Investigating the digital divide in OpenStreetMap: spatio-temporal analysis of inequalities in global urban building completeness

Benjamin Herfort (benjamin.herfort@heigit.org)
HeiGIT at Heidelberg University https://orcid.org/0000-0001-9738-4060
Sven Lautenbach
HeiGIT at Heidelberg University https://orcid.org/0000-0003-1825-9996
João Porto de Albuquerque
University of Glasgow https://orcid.org/0000-0002-3160-3168
Jennings Anderson
Meta Platforms Inc.
Alexander Zipf
Heidelberg University

Article

Keywords:

Posted Date: August 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1913150/v1

Investigating the digital divide in OpenStreetMap: spatio-temporal analysis of inequalities in global urban building completeness

Benjamin Herfort[1,2,*], Sven Lautenbach[1], João Porto de Albuquerque[3], Jennings Anderson[4], and Alexander Zipf[1,2]

[1]Heidelberg Institute for Geoinformation Technology, 69120 Heidelberg, Germany
[2]GIScience Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany
[3]Urban Big Data Centre, University of Glasgow, United Kingdom
[4]Meta Platforms Inc.
[*]benjamin.herfort@heigit.org

ABSTRACT

OpenStreetMap (OSM) has evolved as a popular geospatial dataset for global studies, such as monitoring progress towards the Sustainable Development Goals (SDGs). However, many global applications turn a blind eye on its uneven spatial coverage. We utilized a regression model to infer OSM building completeness within 13,189 urban agglomerations home to 50% of the global population. Our results reveal that for 1,510 cities OSM building footprint data exceeds 80% completeness. Humanitarian mapping efforts have significantly improved completeness, especially in low SHDI regions. The digital divide in OSM has receded, but a strong spatial bias associated with subnational human development index, city size and World Bank region remains. In consequence, global studies will provide misleading results if the biases in OSM's coverage are not accounted for. We recommend combining completeness maps with socio-demographic information to guide mapping to ensure that "nobody is left behind" as encouraged by the SDGs.

Introduction

Cities across the world are generally growing faster in land area than in population size[1]. Between 2001 and 2018 urban population growth and built-up area expansion accelerated especially in large cities in the low-income and lower-middle-income countries[2]. Projections show that global urban land expansion is expected to experience rapid growth in the upcoming 20 years[3]. Buildings constitute one of the most important physical elements of settlements[4]. However, little is known on a consistent basis about building inventories worldwide; a spatially detailed survey of the distribution and concentration of the building stock does not yet exist[5]. Improving the systematic monitoring of the global urbanization process is a requirement for achieving the United Nation's Sustainable Development Goals (SDGs), e.g. "urban" SDG 11, especially in the low-income countries where the data are usually scarce[2].

There is an estimated gap of $1 billion USD globally in funding for national statistical offices, and consequently, baseline geospatial data that should be provided by these agencies are often not accessible, not up-to-date or not available in standard formats[6,7]. Tackling data scarcity requires moving beyond insufficient traditional data sources to utilizing new, non-traditional sources for measuring the SDGs[8,9]. It has been shown that open data communities — such as OpenStreetMap (OSM) — are not only promising, but already contribute to filling existing data gaps[8,10,11]. OSM is now used widely for applications such as web maps and navigation services at and data about buildings from OSM has been used in domains such as urban planning[12], SDG monitoring[13], disaster management[11,14], public health[15–17], as well as during the COVID-19 pandemic[18]. However, particular attention needs to be paid to data quality, when OSM data is utilized in global studies or to derive global data products, e.g. to derive a "global" dataset on critical infrastructure[19], or when using "big data" in comparing urban morphology across the globe[20]. When unaccounted for, spatial bias can lead analysts and researchers to draw general conclusions which are only valid for well-represented (well-mapped) areas[21]. By inadvertently neglecting less well mapped areas in their analyses and datasets, they are in danger of counteracting the overarching goal of the SDGs ensuring that "nobody is left behind".

In addition to contributions by individual volunteers (mappers), there is an intensifying trend that organized corporate and humanitarian mapping communities contribute to OSM in general[10,22]. As of 2022 OSM is no longer an exclusive community of amateurs, but instead a community built of multiple smaller hobbyist, professional, amateur, and experienced mapping communities, in which professional stakeholders are sharply gaining influence on map data creation.

Describing the elements of spatial data quality is imperative to provide stakeholders with the necessary information to evaluate fitness for use of a dataset for their particular application[23]. Better spatial data quality assessment would promote the adoption and (right) usage of new sources of data such as OSM and data products based on OSM. A large community of researchers has analyzed the quality of OSM data[24,25]. It has been acknowledged that OSM data in general is strongly biased, in part due to a much larger contributor basis in countries in the global North as a consequence of socio-economic inequalities and the digital divide[26,27]. While studies based on authoritative reference data are able to provide detailed insights on urban OSM building completeness for selected cities[4,28], they cannot transfer to other regions for which reference datasets are either missing or unavailable in a format compatible with analysis. To overcome dependencies on sparsely available administrative datasets the use of globally available proxy data such as remote sensing data (e.g. Nightime Light, built-up-area, Sentinel 2 derived spectral indices)[29] or population data[30] have been suggested to assess and predict OSM building completeness. Albeit the manifold usage of OSM building footprints, an adequate investigation into their completeness on the global scale has not been conducted so far.

This paper overcomes this gap by investigating OSM building completeness in regions home to a population of 3.5 billion people (about 50% of the global population). Our spatio-temporal analysis pursues the following two research questions:

1. Is OpenStreetMap building data good enough for urban analysis and SDG monitoring?

2. How does this quality vary within the space of one city, between continents and on the global level?

First, we propose a machine learning regression method based on a random forest to assess OSM building completeness within 13,189 urban centers (as defined by the European Commission[31]). We utilize an extensive collection of open building data from commercial and authoritative sources as training data and utilize OSM full-history data for spatio-temporal data analysis on the global scale[32]. The model further relies on information obtained from remote sensing data (land cover, population distribution, night time lights), subnational human development, and urban road network density as predictors. Second, this paper builds upon the extensive methodological skill set developed to investigate urban segregation and transfers it towards analysing geographic inequalities within OSM. This allowed us – for the first time – to present a comprehensive assessment of the evolution of urban OSM building completeness which encompasses all data contributed to OSM since 2008.

Results

Urban OSM Building Completeness

Our analysis found a total of 183 Million buildings in OSM and a global average urban OSM building completeness of 21% per urban center (see Table 1). Relatively high completeness was estimated for Europe & Central Asia (67%) as well as for North America (56%). Completeness values lower than the global average were observed for the regions Latin America & Caribbean (17%), East Asia & Pacific (16%), Middle East & North Africa (11%), and South Asia (7%). The completeness value for East Asia & Pacific was strongly influenced by the fact that urban centers in China were hardly mapped, very likely because mapping in OSM is prohibited by law. Sub-Saharan Africa completeness (29%) was slightly higher than the global mean. These regional differences illustrate that the global average is of limited explanatory power.

We found that organized humanitarian mapping activities in urban centers contributed an average of about 8% of the building footprints globally. However, humanitarian contributions were focused on specific regions, especially in Africa where about 43% of all building edits in Sub-Sahara Africa were related to organized humanitarian mapping activities. Overall, organized humanitarian mapping activities were expectedly associated with lower subnational human development index values, in line with previous findings[10]. We generally found corporate mapping activity to constitute less than 2% of all building edits globally (and only about 0.1% in urban centers), a significant difference in participation from corporate mappers editing nearly 20% of the global road network as previously found[22].

Distinguishing urban centers by SHDI also revealed dramatic differences in the temporal trajectories of completeness (see Figure 1 (b)). In general, urban centers in regions with very high SHDI had the highest levels of mapped building completeness. Surprisingly, however, there was no positive correlation between SHDI and completeness. The completeness in low SHDI urban centers was higher than the completeness of urban centers with high SHDI. Our results suggest that this was due to the positive impact of organized humanitarian mapping activities since 2015, especially on urban centers located in low and medium SHDI regions (see Table 1).

The size of the urban centers measured by population was positively correlated to completeness (see Table 1), albeit the differences were not as pronounced as for World Bank regions or SHDI classes. OSM building data in large metropolitan areas were considerably more complete compared to small urban areas. However, the temporal evolution of urban building completeness showed very similar patterns for urban centers regardless of their population (figure provided in online material only).

