [17] design a framework to map urban public spaces by sequentially analyzing both relevant tags and geometries of the OSM, then validate and enrich the data via extensive on-site mapping in the city of Vienna, Austria. All these four works emphasize the importance of deriving an up-to-date and accurate assessment of OSM data quality in order to better understand how OSM data can be used.

In addition to the existing literature on data-centric research in the OSM scientific community, a large number of abstracts (7 abstracts) focus on domain-specific applications of OSM data. These include harnessing OSM data to build a baseline for healthcare facilities data [18], 3D city modeling and building height estimation by combining OSM data with remote sensing and street-view imagery [19], and exploring OSM data to measure the level of local proximity to services by walking [20]. Other studies explore the potential for integrating OSM data with other data sources in diverse application scenarios. For instance, Evaz Zadeh et al. [21] highlight OSM potential in improving the accuracy of earthquake risk assessment by estimating building type information from existing OSM building footprints. Vierø et al. [22] relate to the use of OSM for planning bicycle networks, developing the BikeDNA tool which integrates extrinsic and intrinsic data quality analysis procedures to assess bicycle path mapping in OSM (and thus also relate to the theme of data quality). Finally, Passmore et al. [23] assess the accessibility of bike-transit (i.e. the use of bicycle as the first and/or the last-mile mode) using an OSM road network, highlighting the need to improve OSM road tags to enhance bicycle routing algorithms for cycling, walking, and transit use.

More recently, Machine Learning (ML) as an emerging topic has gained quite some momentum in the OSM research community, especially in conjunction with Geospatial Artificial Intelligence (GeoAI). Five abstracts in these proceedings contribute to different aspects of GeoAI with OSM. First, Fila et al. [24] compare the editing behavior pattern of AI-assisted and traditional mapping methods by mining historical OSM user contribution data. Vestena et al. [25] propose a framework that combines image semantic segmentation and monocular depth estimation methods on street view images to conduct micromapping

for small-scale features, enhancing the completeness and readability of the OSM map. Gramacki et al. [26] showcase a Python library providing procedures to train embedding models based on OSM and additional geospatial data, thus pushing forward discussions regarding standardizing GeoAI research through a set of commonly accepted models and evaluation data. Melanda et al. [27] analyze the (im)balance in OSM highway tags of nine countries, motivated by the need to provide balanced OSM training data for ML tools that do not lead to biased results. Last but not least, Khellaf et al. [28] develop an interesting natural language interface based on ChatGPT for efficiently querying OSM data and finding meaningful combinations of objects and public space, which helps (investigative) journalists for accurate geo-location in the context of information verification.

These three topics — data quality, applications, and ML — represent the data-oriented approach within OSM Science, which exists alongside the social perspective exploring the contexts from which OSM data emerge and the project's societal implications [11]. This latter aspect is assessed in two abstracts. First, Geddes [29] considers how familiarization with OSM, emerging through the inclusion of OSM within water point mapping in Malawi, may contribute to this effort generally. Schröder-Bergen [30] inspects recent changes in the project from a political economy perspective, using a mixed method to understand the implications of OSM emergence as a prominent data source and service and the increasing involvement of paid and organized mapping teams.

One claim raised in the last work is that although OSM has reached a significant level of maturity, or even because of this, it still goes through significant changes that may alter the very foundations of the project. Similarly, the growing interdisciplinary research interest of and about OSM is a positive sign showing that the scientific endeavor termed as OSM Science [9] is further developing and maturing. But, as its object of interest, it still has room to grow. Changes within the project are certainly a source of new scientific developments, as seen in the increasing interest in corporate editors [31] or unprecedented volumes of political vandalism [32]. Yet, there are also developments outside OSM that will surely drive OSM Science forward, e.g. integrating generative AI and large language models such as ChatGPT into geospatial procedures [33], the Overture Maps initiative [34] whose relations with and impacts on OSM are only beginning to unfold, and the increasing academic research on digital transformation and geographical information science (GIScience). Such developments mark OSM as much more than a crowdsourced database but rather a unique phenomenon with no equivalent in the geospatial domain. As such, existing tags such as GIScience and Volunteered Geographic Information may not be enough to capture the research endeavors that will take place as part of the "OSM Science". Such endeavors will certainly extend beyond disciplinary borders and may give rise to new conceptual and theoretical frameworks. It is thus with great excitement that we look to the future of research with and about OSM, hoping that this OSM Science 2023 meeting will be the first of many meetings that will help shape this emerging field and its agenda.

References

[1] Grinberger, A. Y., Minghini, M., Yeboah, G., Juhász, L., & Mooney, P. (2022). Bridges and barriers: An exploration of engagements of the research community with the OpenStreetMap community. *ISPRS International Journal of Geo-Information, 11*(1), 54.

[2] Schott, M. (2020). *The future of working with OSM data*. https://giscienceblog.uni-heidelberg.de/2020/09/10/the-future-of-working-with-osm-data

[3] Minghini, M., Kotsev, A., & Lutz, M. (2019). Comparing INSPIRE and OpenStreetMap data: how to make the most out of the two worlds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4/W14*, 167–174.
[4] Humanitarian OpenStreetMap Team (2023) *Home*. https://www.hotosm.org
[5] Neis, P., & Zipf, A. (2008). *Openrouteservice. org is three times "open": Combining OpenSource, OpenLS and OpenStreetMaps.* GIS Research UK (GISRUK 08).
[6] Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (2015). An introduction to OpenStreetMap in geographic information science: Experiences, research and applications. In J. J. Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.) *OpenStreetMap in GIScience* (pp. 1-15). Springer.
[7] Juhász, L., & Hochmair, H. H. (2018). OSM data import as an outreach tool to trigger community growth? A case study in Miami. *ISPRS International Journal of Geo-Information*, 7, 113.
[8] Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., & Zipf, A. (2020). Volunteered Geographic Information Research in the First Decade: A Narrative Review of Selected Journal Articles in GIScience. *International Journal of Geographic Information Science*, *34*(9), 1765–1791.
[9] Grinberger, A. Y., Minghini, M., Juhász, L., Yeboah, G., & Mooney, P. (2022). OSM Science—The Academic Study of the OpenStreetMap Project, Data, Contributors, Community, and Applications. *ISPRS International Journal of Geo-Information*, *11*(4), 230.
[10] *Science Info Page*. https://lists.openstreetmap.org/listinfo/science
[11] Haklay, M. (2015). Foreword: OpenStreetMap studies and volunteered geographic information. In J J. Arsanjani, , A. Zipf, P. Mooney, & M. Helbich (Eds.) *OpenStreetMap in GIScience: Experiences, Research, and Applications* (pp. v-vii). Springer.
[12] OSM Science Scientific Committee (2023). *OSMScience 2023*. https://shorturl.at/beAP9
[13] Grinberger, A. Y., Anderson, J., Mooney, P., Ludwig, C., & Minghini, M. (2021). OpenStreetMap as a multifaceted research subject: the Academic Track at State of the Map 2021. In M. Minghini, C. Ludwing, J. Anderson, P. Mooney, & A. Y. Grinberger. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021* (pp. 1–5). Zenodo.
[14] Oostwegel, L. J. N., Evaz Zadeh, T., Lingner, L. & Schorlemmer, D. (2023). A global and dynamic completeness assessment of the OpenStreetMap buildings. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 6–9). Zenodo.
[15] Andorful, F., Lautenbach, S., Ludwig, C., Herfort, B., Nir, F., & Zipf, A. (2023). Exploring Road and Points of Interest (POIs) Associations in OpenStreetMap, A New Paradigm for OSM Road class Prediction. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 69–72). Zenodo.
[16] O'Brien, O. (2023). Towards an Open High-Resolution Land Use Dataset in Great Britain – Comparing and Consolidating Retail Centre Areas from Open Data Sources. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 29–32). Zenodo.
[17] Scheck, E., Ledermann, F., Binn, A. & Dörk, M. (2023). Mapping Public Space in Urban Neighbourhoods Using OpenStreetMap Data. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 33–37). Zenodo.
[18] Herringer, M., Ndiaye, L., & South, A. (2023). Developing a data validation method with OpenStreetMap Senegal and the Ministry of Health in support of accurate health facility data. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 42–44). Zenodo.
[19] Li, H. & Sun, Y. (2023). Beyond Two Dimensions: Large-Scale Building Height Mapping in OpenStreetMap via Synthetic Aperture Radar and Street-View Imagery. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 38–41). Zenodo.

[20] Olivari, B., & Cimini, A. (2023). Are Italian cities already 15-minute? Presenting a glocal proximity index, based on open data. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 57–60). Zenodo.
[21] Evaz Zadeh, T., Oostwegel, L. J. N., Lingner, L., Shinde, S., Cotton, F., & Schorlemmer, D. (2023). Improving the accuracy of earthquake risk estimates with OpenStreetMap building data. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 18–21). Zenodo.
[22] Vierø, A. R., Vybornova, A. & Szell, M. (2023). Assessing OpenStreetMap bicycle data quality with BikeDNA: a Denmark Case Study. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 22–25). Zenodo.
[23] Passmore, R., Guensler, R., & Watkins, K. (2023). Assessing Bike-Transit Accessibility with OpenStreetMap. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 53–56). Zenodo.
[24] Fila, M., Štampach, R., &Benjamin Herfort, B. (2023). Global and regional level of use of buildings and roads prepared by AI for OSM mapping. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 10–13). Zenodo.
[25] Vestena, K., Camboim, S., & Santos, D. (2023). Fostering OSM's Micromapping Through Combined Use of Artificial Intelligence and Street-View Imagery. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 14–17). Zenodo.
[26] Gramacki, P., Leśniara, K., Raczycki, K., Woźniak, S., & Szymański, P. (2023). Utilizing OSM data in geospatial representation learning. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 45–48). Zenodo.
[27] Melanda, E. A., Herfort, B., Ulrich, V., Andorful, F., Zipf, A. (2023). OpenStreetMap Data for Automated Labelling Machine Learning Examples: The Challenge of Road Type Imbalance. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 65–68). Zenodo.
[28] Khellaf, L., Schlicht, I. B., Bayer, J., Bouwmeester, R., Miraß, T., & Tilman Wagner, T. (2023). Spot: A Natural Language Interface for Geospatial Searches in OSM. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 49–52). Zenodo.
[29] Geddes, A. (2023). Rural water point mapping with/in OSM: implications of recent research in Malawi. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 61–64). Zenodo.
[30] Schröder-Bergen, S. (2023). Social, technical and political transformations in OpenStreetMap – From volunteered geographic information to embedding digital commons in platform capitalism. In M. Minghini, H. Li, A. Y. Grinberger, P. Liu, G. Yeboah, L. Juhász, S. Coetzee, P. Mooney, A. Sarretta, & J. Anderson (Eds.). *Proceedings of the OSM Science 2023* (pp. 26–28). Zenodo.
[31] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, *8*(5), 232.
[32] OSM Contributors (2023). *Vandalism and blocks in Israel*. https://community.openstreetmap.org/t/ vandalism-and-blocks-in-israel/105176
[33] Juhász, L., Mooney, P., Hochmair, H. H., & Guan, B. (20230). *ChatGPT as a mapping assistant: A novel method to enrich maps with generative AI and content derived from street-level photographs*. arXiv. https://arxiv.org/abs/2306.03204
[34]. Overture Maps Foundation (2022). *Overture Maps Foundation.* https://overturemaps.org

A global and dynamic completeness assessment of the OpenStreetMap buildings

Laurens J.N. Oostwegel[1,*], Tara Evaz Zadeh[1], Lars Lingner[1] and Danijel Schorlemmer[1]

[1] Seismic Hazard and Risk Dynamics, GFZ German Research Centre for Geosciences, Potsdam, Germany; laurens.jozef.nicolaas.oostwegel@gfz-potsdam.de; tara.evaz.zadeh@gfz-potsdam.de; lars.lingner@gfz-potsdam.de; danijel.schorlemmer@gfz-potsdam.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

The OpenStreetMap (OSM) building dataset is rapidly expanding at a rate of approximately five million buildings per month, or approx. 2 buildings per second. Contributions range from hobbyists mapping an area, official registries that are imported, to the mapping by humanitarian organizations. The joint efforts lead to a non-uniform result, with some regions well-mapped, while others are lacking basic geographic features. The status quo of building completeness is desirable to identify areas on a global scale where OSM is (almost) complete but also where the most effort is needed. Furthermore, because of the non-uniformity, it can be useful to know for the end user if any buildings are missing at their location of interest.

A map is never complete and the ever-changing dataset of OSM embodies this well. With digital maps, where the map scale becomes a fuzzy concept, more details can be added any time and one can improve the accuracy of geometries endlessly. For us, an area is considered complete, as long as all the buildings have been mapped and the footprints follow the outline of the building roughly. Attributes, such as the height, have been disregarded in this research, but are well researched in [1]. Other factors, such as the time of the last modification of a building have been looked at by [2,3]. Also the quality of the building geometry has been well researched [4–6], but was not considered in this research.

We create a dynamic assessment of the completeness of OSM buildings on a global scale. While completeness assessment datasets exist on a small scale [7,8], near-global scale [9,2], or even global scale [10], human development and the nature of OSM ensures that as soon as such a dataset is published, it is already outdated. For a meaningful completeness assessment, it is crucial to update it the moment that a building is added, modified or removed.

Semi-automatically generated datasets based on earth observation data and Machine Learning have been assessed to be used as reference models. These include the Global Human Settlement Layer (GHSL) of 2022 [11], Microsoft Building Footprints [12], and Google Open Buildings [13]. Contrary to the other datasets, GHSL is raster-based, with pixels indicating whether an area contains built up areas. The raster files were vectorized, so that

Oostwegel, J.N.L., Evaz Zadeh, T., Lingner, L., & Schorlemmer, D. (2023). A global and dynamic completeness assessment of the OpenStreetMap buildings
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443307

they can be used in the same manner as the other datasets. We have carefully examined these datasets and decided that the Google buildings have the most accurate footprints, followed by the GHSL dataset. The GHSL dataset, due to its nature as a 10x10 meter resolution raster dataset, consistently overestimates the built-up area compared to OSM. The Microsoft dataset, while covering most countries in the world, is itself incomplete in many areas. The data is very inconsistent: even within cities there can be well-covered areas as well as areas missing altogether. As the coverage of buildings in OSM in many areas is of much better quality, we deemed the Microsoft dataset unfit for the purpose of completeness estimation. The Google dataset seems to have the most complete estimate for the countries that they provide. Therefore we used the Google buildings for countries when available and otherwise used GHSL.

An initial state of OSM completeness was computed on an OSM planet file generated reflecting the exact building data of 1 January 2022, midnight UTC. The built area of OSM and the proxy datasets have been computed on a Quadtree grid using level-18 tiles. Their size is approximately 100x100 meters in Europe and slightly larger than 150x150 meters near the equator. As there are 64 billion of such tiles, a recursive method is used for calculation. It is initiated on a level-1 grid, which divides the world into four tiles. For each tile, it is checked whether it contains any built area and if it does, it is divided into four new tiles. This process is repeated for all resulting tiles of the next zoom level, until the level-18 has been reached. For this zoom level, the total square meter of built area within each tile is calculated for the OSM and proxy datasets. Finally, they are compared to each other to estimate the completeness. The data have been aggregated to level-15 tiles (between 800x800 and 1200x1200 meter), to examine the differences between aggregation levels. The result of the level-15 tiles can be found in Figure 1.

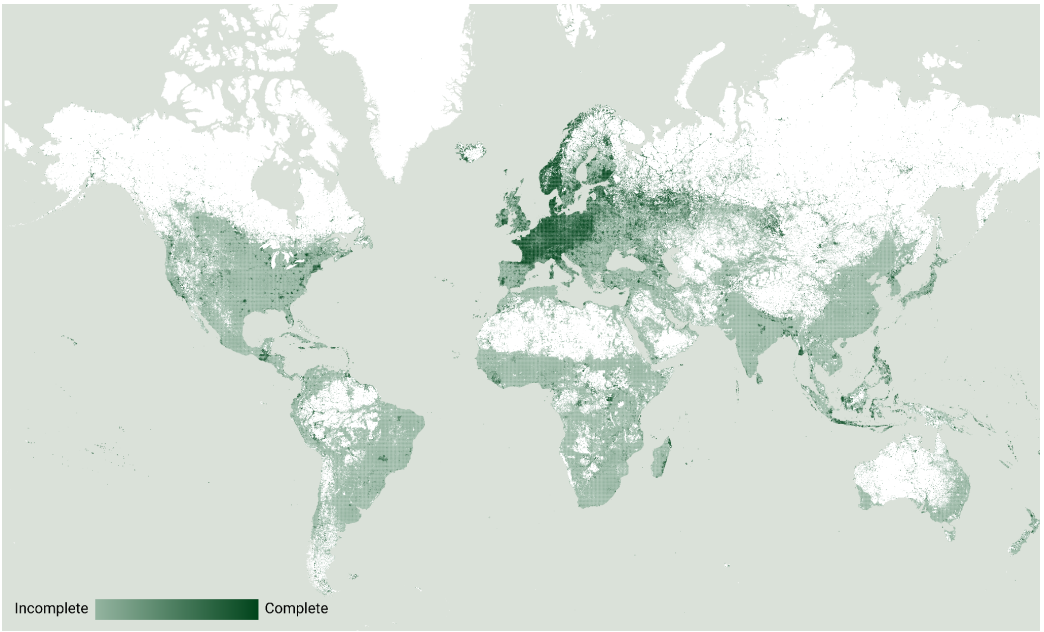

Figure 1. Building completeness of OSM on 1 January 2023, aggregated onto zoom-level-13 Quadkey tiles.

We have identified a list of countries (e.g. France; the Netherlands; Denmark) that we consider to be complete, partly through knowledge of imports and partly through our own

judgment. For these countries we assessed the accuracy of the completeness assessment using the GHSL dataset and we found that the level-15 tiles were more reliable to use, due to the quality of the data in the proxy dataset. As the GHSL dataset has a resolution of 10x10 meter pixels, it is not as detailed as the OSM buildings. Therefore, not only is there a consistent overestimation of built area in the proxy dataset, there may also be false positive built area in some tiles. A level-15 resolution provides a less precise assessment, but a better result. Therefore there is a trade-off between accuracy and precision, where at a larger scale the level-15 assessment is preferred, while at a smaller scale (e.g. a city), the grid of ~1x1 kilometer is a bit coarse. Unfortunately, none of the countries within the coverage of the Google dataset are considered complete in OSM, therefore such validation of the results cannot be made.

Building changes are detected on a minute-by-minute basis. A change in a building geometry triggers a recalculation of the built area of the tile(s) intersecting the building. Also the comparison between the proxy dataset and the building dataset is recalculated. In this way, each change to buildings in OSM is detected and processed to produce a completeness map that is always up-to-date with the latest changes.

The project was built using the Overpass API [14], with the augmented diffs tracking the minutely changes. The Osmium tool [15] has been used to replicate the complete OSM building dataset, for fast processing. All code that is used in this project is available on our GitLab under the GNU AGPLv3 license (https://git.gfz-potsdam.de/globaldynamicexposure). The data that are continuously produced for this project are made available under ODbL and can be found at https://www.openbuildingmap.org/. GHSL, Google Open Buildings and OSM are all open datasets, with the latter two also found under the ODbL license.

The OSM community has always been divided on whether large datasets should be imported into OSM [16]. Importing them directly has many disadvantages. For example, the footprints of the buildings can be crooked and not represent the actual shapes of the building. However, both datasets by Google and Microsoft contain many buildings not present in OSM at all. This research brings us one step closer to a practical purpose, where the tiles that are considered incomplete, compared to the reference datasets can be highlighted per region or country. One of the main applications can be found in the addition of buildings to OSM, even more so in cases when it needs to be decided quickly where the most buildings need to be added geographically, such as humanitarian mapping efforts in the event of a disaster. Our work can be used complementary to the efforts in MapSwipe, where one can manually assess the completeness of the OSM building dataset, because MapSwipe uses the same tiling system [17]. An up-to-date completeness assessment can also greatly benefit users of OSM data for various applications. While a complete dataset does not automatically result in data of good quality, we know for certain that there is a lack of quality when the dataset is not complete. On top of that we can identify not only the most incomplete areas, but also the most incomplete areas that do not have a lot of mapping activity over time.

Further effort is needed to validate the completeness estimation, for example by using up-to-date registry datasets of areas that are not yet complete in OSM. The completeness estimation could be improved, by using other data sources, such as the proximity to OSM street networks, that could indicate the likelihood of a building existing in a proxy dataset.

References

[1] Biljecki, F., Chow, Y. S., & Lee, K. (2023). Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. *Building and Environment*, *237*, 110295. https://doi.org/10.1016/j.buildenv.2023.110295

[2] Herfort, B., Lautenbach, S., Porto De Albuquerque, J., Anderson, J., & Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. *Nature Communications*, *14*(1), 3985. https://doi.org/10.1038/s41467-023-39698-6

[3] Minghini, M., & Frassinelli, F. (2019). OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date? *Open Geospatial Data, Software and Standards*, *4*(1), 9. https://doi.org/10.1186/s40965-019-0067-x

[4] Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

[5] Brovelli, M., & Zamboni, G. (2018). A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. *ISPRS International Journal of Geo-Information*, *7*(8), 289. https://doi.org/10.3390/ijgi7080289

[6] Jacobs, K. T., & Mitchell, S. W. (2020). OpenStreetMap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, *24*(5), 1280–1298. https://doi.org/10.1111/tgis.12680

[7] Biljecki, F., & Ang, L. M.. (2020). *Assessing global OpenStreetMap building completeness to generate large-scale 3D city models*. https://doi.org/10.5281/ZENODO.3922285

[8] Orden, A., Flores, R. A., Faustino, P., & Samson, M. S. (2020). *Measuring OpenStreetMap building footprint completeness using human settlement layers*. https://doi.org/10.5281/ZENODO.3923033

[9] Zhou, Q., Zhang, Y., Chang, K., & Brovelli, M. A. (2022). Assessing OSM building completeness for almost 13,000 cities globally. *International Journal of Digital Earth*, *15*(1), 2400–2421. https://doi.org/10.1080/17538947.2022.2159550

[10] Oostwegel, L. J. N., Garcia Ospina, N., Evaz Zadeh, T., Shinde, S., & Schorlemmer, D. (2023). *Automatic global building completeness assessment of OpenStreetMap using remote sensing data*. https://doi.org/10.5194/egusphere-egu23-13160

[11] Pesaresi, M., & Politis, P. (2022). *GHS-BUILT-C R2022A - GHS Settlement Characteristics, derived from Sentinel2 composite (2018) and other GHS R2022A data*. European Commission, Joint Research Centre (JRC). https://doi.org/10.2905/DDE11594-2A66-4C1B-9A19-821382AED36E

[12] Microsoft. (2023). *Microsoft Building Footprints* [dataset]. https://github.com/microsoft/GlobalMLBuildingFootprints

[13] Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). *Continental-Scale Building Detection from High Resolution Satellite Imagery*. arXiv. http://arxiv.org/abs/2107.12283

[14] Olbricht, R. M. (2015). *Data Retrieval for Small Spatial Regions in OpenStreetMap*. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience* (pp. 101–122). Springer International Publishing. https://doi.org/10.1007/978-3-319-14280-7_6

[15] osmcode (2023). *Osmium Command Line Tool*. https://github.com/osmcode/osmium-tool

[16] Witt, R., Loos, L., & Zipf, A. (2021). *Analysing the Impact of Large Data Imports in OpenStreetMap. ISPRS International Journal of Geo-Information*, 10(8), 528. https://doi.org/10.3390/ijgi10080528

[17] Ullah, T., Lautenbach, S., Herfort, B., Reinmuth, M., & Schorlemmer, D. (2023). Assessing Completeness of OpenStreetMap Building Footprints Using MapSwipe. *ISPRS International Journal of Geo-Information, 12*(4), 143. https://doi.org/10.3390/ijgi12040143

Global and regional level of use of buildings and roads prepared by AI for OSM mapping

Milan Fila[1,*], Radim Štampach[1] and Benjamin Herfort[2]

[1] Department of Geography, Faculty of Science, Masaryk University, Brno, Czechia;
fila.milan@mail.muni.cz, stampach@mail.muni.cz
[2] Heidelberg Institute for Geoinformation Technology, Heidelberg, Germany;
benjamin.herfort@heigit.org

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

OpenStreetMap (hereinafter OSM) is created by volunteer mappers, either individuals or organized groups. Most of them map individually, and some of them during organized mapping parties at so-called mapathons. The motivation of these volunteers is altruism and the effort to be beneficial to the community [1].

