Building footprint data for countries in Africa: to what extent are existing data products comparable?

Heather R Chamberlain (h.chamberlain@soton.ac.uk)
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0003-0828-6974

Edith Darin
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0002-8176-092X

Ademola Adewole
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0002-7538-9781

Warren Christopher Jochem
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0003-2192-5988

Attila N Lazar
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0003-2033-2013

Andrew J Tatem
Geography and Environmental Science, University of Southampton, Southampton, UK
https://orcid.org/0000-0002-7270-941X

Research Article

Keywords: Building footprints, urban form, built environment, spatial, settlement

Posted Date: December 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3334423/v2

Additional Declarations: The authors declare no competing interests.

Building footprint data for countries in Africa: to what extent are existing data products comparable?

[PREPRINT v02]

Heather R. Chamberlain *[1], Edith Darin [1,2], Ademola Adewole [1], Warren C. Jochem [1], Attila Lazar [1], Andrew J. Tatem [1]

[1] WorldPop, Geography and Environmental Science, University of Southampton, Southampton, UK
[2] Leverhulme Centre for Demographic Science, University of Oxford, Oxford, UK

*Corresponding author: h.chamberlain@soton.ac.uk

Abstract

Growth and developments in computing power, machine-learning algorithms and satellite imagery spatiotemporal resolution have led to rapid developments in automated feature-extraction. These methods have been applied to create geospatial datasets of features such as roads, trees and building footprints, at a range of spatial scales, with national and multi-country datasets now available as open data from multiple sources. Building footprint data is particularly useful in a range of applications including mapping population distributions, planning resource distribution campaigns and in humanitarian response. In settings with well-developed geospatial data systems, such datasets may complement existing authoritative sources, but in data-scarce settings, they may be the only source of data. However, knowledge on the degree to which building footprint data products are comparable and can be used interchangeably, and the impact of selecting a particular dataset on subsequent analyses remains limited. For all countries in Africa, we review the available multi-country building footprint data products and analyse their similarities and differences in terms of building area and count metrics. We explore the variation between building footprint data products across a range of spatial scales, including sub-national administrative units and different settlement types. Our results show that the available building footprint data products are not interchangeable. There are clear differences in counts and total area of building footprints between the assessed data products, as well as considerable spatial heterogeneity in building footprint coverage and completeness.

Keywords:

Building footprints, urban form, built environment, spatial, settlement

1. Introduction

Data on the locations and extents of buildings and settlements are essential for many applications, from monitoring urban development (Domingo et al., 2023) and contingency planning (Nirandjan et al., 2022), to humanitarian response efforts (Herfort et al., 2021, Ullah et al., 2023) and in conducting national population and housing censuses (Darin et al., 2022, Sanchez-Cespedes et al., 2023). Over the last decade, with the growth of new sources of data, there has been a shift away from spatially detailed data on settlement extents and building locations being limited to countries with well-developed geospatial systems. In particular, the development of global, remotely-sensed settlement datasets at high spatial resolution (Esch et al., 2017, Corbane et al., 2019) and the growth of

volunteered geographic information (VGI) (Goodchild, 2007) have expanded the availability of data on settlements and the built environment. Most recently, highly-detailed data on buildings has become available with the publication of multi-country datasets of building footprints, providing data on the location, size and shape of individual buildings in both rural and urban settings.

High-resolution maps of settlements, down to the level of individual buildings, were originally created using manual cartography from information collected through surveying. The development of remote sensing in the 20th Century, provided an additional perspective for mapping settlements. Aerial photographs provided a bird's-eye view and an opportunity for manual identification of features such as buildings. Satellite remote sensing enabled regular data collection over larger areas, at spatial resolutions suitable for classification of land cover types (e.g. urban or water) (Barnsley and Barr, 1996, Fugate et al., 2010). Over the past decades, the number of remote-sensing satellites has grown rapidly, and the spatial and temporal resolution of the imagery captured has increased (Donnay et al., 2000, Jensen and Cowen, 1999). Alongside this there has been a massive growth in computing power, which together with the development of machine and deep learning algorithms have resulted in increasingly detailed datasets on settlements (Esch et al., 2017, Tiecke et al., 2017), providing data on individual buildings, with coverage spanning multiple countries and even continents (Sirko et al., 2021).