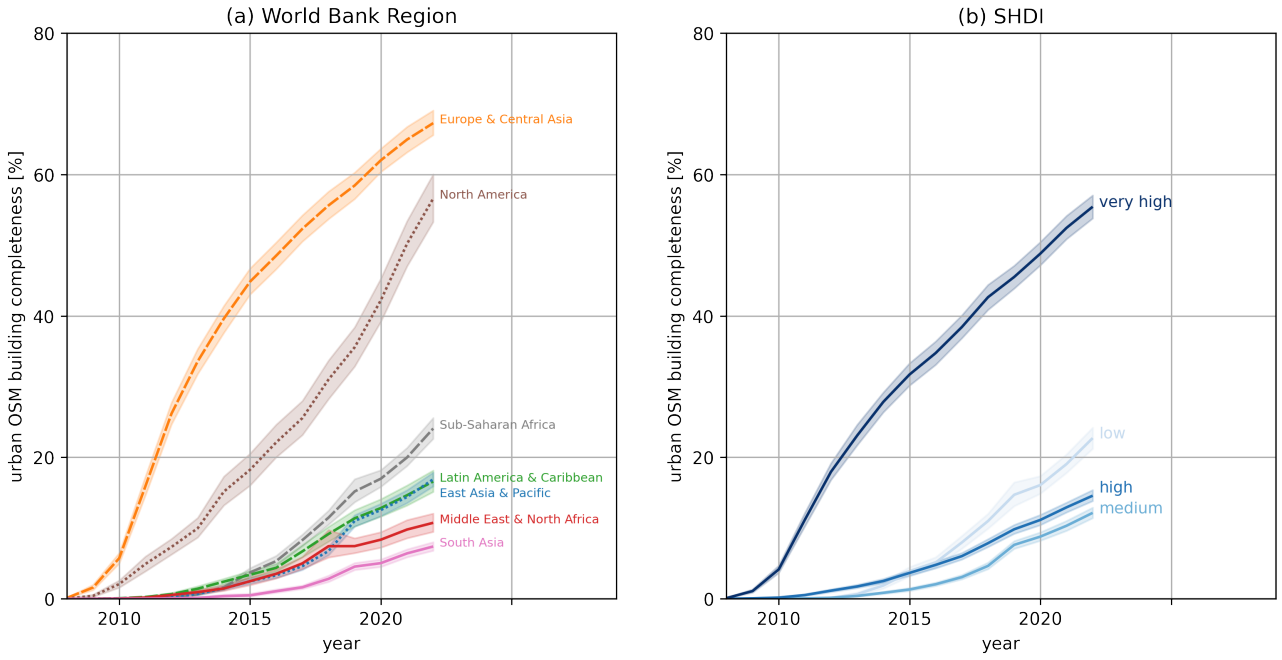

Figure 1. Temporal evolution of average urban OSM building completeness by (a) World Bank regions and (b) Subnational Human Development Index (SHDI) group. Completeness was derived by aggregating building area predictions based on a Random Forests model and monthly OSM building area per urban center. The shaded areas represent the 95% confidence interval for each line. OSM data from 2008-01-01 to 2022-01-01. Created using Matplotlib 3.3. in Python 3.7.5 (https://www.python.org/).

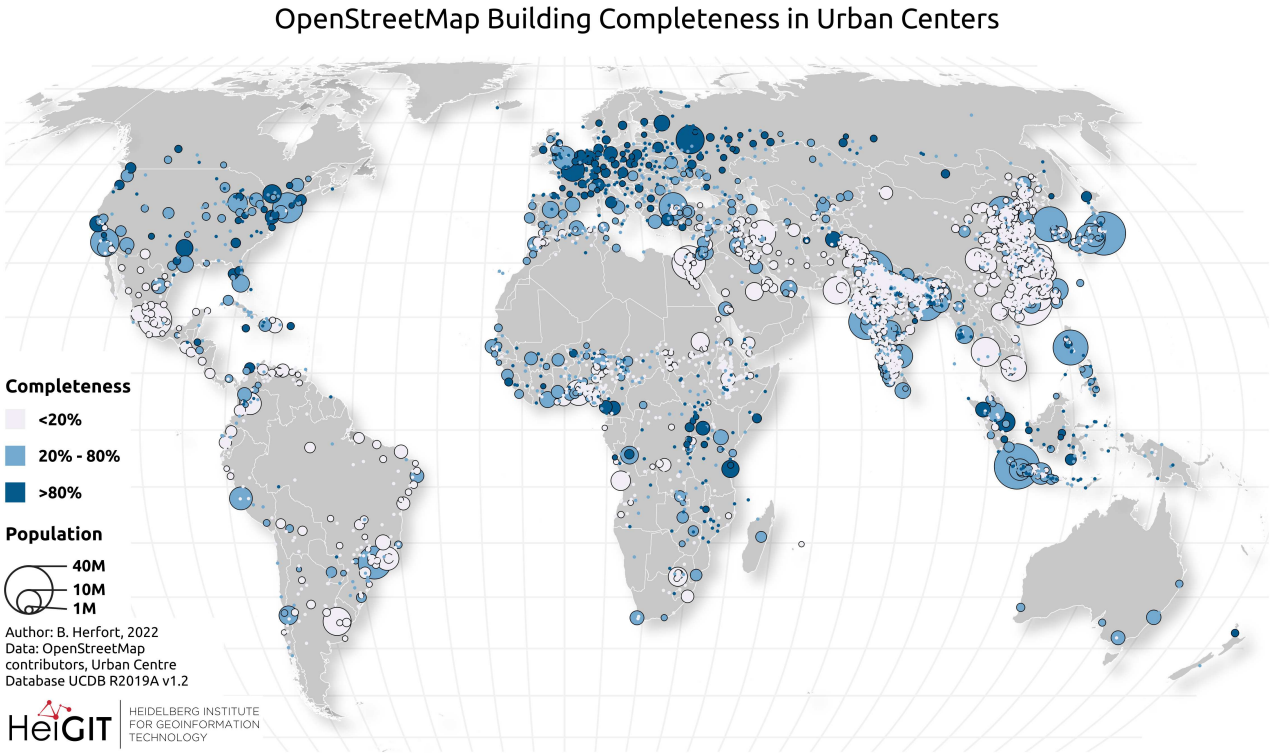

Figure 2. Spatial distribution of OSM building completeness in 13,189 urban centers. OSM data as of 2022-01-01. Created using QGIS 3.22.3 (https://www.qgis.org/en/site/).

Table 1. OSM building completeness in urban centers on the global scale and grouped by World Bank regions, Subnational Human Development Index class and city size class measured by population. Completeness was computed as the average of the individual OSM building completeness values per urban center. Humanitarian mapping and corporate mapping were quantified by their share on the overall map data.

	n	Completeness [%]	Humanitarian Mapping [%]	Corporate Mapping [%]
Global	13,189	21	7.6	0.1
World Bank Regions				
East Asia & Pacific	3,015	16	11.2	0.2
Europe & Central Asia	1,329	67	1.4	0.1
Latin America & Caribbean	1,062	17	11.9	0.4
Middle East & North Africa	893	11	7.4	0.2
North America	368	56	0.4	<0.1
South Asia	3,987	7	17.9	<0.1
Sub-Saharan Africa	2,414	24	42.7	0.1
Subnational Human Development Index				
Low	2,289	23	46.8	0.1
Medium	4,960	12	23.3	0.3
High	3,883	15	12.7	0.2
Very High	1,967	55	1.2	<0.1
City Size by Population				
Small Urban Areas	10,271	20	6.2	0.1
Medium-Size Urban Areas	1,922	25	6.9	<0.1
Metropolitan Areas	687	30	7.6	0.1
Large Metropolitan Areas	309	36	8.6	0.1

The spatial distribution of building completeness across urban centers shows a strong regional variability across that global trend: numerous cities in any region were mapped with a very high completeness regardless the overall completeness or mapping activity in that region. Our results reveal that for 1,510 cities OSM building footprint data exceeded 80% completeness. For instance, within Africa, we found urban centers in Egypt and Ethiopia with particularly low OSM building completeness, whereas cities in Tanzania, Uganda and western African countries achieved much higher completeness. Similarly, building completeness values in Indonesia and the Philippines were notably higher than for other countries of Southeast Asia. In contrast, most urban centers in India and China were hardly mapped with regard to building footprints. Strikingly, the spatial distribution of OSM building completeness for urban centers was characterized by spatially clustered patterns at various scales. This highlights the complexity of OSM mapping activities and the challenge for users of OSM derived data.

The uneven building completeness between urban centers was also indicated by a global Gini coefficient of 0.8. This characteristic was observed across all regions and was most pronounced in South Asia and Sub-Saharan Africa (c.f. Figure 3 a). Slightly higher evenness compared to the other regions was detected for urban centers across North America and Latin America & Caribbean. The temporal evolution of the Gini coefficient indicated that both globally and regionally, OSM building data distribution has become slightly more even over time. This suggests that continuous mapping activity trends towards reducing overall segregation in OSM building data.

As a non-spatial measure, however, the Gini coefficient cannot reveal the spatial arrangement of these well-mapped and unmapped urban centers. Therefore, spatial clustering of building completeness levels was analyzed based on Moran's I[33] as a measure of global spatial autocorrelation. We find that global spatial inequality in OSM building completeness sharply increased between 2008 and 2014 (c.f. Figure 3 b). During that time—although overall OSM building completeness became more even (as measured by the Gini coefficient)—mappers favoured cities near already well-mapped cities. We also find that until 2014, the expansion of OSM mapping to distant and un-mapped regions (likely to be located in the Global South) did not happen at a significant scale.