However, despite their great efforts, many world areas are still insufficiently mapped. In 2020, regions with low and medium human development accounted for 28% of the buildings and 16% of the roads mapped in OSM, although they were home to 46% of the global population [2]. Using AI-assisted tools was seen as one possibility to fill data gaps. The main aim of our study is to research the actual level of AI-assisted mapping in OSM. We show the overall global level and highlight differences between countries.

The most developed projects using AI objects for OSM mapping are the Rapid editor and mapwith.ai plugin for the JOSM editor. These tools allow mappers to use prepared AI buildings generated by Microsoft [3] and Google in cooperation with ESRI [4, 5] and prepared AI roads generated by Meta [6]. Microsoft organized mapping in Kenya and Nigeria using their global AI dataset of buildings [7]. Meta similarly uses their global AI dataset of roads for mapping in Thailand, Indonesia, Malaysia, India, Tanzania, and Vietnam [8]. The Humanitarian OpenStreetMap Team is working on a fAIr project that could produce local models generating AI data for specific regions [9].

During the crisis after the Turkey-Syria Earthquake in 2023, these technologies played a pivotal role in disaster response, illustrating the potential for AI in supporting humanitarian mapping [10]. However, the OSM community's reactions to the results of the AI-assisted mapping campaign were often mixed [10, 11].

The main research question of our study was, "What are the differences between AI-assisted mapping and traditional mapping concerning editing behavior?" The question deals with the editing pattern of AI and non-AI data. There are concerns within the OSM community that AI-assisted mapping will lead to less community engagement and could undermine the contributor base of OSM in the long run. OSM community attitudes toward

Fila, M., Štampach, R., & Herfort, B. (2023). Global and regional level of use of buildings and roads prepared by AI for OSM mapping
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443308

imports and automated edits can be negative and cautious [12]. Human mappers can be mistrustful of the result of AI algorithms that often remain in a black box for them [13], or they can even feel that they are being replaced by AI algorithms for OSM mapping.

We used an extract of all contributions made to OSM during 2021-2023 derived from a combination of the OSHDB developed by the Heidelberg Institute for Geoinformation Technology (hereinafter HeiGIT) and the OSM changeset database. There are three types of operations with objects in OSM: creating, editing (=modifying), and deleting. We focused on creating and editing and searched for differences between buildings and roads mapped using AI and buildings and roads in general.

Buildings were defined for our study as objects with any value in tag *building=**. Buildings mapped with the use of AI-assisted tools (hereinafter AI buildings) are subset from buildings identified by element tag source with values "*microsoft/BuildingFootprints*" (prepared by Microsoft) and "*esri/Google_Africa_Buildings*" (prepared by Google and ESRI).

Only the most important types of roads were used in our study to analyze the roads because they form the transportation network's backbone. They are identified by the following values of tag *highway*: *motorway, motorway_link, trunk, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, unclassified, residential* [14]. Roads mapped using AI-assisted tools (hereinafter AI roads) are subset from roads identified by the following values in changeset tag hashtags: #nsroadimport, #mapwithai, #MapWithAI. Also, the value "*RapiD*" in the *element* tags can identify the AI road.

The numbers of created and edited AI buildings in 2021-2023 were extracted to analyze the buildings. They were divided by the number of all created and all edited buildings. Lengths of created and edited AI roads in 2021-2023 were extracted to analyze the roads. They were divided by the total lengths of created and edited roads.

The relative share of AI buildings from all newly created buildings was 5.2% in 2021, 12% in 2022, and 14.8% in June 2023. The relative share of AI buildings from all edited buildings was 0.7% in 2021, 3.2% in 2022, and 4.2% in June 2023.

The relative share of AI roads from all newly created important roads was 14.7% in 2021, 15.1% in 2022, and 7.7% in June 2023. The relative share of AI roads from all edited important roads was 8.3% in 2021, 5.7% in 2022, and 3.1% in June 2023.

The results show that the global level of using AI buildings and AI roads for creating new buildings and roads is always under 20%. It is increasing year to year in the case of buildings, but not in the case of roads. Portion of AI buildings and AI roads is higher in the case of creating new objects than in the case of editing existing objects. It means that AI buildings and AI roads are (after their creation) only rarely actualized.

We also searched for differences in using AI-assisted mapping between countries. We wanted to know the absolute numbers of AI buildings and the relative portion of AI buildings on all buildings in every country. We used ohsome API developed by HeiGIT to reach the results, allowing access to the entire OSM history. This analysis was prepared only for buildings because ohsome API allows to search only element tags, not changeset tags, so identification of AI roads was not possible.

The result is in Figure 1. The highest absolute and relative numbers of AI buildings can be found in the USA, Turkey, and Kenya. A high absolute number is also in Nigeria. High relative numbers are also found in Morocco, Venezuela, and Australia, but the absolute numbers of AI buildings are low.

This result shows that the higher relative share of AI buildings and roads exists almost only in countries where Microsoft coordinates mapping using AI datasets they created (USA, Kenya, Nigeria) or where a special humanitarian mapping project using AI-assisted tools was held by The Humanitarian OpenStreetMap Team (Turkey).

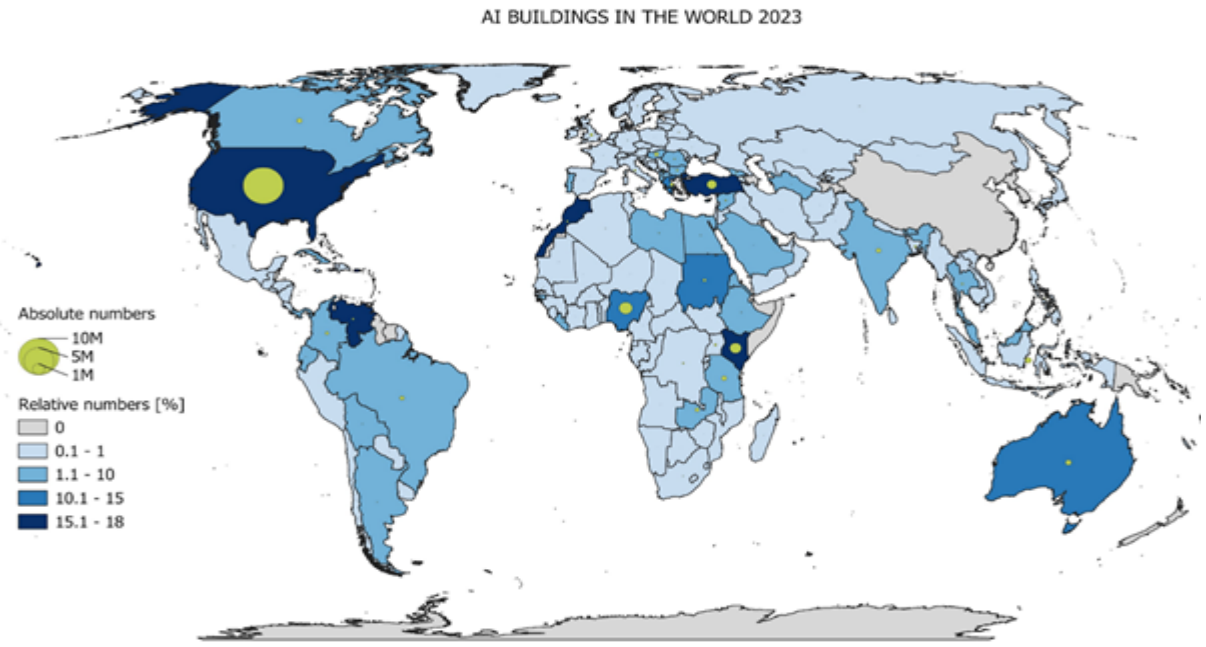

Figure 1. Numbers of AI buildings and portion (in %) of AI buildings on all buildings.

References

[1] Štampach R., Herman L., Trojan J., Tajovská K., & Řezník T. (2021). Humanitarian Mapping as a Contribution to Achieving Sustainable Development Goals: Research into the Motivation of Volunteers and the Ideal Setting of Mapathons. *Sustainability, 13*(24), 13991.

[2] Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific Reports, 11*, 3037.

[3] Microsoft (2022). *microsoft/GlobalMLBuildingFootprints [dataset]*. https://github.com/microsoft/GlobalMLBuildingFootprints

[4] Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). *Continental-Scale Building Detection from High Resolution Satellite Imagery*. arXiv. http://arxiv.org/abs/2107.12283

[5] Kensok, D. (2021). *Africa Buildings*. https://openstreetmap.maps.arcgis.com/home/item.html?id=660457fac76344b195c555e0dff386ff

[6] Facebook (2019). *Open-Mapping-At-Facebook.* https://github.com/facebookmicrosites/Open-Mapping- At-Facebook

[7] Humanitarian OpenStreetMap Team (2022). *Satellite Imagery for Social Good: Kenya and Nigeria.* https://www.hotosm.org/projects/satellite-imagery-for-social-good

[8] OpenStreetMap (2021). *Facebook AI-Assisted Road Tracing.* https://wiki.openstreetmap.org/w/index.php?title=Facebook_AI-Assisted_Road_Tracing&oldid=2115828

[9] Najjar, O. (2022). *hot_tech_talk fAIr: AI-assisted mapping.* https://www.hotosm.org/tech-blog/hot-tech-talks-fair

[10] Ngumenawe, S. (2023). *HOT's Approach to OSM Data Validation for Earthquake Response Mapping.* https://www.hotosm.org/updates/hot-approach-to-osm-data-validation-to-eq-mapping-projects

[11] Humanitarian OpenStreetMap Team (2023). *Session 2: What are your perspectives on AI assisted mapping in OSM? [Video]*. YouTube. https://www.youtube.com/watch?v=6pfdDV9xSoo
[12] Huck, J. J., Perkins, C., Haworth, B. T., Moro, E. B., & Nirmalan M. (2021). Centaur VGI: A Hybrid Human–Machine Approach to Address Global Inequalities in Map Coverage. *Annals of the American Association of Geographers, 111*(1), 231-251.
[13] Franzen, M., Kloetzer, L., Ponti, M., Trojan, J., & Vicens, J. (2021). Machine Learning in Citizen Science: Promises and Implications. In: K. Vohland, L. Ceccaroni, J. Perelló, R. Samson, & L. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 183-198). Springer.
[14] OpenStreetMap (2023) *Key:highway.* https://wiki.openstreetmap.org/w/index.php?title=Key:highway&oldid=2524607

Fostering OSM's Micromapping Through Combined Use of Artificial Intelligence and Street-View Imagery

Kauê Vestena[1,*], Silvana Camboim[1] and Daniel Santos[2]

[1] Department of Geomatics, Federal University of Paraná, Curitiba, Brazil; kauemv2@gmail.com, silvanacamboim@gmail.com
[2] Cartographic Engineering Sector, Military Engineering Institute, Rio de Janeiro, Brazil; daniel.rodrigues@ime.eb.br

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Map scale is fundamental to cartography. The International Cartographic Association's definition of a map [1] already explicitly emphasizes the selection of specific features, pointing to the process of cartographic generalization. This operation simplifies the representation of geographic data to produce a map at a given scale [2]. In the past, geospatial data was collected in a standardized way. Representations at larger scales could then be derived. However, OpenStreetMap (OSM) has changed this approach. Each object can be captured individually, resulting in digital representations of varying and distinct accuracies. Nevertheless, products such as OSM's Slippy Map Tiles are designed for consistent scales, maintaining a uniform scale for each tile. This characteristic gives users a seemingly seamless view of the map. The tile layers on the OpenStreetMap website range in maximum scale from around 1:2000 to 1:250 [3,4], suitable for mapping urban detail. In the context of collaborative mapping, the term Micromapping was coined as the "mapping of small geographic objects" [5] and appears as a topic of growing interest among the OpenStreetMap and general Volunteered Geographic Information community [5,6,7,8], can be helpful in many applications like mapping large-scale infrastructure [5]; pedestrian security and flow prediction [6]; detailed 3D model generation and indoor mapping [7]; assistive technologies like tactile maps generation [8]; and also general-purpose micro mapping rendering [9]. There is also some discussion among the OSM community about their idiosyncrasies, comprising issues that may arise when there is a bigger density of features [10]. However, it is important to remember that omitting some elements at smaller scales can improve the readability of the map. On the other hand, many street-level editors, such as MapComplete [11] emerged, bringing micro-mapping with a central role.

As of August 2023, a tiny share of mapped features can still be categorized as micromapping. According to Taginfo [12], while there are almost 572 million buildings and more than 221 million roads, the most common micromapping objects are 22 million natural=tree nodes, 12 million power=pole nodes, and 8 million highway=crossing nodes.

Vestena, K., Camboim, S., & Santos, D. (2023). Fostering OSM's Micromapping Through Combined Use of Artificial Intelligence and Street-View Imagery
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443310

Other important and very common urban features are also relatively ill-represented, e.g. 1.2 million benches, 1.8 million fire hydrants, and 1.2 million traffic signals. In these 3 cases, all features tend to be concentrated in Europe and the US. The literature also points out that OSM is not accumulating many contributors outside major urban areas, resulting in less dense data in those scenarios [13].

In spite of this, there is a need for quicker methods to generate data belonging to this category, which can contain objects such as those mentioned before but also properties like the width of sidewalk lanes and the presence of access ramps. Among the data freely available for extracting features for OSM, georeferenced street view imagery is one of the most valuable assets, highlighting Mapillary and KartaView platforms [14]. Terrestrial images possess a wealth of visual data that can be transformed into geospatial information. This informal description aligns with the well-established scientific discipline of photogrammetry, which has employed traditional methods like triangulation since the 19th century in conjunction with the advent of photography. These techniques have enabled the mapping of entire cities [15] and the precise capture of 3D points through multiple views [16]. Nevertheless, traditional photogrammetric procedures rely on good triangle geometry, which can be tricky considering terrestrial exposures [17] and depends on good correspondences that can fail or be challenging to achieve on surfaces without texture [18].

The advent of Artificial Intelligence (AI) represents a paradigm shift, introducing novel methods that deliver unparalleled performance in areas such as general automation and predictive analysis powered by big data [19]. In the present work, we are concerned with two tasks with massive recent improvements due to AI's use: Semantic segmentation [20,21,22], which is the capability to segment meaningfully digital images, giving a label for each pixel, and Monocular Depth [23,24,25] which enables to calculate 3D coordinates for any pixel. We propose combining them to harness AI's potential to speed up data generation for OSM. The proposed methodology will use street-view imagery (SVI) and metadata as primary input.

We chose the SAM - Segment Anything Model [26] and Paddleseg [27], both available under the OSM-friendly Apache-2.0 Licence, for the semantic segmentation task. We applied the MIT-Licensed ZoeDepth model [28] for Monocular Depth. The fusion process works as follows: the user selects an image whose depth map is created using ZoeDepth, then prompts for a category like "pole" or "bench", and a list of matching segments (2D pixel patches) is returned by the Paddleseg using SAM backend; then the same image regions are clipped from the depth image. In the next step, through the pixel-wise use of the projective equations alongside the known depth, 3D coloured patches are generated.

The second part is to translate those 3D patches into OSM data. This work comprises many possible features, but we will explain through an example of a street light. From the detection, we extract the base tag "highway=street_lamp". From the 3D patch, we pick the lowest point to create the geometry (latitude and longitude), the absolute elevation of the point on the ground for the tag "ele:wgs84=*" and its difference with the highest point to provide the "height=*" tag. Then, from both 2D and 3D data, other tags can be inferred, such as support=*, lamp_mount=*, light:direction=* and so on. For each category, a different set of steps shall be taken alongside the inferable tags for additional characteristics. Some patches shall provide only additional information, e.g. width=* and incline:transversal=* in the example of already-existent sidewalks mapped as separate geometries.

Up to the present, the methodology has already been tested only to the point of the 3D patch generation with a single test of the extraction of the slope of a ramp that was

measured in the cloud (angle between the normals of ramp and street-section least-squared fitted planes) as 9.13° against a smartphone-made field measurement of 9°. The forthcoming months shall follow with tests regarding the obtainable accuracy and possible improvements like the use of track-anything [29] that expand SAM to the capability of tracking the same object among many frames, enabling redundancy and thus uncertainty estimation. Another topic of concern will be how to properly fuse this data with OSM, considering that both data sources may have inconsistent spatial references.

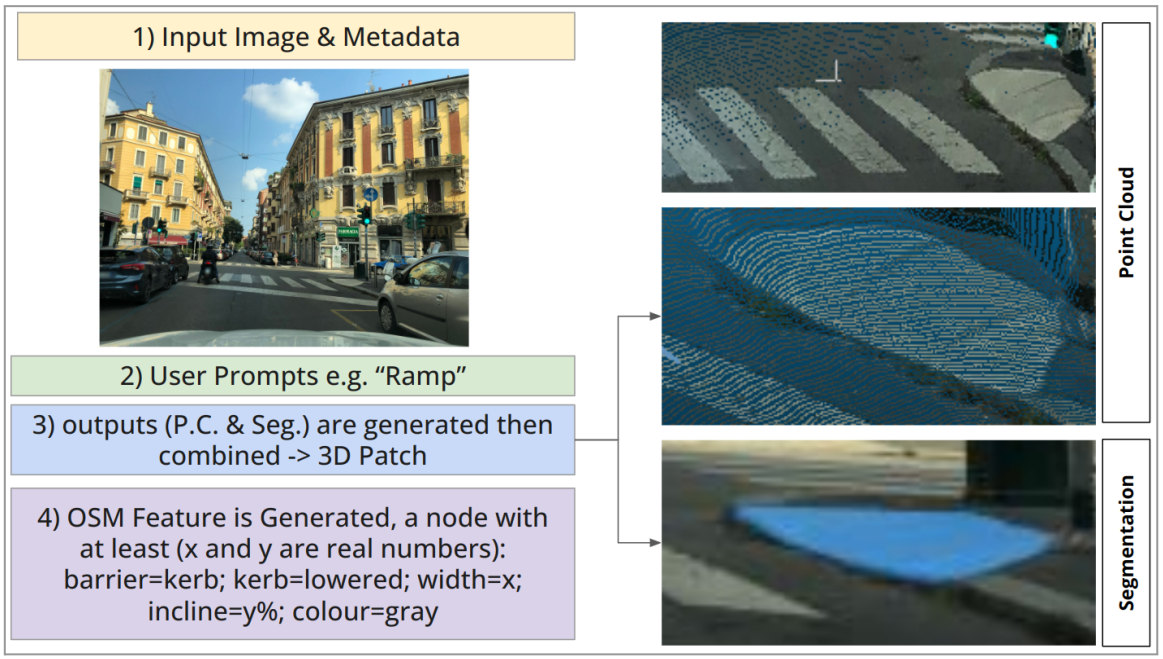

Figure 1. The Workflow Shown Through the example of a Ramp

References

[1] International Cartographic Association (2023). *Mission.* https://icaci.org/mission

[2] Ruas, A. (2008). Map Generalization. In S. Shekhar & H. Xiong (Eds.), *Encyclopedia of GIS* (pp. 631–632). Springer. https://doi.org/10.1007/978-0-387-35973-1_743

[3] Touya, G., & Reimer, A. (2015). Inferring the Scale of OpenStreetMap Features. In Lecture Notes in Geoinformation and Cartography (pp. 81–99). Springer International Publishing. https://doi.org/10.1007/978-3-319-14280-7_5

[4] OSM Contributors (2023). *Zoom levels*. https://wiki.openstreetmap.org/wiki/Zoom_levels

[5] Schmid, F., Cai, C., & Frommberger, L. (2012). A new micro-mapping method for rapid VGI-ing of small geographic features. *Proceedings of the 7th International Conference on Geographic Information Science, USA*, 18–21.

[6] Cohen, A., & Dalyot, S. (2019). Pedestrian Traffic flow prediction based on ANN model and OSM data. *Proceedings of the International Cartographic Association, 2*(20).

[7] Goetz, M. (2013). Towards generating highly detailed 3D CityGML models from OpenStreetMap. *International Journal of Geographical Information Science, 27*(5), 845–865.

[8] Fillières-Riveau, G., Favreau, J. M., Barra, V., & Touya, G. (2020). Génération de cartes tactiles photoréalistes pour personnes déficientes visuelles par apprentissage profond. *Revue Internationale de Géomatique, 30*(1-2), 105–126.

[9] OSM Contributors. (2023). Straßenraumkarte Neukölln. OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/Stra%C3%9Fenraumkarte_Neuk%C3%B6lln

[10] OSM Contributors (2023). *Micromapping*. https://wiki.openstreetmap.org/wiki/Micromapping

[11] MapComplete (2022). *MapComplete - editable, thematic maps with OpenStreetMap.* https://mapcomplete.osm.be
[12] OSM contributors (2023). *Taginfo*. https://taginfo.openstreetmap.org
[13] Quinn, S. (2015). Using small cities to understand the crowd behind OpenStreetMap. *GeoJournal, 82(3)*, 455–473. https://doi.org/10.1007/s10708-015-9695-6
[14] Alvarez Leon, L. F., & Quinn, S. (2018). The value of crowdsourced street-level imagery: examining the shifting property regimes of OpenStreetCam and Mapillary. *GeoJournal, 84*(2), 395–414. https://doi.org/10.1007/s10708-018-9865-4
[15] Grimm, A. (2007). New Products and Services from IGI. In D. Fritsch (Ed.), *Photogrammetric Week 2007* (pp. 1-10). Wichmann Verlag. https://phowo.ifp.uni-stuttgart.de/publications/phowo07/080Grimm.pdf
[16] Luhmann, T., Robson, S., Kyle, S., & Harley, I. A. (2007). *Close Range Photogrammetry: Principles, Techniques, and Applications*. https://api.semanticscholar.org/CorpusID:107005019
[17] Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811685
[18] Chen, Y.-S., & Chen, B.-T. (2002). A solution of correspondence problem for measuring 3D surface. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 4*, IV-3553-IV-3556. https://doi.org/10.1109/ICASSP.2002.5745422
[19] Lu, M. (2023). The industrial revolution brought by AI: Opening the Gateway to the Future. *Geographical Research Bulletin, 0*(2), 166–168. https://doi.org/10.50908/grb.2.0_166
[20] Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval, 7*, 87-93. https://doi.org/10.1007/s13735-017-0141-z
[21] Hao, S., Zhou, Y., & Guo, Y. (2020). A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing, 406*, 302–321. https://doi.org/10.1016/j.neucom.2019.11.118
[22] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., & Cottrell, G. (2018). Understanding Convolution for Semantic Segmentation. *In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. https://doi.org/10.1109/wacv.2018.00163
[23] Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences, 63*(9), 1612–1627. https://doi.org/10.1007/s11431-020-1582-8
[24] Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing, 438*, 14–33. https://doi.org/10.1016/j.neucom.2020.12.089
[25] Bhoi, A. (2019). *Monocular Depth Estimation: A Survey*. arXiv. https://doi.org/10.48550/ARXIV.1901.09402
[26] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything (Version 1)*. arXiv. https://doi.org/10.48550/ARXIV.2304.02643
[27] Liu, Y., Chu, L., Chen, G., Wu, Z., Chen, Z., Lai, B., & Hao, Y. (2021). *PaddleSeg: A High-Efficient Development Toolkit for Image Segmentation (Version 1)*. arXiv. https://doi.org/10.48550/ARXIV.2101.06175
[28] Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., & Müller, M. (2023). *ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth (Version 1)*. arXiv. https://doi.org/10.48550/ARXIV.2302.12288
[29] Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., & Zheng, F. (2023). *Track Anything: Segment Anything Meets Videos (Version 2)*. arXiv. https://doi.org/10.48550/ARXIV.2304.11968

Improving the accuracy of earthquake risk estimates with OpenStreetMap building data

Tara Evaz Zadeh[1,*], Laurens J.N. Oostwegel[1], Lars Lingner[1], Simantini Shinde[1], Fabrice Cotton[1] and Danijel Schorlemmer[1]

[1] Seismic Hazard and Risk Dynamics, GFZ German Research Centre for Geosciences, Potsdam, Germany; tara.evaz.zadeh@gfz-potsdam.de; laurens.jozef.nicolaas.oostwegel@gfz-potsdam.de; lars.lingner@gfz-potsdam.de; shinde@gfz-potsdam.de; fabrice.cotton@gfz-potsdam.de; danijel.schorlemmer@gfz-potsdam.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Seismic risk models provide an estimation of human and building loss and damage due to earthquakes. They help decision makers to better prepare and plan for earthquakes as well as to better understand and manage the post-disaster phase. Thus the more certain and accurate the risk assessments, the more likely they can contribute to a successful disaster management after an event.