Building footprint datasets consist of 2-dimensional outlines of buildings, created by manual digitisation or through automated feature extraction. The 2D geometry of a building may also be accompanied by attributes, such as the building type, age or height/number of storeys. Outlines of individual buildings have long been included in large scale maps of urban areas, but the growth of GIS and spatial data, spawned the first digital building footprint datasets. Increasingly, the push for open data has meant that more building footprint datasets are becoming publicly available. In terms of the current data landscape of building footprint dataset providers, there are three main sources for such data: (1) authoritative datasets from government agencies, (2) data from VGI-initiatives and (3) datasets produced using automatic feature extraction methods by commercial companies.

Authoritative building footprint datasets, developed by government authorities, have traditionally been the sole source of spatial data on buildings. The geographic coverage of these datasets spans from city-level to national, with mapping agencies as the main producer and maintainer of data for cadastral and topographic mapping purposes (Biljecki et al, 2021). In the mid-2010s, authoritative building footprint datasets were generally only openly available with complete coverage for individual cities in high-income countries, including cities in Canada, New Zealand, Australia and the USA (Rae, 2015). In March 2015, Ordnance Survey (Britain's national mapping agency) released their OS OpenMap Local product, which included building footprints for the whole of Great Britain (https://wiki.openstreetmap.org/wiki/Ordnance_Survey_OpenData). The number of authoritative building footprint datasets that are openly available has since grown considerably, particularly in Europe and North America, however no national datasets are openly available for African countries (Biljecki et al., 2021).

In contrast, VGI-initiatives, such as OpenStreetMap (OSM), can provide highly detailed spatial data on buildings and other features, that are independent of national mapping agencies or other authoritative sources. Features, such as buildings, are primarily added to OSM through manual digitisation from high resolution satellite imagery, with contributions from a global community of thousands of mappers and organisations (Anderson et al., 2019). OSM has developed rapidly over the past decade and now contains data on billions of features globally. In particular, the addition of buildings in OSM started increasing notably in 2015 (Herfort et al., 2021) with over half a billion buildings now in OSM (Biljecki et al., 2021). Over the last decade, following the establishment of the Humanitarian OpenStreetMap Team and the Missing Maps Project, there has been a particular growth in mapping to support

humanitarian activities, initially for disaster response and increasingly in terms of preparedness (Herfort et al., 2021).

A relatively new and rapidly developing source of building footprints are those published by commercial companies, such as Microsoft, Google and Ecopia. These are predominantly created using feature extraction algorithms with large quantities of very high-resolution (VHR) satellite imagery. Microsoft, with Bing Maps, was the first, to release building footprint datasets at scale, initially for the USA (Microsoft, 2018), followed by national datasets for Canada (Microsoft, 2019a), Uganda and Tanzania (Microsoft, 2019b). Since then, Google, Ecopia and Microsoft have all released versions of building footprint data products with expanded geographic coverage, to include data for most countries in Africa. The specifics of these are described in more detail in Section 2.

The growth in building footprint data has been accompanied by an expansion in their application to an increasingly diverse range of uses and contexts. In urban settings, building footprints are key for understanding building infrastructure and have been widely used, with increasing importance in the context of urban sustainability (Biljecki et al., 2021). The geometry of building footprints provides insights on urban morphology (Wang et al., 2022, Arribas-Bel and Fleischmann, 2022, Biljecki and Chow, 2022), enabling classification of settlement types (Jochem et al., 2021, Fleischmann and Arribas-Bel, 2022, Wang et al., 2023), characterisation of urban expansion, densification and changes in urban form and use (Domingo et al., 2023, Cao et al., 2023). With the addition of attributes related to 3-dimensional building form, building footprints can be used to develop digital twins (Park and Guldmann, 2019, Dukai et al., 2019). The development of new building footprint data products has increased the sources of data available for urban areas, but also expanded data availability for rural areas and other data-sparse settings, such as informal settlements (Wang et al., 2022). These developments have led to data on building footprints being increasingly used in contexts relevant to humanitarian response, including for population estimation (Checchi et al., 2013, Boo et al., 2022, Wardrop et al., 2018, Leasure et al., 2020), post-disaster damage assessment (Robinson et al., 2023), disaster preparedness (Huang and Wang, 2020, Lazarus et al., 2018), survey sample design (Boo et al., 2020) and identification of vulnerable populations (Buchanan et al., 2020, Gibson et al., 2022, Chamberlain et al., 2022).