Nevertheless, since 2014, Moran's I as a measure of global spatial autocorrelation declined from 0.71 till 0.56 as of 2022. This indicated that the spatially clustered completeness pattern became less intense, albeit still clearly visible. Combined with the decrease of the Gini coefficient in the same period, our results suggests that OSM building completeness became more even

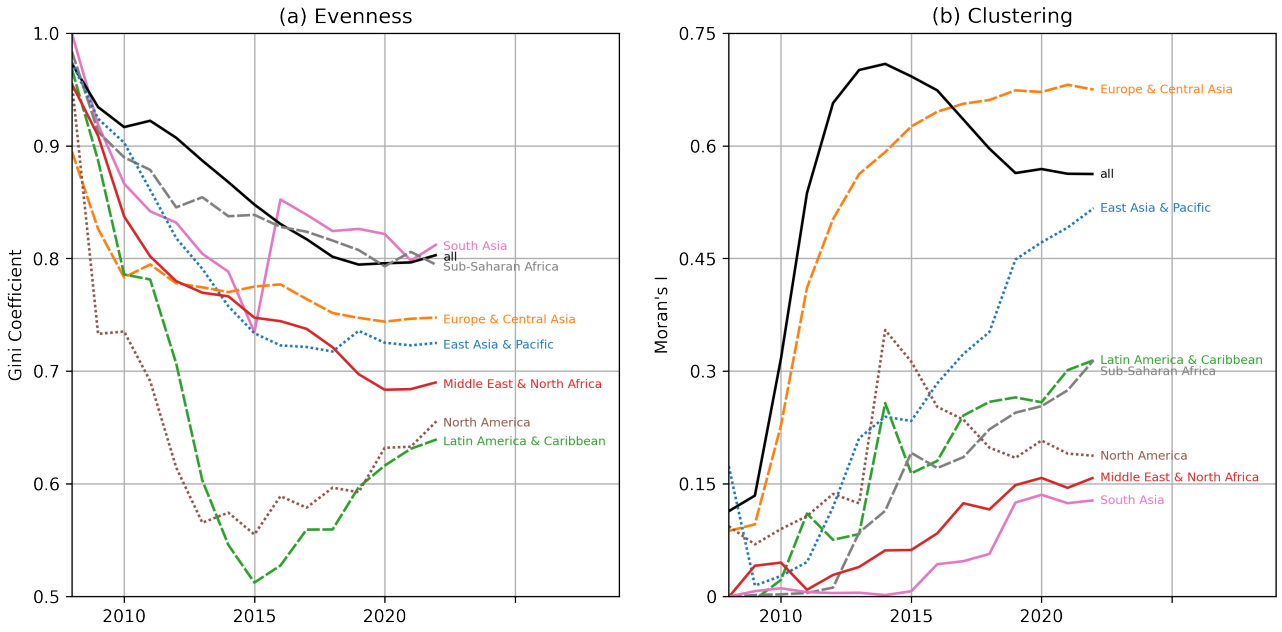

Figure 3. Non-spatial and spatial inequality measures of completeness. Temporal evolution of (a) evenness and (d) clustering of urban OSM building completeness per World Bank region. Moran's I measures spatial autocorrelation, positive values indicate spatial clustering. Values for Moran's I in practice often range between -0.5 and 1.15 with zero indicating absence of global spatial autocorrelation. OSM data from 2008-01-01 to 2022-01-01. Created using Matplotlib 3.3. in Python 3.7.5 (https://www.python.org/).

as mapping activity expanded to regions which previously saw less mapping activity. In that regard, OSM building data in 2022 was much less segregated in both evenness and clustering compared to the state-of-the-map in 2014.

We also found spatial autocorrelation within regions to increase steadily over time regardless of overall map completeness. Europe & Central Asia reveal a moderate spatial clustering (Moran's I: 0.24) in 2010, but a very high spatial clustering (Moran's I: 0.7) in 2022 (see Figure 4 (a) and (b)). Although this region was previously mapped better than any other region, it also constituted the region where spatial inequality in completeness was most pronounced. In 2022, urban centers with high completeness in Europe & Central Asia were surrounded by other urban centers with high completeness and this effect was much stronger than in 2014.

An analogous process was observed in all other regions, e.g. as shown for Sub-Saharan Africa in Figure 4 (c) and (d). Within most regions, we observed steadily increasing spatial autocorrelation over the past years (c.f. figure 3). The only exception to this is North America, where spatial clustering of building completeness has decreased since 2014. It seems that the expansion of OSM mapping activity could be interpreted nearly everywhere as an spatial expansion of an existing mapping community to nearby regions.

With these findings, we draw the following conclusions about the spatial pattern observed in Figure 2: Besides the well-examined Global North - Global South bias in OSM, our results demonstrate that the OSM building stock as of 2022 showed a much more spatially diverse spread pattern across several scales, which was not considered previously. The trends observed in the temporal trajectories suggest that this multi-scale spatially clustered characteristic of inequality in OSM building data completeness is likely to continue to grow.

Intra-Urban OSM Building Completeness

Having thus far focused on completeness between urban centers, we now investigate how segregated OSM building completeness has evolved within urban centers. Accordingly, this section builds upon the intra-urban spatial heterogeneity of OSM building completeness estimated at a resolution of 1x1 kilometers. Here we will refer to each 1x1 km section of the map as a "grid cell". These grid cells were used to calculate the Gini coefficient and Moran's I for individual urban centers. Based on both those indicators and OSM building completeness, urban centers were classified into five different types utilizing an agglomerative clustering approach (c.f. Figure 5 and c.f. Figure 6).

Urban centers of type (a) showed low completeness combined with a high Gini coefficient and low Moran's I. Urban centers

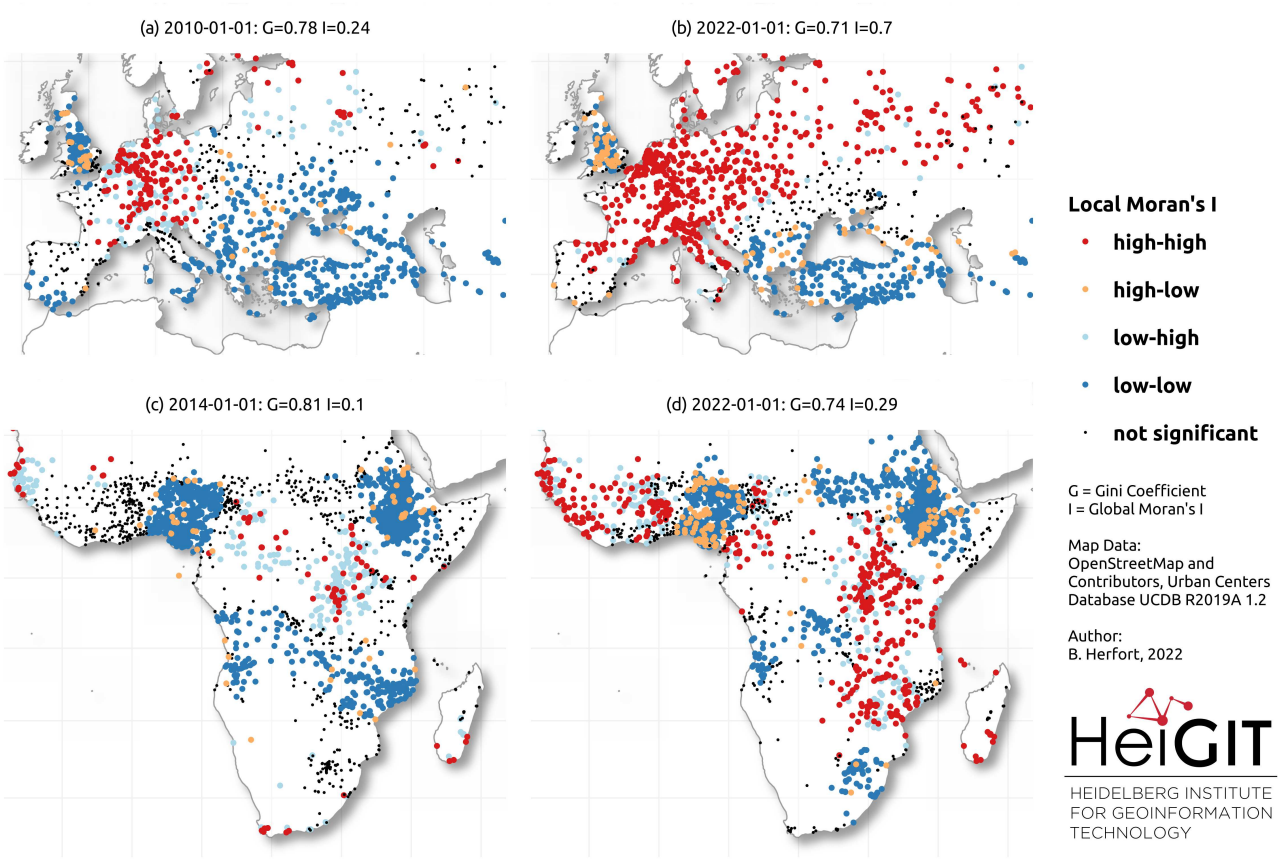

Figure 4. A comparison of local spatial autocorrelation of completeness at two points in time for urban centers within (a)-(b) Europe & Central Asia and (c)-(d) Sub-Saharan Africa. Each urban center was classified according to whether its building completeness value was above (high) or below (low) the global mean and if the weighted mean across its neighbors was above or below the global mean. Based on this four quadrants are defined: high-high (HH), low-high (LH), low-low (LL) and high-low (HL). High-high and low-low describe clusters of high and clusters of low values while low-high and high-low indicate spatial outliers in the sense that the completeness value of the urban area was unexpected in their neighborhood. Significance levels were adjusted for multiple testing. For each region and point in time we provide the Gini coefficient (G) and Moran's I for the region shown in the sub-plot. Created using QGIS 3.22.3 (https://www.qgis.org/en/site/).