The famous statement “Earthquakes don’t kill, buildings do!” refers to the fact that knowing the location and earthquake-relevant properties of buildings is crucial. Depending on the distance from the epicenter, the main construction material, the height, etc., each building reacts differently when exposed to the shaking created by an earthquake, called the earthquake hazard. Exposure models describe the built environment, usually as the number and distribution of building types and are created for entire countries from census data in collaboration with engineers. By combining such models with OpenStreetMap (OSM) data, we create exposure models that also provide the location of buildings. OSM has been been used in the past for augmenting exposure models, for example, Figueiredo and Martina [1] used OSM to improve the quality of the exposure models and Sousa et al. [2] used OSM to propose a method to create a building-specific exposure model for industrial buildings in Europe. We also assign vulnerabilities to each building type describing the expected damage grade depending on the level of shaking. Combining hazard, exposure and vulnerability leads to seismic risk, a description of the expected damage and losses due to earthquakes. For damage and loss assessments done after an earthquake, with the hazard being reliably measured in terms of shaking intensity, the exposure model is the element with the largest uncertainty. This confirms the need to know as much as possible about the distribution and types of buildings in hazard-prone areas.

Thus to increase the accuracy of the risk assessments, we need to increase the accuracy of the exposure model. The exposure models used in this study are building-by-building exposure models covering various countries in Europe (Albania, Andorra,

Evaz Zadeh, T., Oostwegel, L.J.N., Lingner, L., Shinde, S., Cotton, F., & Schorlemmer, D. (2023). Improving the accuracy of earthquake risk estimates with OpenStreetMap building data
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443311

Austria, Belgium, Switzerland, Czech Republic, Greece, Italy, and Turkey), created by taking the building locations from OpenStreetMap (OSM) and the possible types for each building from census data. These types were published by *The European Facilities for Earthquake Hazard and Risk* (EFEHR, www.efehr.org) for Europe in the framework of the European Seismic Risk Model 2020 [3]. Here, we focus on studying the building information taken from OSM that can improve the accuracy of detecting building types in the exposure model. Knowing the type of buildings is an important aspect of exposure models, because the risk-related behavior of buildings is dependent on it.

Each building type is defined using the Building Taxonomy of the Global Earthquake Model [4] describing all relevant features of buildings in a systematic way, such as the main construction material, lateral load resistance system, occupancy and number of stories. Using the census data, each building in the exposure model is assigned in a probabilistic approach with numerous possible types due to lack of information about the respective building. This means that having information about particular features of individual buildings can help us narrow down the possible types assigned to a building by eliminating some of the types that do not match with the building properties and gets us closer to the exact type of the building. This results in increasing the accuracy on the building-type detection aspect of the exposure model.

From OSM we obtain the number of stories and derive occupancy information for each individual building where available. This way, as pointed out above, for each building we can omit the building types assigned to that building that do not match the number of stories and occupancy information taken from OSM and be left with fewer numbers of possible types. The fewer the number of possible types for a building, the more certain we are about the building type.

To study the effect of each of the OSM building properties (number of stories and occupancy), we create the building-specific exposure model four times, using different sets of information. The first model, named the *Basic* model, consists of building polygons, taken from OSM and the possible building types taken from the census data. In this model, each building comprises the full probability distribution of all possible types as reported in the census data for the region the building is located in. This model does not increase the knowledge about building types compared to a classical exposure model but adds exact building locations to the model. The next model, named the *Stories* model, is the *Basic* model enriched with the information about the number of stories per building. This means that the possible types for each building are reduced to the types matching the number of its stories. Another model, named the *Occupancy* model, is the *Basic* model enriched with the occupancy information and again only using the possible types per building that match its occupancy. Finally, the last model, named the *Standard* model, is the *Basic* model but using the full set of information, both occupancy and number of stories.

We observe that the total number of possible types for all the buildings decreases significantly when using OSM information (in the *Occupancy*, *Stories* and *Standard* models) compared to the *Basic* model, among which, unsurprisingly, the *Standard* model contains the lowest number of types. The reduction observed in the *Occupancy* model compared to the *Stories* model is larger due the fact that the occupancy is by far a more available property compared to the number of stories, which can result in a judgment bias.

To remove the bias resulting from the unequal number of buildings carrying information about occupancy and number of stories, we introduce the Reduction Factor (RF).

The RF is defined as the total number of types for the buildings that include a specific information (either number of stories or occupancy or both) divided by the total number of types for the same buildings in the *Basic* model, ignoring the buildings for which no further information is available, to better quantify the effect of adding these properties to OSM. As an example, the RF value for Greece is 0.3, 0.53, 0.15 for the *Stories*, *Occupancy* and *Standard* model, respectively. An RF value of 0.15 for the *Standard* model for Greece means that the buildings carrying occupancy and number of stories in Greece are on average left with only 15% of the types they are initially assigned with, in absence of the occupancy and number of stories (as in the *Basic* model).

Therefore, the number of stories reduces the possible types for a building more than the occupancy information on average per building in Greece. Thus adding the number of stories to buildings can significantly increase the accuracy of the type compared to adding occupancy information. However, both number of stories and occupancy lead to the largest reduction and thus to a significant improvement of exposure models. We observed and report on similar results for all the countries studied except Andorra and the Czech Republic. The highest RF for the *Standard* model among the countries studied belongs to Switzerland with 0.31 and the lowest belongs to Croatia, with 0.10.

The accuracy of an earthquake loss assessment depends on the accuracy of the measured hazard intensity, the exposure and the vulnerability model. Addition of building information to the exposure model, results in more accurate choice of vulnerability functions. Therefore, with a better prediction of the vulnerability class of the buildings, we move from an average to a more accurate, building-specific loss assessment (see Figure 1).

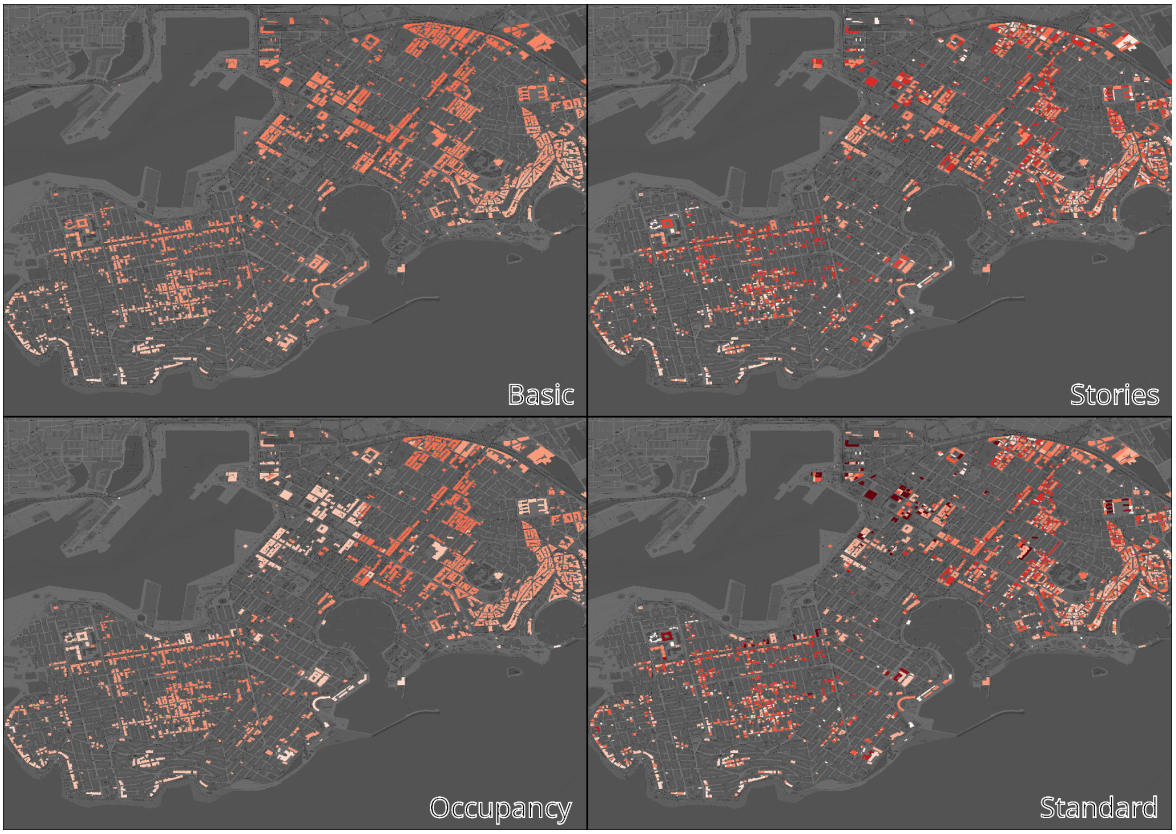

Figure 1. The probability of slight damage computed for a scenario similar to the M6.0 1999 Athens earthquake. The frames cover the area of Piraeus, Greece, and show the damage probability only for the

buildings that carry the number of stories and occupancy properties. Each frame represents one of the four models, *Basic* (top left), *Stories* (top right), *Occupancy* (bottom left) and *Standard* model (bottom right). The darker the color, the higher the probability of slight damage. Please note that the exact values are not provided to protect personal identifiable information.

It should be mentioned that information about the age of a building, its construction material and other building properties would benefit exposure models even further. However, they are usually not found in OSM. This is why we have limited this study to the most common building features that can either be found in OSM (number of stories) or derived from other pieces of information (occupancy).

This study shows the power of Volunteered Geographic Information (VGI), specifically OSM, in increasing the accuracy of detecting building types that results in improving the accuracy of resulting risk assessments that are of great importance for both preparing for and coping with earthquake disasters. Besides, this study points out that activities towards mapping the number of stories or the building occupancy can help significantly improve the quality of exposure data and subsequently the loss and damage assessment of an earthquake disaster. New tools like StreetComplete (https://streetcomplete.app) or Every Door (https://every-door.app) make the collection of building properties very easy and have already contributed significantly to the growth of the number of buildings with more properties provided. A possible reason for the overall lack of building properties is the fact that these are not visible on the standard map provided by OSM and thus not providing a clear incentive to collect such data. Specific maps that highlight more building properties could provide such an incentive. Because exposure models are not used solely for earthquake loss assessments, Evaz Zadeh et al. [5] showed first applications of such a detailed exposure model in the scope of multi-hazard risk assessment. In particular, for floods and tsunamis the building locations are very important and exposure models not providing the exact locations of buildings are prone to produce wrong assessments. Thus, the OSM building dataset is already a great augmentation to exposure models and will get better with more building properties being collected.

References

[1] Figueiredo, R., & Martina, M. (2016). Using open building data in the development of exposure data sets for catastrophe risk modelling. *Natural Hazards and Earth System Sciences, 16*(2), 417-429.
[2] Sousa, L., Silva, V., & Bazzurro, P. (2017). Using open-access data in the development of exposure data sets of industrial buildings for earthquake risk modeling. *Earthquake spectra, 33*(1), 63-84.
[3] Crowley, H., Dabbeek, J., Despotaki, V., Rodrigues, D., Martins, L., Silva, V., Romão, X., Pereira, N., Weatherill, G., & Danciu, L. (2021). European seismic risk model (ESRM20). *EFEHR Technical Report, 2.*
[4] Brzev, S., Scawthorn, C., Charleson, A. W., Allen, L., Greene, M., Jaiswal, K., & Silva, V. (2013). *GEM building taxonomy (Version 2.0) (No. 2013-02)*. GEM Foundation.
[5] Evaz Zadeh, T., Jozef Nicolaas Oostwegel, L., Lingner, L., Shinde, S., Cotton, F., & Schorlemmer, D. (2023). 'risk-calculator': a tool for multi-hazard risk assessments. In *EGU General Assembly Conference Abstracts* (pp. EGU-13001).

Assessing OpenStreetMap bicycle data quality with BikeDNA: A Denmark Case Study

Ane Rahbek Vierø[1,*], Anastassia Vybornova[1] and Michael Szell[1,2,3]

[1] NEtwoRks, Data, and Society (NERDS), Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark; anev@itu.dk, anvy@itu.dk, misz@itu.dk
[2] ISI Foundation, 10126 Turin, Italy
[3] Complexity Science Hub Vienna, 1080 Vienna, Austria

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Getting more people to commute by bike is increasingly accepted as a central element of making the transport system more sustainable. However, a modal shift towards more cycling requires vastly better bicycling conditions. At the moment, ambitious plans for bicycle infrastructure investments are being set up by numerous cities across the world (e.g., in Paris, Milan, Auckland, Bogotá). This uptick in investments is matched by a promising growth in data-driven cycling research. Due to its global coverage, OpenStreetMap (OSM) data play a central role in this context, both for research projects on bicycle infrastructure and for cyclist routing applications.

The quality of volunteered geographic information (VGI) and OSM data is well-studied, and most OSM data are shown to be of high quality. However, we know that errors and lower data quality are not randomly distributed in OSM, but instead correlate with lower population densities and vary both from country to country and within cities [1-4]. When it comes to bicycle infrastructure, lanes and dedicated cycling paths often are among the latter features to be mapped [5,6]. Moreover, the topological quality relevant for bicycling routing lags behind the quality of other road network data [5,6]. Even though these tendencies have profound implications for cycling research using OSM data, to date only few studies looked specifically at the quality of bicycle infrastructure data in OSM. Previous research on the quality of bicycle infrastructure data have shown that data quality issues are prominent in both administrative, commercial, and crowdsourced data sets, and that OSM data often are of a comparable or higher quality [3,4]. OSM data can therefore be an attractive data source – not just for researchers, but also for bicycle planners. Providing tools for reproducible quality assessment can help build the necessary trust in OSM data among people and agencies new to OSM data.

To address the current knowledge gap about bicycle infrastructure data in OSM, and to help others gain insights into the quality of OSM data in their local area, we have developed the tool BikeDNA [7]. BikeDNA is developed for reproducible and rigorous quality assessments of bicycle infrastructure and network data from OSM or elsewhere, based on a

Vierø, A.R., Vybornova, A., & Szell, M. (2023). BikeDNA: A Tool for Bicycle Infrastructure Data & Network Assessment
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443316

series of easy to use Python Jupyter notebooks. BikeDNA is particularly useful for studying the entirety of the bicycle network at a city/regional scale. BikeDNA builds on existing methods for spatial data quality assessment, but is tailored specifically to the peculiarities of bicycle infrastructure data, such as a high network fragmentation and inconsistent data models within and between different data sets.

The tool incorporates both intrinsic and extrinsic methods, i.e., both a standalone evaluation of OSM data properties and a comparison with another reference dataset. BikeDNA accounts for different data availability scenarios and is applicable even if no reference dataset is available. Moreover, for the extrinsic part of the analysis, our approach does not assume that the reference dataset is of higher quality, but is rather focused on identifying where OSM and the reference data differ. BikeDNA can thus be used for extrinsic data quality assessment even when there is no high-quality data on cycling infrastructure available.

The methodology is oriented towards data usages for analyses of e.g. accessibility, routing, and other network-based approaches, and therefore supplements evaluation of data completeness with analysis of topology errors and variations in connectivity, and overall network structure. Data completeness is evaluated based on differences in infrastructure density between OSM and reference data, taking into account differences in data models and how the use of a center line mapping can undercount the amount of infrastructure compared to a data set mapping the exact location of bicycle tracks and lanes. Completeness is furthermore evaluated with a matching of corresponding features in OSM and reference data [8,9]. Data topology and network structure is assessed through an identification of, for example, over and undershoots and the location and size distribution of disconnected components. Finally, we evaluate the local completeness of OSM tags relevant to cycling. BikeDNA is based on existing libraries for working with spatial networks such as OSMnx, pyrosm, NetworkX, and Momepy [10-13]. The tool is free to use and open-sourced under the AGPL-license.

Although OSM data quality is generally high, it is important to keep in mind how quality might deviate not just in some areas, but also for specific subsets of the data. Likewise, despite the frequent use of OSM data, localized data evaluation and assurance is still necessary to support e.g. public administrators and planners to adopt OSM in their work. Through this study and by providing BikeDNA as an open-source tool, we aim to make it easier for the cycling research and planning community to quickly assess the fitness for use of OSM data in their area, and to support and motivate the use of open and crowdsourced geospatial data when working with cycling and other modes of active mobility.

We demonstrate the use case of BikeDNA through a case study of the bicycle infrastructure data in Denmark from respectively OSM and a local administrative data set (Figure 1a and 1b). Through the case study we show how BikeDNA for example can identify differences in data completeness (Figure 1c); detect topological errors and unwanted network fragmentation (Figure 1d and 1e), and help users choose the data set most suitable for bicycle routing and accessibility analysis (Figure 1f).

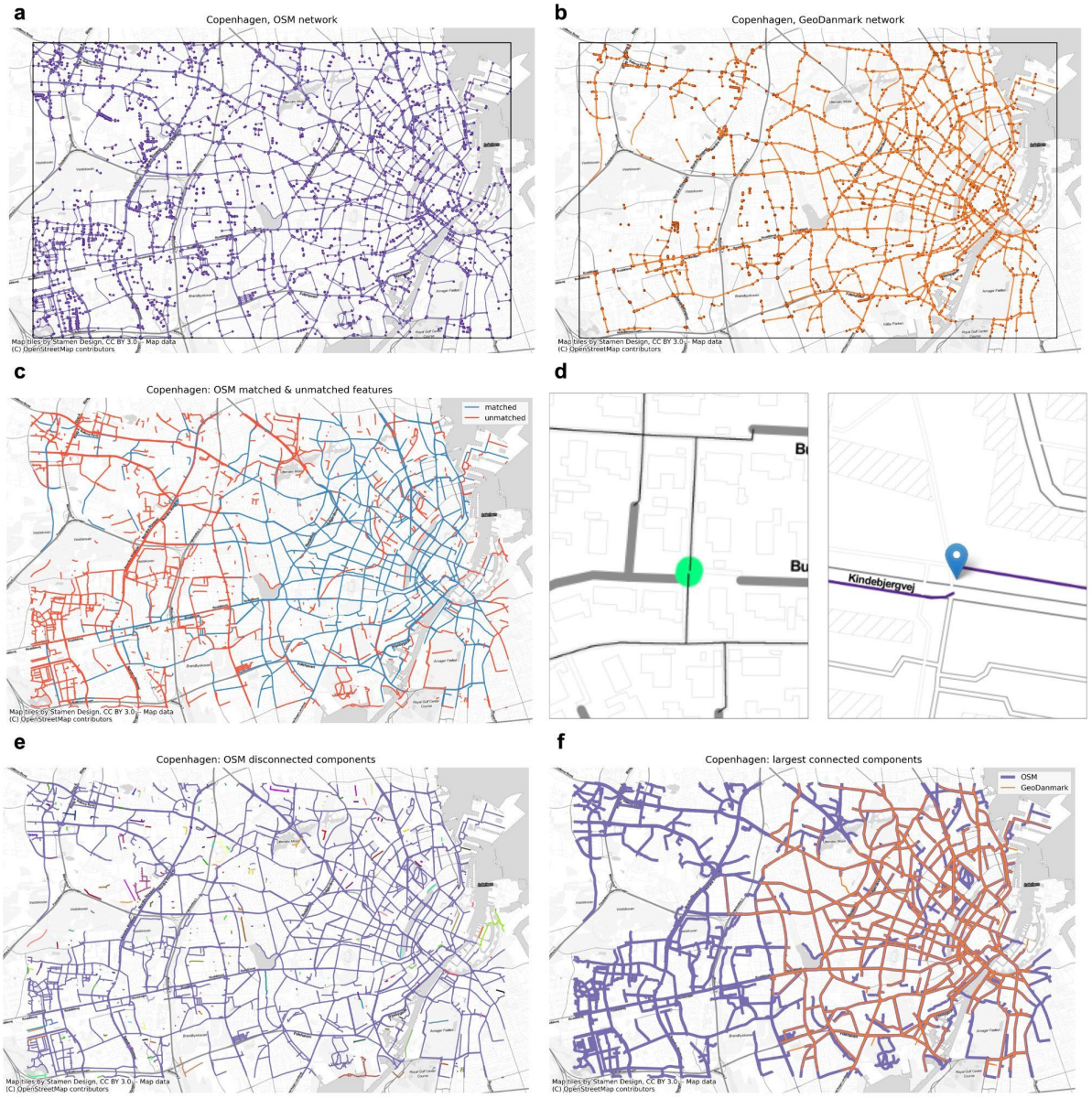

Figure 1. Example outputs from running BikeDNA for a demonstration area in Greater Copenhagen with data on dedicated bicycle infrastructure from OSM and a local reference data set (GeoDanmark): Input data from a) OSM and b) GeoDanmark. c) Outcome of feature matching: Matched (blue) and unmatched (red). d) Examples of identified undershoots. e) Disconnected components in OSM data f) Largest connected components.

References

[1] Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PloS One, 12*(8), e0180698. https://doi.org/10.1371/journal.pone.0180698

[2] Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design, 37*(4), 682–703. https://doi.org/10.1068/b35097

[3] Ferster, C., Fischer, J., Manaugh, K., Nelson, T., & Winters, M. (2020). Using OpenStreetMap to inventory bicycle infrastructure: A comparison with open data from cities. *International Journal of*

Sustainable Transportation, 14(1), 64–73. https://doi.org/10.1080/15568318.2018.1519746
[4] Hochmair, H. H., Zielstra, D., & Neis, P. (2015). Assessing the Completeness of Bicycle Trail and Lane Features in OpenStreetMap for the United States: Completeness of Bicycle Features in OpenStreetMap. *Transactions in GIS, 19*(1), 63–81. https://doi.org/10.1111/tgis.12081
[5] Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS, 18*(6), 877–895. https://doi.org/10.1111/tgis.12073
[6] Neis, P., Zielstra, D., & Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet, 4*(1), 1. https://doi.org/10.3390/fi4010001
[7] Vierø, A. R., Vybornova, A., & Szell, M. (2023). BikeDNA: A tool for bicycle infrastructure data and network assessment. *Environment and Planning B: Urban Analytics and City Science*, 23998083231184471. https://doi.org/10.1177/23998083231184471
[8] Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS, 16*(4), 477–498. https://doi.org/10.1111/j.1467-9671.2012.01304.x
[9] Will, J. (2014). Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network: A case study in Göteborg, Sweden. *Student thesis series INES*. Lund University Press. https://www.semanticscholar.org/paper/Development-of-an-automated-matching-algorithm-to-%3A-Will/b3b77d579077b967820630db56522bef31654f21
[10] Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126–139.
[11] Tenkanen, H. (2021). *HTenkanen/pyrosm: V0.6.1 [Computer software]*. Zenodo. https://doi.org/10.5281/zenodo.5561232
[12] Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference*, 11–15. Los Alamos National Lab, Los Alamos, NM.
[13] Fleischmann, M. (2019). momepy: Urban Morphology Measuring Toolkit. *Journal of Open Source Software, 4*(43), 1807. https://doi.org/10.21105/joss.01807

Social, technical and political transformations in OpenStreetMap – From volunteered geographic information to embedding digital commons in platform capitalism

Susanne Schröder-Bergen[1,*]

[1] Institute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; susanne.schroeder-bergen@fau.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Since its early days OpenStreetMap (OSM) has developed from a small mapping project carried out mainly by mapping enthusiasts with a vision to create an open map of the world to a project that actually can compare its coverage to proprietary providers of mapping data, including Google [1]. Against the background of a world that increasingly relies on geospatial data, the role of OSM is changing in multiple ways. This study focuses from a social and digital geography perspective on the changing social, technical and political conditions underlying OSM and aims to argue how OSM, along with the geospatial industry as a whole, is approaching a stage of maturity, yet remains true to itself at its core. The larger research question behind this paper is how the OSM project is addressing the ever-increasing economic and geopolitical tensions in which it finds itself.