With a growing number of building footprint datasets becoming available globally, data users need to understand the extent to which seemingly similar datasets are interchangeable and their potential advantages, disadvantages and limitations. In this paper, we address this knowledge gap by reviewing available building footprint datasets. We focus on countries in Africa, where previously building footprint data has been very limited, with no national datasets from authoritative sources having been openly released (Biljecki et al., 2021). We review the available building footprint datasets in terms of their spatial and temporal coverage, the methods used in producing the datasets, the input data sources, as well as the licensing terms, file formats and data accessibility. We then compare and contrast simple metrics calculated for each dataset at a range of spatial scales and across the rural-urban continuum.

2. Data

Building footprint datasets that covered the majority of countries in Africa at the time of writing were included in this review. These datasets came from Google, Microsoft, OpenStreetMap and Ecopia (Table 1). Each dataset provides polygons representing building locations and extents which have been digitised or automatically extracted from satellite imagery. Building footprints from Google (https://sites.research.google/open-buildings), Microsoft (https://www.microsoft.com/en-us/maps/building-footprints) and Ecopia (https://www.ecopiatech.com/global-feature-extraction) have all been produced using automated feature extraction algorithms with high-resolution satellite imagery. In contrast, OSM building footprints have been primarily created through manual interpretation and digitisation from satellite imagery by OSM contributors. In some cases, manual interpretation may also have been supported by automated feature-detection algorithms, as is possible using RapiD (Facebook, 2019). Given the lack of openly available authoritative datasets with national coverage for countries in Africa (Biljecki et al., 2021), no such datasets could provide a benchmark or be included in this review.

Since Microsoft published their initial building footprint dataset for the USA in 2018, there has been significant expansion in terms of the geographic coverage of their building footprints. After publishing open, national datasets for Canada, Uganda and Tanzania in 2019, datasets for Australia followed in 2020, with Kenya and Nigeria published in 2021 (Microsoft, 2021a). Also in 2021, Microsoft published building footprint data for South America (Microsoft, 2021b), but with coverage limited to major cities. In 2022, Microsoft further scaled up their building footprint data, making available an additional 856 million building footprints for countries worldwide (Microsoft, 2022), although the coverage was not global.

After developing national building footprint datasets for Australia and the USA, Ecopia partnered with Maxar (formerly Digital Globe) to produce their first multi-country data product under the Digitize Africa project, which was specifically intended for humanitarian purposes. This consisted of building footprints for 51 countries in sub-Saharan Africa (Price and Hallas, 2019), published first in 2019. Google's Open Buildings data product was also initially focussed on Africa, with development led by the AI Research Center in Accra, Ghana (Sirko et al., 2021). At the time of writing, two versions of Google's Open Buildings dataset have been released with v1 published in 2021 and v2 in 2022. The Open Buildings v2 data product expanded the coverage to include data for countries in south/south-east Asia also.