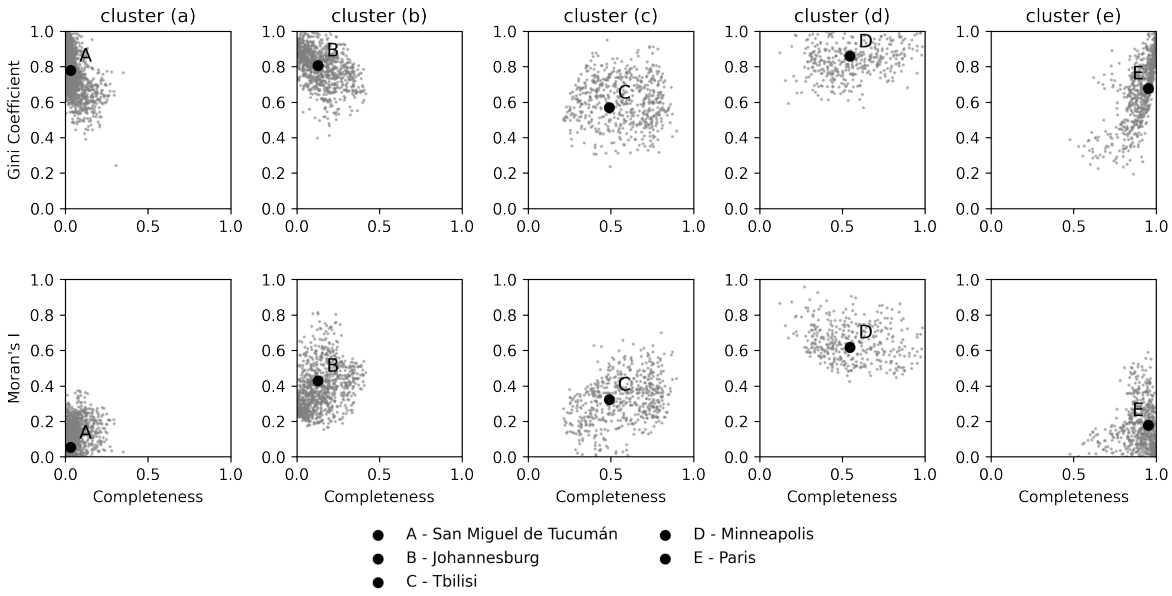

Figure 5. Agglomerative clustering of urban centers based on OSM building completeness, Gini coefficient *G* and Moran's *I*. Each point represents a single urban center with a minimum area of 25 square kilometers. Smaller urban centers were ignored as Gini coefficient and Moran's I could not be reliably estimated. For each of the clusters (a)-(e) a single representative example A-E was selected out of the 4,722 urban centers considered in this analysis. OSM data as of 2022-01-01. Created using Matplotlib 3.3. in Python 3.7.5 (https://www.python.org/).

of this type could appear as white spots on the map, more aptly described as the "unmapped" cities. There was no particular spatial pattern defining where the small number of eventually mapped buildings would be located within the city. Among 1,469 urban centers, the city of San Miguel the Tucumán, Argentina is shown in Figure 6 (a) as an example of this category.

Type (b) urban centers contain more pronounced spatial clustering and slightly higher evenness. Common attributes for these cities are low to medium overall completeness and slightly lower unevenness or stronger spatial clustering. Albeit these cities could be considered hardly mapped in general, there were often a few grid cells which have been mapped with a much higher completeness in regard to buildings. Such mapped grid cells were not distributed randomly, but tended to cluster spatially. The urban agglomeration of Johannesburg, South Africa (see Figure 6 (b)) exemplifies that there were several distinct mapping hot spots surrounded by a larger number of unmapped grid cells. As depicted by the cluster dendrogram in Figure 6 urban centers for the two clusters (a) and (b) were more similar to each other and clearly distinct from urban centers in the other three clusters.

A smaller number of urban centers are characterized by relatively low Moran's I and moderate to high evenness. These urban centers of type (c) such as Tbilisi, Georgia (see Figure 6) could be considered as an in-between case with overall completeness around 50% or slightly higher and mapping spread all over the city with a lower tendency for spatial clustering. As a consequence, the Gini coefficient was lower in these cities, indicating that overall completeness better represented local completeness compared to the other types. This might indicate that a local community already formed which adds data to OSM for the entire city.

A highly segregated spatial distribution of OSM building completeness combined with medium to high completeness and medium to high Gini coefficient was characteristic for urban centers of type (d). For these cities, there exists large blocks of completely mapped grid cells adjacent to large blocks that were not mapped at all. Minneapolis, USA is representative of this type: a divided city from the perspective of mapped building in OSM. Minneapolis depicts a segregated distribution that is uneven and strongly spatially autocorrelated . For urban centers of type (d), the overall completeness value hardly reflected the local completeness values.

Finally, urban centers with the highest overall completeness and very low spatial clustering were most likely to get classified as type (e). As shown in Paris, France (see Figure 6 (e)), almost all parts of the city may be considered completely mapped. Only a few grid cells remained unmapped and these are often not strongly spatially clustered. As opposed to type (a), the urban

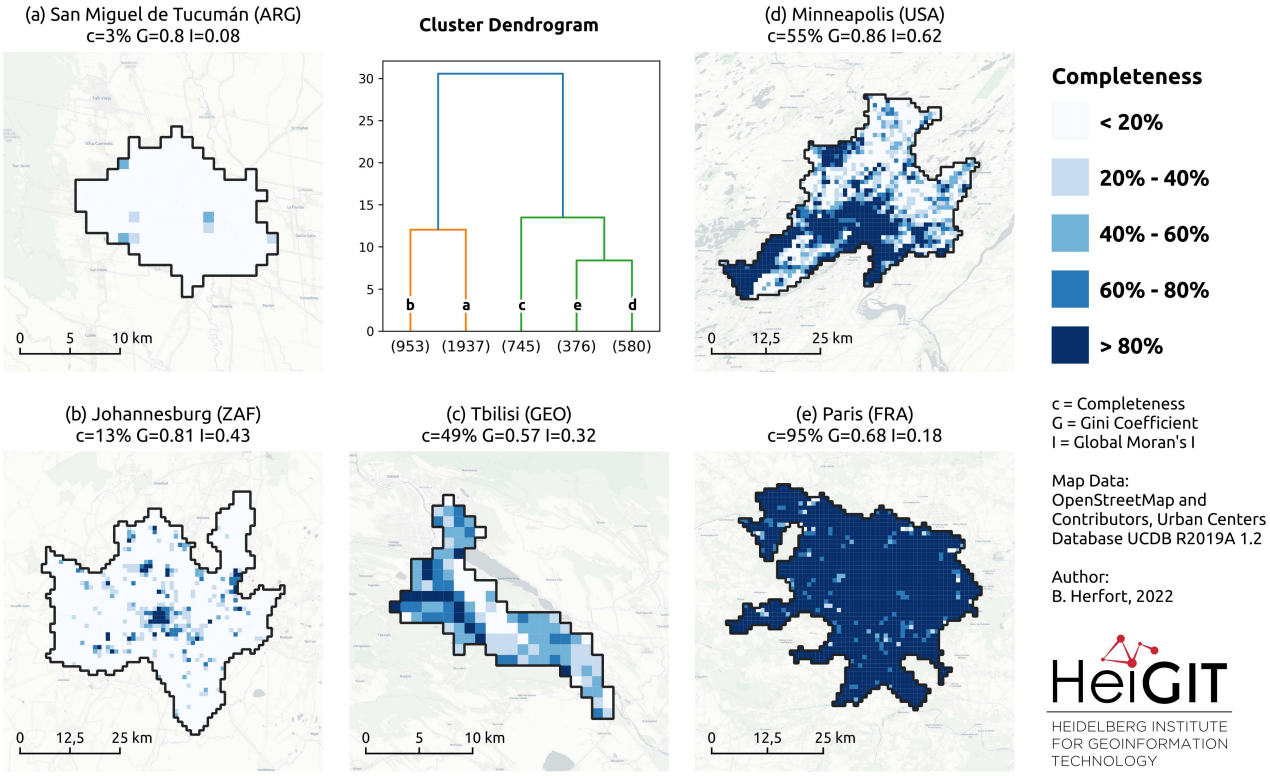

Figure 6. Spatial distribution of intra-urban OSM building completeness for selected urban centers. For each urban center we report on overall OSM completeness c , Gini coefficient G and Moran's I. Cell size is always 1x1 kilometers for any urban center. The clusters (a-e) are the same as in figure 5. The number of urban centers in each cluster is indicated in the dendrogram. OSM data as of 2022-01-01. Created using QGIS 3.22.3 (https://www.qgis.org/en/site/) and Matplotlib 3.3. in Python 3.7.5 (https://www.python.org/).

centers of type (e) could be described as the "well-mapped" cities and represent some finality of mapping building footprints in OSM.