The empirical results of this study focus mainly on developments of OSM over the last four years. Quantitative analysis of OSM data (OSM full history dump and changeset dumps) and qualitative analysis of other online sources (OSM wikis, blogs, forums) were used to explore the growing interest or role of institutional actors in OSM. To gain a deeper understanding of the OSM ecosystem, this research involved attending conferences both online and in the field, and conducting around 30 expert interviews with members of the OSM community, as well as with individuals in the geospatial data industry. These interviews, which were analyzed using methods of qualitative content analysis, gave intriguing insights into the geospatial industry and the embedding of open mapping projects like OSM into it.

The results of this study are categorized into 1) social, 2) technical, and 3) political transformations in OSM, which are conceptualized in a larger context.

1) In recent years, there has been social shifts in the participating groups of OSM, as well as the spatial distribution of contributors. In total OSM became a much more diverse and global project. Not least, organized editing groups – humanitarian mapping teams like

Schröder-Bergen, S. (2023). Social, technical and political transformations in OpenStreetMap – From volunteered geographic information to embedding digital commons in platform capitalism
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443321

HOT and commercial actors – have made the project more known in parts of the world that have not been the main core of OSM. It can be argued that this development has both an empowering effect for new participants from formerly underrepresented regions and might be suppressing local voices when external worldviews are reproduced [2].

2) This is also reflected in technical shifts within OSM. With the greater engagement of organized mapping groups in OSM, a distinction between local knowledge versus remote data makes less sense. The attention has instead changed towards the difference between volunteered and paid or corporate mapping [3]. These new more institutionalized mapping groups also introduce new "modes" of mapping into the OSM ecosystem, including, for example, dealing with machine learning generated datasets [4].

3) Tentatively and only partly, there are, within the OSM community, also trends towards a greater discussion and recognition of the potential social (and political) responsibility that OSM entails as a large, global project [5,6]. This comes also in light of the recent creation of the Overture Maps Foundation. The increasing significance of (open) geospatial data for a multitude of actors – both private and public – positions the project, which has always seen itself as apolitical and global, in a political arena that goes far beyond concrete data conflict issues, such as where to draw a boundary line in a crisis region, to the handling and protection of the OSM project, the data, and the community as a whole. Recent developments and the handling within OSM, however, suggest that the way OSM works may also be the reason that OSM is so resistant and independent to other interested parties (of the broader geospatial industry) who would like to steer the project in their desired direction.

These transformations are symptomatic of larger trends in OSM, which can be conceptualized as an evolution of OSM from a volunteered geographic information (VGI) project to OSM as a project embedded in platform capitalist logics [7]. In addition, the success of OSM can be seen as particularly rooted in internal socio-technical elements of OSM that follow the principles of digital commoning. The concept of digital commons focuses on alternatives to capitalist political economies that remain within market-structured environments, but that emphasizes the openness and resistance of digital domains that are structured through specific practices of regulation [8,9]. It seems analytically useful to frame the socio-technical practices of data production and use in OSM with its specific formal and informal regulations as a form of digital commoning.

This paper situates the current state of OSM in the larger context of a world that is increasingly reliant on geospatial data. OSM is becoming more and more global and also increasingly important for the entire geospatial domain. This means that it is also becoming interesting for highly resourced and powerful actors. OSM must consequently deal more and more with the economic and geopolitical areas of tension of the digital transformation. Questions arise whether the existing structures and rules within OSM are sufficient to further develop the character of OSM as a digital commons. In this context, it also remains to be seen whether OSM will become more involved and cooperate with governmental actors. However, this must be the community's intention, as it could turn essential elements of the project structure upside down.

References

[1] Kilday, B. (2018). *Never lost again: The Google mapping revolution that sparked new industries and augmented our reality*. Harper Business, New York, NY.

[2] Schröder-Bergen, S., Glasze, G., Michel, B., & Dammann, F. (2022). De/colonizing OpenStreetMap?. *GeoJournal*, *87*(6), 5051–5066.
[3] Schröder-Bergen, S. (2020). Analyzing the localness of OSM data. In M. Minghini, S. Coetzee, L. Juhász, A. Y. Grinberger, P. Mooney, & G. Yeboah (Eds.), *Proceedings of the Academic Track at the State of the Map 2020* (pp. 19-20). Zenodo.
[4] Madubedube, A., Coetzee, S., & Rautenbach, V. (2021). A contributor-focused intrinsic quality assessment of OpenStreetMap in Mozambique using unsupervised machine learning. *ISPRS International Journal of Geo-Information, 10(3)*, 156.
[5] Mustard, A. (2020). *Winds of Change in OpenStreetMap*. State of the Map 2020. https://2020.stateofthemap.org/sessions/RRVNAM/
[6] Steele, A. L. (2022). Mapping crises, communities and capitalism on OpenStreetMap: situating humanitarian mapping in the (open source) mapping supply chain. In M. Minghini, P. Liu, H. Li, A. Y. Grinberger, & L. Juhász (Eds.), *Proceedings of the Academic Track at State of the Map 2022* (pp. 10–12). Zenodo.
[7] Michel, B., & Schröder-Bergen, S. (2022). The Politics of Geodata in Urban Platform Capitalism. In A. Strüver, & S. Bauriedl (Eds.), *Platformization of Urban Life - Towards a Technocapitalist Transformation of European Cities* (pp. 73-84). transcript Verlag. https://doi.org/10.14361/9783839459645-005
[8] Dulong de Rosnay, M., & Le Hervé Crosnier, H. (2012). An Introduction to the Digital Commons: From Common-Pool Resources to Community Governance. *Building Institutions for Sustainable Scientific, Cultural and Genetic Resources Commons* (pp. 1–18). https://shs.hal.science/halshs-00736920
[9] Linebaugh, P. (2007). *The Magna Carta Manifesto: Liberties and Commons for All*. University of California Press.

Towards an open high-resolution land use dataset in Great Britain – Comparing and consolidating retail centre areas from open data sources

Oliver O'Brien[1,*]

[1] Consumer Data Research Centre, University College London, London, United Kingdom; o.obrien@ucl.ac.uk

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Great Britain does not have a comprehensive and openly licenced high-resolution land use dataset that includes detail on building usage, but OpenStreetMap (OSM) has potential as a good base for creation of such a dataset [1]. OSM's quality and completeness is highly variable, but often good and improving, including for land use mapping [2,3]. This research evaluates use of separate open datasets to augment OSM for Great Britain. The research focuses on retail areas as these have recently been impacted both by internet shopping and the COVID pandemic [4].

This paper evaluates three generally openly available datasets showing retail centre extents across Great Britain, analysing each by areal footprint and, where available, premises counts. Firstly, the Consumer Data Research Centre (CDRC)'s Retail Centres Boundaries 2022 product, secondly non-domestic Energy Performance Certificates (EPCs) geolocated with Unique Property Reference Numbers (UPRNs), filtered for retail categories, and finally OSM land use retail polygons on their own.

The CDRC has produced several "retail centres" products since its inception in 2014 [5,6]. The initial product was derived from English/Welsh Town Centres as defined in 2004 by a UK government department, to which kernel density estimation was applied [5]. This identified 1312 areas in total in England/Wales.

In 2017 the CDRC produced a new set of retail centre polygons, this time based on commercial data from the Local Data Company (LDC) [6]. The higher-resolution data, ground-truthed by the LDC, was aggregated using a DBSCAN.

In 2022, CDRC published a new version, based entirely on open data sources. For England/Wales, Valuation Office Agency (VOA) data was geocoded. For Scotland, a snapshot of OSM points and polygons using the *shop* key with any value, or *amenity* key with a set of relevant values, was used (this was also used where VOA data was missing in

O'Brien, O. (2023). Towards an Open High-Resolution Land Use Dataset in Great Britain – Comparing and Consolidating Retail Centre Areas from Open Data Sources
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443323

certain areas in England/Wales). OSM polygons with *landuse* key and *retail* value were used to improve the areal geometry of some of the resulting clusters [7]. This identified 6344 areas in Great Britain.

EPCs issued since 2008 in England/Wales have been published in tabular format as open data by the UK Government, and regularly updated. The devolved Scottish government has similarly published EPCs covering Scotland going back to 2013. Non-domestic EPCs cover various categories of buildings, including schools, prisons and hotels, and typically include data on floor area. They include UPRN references, which can then be combined with the National Statistics UPRN Lookup file to reveal the location, typically of the retail unit's building centroid or front entrance. For this evaluation, EPCs corresponding to UK planning classes A1/A2 (retail) and A3/A4/A5 (food outlets) were included.

By performing spatial operations in GIS software, clusters of EPCs can be used to crudely reveal the spatial extent, size (floor area) and unit count of retail areas (see Figure 1). Some caveats apply, for example EPCs listed for subsequently defunct retail units, units undergoing EPC assessments at different frequencies based on changes of tenant, and that the locations of each EPC's UPRN don't reflect the spatial extent of the unit or its building.

EPCs for Great Britain were obtained, cleaned and deduplicated, and their property type field was normalised and filtered to only include the aforementioned planning classes. 322425 EPCs were obtained using this method. Scotland had only 14600 (4.5%), compared with 8.4% of Great Britain's residential population. This could be due to pre-2013 data not being available, a cleaner upstream dataset with replaced units removed at source, and a less dynamic retail sector outside of very large cities like London and Birmingham, not present in Scotland.

QGIS was used to buffer the EPC points by 70m (reaching across typical building extents and roads) and dissolve the resulting polygons into areas of one or more EPC circles. A point-in-polygon operation was applied to count the number of EPCs in each polygon. Polygons with less than five EPCs, or less than 40000sqm, were excluded. 5877 polygons remained, including 298 in Scotland. Floor area is available but was not used in this simple analysis as it was considered that multiple-unit areas were more important to defining "retail centres" than single-tenant large warehouse stores, but is a possible future modification.

QGIS was also used to process the data directly from OSM (via the GeoFabrik service), with OSM retail land use polygons buffered by 50m (reaching across roads). Areas smaller than 40000sqm were removed. This left 5009 polygons, including 384 in Scotland.

Finally, the CDRC product was examined. This was used unmodified, with areal/unit cleaning and manual corrections having been applied during product creation, a threshold of ten retail units having been used for England/Wales (from the VOA data), which corresponds well with the five record threshold used for the EPC dataset because of the incomplete nature of the latter's source data. The CDRC product shows 6344 areas, 392 in Scotland.

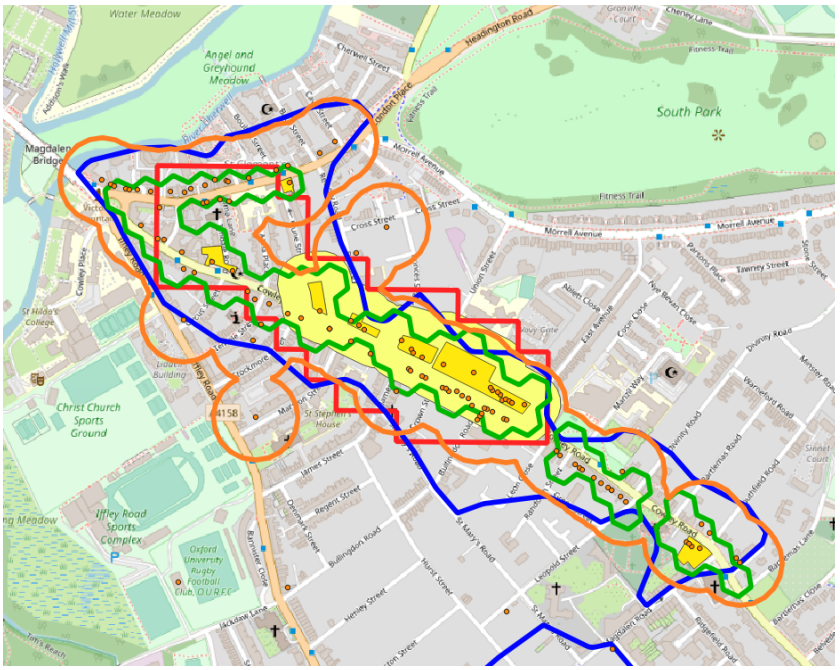

Figure 1: Retail area data used in this study, superimposed on an OpenStreetMap map, in QGIS, for the Cowley Road area in Oxford, UK, showing potential for extension of the retail area defined in OpenStreetMap based on EPC data. Red line: 2004 "Town Centres" VOA/ONS data. Blue line: CDRC 2017 LDC data. Green lines: CDRC 2022 VOA data, H3 geography. Orange points: EPC locations for retail. Orange line: Buffered EPC locations for retail. Dark yellow: OpenStreetMap retail polygons. Light yellow: Buffered OSM retail polygons. Base map © OpenStreetMap contributors.

To determine the quality of each source, identify the most important missing data in the OSM land use data (OSMLU), and evaluate the EPC and CDRC product (CDRC2022) approaches, the polygons from each pair of sources were compared, on a "largest area in X not in Y" basis, using a QGIS "join attributes by location" spatial operation with an intersect geometric predicate and displaying unjoinable features.

The top five are presented, with observations with Google Street View imagery or OSM data to determine the reason for the omission. Where the area within a town/city is not named, the commercial centre is being referred to.

Largest EPC not in OSMLU: Edinburgh Old Town/Southside (*landuse* key missing), Bath (missing), Doncaster (shown as *commercial*), Burnley (missing), Dundee (shown as *residential*). Largest CDRC2022 not in OSMLU: Edinburgh Old Town/Southside, Cardiff (missing), Bath, Edinburgh Greenside (missing), Bishop Auckland (shown as *commercial*).

Largest OSMLU not in EPC: Edinburgh Fort Kinnard, Warrington Gemini, Edinburgh Straiton, Shrewsbury Harlescott, Guildford Slyfield. The first four are low-density retail parks, and the last is mistagged in OSM as an industrial/commercial estate with only a limited retail offering, mainly large car showrooms. Largest CDRC2022 not in EPC: Edinburgh Fort Kinnard, Newcastle Whitley Road, Warrington Gemini, Ashford Orbital, Edinburgh Pentland. All five of these are low density retail parks on the edge of their respective towns and cities. The EPC threshold of five stores and 70m buffer is too constraining to properly include "big box" retail parks.

Largest EPC not in CDRC2022: Kilmarnock, Greenock, Motherwell, Livingston, Bradford Greengates. The first four are all in Scotland, this suggests there are a lot of

missing buildings or retail points in the OSM data, as this was the principal data source for CDRC2022 in the absence of VOA data. Bradford Greengates is a relatively small and spatially dispersed retail centre on a major road, with the larger distances between units likely meaning the method will not have seen them forming a single retail area. Largest OSMLU not in CDRC2022: Livingston, Sutton Scotney, Bridgemere Garden World, Motherwell, Errol Sunday Market. Livingston and Motherwell are both missing the shop points in OSM that was used as the main data source for CDRC2022 for Scotland (In Motherwell's case, a contributor has subsequently added them). The other areas have the OSM retail *landuse* key covering much too large an area.

In conclusion, this analysis has shown there is considerable potential for the EPC data to help target improvements to the retail land use indication in OpenStreetMap for Great Britain, and potentially it could also help refine future iterations of the CDRC retail area product, as a supplementary data source. It also shows considerable promise as a more reliable base dataset for CDRC for defining retail areas in Scotland, as it avoids the variations in the way retail areas are currently presented in OSM.

References

[1] Schultz M., Voss J., Auer M., Carter S., & Zipf A. (2017). Open land cover from OpenStreetMap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *63*, 206–213.

[2] Yang D., Fu C., Smith A., & Yu Q. (2017). Open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations. *Geo-spatial Information Science*, *20*(3), 269–281.

[3] Vu T., Vu N., Phung H., & Nguyen L. (2021). Enhanced urban functional land use map with free and open-source data. *International Journal of Digital Earth*, *14*(11), 1744–1757.

[4] Cook C., Clark D., Heal A., Hollowood E., Elliott O., Joiner S., & Nevitt C. (2022). *How the UK high street was hit by the pandemic: look up your area.* Financial Times. https://www.ft.com/content/9348c644-288f-42e7-9f4b-edea8b71be5b

[5] Dolega L., & Pavlis M. (2016). Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *Journal of Retailing and Consumer Services, 28*(6), 78–90.

[6] Dolega L., Reynolds J., Singleton A., & Pavlis M. (2019). Beyond retail: New ways of classifying UK shopping and consumption spaces. *Environment and Planning B: Urban Analytics and City Science*, *48*(1), 138–150.

[7] Macdonald J., Dolega L., & Singleton A. (2022). An open source delineation and hierarchical classification of UK retail agglomerations. *Scientific Data, 9,* 541.

Mapping public space in urban neighbourhoods using OpenStreetMap data

Ester Scheck[1]*, Florian Ledermann[1], Andrea Binn[1] and Marian Dörk[2]

[1] Research Unit Cartography, TU Wien, Austria; ester.scheck@tuwien.ac.at, florian.ledermann@tuwien.ac.at, andrea.binn@tuwien.ac.at
[2] UCLAB, FH Potsdam, Germany; marian.doerk@fh-potsdam.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

OpenStreetMap (OSM) enriches the exploration and study of urban landscapes. In this research project, we aim to use OSM data to investigate urban public spaces from a distributional justice perspective. While public spaces are acknowledged as an important resource for urban society, it becomes important, in light of ongoing trends towards privatization, commercialization, and festivalization, to critically observe and reflect on the extent to which resources, rights and opportunities regarding public space are distributed equally. The amount, accessibility, and character of public space can differ between cities and neighbourhoods. A quantitative analysis of public space could offer insights into the distribution and availability of public space. To this end, we propose a framework for the identification and categorization of these spaces based on OSM data. The framework aims to enable both the mapping of public spaces as well as an evaluation of the share of public space. We also hope to investigate the potential of OSM data. Some preliminary findings and an introductory overview of the research process, with an emphasis on its cartographic aspects, were presented in a previous publication [1].

The inspiration for this research is the so-called Nolli map, a map of Rome dating back over 250 years. Giovanni Battista Nolli, an Italian architect, engineer and cartographer, analysed the urban fabric beyond the structure of roads and buildings. In his work, titled 'La Nuova Topografia di Roma', Nolli mapped the interior and exterior spaces of Rome in high detail as a figure-ground map with contrasting dark and light sections. This distinction is commonly interpreted as a demarcation between private and public space. Since Nolli's time, the character and diversity of public spaces, as well as cartographic techniques have evolved significantly. In this research, we adapt some ideas behind Nolli's map to contemporary conditions based on open data and semi-automated geoinformation processing methods.

For this purpose, we operationalize the concept of public space as public accessibility. Public space research offers various approaches, definitions, and models, often considering multiple dimensions to define 'publicness,' with accessibility being one such dimension [2]. As one sub-aspect of this, the term 'public accessibility' refers to

Scheck, E., Ledermann, F., Binn, A., & Dörk, M. (2023). Mapping Public Space in Urban Neighbourhoods Using OpenStreetMap Data In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443326

whether a space is open and available for use by the general public. For feasibility reasons, we ignore other accessibility aspects like visual access, reachability or accessibility specifically for people with disabilities in this research. We chose public accessibility as the defining criterion because it is the basic prerequisite for a potential use of public space, it aligns with the common interpretation of the historical Nolli map and it is well represented in the OSM dataset. As a matter of practicality, the analysis excludes buildings and focuses exclusively on elements on the ground-floor level.

As part of the framework, which can be understood as a strategic approach, we develop multiple methods for processing OSM data to prepare and clean the data and to identify, categorize, and ultimately visualize public spaces as a map. The methods are implemented as Python functions which are made available as an open-source script [3]. With the map visualization in mind, the goal of the data processing is an overlap- and hole-free dataset with polygons only. This is a vital consideration as OSM data often involves undefined areas, overlapping elements, and the use of not only polygons but also line and point geometries. Key steps in the data preparation and cleaning therefore include filtering out insignificant point and line geometries, converting road and path line geometries into polygons, and resolving overlapping polygons. Several assumptions are made in this process. For instance, road lines are 'polygonized' by buffering the line geometries with the road width. In cases where no width is specified as a tag, which is quite common, a standard width is assumed based on the highway type and adjusted according to other tag keys such as direction, lanes or cycleway. Towards the end of the data processing, overlapping polygons are refined by merging similar spaces into space types and by cropping the polygons based on the specificity of the space type, visualization requirements and the inherent relationships between the elements - specifically, which elements are usually located within another element in OSM.

Space type as an additional attribute alongside public accessibility serves as a source of information for the accessibility but also allows a more differentiated picture of the areas. The identification of the public accessibility of areas is conducted on three levels, sequentially analysing both tags and geometries of the OSM dataset in the following order:

1. Analysis through explicit public access tags: Tags indicating public access, such as *access*, *foot*, *opening_hours* or *fee*, are initially assessed. For instance, areas with restricted opening hours or those requiring an entrance fee are interpreted as having restricted public access while areas tagged as *public* under the *access* or *foot* key, where the *foot* tag is describing access permissions for pedestrians, are considered publicly accessible.
2. Analysis through tags and geometries representing barriers: Tags and geometries are analysed in order to identify barriers (e.g. fences, hedges, rails, buildings) and therefore inaccessible enclosed areas. For example a playground enclosed by a fence with a locked gate is interpreted as publicly inaccessible, even though this information might not be saved as an explicit tag with the playground feature.
3. Assumption based on the space type: In the absence of explicit access information on level 1 and 2, the public accessibility is assumed based on the space type which is derived from tags like *leisure*, *landuse* or *amenity*. For instance, parks are presumed to be publicly accessible, while roads are

deemed inaccessible as they can not be used as public space in terms of its social and political function.

For the purpose of testing and further development, we apply the framework to two case studies, each covering an area of 500 x 500 m in Vienna, Austria. While for urban areas a good data quality of the OSM dataset can generally be assumed [4], an examination with the ohsome quality analyst (https://heigit.org/big-spatial-data-analytics-en/ohsome/ohsome-quality-analyst-oqt) reveals a limited fitness-for-use for this specific application. According to the ohsome quality analyst models, the primary tags underlying the analysis - *access*, *foot* and *barrier* - are not sufficiently saturated in the city of Vienna. This assessment is confirmed by a preliminary application of the framework in the case studies, which uncovers wrong results in some places during a ground truth verification. Consequently, the existing OSM data within the case study areas are not adequate for conducting a comprehensive public accessibility analysis.

In response, we verified and expanded the OSM data in the case study areas through extensive on-site mapping. The data enhancement took place in two phases: an initial preparatory step that utilized orthophotos and the multipurpose area map of the City of Vienna, which is integrated in the iD editor, followed by on-site field inspections. During the preparatory mapping, we corrected primarily building geometries and delineated space types, particularly green areas. Field inspections, in contrast, concentrated on identifying relevant barriers and access points (such as fences, walls, hedges, gates and building passages), and validating previously mapped areas. We used geolocated photos and map-based notes to document on-site mapping activities. An initial map of publicly accessible space, based on existing OSM data, served as a baseline and reference for comparison. Alterations and additions to geometries and tags were implemented in the iD editor. Apart from data gaps, the field visits revealed that public access is not always clearly evident, for example, when signs designating private property and prohibiting access lacked clear spatial reference or when a fence had an opening along a trail. It becomes evident that the lived and the legal sense of public space can differ from one another. Following OSM's ground truth paradigm [5], we mapped objects as they were observed on-site.