	Data product	Geographic coverage / Release date	Method / Source satellite imagery	Data format	Availability / licensing	Notes
Ecopia	DigitizeAfrica building footprints (year 1)	- 51 countries/ territories/ dependencies in sub-Saharan Africa - Release date: 2019	- Automated feature extraction from satellite imagery - Maxar 30-50cm imagery (2005 - 2020; imagery date is provided for each b. footprint)	- Polygons (not rectilinear), no overlaps. Shapefile format. - Local UTM zone projections	Custom, commercial license which restricts usage predominantly to humanitarian applications	- A later release (termed "year 2") also exists, which has updated imagery for some locations (See further details in section 2.4 'Updates and versioning')
Google	Open Buildings v2	- Africa and south/south-east Asia - Release date: 2022	- Automated feature extraction from satellite imagery - Google Maps imagery, 50cm resolution (years of imagery not stated)	- Polygons (mostly rectilinear) with geometry as WKT, stored in CSV. - Overlap between polygons possible - EPSG 4326 (WGS84)	CC BY-4.0 license ODbL v1.0 license	- Includes data at building level on precision confidence score. - Has greater coverage and more recent imagery than Open Buildings v1
Microsoft	Kenya & Nigeria building footprints	- Kenya and Nigeria only - Release date: 2021	- Automated feature extraction from satellite imagery - Maxar imagery (2020-2021)	- Polygons (mostly rectilinear), geoJSON. - Overlap between polygons possible - EPSG 4326 (WGS84)	ODbL	- Some geographic overlap with Microsoft "global" building footprint dataset
	Uganda & Tanzania building footprints	- Uganda and Tanzania only - Release date: 2019	- Automated feature extraction from satellite imagery - Maxar imagery (years of imagery not stated)	- Polygons (mostly rectilinear), geoJSON. - Overlap between polygons possible - EPSG 4326 (WGS84)	ODbL	
	"Global" building footprints	- The "global" dataset includes countries across all populated continents but is not global. No coverage for Cabo Verde or Nigeria, with very limited data for Kenya, Tanzania and Uganda - Release date: 2022	- Automated feature extraction from satellite imagery - Bing Maps imagery (including Maxar and Airbus) dating from 2014-2022	- Polygons (mostly rectilinear), geoJSON. - Overlap between polygons possible - EPSG 4326 (WGS84)	ODbL	- Microsoft provide an accompanying file to show extent of coverage
OSM	OSM features tagged building=*	- Global, but sub-national coverage varies - Data extracted from OSM on 03 January 2023	- Predominantly manual digitization from satellite imagery, aided by some automated feature detection - Imagery from various sources	- Polygons (mostly rectilinear) - Overlap between polygons possible - Extracted from OSM	ODbL	- Constantly being updated given VGI-natures of OSM

Table 1: Details of the four multi-country building footprint data products covering the African continent as of January 2023.

2.1 Data attributes

Whilst the building footprint data products from Google, Microsoft, OpenStreetMap and Ecopia for countries in Africa all consist of 2-dimensional outlines of buildings, they differ in terms of their attribute information. Ecopia includes attributes on the date of the satellite imagery from which the building footprints were extracted. Google provide a confidence score for each building footprint and the location of its centroid. The confidence score reflects the certainty of a detected feature being a building, with only features with confidence scores exceeding 0.60 included in the dataset (Sirko et al., 2021). No attribute information is provided for Microsoft building footprints. None of the Google, Microsoft or Ecopia building footprints include attributes on building height or number of storeys.

OSM's tagging system enables attributes to easily be added to OSM features such as buildings. However, apart from the tag specifying a feature's type (e.g. building), additional tag values (attributes) are optional and hence vary considerably in their prevalence and quality. Biljecki et al. (2023) explored the completeness and quality of OSM building attributes globally, and found that building type, number of storeys and height were the most common attributes, available for 10.5%, 4.6% and 2.9% of buildings respectively. Similarly, Biljecki et al. (2021) investigated attribute information in openly available building footprint datasets from authoritative sources and found that over half of datasets had no building attributes included. For the remaining datasets, the most common attribute to be included was the type of building. Although no openly available national building footprint datasets from authoritative sources are available for countries in Africa (Biljecki et al., 2021), a few other building footprint datasets with attribute information have been published, although with very limited geographic extents. For example, a building footprint dataset for the city of Lusaka, Zambia (Chiwele et al., 2022) includes information on settlement types.

2.2 Dataset production – methods

For building footprint products from Ecopia, Google and Microsoft, the geometry of building polygons is determined by the feature extraction algorithms, model training datasets and post-processing methods. These data products have been developed using similar approaches insofar that they use deep learning models for semantic segmentation of VHR satellite imagery, and then apply post-processing methods to convert clusters of grid cells into building polygons. Methods for detection of buildings from VHR imagery have been reviewed by Li et al. (2022). Each of the data producers provide varying details of their methods, and none of their models have been published. Microsoft describe using deep convolutional neural networks for semantic segmentation and then applying a polygonization method (Microsoft, 2022). Google state they use a convolutional neural network based on the U-Net architecture, followed by a contouring algorithm to create building polygons (Sirko et al., 2021). The post-processing methods largely determine the geometry of building polygons (e.g. if polygons are coerced to be rectilinear), and implement any constraints in regards to polygon overlaps, minimum area thresholds and the delineation of courtyards within buildings. Model performance statistics (as provided by the data producers) are summarised in Supplementary Table B.2 and Figure B.1.