Discussion

Mapping efforts of communities in OSM over the previous decade have made OSM a unique global database of building footprint data, which is accessible to all with no licensing costs. Our results reveal that for 1,510 cities home to a population of more than 400 Million people, OSM building footprint data is more than 80% complete and can provide an alternative to otherwise complex approaches utilized to derive authoritative and/or automated building datasets. We further showed that humanitarian mapping efforts have significantly contributed to improve OSM's completeness. Especially in low SHDI regions this type of mapping accounted for almost 50% of all map edits. This confirms the potential of using OSM for humanitarian activities e.g. for disaster risk reduction[11] or monitoring progress towards the SDG's[13]. OSM enables diverse communities (amateurs, humanitarian organisations, corporations) to contribute information across the globe and provides baseline geospatial data that is often not available from other authoritative sources[6,7,34]. In light of the critical challenges to finance data systems for SDG monitoring in both low and middle income countries—despite heightened demand[34]—the creation and usage of OSM data should be promoted further as a cheap and good alternative which in addition also allows local communities to be directly involved in the data production process.

The digital divide in OSM has receded over the past decade, but still exists. As such, OSM data completeness improved, but was still strongly biased by regional, socio-economic and demographic factors on several scales. This echoes the highly uneven geographies of participation observed in Wikipedia[35] and stands in contrast to the relatively higher and more evenly distributed completeness for OSM's road network[36]. If this trend continues, OSM will become more complete, but will still not evolve towards a truly global inclusive map. As a consequence, global studies and global frameworks (such as SDGs) which use OSM data will draw wrong conclusions and will provide misleading recommendations for decision makers when the biases in OSM's coverage are not accounted for.

Our findings can help practitioners and researchers to pick the best strategies to cope with OSM's uneven spatial coverage. Because spatial clustering of completeness is inherent to OSM in general, most analyses and modelling approaches will require methods which can incorporate geography into the analysis, e.g. by dealing with spatial autocorrelation[21]. For global studies utilizing OSM data, such as when modelling human population distributions[37], spatial sampling design should account for bias at continental scales (global north vs. global south), regional (clusters of high/low community activity) and intra-urban variations. When using OSM data for assessing urban resiliency and risk, emergency planning[14], authors should utilize extrinsic and intrinsic quality assessment methods or proxy based approaches to identify regions where OSM data coverage is very poor. Some analyses are prone to spatial clustering of completeness, such as sampling protocols utilizing OSM building footprints to randomly select households[15,17]. Uneven spatial coverage might lead to entire communities being neglected because they are simply not visible on OSM. In data scarce environments, often encountered when working in informal settlements, OSM data is sometimes the only available data source[13]. In such situations, explicit spatial validation procedures should be employed, e.g. based on expert review or making use of other crowdsourcing applications such as LACO-Wiki or MapSwipe[8].

Our analysis comes with several unavoidable limitations that need to be considered to put our findings into context. A major limitation is that our analysis only investigates buildings mapped in OSM within urban centers. Whereas the study encompasses about 50% of the global population, one should be careful to transfer our findings to rural areas. Researchers have shown that there is a tendency of OSM data to be of higher quality in cities[4].

Machine learning models, such as the random forest model used here to derive building area predictions, are suspect to potential biases present in the data used or biases that arise from the algorithms[38]. For the authoritative data utilized to train the model, a high geometric accuracy was assumed, however, these datasets might be outdated depending on the publisher's update cycle. We showed that quality of building footprint data from Microsoft (c.f. Table 3) can be prone to low recall values in some areas. These biases are also reflected in the results reported here and might lead to too low building area predictions and consequently too high building completeness estimates. To disclose general uncertainty, we reported on the model performance utilizing a spatial cross validation procedure (c.f. Table 5). Whereas the completeness estimation performed well with a global r^2 score of 0.84, slightly higher uncertainty was observed for East Asia & Pacific and South Asia. This should be critically assessed as another potential limitation of our study especially as these two regions account for more than half of all urban centers globally. However, this also constitutes a starting point for local communities and researchers to design local completeness models which can overcome the limits of the global modelling approach utilized in this study, albeit these local models might not be easily transferred to other regions.

Geospatial data quality is comprised of dimensions beyond measuring completeness[23]. For some sectors, such as public health programs, assessment of completeness is only the first step, and information on building usage is also required, but often only available for a small subset[16]. Future work should further investigate the potential of a harmonious ensemble dataset that combines the best of OpenStreetMap buildings with additional building coverage from deep learning based datasets such

from Microsoft Buildings[39]. For some places, the building footprints in OSM might already come from a derived dataset, having been accepted by a human-in-the-loop import process via the mapwith.ai editor[40] or similar tools. It should be critically assessed if and how novel data products such as Meta's Daylight Distribution[41] of OpenStreetMap, which also provides a subset of the Microsoft building footprints that do not overlap with the buildings already present in OSM, can empower local communities to fill the existing gaps in data coverage identified in this paper.

Finally, we want to to emphasize that maps and geospatial databases such as OSM are products of social interactions and represent the specific cultural perspectives of the mappers. It has been widely discussed that there is no unquestionable "objective" form of spatial knowledge creation[42]. However, the cultural openness and social nature of OSM could also be considered a strength especially when comparing to building footprint datasets derived using proprietary, black-box machine learning approaches for which bias and fairness are often still unknown[38].

In conclusion, we want to highlight two important recommendations for OSM data users and producers to promote sustained impact of future mapping efforts.

First: To assess the potential negative impact of missing data, as a OSM data user you should investigate if your study is subject to spatial bias caused by OSM's uneven spatial coverage at multiple scales. In a subsequent step, you should account for spatial bias explicitly, either by applying appropriate sampling schemes or by clearly stressing for which regions the analysis results may not be reliable due to low or unknown completeness.

Second: As a OSM data producer, you should use completeness maps to decide where future mapping activities should take place to ensure that "nobody is left behind" as encouraged by the SDGs. By combining the completeness maps with socio-demographic characteristics of the areas of interest you can ensure a fair and balanced selection of project regions to reduce existing inequalities within OSM.

The OSM community started its journey in the early 2000's in western European cities and this history was still clearly visible in the unbalanced spatial distribution of map data in 2022. Nevertheless, there are numerous successful examples of local mapping communities that overcome structural barriers which exclude others from participating in OSM. We believe that by empowering these communities, OSM will further evolve into the most comprehensive open geographic data base which is needed to help achieving the SDGs.

Methods

Building Data

The analysis was carried out for 13,189 urban centers on a global scale. To delineate our study areas we used the Global Human Settlement Layer Urban Centred Database (GHS-UCBD) which has been developed by the European Commission[31]. Accordingly urban centers have been characterized as "high-density clusters of contiguous grid cells of 1 km^2 with a density of at least 1500 inhabitants per km^2 and a minimum population of 50000"[31]. Each urban center was spatially disaggregated using a 1x1 kilometer grid based on the equal-area Mollweide projection. For each of the resulting 665,641 grid cells we aggregated both the reference data sets (if available) and the datasets utilized as predictors in the model.

For each grid cell we derived the overall OSM building footprint area in square kilometers using the ohsome API which relies on the OSHDB framework for spatio-temporal analysis of OSM history data[32]. We included buildings which have been mapped in OSM as of 2022-01-01. We considered all OSM objects that have been tagged with 'building=*' and were of the geometry type 'Polygon'.

As no single reference building data set on the global scale exist, we combined a set of external datasets (c.f. Table 2) which have been obtained either from authoritative or commercial sources building upon the great work by Biljecki et al. (2021)[43]. In total these reference datasets covered 6,737 urban centers (417,253 grid cells) among 140 countries. For some regions data was not available for the entire country, but only selected cities. The data obtained from Microsoft was derived using a deep learning based building detection approach[39]. Microsoft reports precision and recall for subsets of countries (c.f. Table 3). As we only used building footprints for the urban centers and classification quality presumably differed between urban and rural areas, this provides only a rough guidance.

Explanatory Variables

As explanatory variables we used the following datasets: the Global Human Settlement Layer Population (GHS-POP) is provided by the European Commission[44] and based on a disaggregation of CIESIN's Gridded Population of the World (GPWv4.10). To characterize regions based on their socio-economic status we utilized the Subnational Human Development Database[45]. In our analysis we relied on the aggregated SHDI and did not further consider its individual components (education, standard of living, health). Information on night-time lights was obtained as the annual average of 2020 and aggregated by summing up all values per grid cell[46]. Land cover information at 10 meter resolution was utilized from the ESA WorldCover 2020 dataset which has been derived from Sentinel-1 and Sentinel-2 data[47]. For each grid cell we derived the overall area per land cover class in square kilometers.

Table 2. Reference datasets used for training the machine learning model.