The outcomes of the case study applications and data analysis provide an impression of the extent and nature of public spaces in the case study areas. Presented in the form of a map (see Figure 1), inspired by the cartographic design of the historic Nolli map, it shows where and what type of space is publicly accessible. For instance, more than half of the inaccessible space is traffic area. These inaccessible spaces can be considered as opportunities to increase and create attractive public space. In addition to the map, a quantitative comparison of areas, for example visualized through a bar chart, offers insights to the ratio of publicly accessible and inaccessible areas. These values can be compared between different neighbourhoods and the result can serve as an indicator and basis for identifying needs in urban planning.

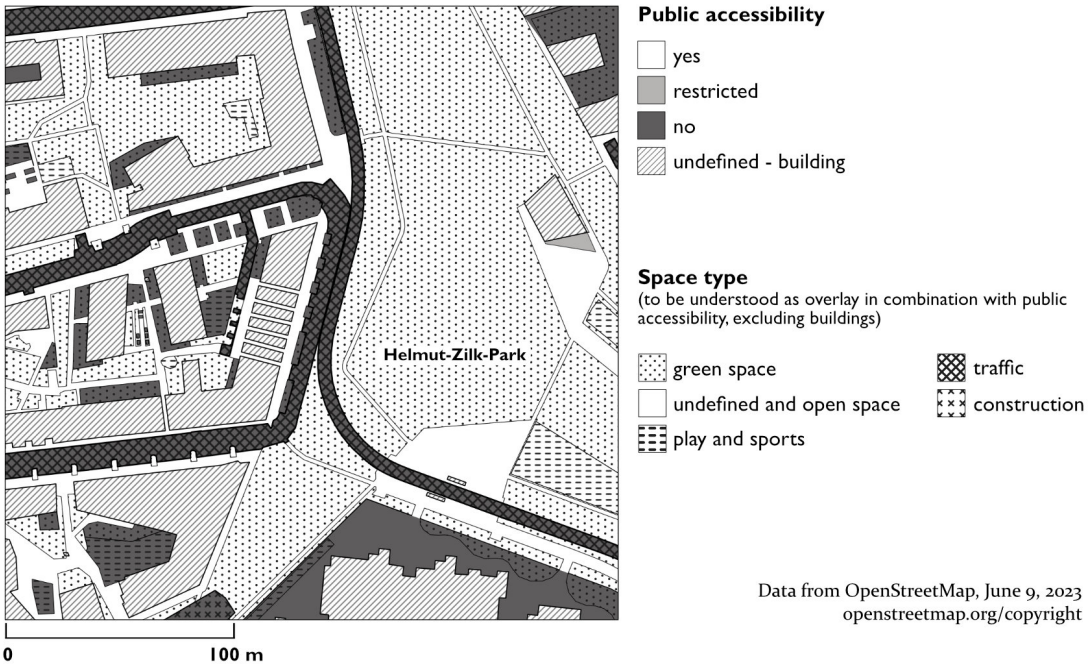

Figure 1. Map section of the case study Sonnwendviertel, Vienna, showing the public accessibility and space type in the area.

The case study applications not only offer valuable insights into public spaces within the selected areas but also provide an opportunity to reflect on the data analysis process and the reliability of the results. In particular, the importance of the on-site data collection has to be emphasized. In both case studies, the number of OSM elements in the areas increased by almost 50% during the on-site mapping phase. While it is possible that some edits may have originated from other OSM contributors, it is reasonable to assume that the majority of these new elements were added as part of this research project. These additions are crucial in achieving differentiated and, as validated through ground-truth checks, accurate results. Another interesting revelation is that most of the area in the resulting map had its public accessibility identified through the space type (level 3 of the analysis). Although this is the most generalized analysis level, we consider the assumptions to be reliable, because more specific information (like access permissions or fences) would have been added during on-site mapping and included on level 1 and 2 of the analysis.

Furthermore, the on-site mapping unveiled some qualitative aspects and limitations of the analysis. For example it became evident that public accessibility alone does not necessarily make a space truly public, as it may still be difficult to access or feel unwelcoming for individuals. Nevertheless, this analysis can be a foundation for deeper research in the realm of public spaces. Given its open-source code and the availability of the global OSM dataset, this framework can be applied to diverse neighbourhoods and locations. However, it is essential to bear in mind that the framework is developed with Central and Western European cities in mind. The variables and tags used in the analysis may need to be adapted and it should be noted that the underlying understanding of public space is not easily transferable to fundamentally different cultural and spatial contexts. On a larger scale, the examination of extensive areas or entire cities would provide valuable insights, but the developed framework currently faces performance limitations and gaps of

the relevant information in the OSM dataset, making it less suitable for extensive application in its current state. Subsequent research should delve into how the OSM dataset can be further leveraged and which other open datasets might be suitable to enhance the analysis. The work at hand can be the base for research in this direction and enrich the social and political debate about the distribution of urban space. It also highlights the potential of OSM data for social study of spatial issues. Furthermore, it demonstrates the adaptability of a cartographic classic to contemporary circumstances and the opportunities afforded by GIS and open data. Through semi-automated data processing, the framework extends the previous approaches to create a modern-day Nolli map.

References

[1] Scheck, E., Binn, A., Dörk, M., & Ledermann, F. (2023). A Contemporary Nolli Map: Using OpenStreetMap Data to Represent Urban Public Spaces. *Abstracts of the International Cartographic Association*, *6*, 223.

[2] Li, J., Dang, A. & Song, Y. (2022). Defining the ideal public space: A perspective from the publicness. *Journal of Urban Management*, *11*(4), 479–487.

[3] Scheck, E. (2023). *OSM Public Space Mapper*. https://github.com/ester-t-s/_osm-public-space-mapper

[4] Fonte, C. C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J. J., Laso Bayas, J.-C., See, L., & Vatseva, R. (2017). Assessing VGI Data Quality. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte, & V. Antoniou (Eds.) *Mapping and the Citizen Sensor* (pp. 137–163). Ubiquity Press.

[5] OSM Contributors (2023). *Good Practice*. https://wiki.openstreetmap.org/wiki/Good_practice#Map_what's_on_the_ground

Beyond Two Dimensions: Large-Scale Building Height Mapping in OpenStreetMap via Synthetic Aperture Radar and Street-View Imagery

Hao Li[1,*] and Yao Sun[2]

[1] Professorship of Big Geospatial Data Management, Technical University of Munich, Munich, Germany; hao_bgd.li@tum.de
[2] Chair in Data Science in Earth Observation, Technical University of Munich, Munich, Germany; yao.sun@tum.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

In the past decades, the world has been comprehensively mapped in 2D, however, a vertical dimension remains underexplored despite its huge potential. For instance, as of August 2023, more than 571 million buildings are mapped in OpenStreetMap (OSM) according to statistics from Taginfo, but less than 3% of them are associated with height values via the key/value pairs *heights=**. Though one can often estimate the height information via OSM key/value pairs such as *building:levels=** and *stories=**. Mapping human settlements as a 3D representation of reality requires an accurate description of vertical dimensions besides the 2D footprints and shapes. A 3D representation of human settlement is important in many aspects, including public health, urban planning, and environment monitoring, disaster management, etc. In this context, a list of the most relevant 3D building attributes mainly includes but is not limited to building height, building floor, and roof type [1,2]. For instance, building height is a key and fundamental factor in post-disaster (e.g., earthquake and flood) damage and situation assessment. Similarly, the roof type information is beneficial in estimating photovoltaic electricity potential at scale. As defined in CityGML 2.0 [3], 3D building models are divided into five levels of detail (LoDs) [4]. In LoD0, only the 2D footprint information is involved in the model. In LoD1, the LoD0 model is extruded by their building heights, and the obtained cuboid after extrusion is the LoD1 model. In LoD2, the 3D roof structure information is added to the LoD2 model. The LoD3 model further contains facade elements such as windows and doors. The LoD4 model is more complicated and contains both external and internal building elements. However, there are still open questions about how to automatically generate large-scale and fine-grained 3D building attributes in a fully reproducible and generalizable manner. Moreover, how to integrate such information into OSM or existing mapping tools.

Li, H., & Sun, Y. (2023). Beyond Two Dimensions: Large-Scale Building Height Mapping in OpenStreetMap via Synthetic Aperture Radar and Street-View Imagery
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443329

Traditional studies on building height retrieval in the Remote Sensing (RS) domain primarily employ high-resolution optical images and airborne LiDAR data [5]. However, optical data acquisition requires cloud-free weather, and airborne or terrestrial data are too expensive to collect globally. This proposal reports recent advances in supplementing OSM data with large-scale building height, specifically using Synthetic Aperture Radar (SAR) data and Street-View Images (SVI) data.

SAR data have been employed for modeling buildings due to their imaging capability regardless of the time or weather conditions. Since the launch of TerraSAR-X in 2007, modern SAR satellites, e.g., TerraSAR-X, TanDEM-X, and CosmoSky-Med, have been providing meter or even sub-meter resolution images, making it possible to extract and reconstruct man-made objects from spaceborne SAR data. Complete global coverages of TerraSAR-X/TanDEM-X stripmap mode data have been acquired since 2012, offering significant potential as a data source for worldwide building reconstruction [6]. Nonetheless, owing to the side-looking geometry and employment of one-band radar sensors, urban structures appear distinct in SAR images but pose challenges in differentiation. Researchers have investigated building height retrieval from a single SAR image, InSAR data, and multi-aspect SAR or InSAR data or even circular SAR to overcome the limitations resulting from the side-looking geometry, especially occlusions [7-9].

While building reconstruction with SAR data has been extensively researched, large-scale studies are limited. Deep neural networks, gaining prominence for breakthroughs across various domains including remote sensing, face a key hurdle in urban SAR analysis: the scarcity of annotated data. To address this issue, the authors annotated individual buildings in a TerraSAR-X spotlight image employing a highly accurate Digital Surface Model (DSM) and proposed a segmentation network for predicting building areas in the SAR image [10]. The segmentation results are then applied to reconstruct building heights. The extracted building segments are then employed for LoD1 model reconstruction and achieved the mean height error of 2.39 m in the study site of Berlin. To reconstruct buildings in larger areas, more training data are needed. However, accurate DSMs are unavailable in most cases. Therefore, the LoD1 building reconstruction problem is reformulated as a bounding box regression problem in [11] so that height data from multiple sources can be employed to generate bounding boxes of buildings. A regression network is proposed and examined for four study sites using TerraSAR-X spotlight and stripmap mode images. The absolute mean height error achieved in the four sites ranges from 4.3m to 5.7m, which is significant, given that they are achieved from a single stripmap/spotlight TerraSAR-X image.

Compared with photogrammetry and RS data, SVI data and 2D building footprint data are easier and cheaper to be collected and processed (e.g., Mapillary and OSM). There have been some early efforts to estimate building height based on these new data sources. More specifically, one can extract multi-level morphological features (or urban-form features) in predicting key attributes (e.g., height, function, energy consumption, etc.) of buildings and streets from an urban analytic perspective, then use these features with a machine learning (ML) model for supervised or semi-supervised building height regression.

In [12], a novel semi-supervised method was developed for building floor estimation based on automatic facade parsing and urban architecture rules. In short, the authors seek to generate the estimation of building floor or height (by multiplying an average floor height) as the "pseudo label" to guide ML regression models with OSM morphometric features [2] as covariates. The developed method consists of three main steps: 1) SVI and OSM building

alignment: we used the compass angle and geotagged coordinates of the camera of SVI and applied a ray-tracing method to determine their relationship. 2) Facade object detection: we use a pre-trained one-stage object detection network (YOLO v3) to detect key facade features (e.g., window, balcony, and door) [13]. 3) Building floor estimation: based on facade object detection results, one can then apply a rule-based approach to determine the floor number in order to estimate the height of corresponding OSM buildings, which is used as "pseudo labels" to train an ML-based building height estimation model. As a case study, the authors validate the proposed method in the city of Heidelberg, Germany, and the preliminary result looks very promising with a Mean Absolute Error (MAE) of around 2.1 meters for 308 OSM buildings.

As lessons learned, the potential of SVI in mapping large-scale 3D building attributes are acknowledged, while it still has its own limitation regarding data quality and estimation accuracy. Therefore, a potential solution is to conduct data and model fusion of RS data, especially with SAR imagery, and crowdsourcing data (e.g., OSM and SVI) in large-scale real-world mapping and data-producing scenarios. One can imagine a conceptual framework, which can include multiple data sources and data views (street-level and overhead) and explore their synergies w.r.t granularity, scalability, availability, and computational efficiency for automatic 3D building attribute mapping, which shall then benefit the OSM community by enriching the OSM building attribute data quality in a longer term [14,15].

To sum up, we share the latest developments in large-scale 3D building attribute mapping in OpenStreetMap with two solid case studies using SAR and SVI, which sheds important light on the future research direction in this emerging research area. We expect this research stream will make a long-term and sustainable impact in numerous downstream applications within urban planning, disaster response, and global environmental monitoring and mapping.

References

[1] Biljecki, F., & Chow, Y. S. (2022). Global building morphology indicators. *Computers, Environment and Urban Systems*, *95*, 101809.

[2] Milojevic-Dupont, N., Hans, N., Kaack, L. H., Zumwald, M., Andrieux, F., de Barros Soares, D., Lohrey, S., Pichler, P.-P., & Creutzig, F. (2020). Learning from urban form to predict building heights. *PLOS one*, *15*(12), e0242010.

[3] Kolbe, T. H., Gröger, G., & Plümer, L. (2008). CityGML–3D city models and their potential for emergency response. In S. Zlatanova, & J. Li (Eds.), *Geospatial information technology for emergency response* (pp. 257–274). Taylor & Francis.

[4] Fan, H., & Meng, L. (2012). A three-step approach of simplifying 3D buildings modeled by CityGML. *International Journal of Geographical Information Science*, *26*(6), 1091–1107.

[5] Alobeid, A., Jacobsen, K., & Heipke, C. (2009). Building height estimation in urban areas from very high resolution satellite stereo images. In *ISPRS Hannover Workshop* (pp. 2–5).

[6] Zhu, X. X., Sun, Y., Shi, Y., Wang, Y., & Ge, N. (2018). Towards global 3d/4d urban modeling using tandem-x data. In *EUSAR 2018; 12th European Conference on Synthetic Aperture Radar* (pp. 1-6). VDE.

[7] Soergel, U., Michaelsen, E., Thiele, A., Cadario, E., & Thoennessen, U. (2009). Stereo analysis of high-resolution SAR images for building height estimation in cases of orthogonal aspect directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 64*(5), 490–500.

[8] Dubois, C., Thiele, A., & Hinz, S. (2016). Building detection and building parameter retrieval in InSAR phase images. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 228–241.

[9] Xu, F., Jin, Y.-Q. (2007). Automatic reconstruction of building objects from multiaspect meter-resolution SAR images. *IEEE Transactions on Geoscience and Remote Sensing 45*(7), 2336–2353.
[10] Sun, Y., Hua, Y., Mou, L., & Zhu, X. X. (2021). CG-Net: Conditional GIS-Aware network for individual building segmentation in VHR SAR images. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-15.
[11] Sun, Y., Mou, L., Wang, Y., Montazeri, S., & Zhu, X. X. (2022). Large-scale building height retrieval from single SAR imagery based on bounding box regression networks. *ISPRS Journal of Photogrammetry and Remote Sensing, 184*, 79-95.
[12] Li, H., Yuan, Z., Dax, G., Kong, G., Fan, H., Zipf, A., & Werner, M. (2023). Semi-Supervised Learning from Street-View Images and OpenStreetMap for Automatic Building Height Estimation. In *12th International Conference on Geographic Information Science (GIScience 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
[13] Kong, G., & Fan, H. (2020). Enhanced facade parsing for street-level images using convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(12), 10519-10531.
[14] Sun, Y., Kruspe, A., Meng, L., Tian, Y., Hoffmann, E. J., Auer, S., & Zhu, X. X. (2023). *Towards Large-scale Building Attribute Mapping using Crowdsourced Images: Scene Text Recognition on Flickr and Problems to be Solved*. arXiv. https://doi.org/10.48550/arXiv.2309.08042
[15] Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. *Nature Communications*, *14*(1), 3985.

Developing a data validation method with OpenStreetMap Senegal and the Ministry of Health in support of accurate health facility data

Mark Herringer[1,*], Lamine Ndiaye[2] and Andy South[3]

[1] Open Healthsite Consulting, Hoorn, Netherlands; mark@healthsites.io
[2] OpenStreetMap Senegal, Dakar, Senegal; lamineyasey@gmail.com
[3] Afrimapr, Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, UK; andy.south@lstmed.ac.uk

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

This research examines the collaboration between a local OpenStreetMap chapter and health authorities to improve health facility data accuracy. By utilizing open data and statistical methods, communities can empower Ministries of Health, address Sustainable Development Goals (SDGs) indicators, and enhance emergency response.

The healthsites.io Digital Public Good [1] has been working with OpenStreetmap Senegal [2] since 2017. We have established a data collaborative focused on health facility data that lives in OpenStreetMap. The collaborative is a semi-formal network that identifies and shares geospatial data on health to OpenStreetMap. It works to identify gaps and barriers to sharing, defines methodologies and data models for sharing and supports stakeholders with sharing and the use of data especially for decision-making. Crucially, the collaborative saves validated data to OpenStreetMap which means that successive projects are able to benefit from the work even when programs end.

Accurate health data plays a vital role in effective healthcare planning, resource allocation, and emergency response. However, existing data sources often suffer from inaccuracies and limited sharing, hindering the potential for informed decision-making and comprehensive health interventions. In response, we have developed an Emergency Health data validation method [3]. The method involves local stakeholders and the healthsites.io open data platform as a means to enhance data quality and accessibility.

The Global Fund's COVID-19 response mechanism underscores the significance of accurate health facility data [4]. This mechanism relies on a robust Health Facility Registry to guide resource allocation, emergency response strategies, and pandemic management. The imperative to understand health capacity for effective emergency responses highlights the critical need for accessible, reliable health facility data.

Herringer, M., Ndiaye, L., & South, A. (2023). Developing a data validation method with OpenStreetMap Senegal and the Ministry of Health in support of accurate health facility data
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443334

Acknowledging the need for accurate health data, the healthsites.io collaborative demonstrates how citizens are supporting the Ministry of Health to improve data accuracy and maintenance through a citizen-Ministry of Health (MoH) collaboration for data enhancement. The Emergency Health data validation method establishes cooperation between OpenStreetMap Senegal, the Direction of Planning Research and Statistics (DPRS/MoH) [5] and the Centre des Opérations d'Urgence Sanitaire (COUS/MoH) [6]. Recognizing citizens as valuable stakeholders, the method integrates input via OpenStreetMap's collaborative mapping. By involving Ministry of Health entities, the method endorses the legitimacy of citizen-contributed data and streamlines its integration into official health systems. This merger enhances data accuracy and amplifies its accessibility, rendering it a resourceful foundation for health interventions and emergency responses. Such collaboration not only empowers communities by valuing their input but also cultivates a sustained framework for reliable health data, capable of benefiting successive endeavors. Ultimately, this operationalized collaboration bridges the citizen-Ministry of Health gap, supporting accurate health data and informed decision-making.

Research questions include 'How can OpenStreetMap data and open statistical methods empower communities to collaborate with MoH and address SDG 3.8.1 indicators?' [7] and 'What is the business case for sharing baseline health facility data, and how does it impact health outcomes and emergency response?'

The method adopts a human-centered design approach to ensure that the collaboration between communities and health authorities is rooted in the needs and perspectives of all stakeholders. This encourages active participation, transparency, and inclusivity in the collaborative process.

A thorough data audit is undertaken to evaluate the accuracy and comprehensiveness of health facility data. This stage highlights prevailing gaps and obstacles that require resolution for productive collaboration. Employing an R building block of reusable code streamlines the process of enhancing health facility location data by juxtaposing information from diverse sources. By comparing datasets, this approach harmonizes naming conventions between Centre des Opérations d'Urgence Sanitaire (COUS), Department of Public Health and Statistics (DPRS), and OpenStreetMap, fostering data consistency and collaborative efficiency (see Figure 1).

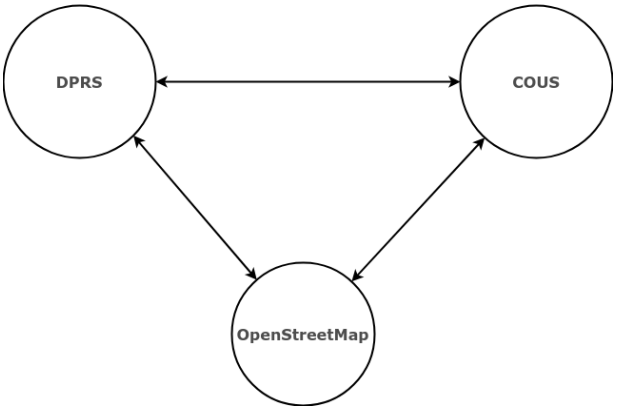

Figure 1. Harmonized health facility names between OpenStreetmap the Centre des Opérations d'Urgence Sanitaire (COUS) and the Department of Planning Research and Statistics (DPRS)

Field validation is carried out to verify the accuracy of health facility data on the ground. This involves on-site assessments, interviews, and data cross-referencing to ensure the reliability of the information.

The validated health facility data is shared to the OpenStreetMap platform, a widely used open mapping tool. This enables real-time access to accurate data by various stakeholders and supports evidence-based decision-making.

The approach underscores the significance of open development and open science principles in promoting effective collaborations. It emphasizes the use of open data and statistical methods to enable communities to actively work together with health authorities, leveraging local knowledge to improve health outcomes and address SDG 3.8.1 indicators. This collaborative approach not only strengthens health facility data but also involves local stakeholders in decision-making, fostering a sense of ownership and commitment. Sharing baseline health facility data forms a solid foundation for informed decision-making, resource allocation, and emergency responses, demonstrating the benefits of data sharing in enhancing health interventions. The OpenStreetMap-based approach supports transparency, trust, and ongoing engagement between communities and health authorities, illuminating the potential of collaborative mapping and open data initiatives in bridging the gap between communities and Ministries of Health. This symbiotic relationship can lead to improved health data accuracy, empowered communities, and enhanced health outcomes, contributing to the advancement of SDGs, resilient healthcare systems, and equitable health coverage through open development and open science principles.

References

[1] The Digital Public Goods Alliance (2023). *Unlocking the potential of open-source technologies for a more equitable world*. https://digitalpublicgoods.net

[2] OSM Contributors (2023). *WikiProject Senegal*. https://wiki.openstreetmap.org/wiki/WikiProject_Senegal

[3] Healthsites (2023). *Emergency health mapping campaign*. https://github.com/healthsites/emergency-health-data/wiki/Emergency-health-mapping-campaign

[4] The Global Fund (2023). *Technical Note on Leveraging COVID-19 Response Mechanism Investments for Geographic Information Systems*. https://www.theglobalfund.org/en/covid-19/news/2021-06-01-technical-note-on-leveraging-covid-19-response-mechanism-investments-for-geographic-information-systems

[5] Republic of Senegal, Ministry of Health and Social Action (2023). *La Direction de la Planification, de la Recherche et des Statistiques*. https://www.sante.gouv.sn/les-directions/la-direction-de-la-planification-de-la-recherche-et-des-statistiques-0

[6] Republic of Senegal, Ministry of Health and Social Action (2023). *Centre des Opérations d'Urgence Sanitaire*. https://www.sante.gouv.sn/les-services-rattaches/le-centre-des-op%C3%A9rations-durgence-sanitaire-cous

[7] World Health Organization (2023). *Coverage of essential health services (SDG 3.8.1)*. https://www.who.int/data/gho/data/themes/topics/service-coverage

Utilizing OSM data in geospatial representation learning

Piotr Gramacki[1,*], Kacper Leśniara[1], Kamil Raczycki[1], Szymon Woźniak[1] and Piotr Szymański[1]

[1] Department of Artificial Intelligence / Kraina.AI Lab, Wroclaw University of Science and Technology, Wrocław, Poland; piotr.gramacki@pwr.edu.pl, klesniara@kraina.ai, kraczycki@kraina.ai, swozniak@kraina.ai, piotr.szymanski@pwr.edu.pl

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Representation learning has been proven to be a very capable approach in many domains of artificial intelligence like Natural Language Processing (NLP) [1,2], Computer Vision (CV) [3,4] or Network Science (NS) [5]. Representation learning of spatial and geographic data is a rapidly developing field that allows for similarity detection between areas and high-quality inference using deep neural networks. Existing approaches concentrate on embedding raster imagery (maps, street or satellite photos), mobility data or road networks. We propose methods for learning vector representations of OpenStreetMap regions concerning urban functions, land use, POI location and road network with additional features. We also propose an approach to include public transport availability in the representation learning approach. We have designed our representation learning methods operating on various types of OpenStreetMap (OSM) data, one of the largest open databases with structured geospatial data. All our methods make use of a hierarchical spatial index to increase reproducibility. We utilize Uber's H3 geospatial indexing system [6] that partitions the space into uniquely identifiable hexagonal regions. We presented our results at different workshops accompanying the SIGSPATIAL conference.