2.3 Source satellite imagery

As all the building footprint datasets are extracted or digitised from satellite imagery, details of the imagery, and in particular its recency, are important for data users to know. Ecopia's Digitize Africa building footprints are extracted from 30-50cm Maxar imagery and include an imagery date attribute for each individual building footprint (Supplementary Figure A.1). Very limited information on the recency of satellite imagery is provided by the other data producers (Table 1). As of January 2023, the "global" Microsoft building footprints were extracted from Bing Maps imagery (sources included Maxar, Airbus and IGN France) dating from 2014-2022 (Microsoft, 2022). The earlier country-specific

datasets for Kenya/Nigeria and Uganda/Tanzania were based on Maxar satellite imagery. The imagery for Kenya/Nigeria dated from 2020/21, however Microsoft provide no information on the Uganda/Tanzania imagery (although must predate 2020 as the dataset was released in January 2020). The spatial resolution of the imagery used for Google Open Buildings was 50cm (Sirko et al., 2021). The documentation for Google Open Buildings v2 states that v2 is based on more recent imagery than v1, but no information on the imagery dates or sources are provided. Users are advised to utilise the Historical Imagery tool in Google Earth Pro to find imagery dates for a particular location, however this is not feasible for reviewing more than a small area.

2.4 Updates and versioning

As time elapses and data products are updated, documentation of datasets and versioning becomes increasingly important. At the time of writing, Google has released two versions of their Open Buildings data product, named as v1 and v2. Microsoft has released building footprints for various countries at different times, with updates for some locations also. The Microsoft country-specific datasets are distinct from the "global" dataset insofar that they are in separate GitHub repositories, but any plans for versioning or naming such datasets, particularly if further updated, are unclear. The nature of Ecopia's building footprints for sub-Saharan Africa being produced through the Digitize Africa project, resulted in an initial dataset for each country and then an update for a subset of areas in each country. Given the project timeframe, these versions have been referred to as "year 1" and "year 2" respectively, with no further updates expected. Unlike the other three sources of building footprints, OSM is constantly being edited and updated. The date when features are created or modified forms the basis of the OSM versioning system. In this work, we utilise the following versions of data products: Google Open Buildings v2 and Ecopia year 1 building footprints. Microsoft country specific datasets are used where these are available (Kenya, Nigeria, Tanzania and Uganda) and supplemented by Microsoft's "global" product in the version that was available in January 2023. Building footprints were extracted from OSM in January 2023.

2.5 Geographic coverage

Focussing specifically on the countries in Africa (as defined by the UN Africa region), the spatial coverage and extent of the building footprint datasets varies considerably across the region. At the time of writing, none of the data products has complete coverage for all African countries (Supplementary Table A1 and Figure 1). For the data products that are created using automatic feature extraction techniques, the geographic coverage is largely determined by the extent of the satellite imagery (assuming sufficient spatial resolution and quality). Ecopia's Digitize Africa building footprints are available for 51 countries/territories in sub-Saharan Africa. Google Open Buildings v2 dataset includes building footprints for most countries in Africa, although no documentation is provided on exactly which. Exploration of data coverage (Figure 1 and Supplementary Figures A.2-A.5) indicates that data has been excluded for several countries and areas, often in areas experiencing conflict. Building footprints from Microsoft are available for all countries/territories in Africa, except for Cabo Verde. Microsoft's country specific datasets provide data for Tanzania, Kenya, Uganda and Nigeria, while the rest of Africa (and some parts of these four countries) is included in Microsoft's more-recent "global" building footprints (Microsoft, 2022). The coverage of Microsoft's building footprints is somewhat variable; areas with missing buildings can occur due to satellite imagery being too old (prior to 2014) or if it was considered that the area had a low-probability of detection (broadly based on distance to roads and population centres) (Microsoft, 2022). Following the publication of their "global" building footprint dataset, Microsoft released an accompanying file indicating geographic coverage and thus enabling users to identify gaps. However, as such gaps can be quite extensive, the utility of the building footprints for users is potentially somewhat compromised. Critically at the time of downloading the building footprint data products, none of the data producers

provided their national boundary definition, resulting in potentially different geographic extents for national datasets from different producers.