Dataset Name	Countries	Urban Centers	Grid Cells
East Asia & Pacific			
Microsoft Building Footprints	Australia, Cambodia, China, Indonesia, Laos, Malaysia, Mongolia, Myanmar, Philippines, Singapore, Thailand, Vietnam	825	58,707
GSI Basic Map Information Buildings	Japan	108	22,342
NSDI Continuous Numerical Topographic Map Building Data	South Korea	37	5,258
LDS NZ Building Outlines	New Zealand	8	941
Europe & Central Asia			
Microsoft Building Footprints	Albania, Armenia, Austria, Azerbaijan, Belarus, Bosnia and Herz., Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Georgia, Greece, Hungary, Iceland, Italy, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Macedonia, Moldova, Montenegro, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Tajikistan, Turkey, Ukraine, Uzbekistan	823	50,968
OS OpenMap Local	United Kingdom	135	10,845
IGN BD TOPO Bâtiments	France	71	6,292
GUGiK BDOT10k Budynki	Poland	48	3,221
NGR Basisregistraties Adressen en Gebouwen (BAG)	Netherlands	38	2,687
Hausumringe Nordrhein Westfalen, Sachsen, Berlin	Germany[a]	24	4,203
CUZK Budovy (BU)	Czechia	12	717
Estonian Land Board Buildings	Estonia	2	130
Latin America & Caribbean			
Microsoft Building Footprints	Argentina, Bahamas, Barbados, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Curaçao, Dominican Rep., Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Suriname, Trinidad and Tobago, Venezuela	519	31,456
SP Mapa Digital da Cidade	Brazil[a]	3	887
Quito Gobierno Abierto Construcciones	Ecuador[a]	2	340
Construcción. Bogotá D.C	Colombia[a]	1	415
BA Data Tejido Urbano	Argentina[a]	1	187
Middle East & North Africa			
Microsoft Building Footprints	Algeria, Djibouti, Egypt, Iran, Iraq, Israel, Jordan, Lebanon, Libya, Malta, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Syria, Tunisia, United Arab Emirates, Yemen	672	29,880
North America			
Microsoft Building Footprints	Canada, United States	364	95,837
South Asia			
Microsoft Building Footprints	Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Sri Lanka	1,977	72,240
Sub-Saharan Africa			
Microsoft Building Footprints	Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Rep., Chad, Comoros, Congo, Côte d'Ivoire, Dem. Rep. Congo, Eq. Guinea, Eritrea, eSwatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, Somalia, South Africa, S. Sudan, Sudan, Tanzania, Togo, Uganda, Zambia, Zimbabwe	1,554	42,741

[a] Building data did not cover the entire country.

Table 3. Precision and recall for the Microsoft building footprints. Values might differ across the region.

Region	Percision [%]	Recall [%]
Country Level		
USA	98.5	92.4
Canada	98.7	72.3
South America	96.0	68.7
Uganda and Kenya	94.5	61.8
Nigeria	97.5	69.1
Kenya	95.3	69.0
Australia	98.6	65.0
Indonesia, Malaysia, Philippines	88.6	77.5
Region Level		
Africa	94.4	70.9
Caribbean	92.2	76.8
Central Asia	97.2	79.5
Europe	94.3	85.9
Middle East	95.7	85.4
South America	95.4	78.0
South Asia	94.8	76.7

From OSM we extracted the road network length in kilometers per grid cell for main roads. The main roads were selected using the ohsome API[32] using the following filter: 'highway in (primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, unclassified, residential)'. We included data which have been mapped in OSM as of 2022-01-01. We have investigated the spatial variations in the completeness of OSM road data by utilizing an intrinsic quality assessment approach following the 'mapping saturation' methodology proposed by Rehrl & Gröchenig (2016)[48]. Accordingly, the completeness of the road network was estimated for each urban center and aggregated by World Bank region (see Table 4). The intrinsic completeness measure revealed that for a small, but still decent share of the urban centers in South Asia (10.5%), Middle East & North Africa (10.2%) and Sub-Saharan Africa (5.9%) road network mapping could be considered only in the initial stage. For these urban centers it very likely that the majority of the road network was not (yet) completely mapped.

Table 4. Aggregated road network completeness estimation per World Bank region based on the mapping saturation approach. Mapping activity in urban centers was classified into a "start", "growth" and "saturation" phase. Saturation indicates that most of the mapping of the road network has been completed. During the growth phase the majority of features have been added.

Region	Start	Growth	Saturation
East Asia & Pacific	0.022	0.665	0.302
Europe & Central Asia	0.003	0.145	0.847
Latin America & Caribbean	0.004	0.238	0.752
Middle East & North Africa	0.102	0.504	0.384
North America	0.013	0.016	0.947
South Asia	0.105	0.553	0.334
Sub-Saharan Africa	0.059	0.478	0.460

Building Area Prediction Model

We used a Random Forest (RF) regressor[49] to predict the building area per 1x1 kilometer grid cell using the covariates described in the section above. There are various applications of Random Forest regressors for producing spatial predictions in general[50] and to estimate building completeness in particular (e.g. in Haiti, Dominica and St. Lucia[29]). RF constitutes a non-spatial approach to spatial prediction as sampling locations are ignored during the calculation of the model parameters[50]. Hence, in this study we initially also considered generalized additive models (GAMs) as an explicit spatially aware approach (if smooths of coordinates are included) which has been used for geospatial modelling e.g. in the domains of geomorpholgy[51], public

health[52] or for the analysis of social media data[53]. Nevertheless, our results revealed that RF outperformed the GAM approach and we decided to utilize the RF implementation in the Python package 'scikit-learn'[54].

To evaluate the performance of the proposed building area prediction we adopted a spatial cross validation approach based on k-means clustering. Especially for large scale mapping studies, such as this work, but also in the domain of ecological modelling data are almost always spatially autocorrelated and a spatially explicit assessment of machine learning models is required[55]. Due to spatial autocorrelation in the observations (data from nearby locations will not be independent) training samples and validation samples cannot be randomly selected as this would lead to overly optimistic error estimates[56]. Spatial blocking of samples, e.g. through k-means clustering, decreases this spatial dependence and provides more realistic performance scores[56,57]. Our spatial cross validation blocks were derived using a 20-fold k-means clustering based on scikit-learn's python implementation[54].

We investigated the performance of our model in respect to building area prediction for the 1x1 km grid and in respect to OSM building completeness prediction for the urban centers using the following indicators: r^2 score, explained variance, mean squared error and mean absolute error. To report global performance, we first estimated the model performance metrics for each of the seven regions (see Table 2). As each of these regions contained different numbers of samples, we computed the global scores from the weighted average of the regional scores using the total number of samples per region as the weight. Table 5 provides the performance scores. Overall the OSM building completeness model performed with a mean absolute error of 0.067 and achieved a r^2 score of 0.84 and explained variance of 0.85. In addition, we checked for spatial clustering in the residuals of the OSM building completeness prediction utilizing Moran's I as a measure of spatial autocorrelation[33]. The residuals were not distributed entirely random across space, but nevertheless only showed a slight tendency to cluster (Moran's I: 0.29).

Table 5. Global and regional model performance measures based upon 20-fold spatial cross validation.

region	n	r2	exp var	MSE	MAE
Building Area Prediction (1km Grid)					
Global	400,438	0.74	0.75	0.0025	0.034
East Asia & Pacific	74,731	0.74	0.74	0.0027	0.039
Europe & Central Asia	58,431	0.74	0.74	0.0017	0.027
Latin America & Caribbean	31,002	0.72	0.72	0.0051	0.049
Middle East & North Africa	29,524	0.79	0.80	0.0025	0.035
North America	97,041	0.70	0.70	0.0012	0.025
South Asia	110,508	0.81	0.82	0.0026	0.031
Sub-Saharan Africa	52,449	0.70	0.71	0.0022	0.034
OSM Building Completeness Prediction (Urban Centers)					
Global	6,370	0.84	0.85	0.015	0.067
East Asia & Pacific	773	0.81	0.83	0.026	0.101
Europe & Central Asia	797	0.84	0.84	0.012	0.079
Latin America & Caribbean	496	0.87	0.88	0.011	0.048
Middle East & North Africa	640	0.90	0.91	0.004	0.020
North America	364	0.94	0.94	0.006	0.055
South Asia	1,940	0.78	0.78	0.011	0.032
Sub-Saharan Africa	1,360	0.82	0.84	0.022	0.068

For all regions the distribution of raw residuals resembled a normal distribution (c.f. Figure 7). The histogram of raw residuals revealed that for samples located in Sub-Saharan Africa and North America the distribution was slightly skewed to the left and had a weak tendency to predict too low completeness values for urban centers.

Urban OSM Building Completeness

For each urban center we calculated the OSM building completeness using the area ratio method which has been applied by several other researchers in the context of urban areas[4,58]. First, we obtained the overall predicted building area by summing up the values for all grid cells per urban center. OSM building completeness per urban center was derived annually by computing the ratio of OSM building area versus predicted building area. We report on the average monthly OSM building completeness for urban centers globally and distinguished this score further by World Bank region, SHDI class and city size class by population. In addition, 95% confidence intervals have been calculated for each time series. SHDI classes were based on cut-off points defined by the United Nations Development Programme[59]: low human development (SHDI< 0.550), medium human