Our methods are available in an open-source Python library - *Spatial Representations for Artificial Intelligence* (*srai*) [7] designed for working with geospatial data. The library can download geospatial data, split a given area into microregions using multiple algorithms and train an embedding model using various architectures. It includes baseline models as well as more complex methods from published works. Those capabilities make it possible to use *srai* in a complete pipeline for geospatial task solving. An exemplary outcome of one of the available methods is depicted in Figure 1. We hope our library will take the first steps to standardize the geospatial AI domain. The library is fully open-source and published under the Apache 2.0 license.

In our first research paper [8], we utilize a key/tag metadata structure bound to every object in the OSM to create contextual embeddings and use it to predict bicycle-sharing systems' (BSS) stations' location. Laying out a bike-sharing system is critical in addressing the increase in bike demand. We address the problem of cost and time in BSS layout design and propose a new solution to streamline and facilitate the process of such planning by

Gramacki, P., Leśniara, K., Raczycki, K., Woźniak, S., & Szymański, P. (2023). Utilising OSM data in geospatial representation learning In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443338

using spatial embedding methods. Based only on publicly available data from OpenStreetMap and station layouts from 34 cities in Europe, a method has been developed to divide cities into microregions using the Uber H3 discrete global grid system and to indicate regions where it is worth placing a station based on existing systems in different cities using transfer learning. The result of the work is a mechanism to support planners in their decision-making when planning a station layout with a choice of reference cities.

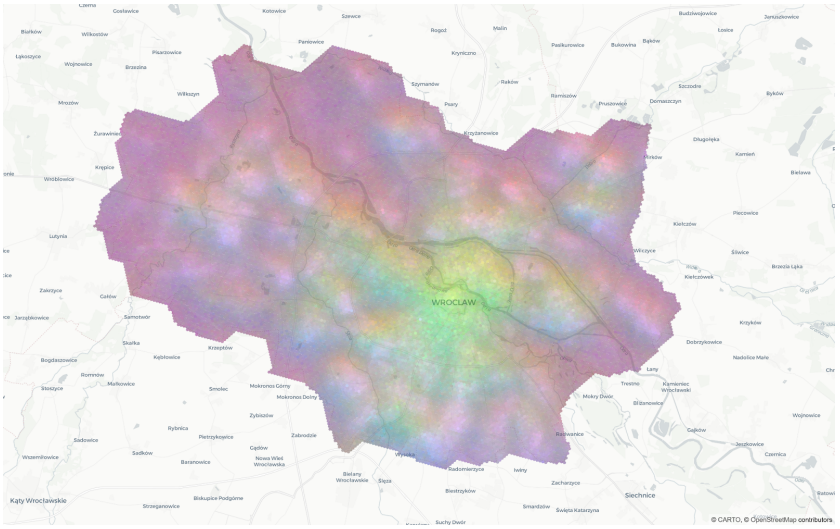

Figure 1. Embeddings of regions represented in RGB colour space in Wrocław, Poland.

Secondly, in hex2vec [9], we worked with the same data type and proposed a word2vec-style [1] embedding method that yields semantic aware embeddings via the Skip-gram model [10] with negative sampling [11]. The resulting vector representations showcase semantic structures of the map characteristics, similar to ones found in vector-based language models. Through manual verification of tagging quality, we selected 36 cities for training region representations. We also present insights from region similarity detection in six Polish cities and propose a region typology obtained through agglomerative clustering.

We also propose a highway2vec method [12] to embed network structures with an autoencoder neural network. This method generates microregions' embeddings concerning their road infrastructure characteristics. We obtained vector representations that detect how similar map hexagons are in the road networks they contain. Additionally, we observe that embeddings yield a latent space with meaningful arithmetic operations. Finally, clustering methods allowed us to draft a high-level typology of obtained representations.

Finally, in gtfs2vec [13], we reach external data sources to represent public transport availability in our embedding methods. We use public transport schedules publicly available in the GTFS format. We selected 48 European cities and gathered their public transport timetables. We utilized Uber's H3 spatial index to divide each city into hexagonal micro-regions. Based on the timetables data, we created certain features describing the quantity and variety of public transport availability in each region. Next, we trained an auto-associative deep neural network to embed each region. Later, we used a hierarchical clustering approach to identify similar regions. Finally, we analyzed the obtained clusters at

different levels to identify some number of clusters that qualitatively describe public transport availability. We showed that our typology matches the characteristics of analyzed cities and allows successful searching for areas with similar public transport schedule characteristics.

We argue that the geospatial artificial intelligence domain needs its *huggingface moment*. Huggingface [14,15] standardized research in Natural Language Processing, providing unified interfaces for models, benchmark datasets and more. Such easy access to models and evaluation data encourages researchers to present their results reproducibly. The geospatial domain needs just this since it is often challenging to reproduce results. It can result from closed-source data or hard-to-reproduce training code. With our library and our models' inclusion, we want to contribute towards standardizing the geospatial artificial intelligence domain.

While Huggingface has set a standardization precedent in the NLP domain, we do not position ourselves as its direct equivalent in geospatial AI. Our primary objective is to highlight the need for a unified approach in this domain. Our contributions aim to initiate a dialogue on GeoAI standardization, emphasizing reproducibility and accessibility.

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems, 26*.

[2] Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., & Zettlemoyer L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. Association for Computational Linguistics.

[3] Feichtenhofer C., Fan H., Xiong B., Girshick R., & He K. (2021). A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3298–3308.

[4] Ghosh R., Jia X., Yin L., Lin C., Jin Z., & Kumar V. (2022). Clustering Augmented Self-Supervised Learning: An Application to Land Cover Mapping. In Renz, M., Sarwat, M., Nascimento, M. A., Shekhar, S., & Xie, X. (Eds.) *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL '22),* 3. The Association for Computing Machinery.

[5] Zhang, D., Yin, J., Zhu, X., & Zhang, C. (2017). Network Representation Learning: A Survey. *IEEE Transactions on Big Data*, 6(1), 3-28.

[6] Uber Technologies Inc. (2018). *H3: Uber's hexagonal hierarchical spatial index | Uber Blog*. https://www.uber.com/blog/h3

[7] Gramacki, P., Leśniara, K., Raczycki, K., Woźniak, S., Przymus, M., & Szymański, P. (2023). SRAI: Towards Standardization of Geospatial AI. In S. Newsam, L. Yang, G. Mai, B. Martins, D. Lunga & S. Gao (Eds.), *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '23)*. 43–52. The Association for Computing Machinery.

[8] Raczycki, K. & Szymański, P. (2021). Transfer learning approach to bicycle-sharing systems' station location planning using OpenStreetMap data. In B. Kar, S. Mohebbi, G. Fu, X. Ye, & O. A. Omitaomu (Eds.), *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities (ARIC '21)*, 1–12. The Association for Computing Machinery.

[9] Woźniak, S. & Szymański, P. (2021). Hex2vec: Context-Aware Embedding H3 Hexagons with OpenStreetMap Tags. In D. Lunga, L. Yang, S. Gao, B. Martins, Y. Hu, X. Deng, & S. Nesam (Eds.), *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GEOAI '21)*, 61–71. The Association for Computing Machinery.

[10] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv. https://arxiv.org/abs/1301.3781
[11] Bojanowski P., Grave E., Joulin A., & Mikolov T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.
[12] Leśniara, K. & Szymański, P. (2021). Highway2vec: representing OpenStreetMap microregions with respect to their road network characteristics. In B. Martins, D., Lunga, S. Gao, S. Newsam, L. Yang, X. Deng, & G. Mai (Eds.), *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GEOAI '22)*, 18-29. The Association for Computing Machinery.
[13] Gramacki, P., Woźniak, S. & Szymański, P. (2021). Gtfs2vec: Learning GTFS Embeddings for comparing Public Transport Offer in Microregions. In G. Cavallaro, B. Heras, D. Lunga, M. Werner, & A. Züfle (Eds.), *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data (GeoSearch'21)*, 5-12. The Association for Computing Machinery.
[14] Lhoest, Q., Villanova del Moral, A., von Platen, P., Wolf, T., Šaško, M., Jernite, Y., Thakur, A., Tunstall, L., Patil, S., Drame, M., Chaumond, J., Plu, J., Davison, J., Brandeis, S., Sanh, V., Le Scao, T., Canwen Xu, K., Patry, N., Liu, S., McMillan-Major, A., Schmid, P., Gugger, S., Raw, N., Lesage, S., Lozhkov, A., Carrigan, M., Matussière, T., von Werra, L., Debut, L., Bekman, S., & Delangue, C. (2021). Datasets: A Community Library for Natural Language Processing. In H. Adel, & S. Shi (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175–184. Association for Computational Linguistics.
[15] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu, & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 38–45. Association for Computational Linguistics.

Spot: A natural language interface for geospatial searches in OSM

Lynn Khellaf[1,*], Ipek Baris Schlicht[1], Julia Bayer[1], Ruben Bouwmeester[1], Tilman Miraß[1] and Tilman Wagner[1]

[1] Research and Cooperation Projects, Deutsche Welle, Bonn, Germany; lynn.khellaf@dw.com, ipek.baris-schlicht@dw.com, julia.bayer@dw.com, ruben.bouwmeester@dw.com, tilman.mirass@dw.com, tilman.wagner@dw.com

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Investigative journalists and fact-checkers have found OpenStreetMap (OSM) [1] to be an invaluable resource for their work due to its extensive coverage and intricate details of various locations, which play a crucial role in investigating news scenes. However, the accessibility and usability of this tool pose significant challenges for individuals without a technical background. This need for simplified access to OSM data has brought attention to the potential of Large Language Models (LLMs). Known to the public through applications like OpenAI's ChatGPT [2], Meta's Llama [3], or Google's Bard [4], these models have showcased remarkable capabilities in tackling a range of natural language processing (NLP) tasks. Nevertheless, their deployment in the context of low-resourced languages, both natural and for programming, comes with its own set of challenges, as is the case with OSM query languages such as Overpass Turbo (OT) [5] - an essential OSM data querying tool.

Notably, individuals have started using systems like ChatGPT to formulate OT queries due to the complexity of the language, highlighting the desire for easier access to OSM data. However, the results of such attempts are often only half-convincing. A critical issue has emerged with these models' tendency to hallucinate, leading to the generation of erroneous and non-executable outputs. Other prior efforts [6,7] were highly limited in their functionality and performance, further underscoring the need for effective solutions in this domain.

The central focus in our KID2 ("KI gegen Desinformation #2") project is therefore to enable a broader audience to seamlessly query the extensive OSM database without the prerequisite of an in-depth understanding of the intricate OSM tagging system or complex OSM query languages like Overpass Turbo. This is achieved through the development of Spot, a robust natural language interface tailored explicitly for querying OSM data, finding Spots, combinations of objects in the public space. Users' natural sentences are translated by a transformer model and turned into OSM database searches. The primary use case within the KID2-project is geo-location in the context of verification, a capability important for (investigative) journalists who often require precise location information for their work.

Khellaf, L., Schlicht, I.B., Bayer, J., Bouwmeester, R., Miraß, T., & Wagner, T. (2023). Spot: A Natural Language Interface for Geospatial Searches in OSM
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443346

The application includes a user-friendly graphical interface where users can effortlessly enter their textual search requests, with the subsequent results being visually displayed on the map. This system architecture takes inspiration from the existing Overpass Turbo; however, a conscious decision was made to develop proprietary database query methods to enhance flexibility and functionality.

To construct the foundation for this interface, the OSM data was transformed into a database format that allows for fast and easy data access. This included the selection and processing of a subset of OSM tags – specifically, those that are both visible and instrumental in describing a scene, resulting in a 20% reduction in data size. The subsequent step involved importing this processed dataset into a Postgres (https://www.postgresql.org) instance. Additionally, the Postgres instance was enhanced with the PostGIS [8] extension to facilitate geospatial queries.

To align with the overarching goal of ensuring accessibility for a wider user base unacquainted with the complexities of OSM tagging, visually similar tags were grouped logically, and the bundles were assigned numerous natural words users might use to describe the corresponding objects. The result is a novel table structure containing information from the OSM database regarding the tag bundles, the natural descriptors, and a list of tags that tend to appear in combination with the current bundle frequently, as extracted from the OSM database.

As part of our training data generation process, we randomly generated tag samples representing objects from the underlying database to construct artificial queries. Each drawn object can be assigned multiple tags based on the previously extracted frequent tag combinations. This process was iterated multiple times to generate an extensive list of random tag combinations, each representing one artificial user query. The selected tags were then stored in a novel intermediate graph-database format. Aside from an area specification (area name or blank if area is defined in the UI), this format encompasses objects with a variable number of tags as nodes, while the edges consist of distance information between the objects where required. This database laid the groundwork for generating effective OSM database queries.

Harnessing the power of ChatGPT, tailored prompts were generated using the selected tag combinations. These prompts were used to instruct ChatGPT to generate natural search sentences emulating user queries. Aside from all the query information from the intermediate format, style instructions were given to improve language variation. The subsequent step focused on training a neural network to translate these natural sentences back into the novel graph-database format. The chosen translation model was a text-to-text transformer which is pretrained on Common Crawl (https://commoncrawl.org) and a variety of supervised task datasets, T5 [9]. The decision to develop a custom machine learning model was due to the necessity to have full control over bias, accuracy and data privacy. Future tests might include the use of open-source LLMs as the translation model. Figure 1 provides a visual representation of the sequential steps within the pipeline through an example sentence.

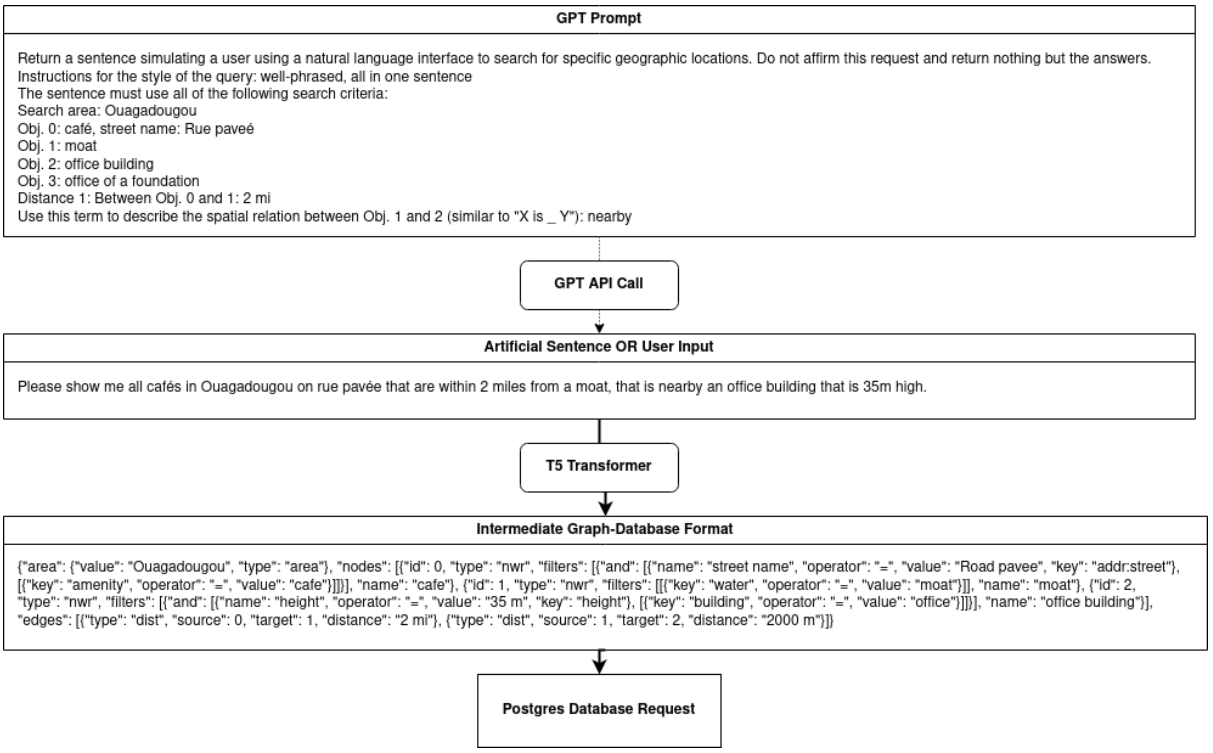

Figure 1. Illustration of the pipeline that transforms a natural sentence into a PostGres request to search the OSM database. The first container shows a GPT prompt that is only used during the generation of artificial training data.

Evaluation criteria of the model output are the validity of the format and the semantic accuracy of the extracted information. Benchmarking, which is going to be performed in the near future, will involve a comparison against established natural language OSM interfaces and common LLMs like GPT-4 [10]. To achieve this objective, we are currently developing a comprehensive, gold-standard natural language query dataset that encompasses all the desired use cases relevant to geolocation verification.

The current project goal is to construct a functional prototype for use in investigative journalism with a broad, but finite set of possible query structures. A first working proof-of-concept exists and is currently undergoing a process of iterative improvements. Future versions can be scaled up for different use cases by adapting the underlying database and prompting templates. Additionally, to foster collaboration and future advancement, all code and generated data is available as an open source repository under the following link: https://github.com/dw-innovation/kid2-spot. A public beta release of a usable demo is planned for February 2024.

In summary, we think this research could mark a further step towards democratizing access to OSM data through the development of a cutting-edge natural language interface. By bridging the gap between intricate query languages and user-friendly interactions, this interface has the potential to greatly improve how investigative journalists and a broader audience interact with geospatial data.

Acknowledgments

This project is led by the Deutsche Welle Research and Cooperation Projects teams and was co-funded by BKM (”Beauftragte der Bundesregierung für Kultur und Medien,” the German

Government's Commissioner for Culture and Media). Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

References

[1] OSM contributors (2017). *Planet dump*. *https://planet.osm.org*
[2] OpenAI (2022). *Openai: Introducing chatgpt*. https://openai.com/blog/chatgpt
[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi`ere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *Llama: Open and efficient foundation language models*. arXiv. https://arxiv.org/abs/2302.13971
[4] Pichai, S. (2023). *A message from our ceo - an important next step on our ai journey*. https://blog.google/technology/ai/bard-google-ai-search-updates/
[5] Raifer, M. (2023). *Overpass turbo github repository*. https://github.com/tyrasd/overpass-turbo
[6] Lawrence, C., & Riezler, S. (2016). Nlmaps: A natural language interface to query openstreetmap. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 6–10). The COLING 2016 organizing committee.
[7] Will, S. (2021). Nlmaps web: A natural language interface to openstreetmap. In M. Minghini, C. Ludwig, J. Anderson, P. Mooney, & A. Y. Grinberger (Eds.) *Proceedings of the Academic Track at State of the Map 2021* (pp. 13–15). Zenodo.
[8] PostGIS Project Steering Committee (2018). *PostGIS, spatial and geographic objects for postgreSQL*. https://access.crunchydata.com/documentation/postgis/2.2.7
[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html
[10] OpenAI. (2023). *Gpt-4 technical report*. https://cdn.openai.com/papers/gpt-4.pdf

Assessing bike-transit accessibility with OpenStreetMap

Reid Passmore[1,*], Randall Guensler[1] and Kari Watkins[2]

[1] Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA; tpassmore6@gatech.edu, randall.guensler@ce.gatech.edu
[2] Department of Civil and Environmental Engineering, University of California at Davis, CA, USA; kewatkins@ucdavis.edu

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Low-density land use, sprawl, and Euclidean zoning (i.e., separation of commercial and residential land-uses) can reduce the effectiveness of public transit by reducing the number of homes, amenities, services, and jobs near transit stops [1,2]. This gives rise to the first-last mile problem, where transit riders must travel long distances to access transit from their origin and from transit to their destination. Bicycles as a first and/or last-mile mode (henceforth referred to as bike-transit) can extend the service coverage area of a transit stop or station by allowing transit users to cover a greater distance in the same amount of time [3]. Not only can people reach transit stops faster on a bicycle than they could by walking; people using bicycles can also reach more transit stops within the same time frame. Lastly, people may be able to avoid bus feeder routes and cycle directly to higher service quality transit routes (such as rail).

Despite bike-transit's potential for shortening travel times, bike-transit is not commonly modeled in traditional travel demand modeling or trip planners. This is because bike-transit trips are computationally intensive to calculate given the number of possible transit stop pairs and departure times. Our solution to this is to use bicycle and transit shortest path algorithms to demonstrate how bike-transit improves public transit's accessibility to destinations by reducing overall travel times, transit waiting times, and the number of transit transfers needed. Previous work has been done in assessing how bike-transit improves the effectiveness of public transit [4-7]; in this case, we will take an in-depth look at three different locations that are varying in distance from bus and heavy rail service in Atlanta, Georgia, USA.

To develop our network for bike routing, the street network for the study area was retrieved from OpenStreetMap (OSM) through Osmnx and Overpass API. The OSM data were filtered to only include public roads and multi-use paths that allowed bicycle travel. Static GTFS data from the Metropolitan Atlanta Rapid Transportation Authority (MARTA) was used for transit shortest path calculation. Supplementary data from the Atlanta Regional Commission's 2022 Regional Bicycle Facility Inventory were used to add cycling

Passmore, R., Guensler, R., & Watkins, K. (2023). Assessing Bike-Transit Accessibility with OpenStreetMap
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443357

infrastructure not present in OSM. The final network was composed of 366,926 links and 323,275 nodes. Traffic analysis zones (TAZ) from the Atlanta Regional Commission's (ARC) 2020 activity-based model run served as potential origins and destinations. There were a total of 2,221 TAZs in the study area. In the case presented here, three TAZs were selected to represent the variability in transit service and land use throughout the study area. Three visualizations on accessibility, travel times, and transit mode(s) utilized are generated for each location to communicate the increase in accessibility from bike-transit for that area. The utilization of the transit network has not been examined in previous studies measuring bike-transit accessibility. Two configurations of bicycle first-last mile travel are considered: bringing the bicycle aboard transit to have the bicycle for biking at both ends of the trip (bike-transit) and leaving the bike at the first stop (bike-transit-walk). The optimal routes to all possible destinations in the transit service area are calculated for walk-transit, bike-transit-walk, and bike-transit.

The walking and biking portions of trips are modeled using Dijkstra's algorithm [8], and the transit portion is modeled using the round-based public transit optimized routing (RAPTOR) algorithm [9]. The walking and biking routing steps only considered travel time as an impedance, leaving out a number of potential attributes that pedestrians and cyclists consider in choosing a route such as elevation, the number of turns, and the presence of sidewalks/cycling infrastructure [10]. MARTA heavy rail stations at the extents of the study area tend to be park-and-ride oriented, and residential/commercial development around these locations is limited. The streets surrounding these stations are large, high speed roads that are not comfortable to cycle on or walk next to. As such, the current model likely overestimates the number of accessible areas. In future studies, these attributes need to be incorporated to more accurately assess bike-transit accessibility.