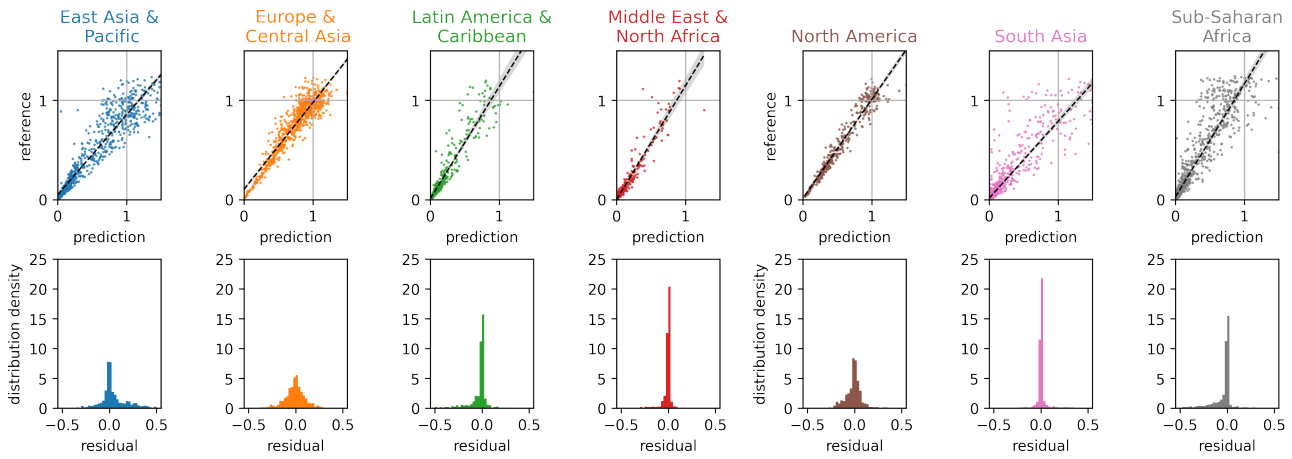

Figure 7. Scatterplot of raw residuals in regard to predicted building density and distribution density of raw residuals. Created using Matplotlib 3.3. in Python 3.7.5 (https://www.python.org/).

development (SHDI: 0.550 - 0.699), high human development (SHDI: 0.700–0.799), very high human development (SHDI> 0.800). City size classes were based on population thresholds defined by OECD[60]: small urban areas (50k–200k), medium-size urban areas (200k–500k), metropolitan areas (500k–1.5M), large metropolitan areas (>1.5M).

We investigated the impact of humanitarian mapping through the HOT Tasking Manager and corporate mapping by Apple, Meta, MapBox, Microsoft and Kaart on overall completeness and inequality measures. OSM contributions have been considered as humanitarian mapping activities following the approach developed by Herfort et al. (2021) which utilizes information obtained from a HOT Tasking Manager database dump[10]. Corporate mapping activities were identified by OSM user ID, expanding on the approach presented in[22] by using a mapper's self-disclosed corporate affiliation in their OSM user bio instead of relying on out-of-date lists on the OSM wiki[61]. Based on this information, we derived the share of humanitarian map edits and corporate map edits on the overall OSM building data.

Several measures have been adopted to describe the temporal evolution of inequality in urban OSM building mapping on the global scale and per World Bank region. This analysis has been conducted for annual snapshots from 2008-01-01 up until 2022-01-01. The Gini coefficient has been utilized to derive the degree of evenness of urban OSM building completeness following an approach proposed by Massey & Denton (1988) to study residential segregation[62]. Analogous to their approach, the Gini coefficient was derived from the Lorenz curve, which plots the cumulative proportion of observed OSM building area against the cumulative proportion of "missing" building area (difference between OSM building area and predicted building area) across urban centers, which are ordered from smallest to largest proportion of observed building area. The Gini coefficient constitutes a non-spatial measure of segregation which provides insights on the evenness dimension, but does not allow conclusion about the spatial structure. Pysal's 'segregation' package has been utilized to calculate the annual Gini coefficient from 2008-01-01 up until 2022-01-01[63].

Moran's I[33] has been selected as a measure of global spatial autocorrelation of urban OSM building completeness. A high Moran's I value describes situations where urban centers and their neighbors showed similar high (or low) values of completeness. It's values are not strictly bound by the interval [-1,1]. The range depends on the largest and the smallest eigenvalue of the spatial weight matrix used, but frequently ranges from -0.5 to 1.15[64]. A Moran's I value close to zero indicates a spatially random pattern, where the completeness of an urban center was not correlated to the completeness of its neighbours. Spatial autocorrelation has been proposed as an explicitly spatial indicator of segregation covering the dimension of clustering[62,65]. Following this approach a high degree of clustering describes a spatial structure where areas with OSM building mapping are contiguous and closely packed, creating a single large block of "mapped" urban centers. In contrast, a low level of clustering implies that the observed OSM building stock is widely scattered around the globe (or within regions)[62]. Moran's I relies on the definition of the spatial weight matrix - here, it was defined based upon the centroid of each urban center using a distance band threshold of 5 degree and an inverse distance weighting with a power of 1. Weights were row standardized. The neighborhood definition led on average to 367 neighbours per urban centers and 13 urban centers were classified as 'islands' for which no neighbours were identified.

We used the same spatial weight matrix to calculate local spatial autocorrelation (Local Moran Statistics[66]) for Europe & Central Asia and Sub-Saharan Africa to compare spatial inequalities within these regions between two timestamps. Pysal's

'esda' package was utilized to calculate the global and local Moran's I statistics from 2008-01-01 up until 2022-01-01[63].

Intra-Urban OSM Building Completeness

To ensure a sufficient sample size, we calculated both inequality measures (Gini coefficient, Moran's I) for the intra-urban assessment only for urban centers with a minimum area of 25 square kilometers respective 25 data points. For the resulting 4,722 urban centers 'local' OSM building completeness was derived using the area ratio method described above for each 1x1 km grid cell. The Gini coefficient and Moran's I global spatial autocorrelation of the 'local' OSM completeness were calculated per urban center as described in the previous section. The spatial weight matrix has been defined based upon the Queen contiguity graph. As such grid cells that share at least a vertex were considered as neighbours. The weight matrix was row-standardized.

The investigation was complemented by an agglomerative hierarchical cluster analysis which defined five groups of urban centers considering evenness and clustering within each city. The number of clusters has been selected based on the hierarchical structure of the full dendrogram. The distance matrix was based on the euclidean distance between OSM completeness, Gini coefficient and Moran's I. Since all variables considered in this analysis already showed a similar range of values between 0–1 normalization was not necessary. The analysis was conducted based on scikit-learn's python implementation[54] using the ward linkage criterion. Results were displayed for each cluster using scatter plots.

Cluster representatives were selected by first calculating the average values for Moran's I, Gini coefficient and completeness across all urban centers per cluster. Based on these cluster centroids the euclidean distance to each sample was derived. Among the 15 samples closest to the cluster centroid one representative was selected per cluster.

Data Availability

The full set of data for training and running the machine learning model and the final results presented in all figures and maps are available on HeiBox: https://heibox.uni-heidelberg.de/f/b2f22e7f341f48a89100/.

Code Availability

All Python code and Jupyter notebooks necessary to calculate the geospatial statistics, create maps and derive figures are available in this GitHub repository: https://github.com/GIScience/global-urban-building-completeness-analysis.

References

1. Gao, J. & O'Neill, B. C. Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways. *Nat. Commun.* **11**, 1–12, DOI: 10.1038/s41467-020-15788-7 (2020).
2. Sun, L., Chen, J., Li, Q. & Huang, D. Dramatic uneven urbanization of large cities throughout the world in recent decades. *Nat. Commun.* **11**, DOI: 10.1038/s41467-020-19158-1 (2020).
3. Boo, G. *et al.* High-resolution population estimation using household survey data and building footprints. *Nat. Commun.* **13**, 1–10, DOI: 10.1038/s41467-022-29094-x (2022). 2106.07461.
4. Hecht, R., Kunze, C. & Hahmann, S. Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS Int. J. Geo-Information* **2**, 1066–1091, DOI: 10.3390/ijgi2041066 (2013).
5. Esch, T. *et al.* World Settlement Footprint 3D - A first three-dimensional survey of the global building stock. *Remote. Sens. Environ.* **270**, 112877, DOI: 10.1016/j.rse.2021.112877 (2022).
6. Braunschweig, K., Eberius, J., Thiele, M. & Lehner, W. The State of Open Humanitarian Data. Tech. Rep. January (2020).
7. Altay, N. & Labonte, M. Challenges in humanitarian information management and exchange: Evidence from Haiti. *Disasters* **38**, 50–72, DOI: 10.1111/disa.12052 (2014).
8. Fritz, S. *et al.* Citizen science and the United Nations Sustainable Development Goals. *Nat. Sustain.* **2**, 922–930, DOI: 10.1038/s41893-019-0390-3 (2019).
9. Sustainable Development Solutions Network. Indicators and a Monitoring Framework for the Sustainable Development Goals: Launching a data revolution for the SDGs. *A report by Leadersh. Counc. Sustain. Dev. Solutions Netw.* 160 (2015).
10. Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J. & Zipf, A. The evolution of humanitarian mapping within the OpenStreetMap community. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-82404-z (2021).
11. Scholz, S., Knight, P., Eckle, M., Marx, S. & Zipf, A. Volunteered Geographic Information for Disaster Risk Reduction—The Missing Maps Approach and Its Potential within the Red Cross and Red Crescent Movement. *Remote. Sens.* **10**, 1239, DOI: 10.3390/rs10081239 (2018).

12. Milojevic-Dupont, N. *et al.* Learning from urban form to predict building heights. *PLoS ONE* **15**, 1–22, DOI: 10.1371/journal.pone.0242010 (2020).

13. Van Den Hoek, J., Friedrich, H. K., Ballasiotes, A., Peters, L. E. R. & Wrathall, D. Development after Displacement: Evaluating the Utility of OpenStreetMap Data for Monitoring Sustainable Development Goal Progress in Refugee Settlements. *ISPRS Int. J. Geo-Information* **10**, 153, DOI: 10.3390/ijgi10030153 (2021).

14. Feldmeyer, D., Nowak, W., Jamshed, A. & Birkmann, J. An open resilience index: Crowdsourced indicators empirically developed from natural hazard and climatic event data. *Sci. Total. Environ.* **774**, 145734, DOI: 10.1016/j.scitotenv.2021.145734 (2021).