However, most of the road attributes that would enhance bicycle routing are not currently available in OSM. For this study, there were 467 unique OSM attributes (keys) present in the network data, yet most of the values for these keys were empty. When just examining public roads (excluding restricted access roads, service roads, and bicycle/pedestrian paths), there were 293 unique attributes. Figure 1 shows that most of these attributes were also empty. Attribute completion was calculated by adding the lengths of all roads with non-empty values for a particular attribute and dividing by the total length of all the roads. Meta attributes such as *highway* and *osmid* were completely filled out, but essential attributes for calculating bike impedances such as the number of lanes, the speed limit, and the presence of street parking had low completion values of 16%, 13%, and 0.1%, respectively. Other key attributes for cycling routing such as percent grade or incline were not present in the data at all.

Since these attributes are not available through OSM for many areas, researchers currently have to seek data from municipal or private sources. In the former case, the availability and quality of municipal data varies widely by jurisdiction. Municipal data on streets with the desired attributes were not available for this abstract’s study area. In the latter case, private data are often cost-prohibitive and cannot be shared widely, which limits the repeatability of research results. Another approach is to impute necessary data based on OSM’s *highway* key. The People for Bike’s Bicycle Network Analysis tool uses this approach to impute the number of lanes and the speed limit [11]. However, this could lead to inaccurate routing if the imputed data is incorrect. Other attributes, such as street parking, are more difficult to impute using just the *highway* key.

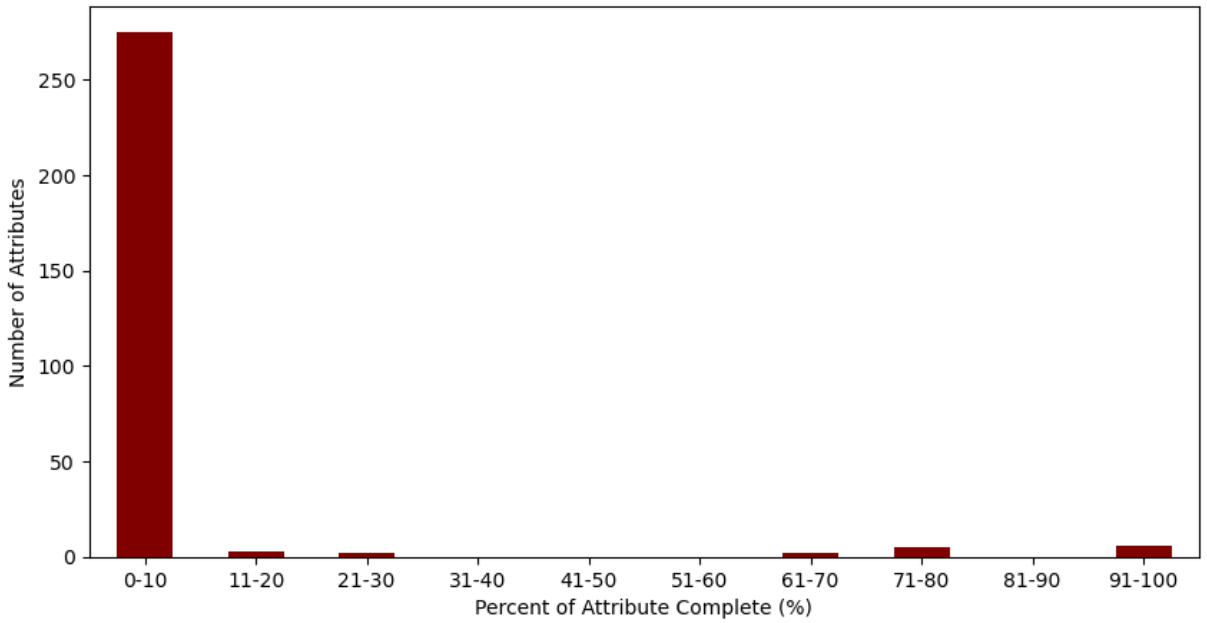

Figure 1. Attribute completion weighted by road length and rounded to the nearest percent (N = 293).

In the future, there should be a more concerted effort by OSM contributors, researchers, and organizations to fill in these attributes. This would allow researchers to develop more nuanced routing algorithms for cycling, walking, and transit use. However, for this abstract, bike-transit accessibility was assessed without these attributes. Current results indicate that bike-transit and bike-transit-walk decrease travel and wait times for transit, and in many cases reduce the number of transfers required compared to walk-transit. Transit services with higher travel speeds or frequencies such as heavy rail, greatly increased the geographic extent of accessible destinations and reduced travel times. Thus, an origin's distance to rail service had a major impact on the number of accessible TAZs.

Again, these results likely overestimate the extent of bike-transit accessible areas due to the lack of attributes relevant for using cycling-specific routing impedances. Future research, in addition to incorporating cycling specific impedances, will need to balance these impedances against public transit-specific impedances (which is also currently set as a time-only impedance). For example, even with two transit routes that provide an equivalent total transit link travel time, the route with the shorter wait time will likely be scored by users as having lower impedance (most users penalize wait time more than travel time in their mode choice decision-making) [4]. Additionally, if the bike route to access the transit route with the shorter waiting time had a higher impedance than the other transit route, some users may endure the higher impedance bike route to access the route with shorter wait time. Cyclists may also have public transit-specific preferences that can also be accounted for in estimating impedance, such as a preference for rail service over bus service and/or avoiding transfers. More research is needed to bring in specific mode impedances for transit and walking into overall impedance calculations.

Despite the limitations of this research, these results suggest that planners and engineers should consider bike-transit-walk and bike-transit in their planning and travel demand modeling processes to maximize the utilization of their existing transit network. Planners and engineers can repeat the analyses presented in this research to identify where bike-transit would be most effective. Additionally, this methodology can be used to assess

how public transit service changes and new cycling infrastructure can impact the accessibility of bike-transit trips. Other first-last mile modes such as bike-share and scooter-share can also be assessed.

However, efforts should be made by researchers, contributors, and other organizations to complete OSM's attribute data. This would allow planners and engineers to more accurately assess where the potential for bike-transit trips is greatest, so that measures that increase accessibility of these trips (secure bicycle storage and bicycle infrastructure) can be introduced. In many cases, adding cycling infrastructure to existing roadways will be less costly and more timely than transit-oriented development or new transit service. In addition, encouraging bike-transit trips could replace automobile trips and reduce transportation greenhouse gas emissions and vehicle miles traveled. Lastly, because bike-transit is a complicated mode due to the large number of potential transit stop and route combinations, transit operators should also work to provide mobile trip planners that offer step-by-step bike-transit routing. Widely used trip-planning apps do not include this option for most areas, which makes it difficult for travelers to plan these trips.

References

[1] Walker, J. (2012). *Human Transit*. Island Press/Center for Resource Economics.
[2] Hall, E. (2007). Divide and Sprawl, Decline and Fall: A Comparative Critique of Euclidean Zoning. *University of Pittsburgh Law Review*, *68*(4), 915-952.
[3] National Academies of Sciences, Engineering, and Medicine (2013). *Transit Capacity and Quality of Service Manual, Third Edition*. The National Academies Press. https://doi.org/10.17226/24766
[4] Zuehlke, K. (2007). *Impossibility of Transit in Atlanta: GPS Enabled Revealed Drive Preferences and Modeled Transit Alternatives for Commute Atlanta Participants*. Master's thesis, Georgia Institute of Technology.
[5] Wu, H., & Watkins, K. E. (2018). Accessibility Disparity Between Transit and Automobile: A Study of Atlanta and Seattle. *Transportation Research Board 97th Annual Meeting*.
[6] Ling, S.W.H. (2021). *An Atlanta-based Analysis on the Feasibility of Employee Commute Options Programs and Switching from Driving Alone to Alternative Commute Modes*. Master's thesis, Georgia Institute of Technology.
[7] Conway, M. W., Byrd, A., & van der Linden, M. (2017). Evidence-Based Transit and Land Use Sketch Planning Using Interactive Accessibility Methods on Combined Schedule and Headway-Based Networks. *Transportation Research Record: Journal of the Transportation Research Board*, *2653*(1), 45–53. https://doi.org/10.3141/2653-06
[8] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271.
[9] Delling, D., Pajor, T., & Werneck, R. F. (2015). Round-Based Public Transit Routing. *Transportation Science*, *49*(3), 591–604.
[10] Fitch, D. T., & Handy, S. L. (2020). Road environments and bicyclist route choice: The cases of Davis and San Francisco, CA. *Journal of Transport Geography*, *85*, 102705. https://doi.org/10.1016/j.jtrangeo.2020.102705
[11] People for Bikes (n.d.). *Bicycle Network Analysis Tool.* https://bna.peopleforbikes.org

Are Italian cities already 15-minute? Presenting a glocal proximity index, based on open data

Beatrice Olivari[1,*] and Angela Cimini[2,3]

[1] Department of Business Innovation & Development, Deda Next srl, Bolonga, Italy; beatrice.olivari@dedagroup.it
[2] Department of Architecture and Project, University of Rome La Sapienza, Piazza Borghese 9, 00186 Roma, Italy; angela.cimini@uniroma1.it
[3] Italian Institute for Environmental Protection and Research (ISPRA), Department of Networks and Environmental Information Systems (SINA), Via Vitaliano Brancati 48, 00144 Rome, Italy

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

In recent years, the concept of proximity has gathered significant attention and the best-known model dealing with this concept is Carlos Moreno's 15-minute city, where citizens can easily reach any essential service through a 15 minutes' walk (or bike ride) [1]. This city model presents numerous advantages, including reductions in car traffic and carbon footprint, improvement in citizens' health and safety, enhancement of the economy in the whole city, improvement of accessibility and so on. However, transitioning to a 15-minute city is not a straightforward undertaking and for this process to succeed it is best to rely on data-driven assessments of its developments. Deda Next developed an index of proximity [2], to help municipalities monitor the accessibility of their territory and develop targeted policies to improve mobility.

The Next Proximity Index (NEXI) is entirely based on OpenStreetMap data and capable of measuring the level of local proximity to services by walking, according to the principles of the 15-minute city. The goal of the index is to identify which of the different areas of a given territory already follow the 15-minute paradigm and its implementation is made available as an interactive map where the index is computed on a hexagonal grid and thematized according to its value.

In the last few years, some proximity indexes have been developed. With respect to the existing solutions that we were able to analyze, the added value of our index is given by three main characteristics:

- Scalability - the index computation algorithm can process runs at different scales without relevant performance issues. It is in fact available for the entirety of Italy.
- Replicability - thanks to OSM data, the index is easy to replicate in different areas and regions. It can also be customized, according to the needs of the municipality.

Olivari, B., & Cimini, A. (2023). Are Italian cities already 15-minute? Presenting a glocal proximity index, based on open data. In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443359

- Interoperability - the index is not only accessible as interactive web map, but the output data are also available through interoperable web protocols, as required by the European Directive 2007/2/CE (INSPIRE), with endpoints offering WMS (ISO19128) and WFS (ISO19142).

Finally, the Next Proximity Index was designed to be glocal: thanks to the world-wide availability of OSM data, it can be replicated everywhere (global), but it is also granular enough to be able to evaluate the proximity at a small scale (local) [3].

The Index calculation begins with the step of data downloading. Particularly, two types of input data are required, both downloaded from OSM, using Pandana, a Python library for network analysis.

- Services data - we identified eight categories of essential services: education, entertainment, grocery, health, posts and banks, green public areas, sustenance, shops. For most of these categories we adopted the categorization provided by OSM wiki.
- Road network data - calculating the accessibility to those services means considering the routes to follow to get there.

After downloading the data, the index is computed. First, we measure the time to reach the closest POI of each category, starting from all the nodes of the network. However, the information about accessibility of nodes is too granular, as the goal of the index is to identify the proximity of areas, not nodes. Therefore, the index provides a hexagonal grid (with 125m-sided hexagons). For each hexagon we aggregate the distances and assign a unique value of proximity.

The output of the Index for Italy is available as an interactive map (https://www.dedanext.it/topic-citta-15-minuti). It provides a view by category (accessible by the top-left menu) that shows the proximity of single categories of services. Otherwise, it shows a general categorization in five levels: 15-minute, 30-minute, 60-minute, low proximity and not measurable. Figure 1 represents the general index in the city of Rome.

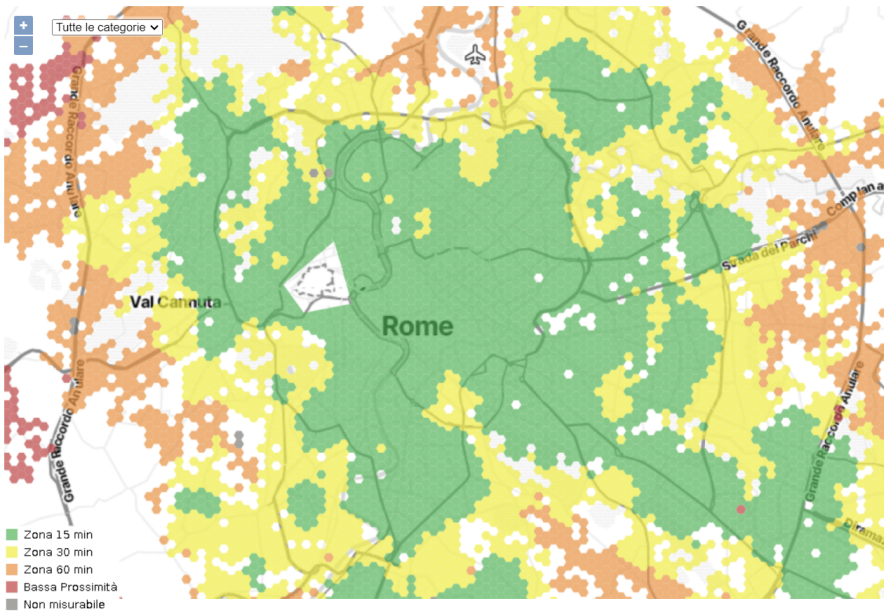

Figure 1. 15-minute index - City of Rome

Custom analyses have been performed in collaboration with municipalities and organizations:

- With the municipality of Ferrara, the data from the index were combined with population data, to highlight those areas where the lack of services was affecting more people.
- The municipality of Bologna asked us to integrate their open data in the index and add two custom categories.
- In partnership with the Italian Institute for Environmental Protection and Research (ISPRA), we are exploiting the index algorithm to develop an indicator of accessibility of green public spaces in Italy. The indicator will also be included in the next edition of ISPRA's Report "Land consumption, territorial dynamics and ecosystem services", aimed at the assessment of the UN Sustainable Development Goal (SDG) 11.7.1 indicator (https://unhabitat.org/sites/default/files/2022/08/sdg_indicator_metadata-11.7.1.pdf), which analyzes the presence of public spaces in urban areas in terms of the "Share of urban population without green urban areas in their neighborhood".

The Next Proximity Index is intended as a support tool for municipalities that are implementing new mobility strategies and to help them make data-informed decisions. The main goal is to develop an Index with world-wide replicability. Getting the required data from local administrations worldwide would be too challenging and it would also require complicated activities to integrate such datasets with mutually inhomogeneous structures. Therefore, the adoption of OSM data is crucial for reaching the goal. Moreover, the size and reliability of OSM is also responsible for a positive feedback effect regarding data coverage: so many public and private actors rely on it that they are stimulated to keep it updated and expand its coverage, so as to maintain the benefits they have from using it [4].

Even though OSM is the largest volunteer geographic information project in the world, OSM data is not always evenly available around the globe and the index is sensitive to the non-uniformity of data. To try to overcome this issue, we partnered with the municipality of Bologna, to integrate their more uniform and up-to-date data into the index, obtaining very promising results. In particular, the municipality of Bologna was interested in analyzing the proximity related to two new categories of services: "community services" (such as services for young people and citizen helpdesks) and "access to public transport" (basically represented by the distribution of public transport access points - stops - within the city). Therefore, these two categories were added alongside the eight categories of the standard implementation of NEXI, so that the corresponding levels of proximity could be analyzed. Another customization of the NEXI was requested concerning the parks category. The municipality was not satisfied by the corresponding OSM data as it included, in addition to the actual public parks, also some very small patches of green that could not be considered as parks. Therefore, OSM parks data were replaced by municipalities data.

During the collaboration with the municipality of Bologna, we found the topic of public urban greenery to be particularly interesting and suitable for experimentation. Therefore we started a partnership on this topic with ISPRA. When considering accessibility to green public spaces, it is essential to exclude private gardens and small portions of greenery that do not fall in the category. Moreover, parks can have multiple access points that are not indicated in OSM and this could affect the results of the index. With this respect, the ITO map classification system is considered, with reference to areas larger than one

hectare [5], but the identification of further appropriate supplementary information for the correct monitoring of the SDGs would provide useful indications to maximize the use of OSM data. This last aspect is the object of the collaboration between Deda Next and ISPRA for the evaluation of the UN SDG 11.7.1 indicator [5–7].

When it comes to walkability (or bikeability), it would be useful to integrate the index with morphological data to consider the slope as a factor of road's accessibility. Unfortunately, at the moment, the OSM road's incline tag is only available for a small percentage of OSM roads (around 0.01%). Therefore, as a future work we intend to integrate an orography dataset (https://tinitaly.pi.ingv.it). It would also be possible to organize campaigns in some partner cities, like Bologna or Ferrara, to involve local volunteers and add this data in OpenStreetMap.

Additionally, we explored the integration of demographic data into the index. We found it particularly relevant to compare data about accessibility to services with population density in the area. Currently, our collaboration with the municipality of Ferrara has furnished us with population data; nevertheless, our aspiration would be to expand this analysis to the entire Italian territory, incorporating a broader demographic dataset (https://data.humdata.org/dataset/kontur-population-dataset?).

References

[1] Moreno, C., Allam, Z., Chabaud, D., Gall, C., & Pratlong, F. (2021). Introducing the "15-Minute City": Sustainability, resilience and place identity in future post-pandemic cities. *Smart Cities, 4*(1), 93–111.

[2] Olivari, B., Cipriano, P., Napolitano, M., & Giovannini, L. (2023). Are Italian cities already 15-minute? Presenting the Next Proximity Index: A novel and scalable way to measure it, based on open data. *Journal of Urban Mobility, 4*, 100057.

[3] Handy, S. L., & Niemeier, D. A. (1997). Measuring accessibility: an exploration of issues and alternatives. *Environment and planning A, 29*(7), 1175–1194.

[4] Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive computing, 7*(4), 12–18.

[5] Konijnendijk, C. C. (2023). Evidence-based guidelines for greener, healthier, more resilient neighbourhoods: Introducing the 3–30–300 rule. *Journal of forestry research, 34*(3), 821–830.

[6] Giuliani, G., Petri, E., Interwies, E., Vysna, V., Guigoz, Y., Ray, N., & Dickie, I. (2021). Modelling accessibility to urban green areas using Open Earth Observations Data: A novel approach to support the urban SDG in four European cities. *Remote Sensing, 13*(3), 422.

[7] Munafò, M. (Ed.) (2022). *Consumo di suolo, dinamiche territoriali e servizi ecosistemici*. Report SNPA 32/22.

Rural water point mapping with/in OSM: implications of recent research in Malawi

Alistair Geddes[1,*]

[1] Division of Energy, Environment & Science, University of Dundee, UK; a.y.geddes@dundee.ac.uk

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

This paper provides a brief review of the state of water point mapping associated with and beyond OSM, with a focus on rural Malawi in south-eastern Africa. The review is offered to stimulate exchange with other researchers who are similarly concerned to address ways to leverage OpenStreetMap (OSM) to enhance management and development of rural water services. It stems from recent research conducted by the author on drinking water quality in southern Malawi, which has also included a new collaboration with a community-based organisation (CBO) with its own record of supporting local community water management developed over several years. As well as drawing on the author's own experiences working with this CBO the review also considers papers from other recent Malawi-focussed research projects plus reflections of a separate sub-Saharan water development specialist.

As in many Least Developed Country contexts the water service infrastructure across rural Malawi is generally at a basic-only level. Typically 'water points' refers to communal infrastructure such as pumped boreholes, protected springs, or gravity flow schemes. In rural Malawi such supply points (especially hand-pumped boreholes) exist in their tens of thousands, many created with funding from the charitable endeavours of non-governmental organisations. Otherwise, infrastructure investment (either public or private) remains limited, while widespread poverty affects many users' ability to pay for water services.

As well as being a key lifeline for rural communities, water points are relatively prominent features, located in the open air within or near rural settlements. However, despite such characteristics, water points in rural Malawi are presently not well mapped in OSM, with a search of the OSM database (for Malawi as a whole) finding a minute fraction of nodes - around just 200 - associated with appropriate tags ('amenity=drinking water'; 'man_made=water_well'). Moreover, this situation is despite several efforts over the last decade to increase OSM data creation within and about Malawi, most notably via Humanitarian OpenStreetMap, Missing Maps and the Youth Mappers initiative.

However, the motivation for this paper lies less with the goal of increasing numbers of water points within OSM than with considering how knowledge of and engagements with OSM may contribute to improved compilation, sharing and use of water point data more broadly. A rather similar perspective to this has been adopted in other recent work,

Geddes, A. (2023). Rural water point mapping with/in OSM: implications of recent research in Malawi
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443367

specifically by a research consortium led by the 510 initiative of The Netherlands Red Cross [1]. This consortium approached the aforementioned challenges in the context of exploring combinations of 'Big' and 'Small' Data to enhance reporting on Target 6.1 of the UN Sustainable Development Goals (SDGs): namely, to achieve, by 2030, 'universal and equitable access to safe and affordable drinking water for all'. In the consortium's study, 'big' data included those extractable from earth observation imagery, and 'small' data include geo-located water point datasets created by various water-sector organisations as part of their own regular activities (further information and links to the latter are available at [2]). The study entailed overlaying these datasets together, to identify and reduce locational inconsistencies among the water point datasets, and to explore enriching them with additional attribute data. Building outlines created in OSM were also included in the overlay process, both to assist the checking of water point locations and to provide a basis for making crude estimates of water demand. Water points were visually detected from custom-flown high-resolution drone imagery, supporting assessment of locational inconsistencies across the other existing datasets. In consequence, the workflow developed including the drone imagery was deemed cost-effective over limited size study areas.

The existing geo-located water point datasets just mentioned contain considerably larger numbers of water points than OSM, albeit varying in their spatial coverage, content, quality and currency. These variations reflect the multiplicity of stakeholders working in Malawi's rural water supply sector, as well as limits of official data and in terms of data coordination, data sharing and wider data literacy. The same 501-led consortium also conducted a 'data ecosystem'-based assessment of this situation, to assess further the prospects for enhanced SDG monitoring [3]. For that assessment the water point datasets (including OSM) were scored according to various standardised data supply characteristics. While the OSM dataset had smaller numbers of water points than others, it nonetheless scored higher for some characteristics, including for data reliability and accuracy, reflecting the processes within OSM whereby experienced users conduct validation checks of new data. However, the OSM data were also more limited in the range of water point attributes they included.

Broadly contemporaneous to these studies, the Scottish Government Climate Justice Fund Water Futures Programme, working in partnership with Government of Malawi, has undertaken to compile a major new rural water supplies asset register, covering all of Malawi [4]. Hosted on a third-party water data management platform (mWater - not connected with OSM), this register is reported to have included field inspections of all water supply points, so as to enable thorough assessment of supply sustainability. The creation of this new register suggests that data-based policy development in respect of water services is being taken seriously. Furthermore, it could also be construed as an initial step away from current community-based water management models, instead towards a more centralised and professionalised approach to water service delivery and management. Such a community-professional transition appears all the more desirable given that, at any given time, a high proportion, almost half, of all survey water supply points are identified as not functioning to standard, suggesting that many communities are incapable of effective management [4]. However, a question is whether such numbers reflect inherent failings of rural water-using communities themselves, or instead whether they reflect inadequacies in support and capacity development provided by organisations favouring community-based management [5].