15. Bhatia, A. *et al.* The Rohingya in cox's bazar: When the stateless seek refuge. *Heal. Hum. Rights* **20**, 105–122 (2018).

16. Sturrock, H. J., Woolheater, K., Bennett, A. F., Andrade-Pacheco, R. & Midekisa, A. Predicting residential structures from open source remotely enumerated data using machine learning. *PLoS ONE* **13**, 1–10, DOI: 10.1371/journal.pone.0204399 (2018).

17. Yeboah, G. *et al.* Analysis of openstreetmap data quality at different stages of a participatory mapping process: Evidence from slums in Africa and Asia. *ISPRS Int. J. Geo-Information* **10**, DOI: 10.3390/ijgi10040265 (2021).

18. Marco Minghini *et al.* Editorial: OpenStreetMap research in the COVID-19 era. *Proc. Acad. Track at State Map 2020* 1–4, DOI: https://doi.org/10.5281/zenodo.3922054 (2020).

19. Nirandjan, S., Koks, E. E., Ward, P. J. & Aerts, J. C. A spatially-explicit harmonized global dataset of critical infrastructure. *Sci. Data* **9**, 1–13, DOI: 10.1038/s41597-022-01218-4 (2022).

20. Boeing, G. Spatial information and the legibility of urban form: Big data in urban morphology. *Int. J. Inf. Manag.* **56**, 102013, DOI: 10.1016/j.ijinfomgt.2019.09.009 (2021).

21. Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208, DOI: 10.1038/s41467-022-29838-9 (2022).

22. Anderson, J., Sarkar, D. & Palen, L. Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS Int. J. Geo-Information* **8**, 232, DOI: 10.3390/ijgi8050232 (2019).

23. Oort, P. *Spatial data quality: from description to application*. Ph.D. thesis, Wageningen Universiteit (2006).

24. Barron, C., Neis, P. & Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions GIS* **18**, 877–895, DOI: 10.1111/tgis.12073 (2014). 9605103.

25. Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C. & Haklay, M. M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **8816**, 1–29, DOI: 10.1080/13658816.2016.1189556 (2016).

26. Sui, D., Goodchild, M. & Elwood, S. Volunteered geographic information, the exaflood, and the growing digital divide. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, vol. 9789400745872, 1–12, DOI: 10.1007/978-94-007-4587-2_1 (Springer, 2013).

27. Neis, P., Zielstra, D. & Zipf, A. Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Futur. Internet* **5**, 282–300, DOI: 10.3390/fi5020282 (2013).

28. Brovelli, M. & Zamboni, G. A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. *ISPRS Int. J. Geo-Information* **7**, 289, DOI: 10.3390/ijgi7080289 (2018).

29. Goldblatt, R., Jones, N. & Mannix, J. Assessing OpenStreetMap Completeness for Management of Natural Disaster by Means of Remote Sensing: A Case Study of Three Small Island States (Haiti, Dominica and St. Lucia). *Remote. Sens.* **12**, 118, DOI: 10.3390/rs12010118 (2020).

30. Zhang, Y., Zhou, Q., Brovelli, M. A. & Li, W. Assessing OSM building completeness using population data. *Int. J. Geogr. Inf. Sci.* 1–24, DOI: 10.1080/13658816.2021.2023158 (2022).

31. Florczyk, A. J. *et al. Description of the GHS Urban Centre Database 2015*. February (2019).

32. Raifer, M. *et al.* OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Softw. Standards* **4**, DOI: 10.1186/s40965-019-0061-3 (2019).

33. Cliff, A. D. & Ord, J. K. *Spatial processes: models & applications* (Taylor & Francis, 1981).

34. Sachs, J., Lafortune, G., Kroll, C., Fuller, G. & Woelm, F. Sustainable Development Report. Tech. Rep. (2022). DOI: 10.1017/9781009210058.

35. Graham, M., Straumann, R. K. & Hogan, B. Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia. *Annals Assoc. Am. Geogr.* **105**, 1158–1178, DOI: 10.1080/00045608.2015.1072791 (2015).

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

37. Lloyd, C. T. *et al.* Using GIS and machine learning to classify residential status of urban buildings in low and middle income settings. *Remote. Sens.* **12**, 1–20, DOI: 10.3390/rs12233847 (2020).

38. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **54**, DOI: 10.1145/3457607 (2021). 1908.09635.

39. Microsoft. Microsoft building footprints. https://github.com/microsoft/GlobalMLBuildingFootprints (2022). Accessed: 2022-06-01.

40. Meta. Map with ai. https://mapwith.ai (2022). Accessed: 2022-06-01.

41. Meta. Daylight map distribution. https://daylightmap.org/ (2022). Accessed: 2022-06-01.

42. Harley, J. B. Deconstructing the map. *Cartogr. The Int. J. for Geogr. Inf. Geovisualization1* **26**, 1–20 (1989).

43. Biljecki, F., Chew, L. Z. X., Milojevic-Dupont, N. & Creutzig, F. Open government geospatial data on buildings for planning sustainable and resilient cities, DOI: https://doi.org/10.48550/arXiv.2107.04023 (2021). Preprint, 2107.04023.

44. Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E. & Mills, J. Development of new open and free multi-temporal global population grids at 250 m resolution. *Agile* 6 (2016).

45. Smits, J. & Permanyer, I. Data descriptor: The subnational human development database. *Sci. Data* **6**, 1–15, DOI: 10.1038/sdata.2019.38 (2019).

46. Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F. C. & Taneja, J. Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019. *Remote. Sens.* **13**, 1–14, DOI: 10.3390/rs13050922 (2021).

47. Zanaga, D. *et al.* ESA WorldCover 10 m 2020 v100. Tech. Rep. (2021). DOI: https://doi.org/10.5281/zenodo.5571936.

48. Rehrl, K. & Gröchenig, S. A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS Int. J. Geo-Information* **5**, 37, DOI: 10.3390/ijgi5030037 (2016).

49. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

50. Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. & Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, DOI: 10.7717/peerj.5518 (2018).

51. Goetz, J. N., Guthrie, R. H. & Brenning, A. Integrating physical and empirical landslide susceptibility models using generalized additive models. *Geomorphology* **129**, 376–386, DOI: 10.1016/j.geomorph.2011.03.001 (2011).

52. Ravindra, K., Rattan, P., Mor, S. & Aggarwal, A. N. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environ. Int.* **132**, 104987, DOI: 10.1016/j.envint.2019.104987 (2019).

53. de Albuquerque, J. P., Herfort, B., Brenning, A. & Zipf, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *Int. J. Geogr. Inf. Sci.* 1–23, DOI: 10.1080/13658816.2014.996567 (2015).

54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

55. Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 1–11, DOI: 10.1038/s41467-020-18321-y (2020).

56. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929, DOI: 10.1111/ecog.02881 (2017).

57. Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *Int. Geosci. Remote. Sens. Symp. (IGARSS)* 5372–5375, DOI: 10.1109/IGARSS.2012.6352393 (2012).

58. Fan, H., Zipf, A., Fu, Q. & Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **28**, 700–719, DOI: 10.1080/13658816.2013.867495 (2014).

59. United Nations Development Programme. *Human development report 2019 : beyond income, beyond averages, beyond today: inequalities in human development in the 21st century.* (2019).

60. OECD. *OECD Regions at a Glance 2016* (2016).

61. OpenStreetMap Contributors. Openstreetmap wiki - organized editing activities. https://wiki.openstreetmap.org/wiki/Organised_Editing/Activities (2022). Accessed: 2022-06-18.

62. Massey, D. S. & Denton, N. A. The dimensions of residential segregation. *Soc. Forces* **67**, 281–315, DOI: 10.1093/sf/67.2.281 (1988).

63. Rey, S. J. & Anselin, L. PySAL: A Python Library of Spatial Analytical Methods. *The Rev. Reg. Stud.* **37**, 5–27 (2007).

64. Griffith, D. A., Chun, Y. & Li, B. *Spatial Regression Analysis using Eigenvector Filtering* (CRC Press, 2019).

65. Morrill, R. L. On the measure of geographic segregation. *Geogr. Res. Forum* **11**, 25–36 (1991).

66. Anselin, L. Local Indicators of Spatial Association - LISA. *Geogr. Analysis* **27**, 93–115, DOI: 10.1111/j.1538-4632.1995.tb00338.x (1995). 1011.1669.

Acknowledgements

The authors would like to thank the OSM contributors for their inspiring work, as well as Matthias Schaub, Levi Szamek, Rafael Troilo and Clemens Langer for their great help with building the computational framework for the analysis. B.H. and S.L. were supported by the Klaus Tschira Stiftung.

Author contributions statement

B.H., J.P.A. and S.L. conceived the analysis. B.H., S.L. and J.A. conducted the experiments and analyzed the results. B.H. and S.L. prepared graphics and tables. B.H. interpreted the data and wrote the paper with contributions from S.L., J.P.A., J.A. and A.Z.

Additional information

Competing interests S.L. and A.Z. declare no competing interests. B.H., J.P.A. and J.A. are unpaid voting members of the Humanitarian OpenStreetMap Team. Voting Members are responsible for voting "on matters affecting the Corporation including, but not limited to, the election of directors and [additional] voting members."