The author's partnership with the CBO suggests community-based water management can actually be effective, provided sufficient planning, monitoring and resourcing are available to communities and supporting organisations. The CBO's own programme encompasses several elements: comprehensive baseline surveys of water points; targeted training for local water user communities; routine monitoring and refresher training; and programmes for sourcing and quality control of replacement pump parts [6,7]. In addition, many of the baseline and monitoring data collected are entered onto the CBO's own public-facing water point data sharing platform, Madzi Alipo (https://www.madzialipoapp.org), which is also integrated with a custom designed mobile data collection app. Through its independence from government, and because of its reputation undertaking other community development projects, the CBO is well-placed to develop trusted relationships with local community water committees. Nonetheless, our experience working with them reveals several challenges with data handling, management and integration which risk undermining the value of community engagement and data collection efforts.

Where does this leave us? There are a number of implications and opportunities to discuss further. These include, first, the possibilities for using higher resolution imagery, in conjunction with algorithmic methods, for more automated and scaled-up rural water point mapping, and for shortening data update cycles. Second, with regard to improving data coordination, consistency and sharing, encompassing both locational data and data on the functional status and other attributes of water points (e.g., water quality). Third, considering also the scope to improve the data management and analysis capacity of CBOs, and likewise, the capacity and involvement of local communities in data collection and monitoring activities.

Acknowledgements

This research is part of the project 'Tackling water security-health connections in rural Malawi', funded with an award from the University of Dundee Global Challenges Research Fund (GCRF) Urgency Call 2020-2021.

References

[1] van den Homberg, M., Crince, A., Wilbrink, J., Kersbergen, D., Gumbi, G., Tembo, S., & Lemmens, R. (2020). Combining UAV imagery, Volunteered Geographic Information, and field survey data to improve characterization of rural water points in Malawi. *ISPRS International Journal of Geo-Information*, *9*(10), 592. https://doi.org/10.3390/ijgi9100592

[2] Netherlands Red Cross (n.d). *Data4SDG: Characterizing the Data Supply Dimension of Data Ecosystems*. https://rodekruis.github.io/Data4SDG_Malawi

[3] van den Homberg, M., & Susha, I. (2018). Characterizing data ecosystems to support official statistics with open mapping data for reporting on Sustainable Development Goals. *ISPRS International Journal of Geo-Information*, *7*(12), 456. https://doi.org/10.3390/ijgi7120456

[4] Kalin, R., Mwanamveka, J., Coulson, A., Robertson, D., Clark, H., Rathjen, J., & Rivett, M. (2019). Stranded Assets as a key concept to guide investment strategies for Sustainable Development Goal 6. *Water*, *11*(4), 702. https://doi.org/10.3390/w11040702

[5] Carter, R. (2023). *Professionalising community management of rural water supply. Navigating the transition in sub-Saharan Africa (Publication 2023-1)*. Rural Water Supply Network. https://rural-water-supply.net/en/resources/1103-

[6] Chichlowski, W., Carnes, S., Norman, J., & Zaunda, H. (2022). *Madzi Alipo supporting waterpoint evaluation, repairs and maintenance in Blantyre Rural with the Community-Based Management-Plus model*. Malawi National Water Conference 2022.
[7] Carter, R. (2021). *Rural community water supply. Sustainable services for all*. Practical Action Publishing. https://practicalactionpublishing.com/book/2556/rural-community-water-supply

OpenStreetMap data for automated labelling of machine learning examples: The challenge of road type imbalance

Edson Melanda[1,2,*], Benjamin Herfort[3], Veit Ulrich[2], Francis Andorful[2] and Alexander Zipf[2,3]

[1] Department of Civil Engineering, GIS Nucleus - NGeo®, Federal University of São Carlos, São Carlos, SP, Brazil; melanda@ufscar.br
[2] GIScience Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany; veit.ulrich@uni-heidelberg.de, francis.andorful@uni-heidelberg.de, zipf@uni-heidelberg.de
[3] HeiGIT - Heidelberg Institute for Geoinformation Technology, 69120 Heidelberg, Germany; benjamin.herfort@heigit.org

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Advances in Artificial Intelligence (AI) and, specifically, in Deep Learning (DL) have fostered geospatial analysis and remote sensing, culminating in the establishment of GeoAI [1,2] and the solidification of research on methodologies and techniques for AI-assisted mapping [3-7]. Nevertheless, a particular challenge lies in the substantial demand for training examples in DL. Manual labelling of these examples is labour-intensive, consuming a considerable amount of time and financial resources. Alternatively, semi or automated labelling of data emerges as a prominent solution, as exemplified by the tool *ohsome2label* [8], which harnesses data from the OpenStreetMap [9] to label satellite images. However, moving from characterising object types (road, river, building) based on geometry to categorising them by attributes might result in an imbalanced class distribution in the utilised Machine Learning (ML) dataset.

Such imbalances are common in numerous practical applications. Learning from skewed datasets can be particularly challenging and often requires non-conventional ML techniques. A comprehensive awareness of the issues associated with class imbalance, as well as strategies for mitigating them, is essential [10]. In the context of spatial data, the distribution of classes can vary from country to country and region to region, adding a new layer of complexity and exacerbating this issue.

In this context, an analysis was conducted on the distribution of road types, defined by the values of the OSM "highway" tag, in diverse-profile nations. The aim was to evaluate the extent of class imbalance and to identify any consistent patterns in the distribution of road types used by mappers.

The initial hypothesis posits the existence of distinctive distribution patterns of road types, which are linked to the unique socioeconomic and cultural aspects of each country.

Melanda, E.A., Herfort, B., Ulrich, V., Andorful, F., & Zipf, A. (2023). OpenStreetMap Data for Automated Labelling of Machine Learning Examples: The Challenge of Road Type Imbalance
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443373

Another consideration is the voluntary nature of OSM data production, frequently carried out by individuals who map remote areas using solely remotely sensed images.

In this study, nine countries with diverse profiles were selected through random stratification, taking into account their unique characteristics across different continents. The goal was to capture a wide range of socioeconomic and cultural scenarios. From the Northern Hemisphere, Germany, Great Britain, Italy, and the United States were chosen, while Brazil, Chile, Ethiopia, India, and Sudan represented the Southern Hemisphere.

To conduct the study, the ohsome platform [11] was employed, facilitating the retrieval of historical statistics from OSM data, in collaboration with the ohsome-Dashboard interface [12], which enables rapid generation of metrics related to the discussed issue.

The experiment was set up in the ohsome-Dashboard with the following parameters, quantifying the overall road length grouped by tag: `OSM_tag: "highway=*"; OSM_type: "object=way"; measure: "length"; results grouping: "by tag"; time interval: "2023-01-01T00:00:00Z - 2023-07-16T20:00:00Z"; Period: "quarterly"`.

From the results, a significant variation in the number of distinct tags adopted by each country is observed initially, namely: Brazil 49; Chile 38; Ethiopia 31; Germany 73; Great Britain 55; India 48; Italy 64; Sudan 22; and United States 73. Due to that, the analysis was restricted to the top 10 tags with the highest percentage extension in each country, identifying four unique tag values as most frequent: u*nclassified, track, residential, and service* (see Figure 1).

The tag *unclassified* emerges as the most frequent of all, particularly prevalent in Brazil, Ethiopia, and Sudan (45.5%, 35.5%, and 37.7%, respectively). That tag value ranks second in countries such as Chile (16.9%), India (24.4%), and Italy (16.4%), while also featuring among the top five tags in three other nations: Germany (5th - 5.5%), Great Britain (5th - 12.0%), and United States (4th - 7.3%). The tag *track* emerges as the dominant one in terms of extension in Germany (45.2%), Chile (31.0%), and Italy (28.7%), also positioning itself among the top five most used tags in other nations. The third most frequent tag is *residential*, with percentages of 34.2% in the United States and 24.4% in India, also placing among the top five tags in other territories. Lastly, the tag *service* manifests as the most frequent in Great Britain, constituting 18.1% of the total road extension in the country. A broad analysis of the results reveals class imbalance in eight out of the nine selected countries. An exception to this scenario is observed in Great Britain, where a balanced distribution is found across the top five highway classes ranging from 18.1% to 12.0%.

The tag *unclassified* stands out from the others by being among the top two most frequent tags in six out of the nine nations considered. On the whole, this tag exhibits lower proportions in Northern Hemisphere countries and higher proportions in Southern Hemisphere countries. However, Chile and Italy deviate from this generalisation, revealing similar percentage proportions.

With the aim of establishing a link between tag frequency and the socioeconomic and cultural aspects of a nation, which could serve as a coarse proxy in an ML model, a comparison was drawn between the proportions of the aforementioned tag and the Human Development Index (HDI). This resulted in a negative correlation with an R^2 value of 0.62. In the graph that illustrates the relationship between HDI and the proportion of the *unclassified* tag, Brazil emerges as an outlier. When Brazilian data is excluded from the analysis, the R^2 value rises to 0.93, indicating a strong correlation for the remaining countries.

Figure 1. Top Ten Highway Tag Distributions by Country and Relationship of *Unclassified* Tag to HDI.

One hypothesis explaining Brazil's non-conformity lies in the country's internal heterogeneity, marked by distinct territorial occupation patterns, such as small and medium-sized agricultural properties in the South and Southeast, as well as vast properties in the Midwest. Another possibility would be an incorrect interpretation of the term unclassified instead of the term "road" (roads with unknown classification). However, a more detailed investigation of this issue is beyond the scope of this work.

To explain the relationship between HDI and the frequency of the tag unclassified two approaches could be employed. Firstly, assuming that the data on OSM reflects the reality, low values of HDI would be related to sparse urbanisation, and therefore a relatively elevated number of small settlements far apart from each other, leading to an increased proportion of roads of unclassified type. Another possibility would be, like for Brazil, a miss-concept use of the term unclassified, that would be related to a lower education level, reflected by a low value of HDI.

In the context of automatic labelling of examples for ML, the results underscore the necessity for careful evaluation of class (tag) distribution, as well as the caution required when utilising pre-trained ML models in diverse geographical regions.

Furthermore, these findings can be extrapolated to other OSM objects, like buildings and Land Use Land Cover (LULC), where class imbalance might arise due to the socio-economic and cultural traits of the region or country.

The investigation of the theme addressed in this study assumes an even greater level of significance when contemplating the growing application of AI-assisted mapping techniques. Specifically, when AI models are trained using OSM data for labelling ML examples, and subsequently the resultant mapped data is integrated back into the OSM database, it has the potential to lead to a vicious cycle.

To address those concerns, several avenues for future research have been identified: broadening the assessment of class imbalance across different nations and object categories, to identify anomalous patterns; developing methods and tools to semantically and topologically evaluate the data coherence and consistency in OSM.

Acknowledgments

This study was partially financed by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES Brazil - Finance Code 001 - Process: *88887.717541/2022-00*.

References

[1] Gao, S., Hu, Y., Li, W., & Zou, L. (2023). Special issue on geospatial artificial intelligence. *GeoInformatica, 27*(2), 133–136. https://doi.org/10.1007/s10707-023-00493-6

[2] Li, W., Hsu, C.-Y., & Hu, M. (2021). Tobler's first law in GeoAI: A spatially explicit deep learning model for terrain feature detection under weak supervision. *Annals of the American Association of Geographers, 111*(7), 1887–1905. https://doi.org/10.1080/24694452.2021.1877527

[3] Ramalingam, S. P., & Kumar, V. (2023). Automatizing the generation of building usage maps from geotagged street view images using deep learning. *Building and Environment, 235,* 110215. https://doi.org/10.1016/j.buildenv.2023.110215

[4] Usmani, M., Napolitano, M., & Bovolo, F. (2023). Towards global scale segmentation with openstreetmap and remote sensing. I*SPRS Open Journal of Photogrammetry and Remote Sensing, 8,* 100031. https://doi.org/https://doi.org/10.1016/j.ophoto.2023.100031

[5] Trento Oliveira, L., Kuffer, M., Schwarz, N., & Pedrassoli, J. C. (2023). Capturing deprived areas using unsupervised machine learning and open data: A case study in são paulo, brazil. *European Journal of Remote Sensing, 56*(1), 2214690. https://doi.org/10.1080/22797254.2023.2214690

[6] Botelho, J., Costa, S. C. P., Ribeiro, J. G., & Souza, C. M. (2022). Mapping roads in the Brazilian amazon with artificial intelligence and sentinel-2. *Remote Sensing, 14*(15), 3625. https://doi.org/10.3390/rs14153625

[7] Mahmoody-Vanolya, N., & Jelokhani-Niaraki, M. R. (2023). Measuring the spatial similarities in volunteered geographic information. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-4/W1-2022*, 411–416. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-411-2023

[8] Wu, Z., Li, H., & Zipf, A. (2020). From Historical OpenStreetMap data to customized training samples for geospatial machine learning. In M. Minghini, S. Coetzee, L. Juhász, G. Yeboah, P. Mooney, & A. Y. Grinberger (Eds.) *Proceedings of the Academic Track at the State of the Map 2020* (pp. 9-10). http://doi.org/10.5281/zenodo.3923040

[9] OSM contributors. (2023). *OpenStreetMap*. https://www.openstreetmap.org

[10] Johnson, J.M., & Khoshgoftaar, T.M. (2019) Survey on deep learning with class imbalance. *Journal of Big Data, 6*, 27. https://doi.org/10.1186/s40537-019-0192-5

[11] Heidelberg Institute for Geoinformation Technology (2023). *ohsome.* https://heigit.org/big-spatial-data-analytics-en/ohsome

[12] Heidelberg Institute for Geoinformation Technology (2023). *ohsome-dashboard.* https://dashboard.ohsome.org

Exploring road and points of interest (POIs) associations in OpenStreetMap, a new paradigm for OSM road class prediction

Francis Andorful[1]*, Sven Lautenbach[2], Christina Ludwig[1,2], Benjamin Herfort[2], Fulman Nir[1] and Alexander Zipf[1,2]

[1] GIScience Research Group, Heidelberg University, Heidelberg, Germany;
francis.andorful@uni-heidelberg.de, nir.fulman@uni-heidelberg.de
[2] HeiGIT - Heidelberg Institute for Geoinformation Technology, 69120 Heidelberg, Germany;
sven.lautenbach@uni-heidelberg.de, christina.ludwing@uni-heidelberg.de, herfort@uni-heidelberg.de, zipf@uni-heidelberg.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the OSM Science 2023 Conference after peer-review.

Completeness is the key metric in the OpenStreetMap (OSM) data quality evaluation space that assesses the number of mapped features in relation to actual objects [1]. Attribute completeness is of utmost importance for many applications. With regard to city planning, environmental risk assessment, routing, and accessibility applications, road class is important. The increasing availability of machine-derived road information, such as Microsoft Global Open Road Data, may lead to a further increase in the completeness of the road network; however, the products miss attribute information. False assumptions regarding capacity, top speed, and road quality are frequently the result of incorrect road classifications, which may necessitate routing errors in applications [2]. Therefore, an additional estimation of road class information is beneficial for many applications.

The strategic placement of Points of Interest (POIs) is not arbitrary but rather a deliberate outcome influenced by a variety of factors. Entrepreneurs and city planners serve as pivotal agents who can identify opportunities, mobilize resources, and generate value within a particular spatial and temporal context. Their deep familiarity with the historical background and specific contextual intricacies equips them with the ability to identify opportunities that might go unnoticed by others [3]. This astute awareness allows for intentional placement of POIs, creating dynamic hubs of innovation and economic activity.

In this study, we explored the answers to the question: Can we gain insights into road types by observing the locations where entrepreneurs choose to place their businesses? We looked at any connections that might exist between various road types and nearby POIs. We analyzed the co-occurrence between amenities such as gas stations, restaurants, and hotels on the one hand side and different road categories such as

Andorful, F., Lautenbach, S., Ludwig, C., Herfort, B., Fulman, R., & Zipf, A. (2023). Exploring Road and Points of Interest (POIs) Associations in OpenStreetMap, A New Paradigm for OSM Road class Prediction
In: Minghini, M., Li, H., Grinberger, A.Y., Liu, P., Yeboah, G., Juhász, L., Coetzee, S., Mooney, P., Sarretta, A., & Anderson, J. (Eds.). Proceedings of the OSM Science at State of the Map Europe 2023, Antwerp, Belgium, 10-12 November 2023. Available at https://zenodo.org/communities/osmscience-2023
DOI: 10.5281/zenodo.10443380

residential, service, and primary roads across 6 cities (Accra, Berlin, Tehran, Rio de Janeiro, Guatemala, Washington). We hope to identify stable relationships between road class and the characteristics of nearby POIs. Although roads are often considered the most complete feature in a region, road class information is frequently missing. At least in some of the cities investigated, the POIs were extensively mapped. Therefore, being able to predict road classes based on POI information seems to be an interesting option that we explore here.

To address the potential biases arising from the dominance or absence of certain road classes, we systematically assessed all road segments based on their labeled classifications. We observed that residential roads were consistently twice as prevalent as the next most common class across all the cities. However, it is noteworthy that certain classes, such as cycleways, were not uniformly distributed across all cities. For instance, cities such as Rio De Janeiro and Berlin exhibited significant cycleway networks, while Accra lacked any designated cycleways, and Washington showed a more random distribution. Additionally, road classifications, such as footways, living streets, and tracks, are relatively scarce. This limited presence may render them susceptible to overshadowing in a frequency-based analysis approach, as utilized in this study. Consequently, we made the decision to exclude road classes that had the potential to either exert undue influence or be insufficiently represented in our analysis. We focused on the six pivotal road classes outlined in the OSM Wiki (motorway, primary, secondary, tertiary, trunk, and unclassified roads).

Through visual exploration, we conducted random assessments of the POIs situated in proximity to roads within these cities. This approach has allowed us to gain a comprehensive understanding of this area. As a result of this process, we identified a set of ten distinct POI categories that collectively encompassed all observed POIs. We linked the OSM road segment with 10 POIs categories, such as amenities, shops, and public transport, based on a spatial join using a 100m (approximately two minutes) perpendicular walk distance threshold with the options of one road segment to many POIs and returned an inner join as the result. The result was a series of road segments with associated POI in the attribute. The road segment may repeat, depending on the number of POIs around it. To unravel the underlying patterns, we adopted a descriptive statistical approach, primarily cross-tabulation, to explore the raw observation frequencies between individual POIs and specific road classes. Within this framework, we estimated the mean and standard deviation for each road class in the observation matrix. The calculated z-scores then revealed the likelihood of encountering a specific road-POI pair under a normal distribution curve.

Considering that the highest probability for a singular POI to align with a particular road type was approximately 85%, we established a threshold at the 90th percentile, equating to 70% probability. This criterion serves as the basis for selecting the final candidate POIs. A hierarchical pyramid structure (see Figure 1) is designed to classify POIs based on their uniqueness in relation to specific roads. This structure comprises three distinct classes. Class 1 encapsulates unique POIs closely associated with one or two road classes. Class 2 encompasses more loosely linked candidate POIs, potentially aligning with three to four road classes. Finally, Class 3 encompassed liberal POIs distributed along five or all of the considered road classes, acknowledging the inherent uncertainty in OSM

labeling. This class pyramid provided a refined framework that was factored into labeling uncertainty.

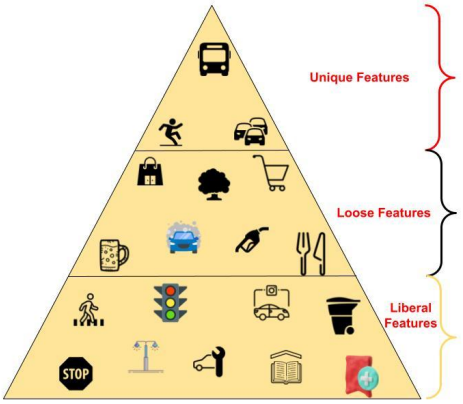

Figure 1. Hierarchical Pyramid Used in classifying POIs based on their uniqueness.

Examining the road class and POIs associations of different cities revealed intriguing patterns in the distribution of POIs along various road types. These insights are organized within a hierarchical framework, following the pyramid structure of the methodology.

In Accra, certain POIs such as motorway junctions and gift shops align exclusively with motorway road classes. Primary roads host unique establishments, such as car dealerships, guesthouses, monuments, and wastewater plants. Secondary roads feature diverse POIs, such as chemists and furniture shops. In Berlin, motorway junctions and vending machines cluster around motorways, whereas primary roads are characterized by supermarkets. Berlin's secondary roads are marked by cafes, vending machines, and more, whereas tertiary roads feature cafes, weirs, and doctors' establishments. Taxis align exclusively with the trunk roads.

Guatemala's data lack motorway classes, which influences the distribution of POIs. Primary roads host various establishments, while secondary roads include bars, worship houses, etc.. Tertiary roads include dentists, embassies, and diverse shops. Trunk roads offer a range of POIs such as car dealerships and fire stations. Unclassified roads showcase unique elements such as artwork and bars.

In Rio de Janeiro, motorways host clinics, police stations, and towers. Primary roads stand out with railway stations, whereas secondary roads feature monuments and travel agencies. Tertiary roads host bakeries and clinics. Washington City's motorways and tertiary roads reveal a sense of uniformity with no singular POIs that distinctly define them. Primary roads encompass beauty shops, mobile phone shops, etc.. Secondary roads are unique to banks, clothing stores and supermarkets. Trunk roads display diversity from gift shops to pharmacies. Unclassified roads have POIs such as peaks and recycling centers.

During our investigation across various cities, we sought patterns within the urban landscapes. It was evident that there was no single distinctive POI exclusively associated with a particular road class across all six cities. Instead, such unique associations were found in two, three, or five of the six cities. The observed difference might be explained by the fact that the arrangement and configuration of street networks worldwide are shaped by a wide range of influences, including cultural, political, historical, technological,

design-related, climatic, and geographical factors which are beyond the scope of this paper [4]. For example, while motorway junctions were a distinct feature of motorways in Accra and Berlin, the cities of Rio de Janeiro and Tehran exhibited motorway junctions across more than two road classes, resulting in a more liberal interpretation of this POI.

In conclusion, the choice of descriptive statistics as a preliminary study over direct predictive modeling algorithms has improved understanding, as it was not only to predict road class, but also to understand how and the dynamics associated with it. Our exploration of the co-occurrence of road classes and their corresponding POIs presents promising insights into specific local contexts. This suggests the potential of predicting road types based on adjacent POIs within a given country. However, a generalized approach remains ambiguous and warrants further investigation. This study introduces a novel framework for establishing road class profiles or signatures through adjacent POIs, which can aid in predicting road class types. This approach has the potential to enrich geographic road data generated by AI models by incorporating attribute information.

The implications of these findings extend to suggest candidate POIs for road class modeling based on their geographic locations. A deeper understanding of the associations will create an established signature that can then be used to guide mappers by suggesting road classes in scenarios where POIs are mapped before the roads themselves, as we see in some places in Guatemala, thereby ensuring simultaneous quality checks as the data are updated by contributors.

Regardless, this is not free from limitations; using osm data to predict osm data is somewhat questionable because the quality is not certain. In addition, as the data continues to be updated, the association between the roads and the POIs may change; hence, this method may not be appropriate for areas where the mapping is incomplete (current work is investigating a threshold for completeness), and hence it might not work for all areas.

References

[1] Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLOS ONE*, *12*, 1–20. https://doi.org/10.1371/journal.pone.0180698

[2] Boeing, G. (2022). Street Network Models and Indicators for Every Urban Area in the World. *Geographical Analysis*, *54*, 519–535. https://doi.org/10.1111/gean.12281

[3] Guth, J., Keller, S., Hinz, S., & Winter, S. (2020). Towards detecting, characterizing, and rating of road class errors in crowd-sourced road network databases. *Journal of Spatial Information Science*, 22, 1-31. https://doi.org/10.5311/JOSIS.2021.22.677

[4] Maryann, P. F. (2014). The character of innovative places: entrepreneurial strategy, economic development, and prosperity. *Small Business Economics*, *43*(1), 9–20. https://doi.org/10.1007/s11187-014-9574-4