Unlike the other data products, OSM building features are predominantly created through digitisation by a community of contributors. Therefore, inevitably the geographic coverage of OSM features is spatially heterogeneous. Nevertheless, as an open, widely used data source with permissive licensing terms, we deemed it important to include in this paper given our focus on reviewing available building footprint datasets for African countries. The coverage and completeness of building features in OSM has been the subject of numerous studies (e.g. Hecht et al. (2013), Ullah et al. (2023), Zhang et al. (2022)). Most studies assessing OSM building completeness are however limited to relatively small areas, with few studies focused on countries in Africa. Recent studies have assessed completeness of OSM building footprints for cities globally, finding that overall, completeness was generally low in most cities but with considerable spatial heterogeneity (Herfort et al., 2023 and Zhou et al., 2022). Specifically, for cities in sub-Saharan Africa, Herfort et al. (2023) estimated completeness of OSM buildings to be 30%, with more than 50% of edits coming from humanitarian-related activities.

2.6 Dataset availability and licensing

Differences also exist between the four building footprint data products (as of January 2023) in terms of their availability and licensing (Table 1). Microsoft's building footprints are made openly available with an Open Data Commons Open Database License (OdbL), and are available to download from GitHub, with separate repositories for country-specific and "global" datasetst. Google building footprints are similarly made openly available with either OdbL or Creative Commons Attribution (CC BY-4.0) licenses. Users can download the data in several different ways: as individual tiles (level 4 S2 cells), by specifying a region based on coordinates or a supplied boundary in a Google Colab Notebook, or using the gsutil tool. Ecopia building footprints for sub-Saharan African countries have licensing terms that restrict their use but are available for humanitarian purposes. To access the data, users are required to submit a request describing the intended data use, for review. OSM building footprint data can be downloaded using the various available OSM download tools, in particular specifying features that have the tag building=*. OSM buildings are included in Shapefile format in the free OSM data extract from Geofabrik (http://download.geofabrik.de/). As with all features from OSM, the license is OdbL.

3. Methods

The methods section is structured in two parts. The first (Section 3.1) describes the pre-processing of building footprint datasets and the subsequent steps taken to calculate metrics related to building footprint size and area, resulting in gridded building datasets, The second part (Section 3.2) outlines the steps taken to calculate summary statistics from the building metrics, and the comparative analysis of these.

	Metric	**Description**
Core metrics	Count of building footprints (centroid-based)	The number of building footprints in a grid cell, where the building footprint location is determined by its centroid.
	Count of building footprints (any part)	The number of building footprints in a grid cell, where any part of a building footprint that intersects with any part of the grid cell is included in the count.
	Total building footprint area [m^2]	The summed area of all building footprints/parts of building footprints in a grid cell.

Table 2: Core metrics related to building footprint count and total area, calculated for each data product.

3.1 Creation of gridded building datasets

All building footprint datasets were downloaded, either in country-specific zipped files or tiles. Those that were in a projected coordinate system were reprojected such that all datasets were in WGS 1984 coordinate system (EPSG 4326). Each country or tile of building footprint data was processed to produce gridded (raster) datasets of metrics related to the count and area of building footprints. A common spatial resolution (3 arc seconds, 0.00083333 decimal degrees, approximately 100m at the Equator) and grid cell alignment was used for the gridded outputs. We adopted the approach of creating gridded outputs at this resolution to provide flexibility in terms of comparing values at different geographic scales, ranging from individual grid cells to national or continental comparisons. In addition, the output gridded datasets themselves can have further utility for other data users.

For each building footprint source, gridded outputs were calculated for six metrics that provide insights into the coverage, similarities and differences in the building footprint datasets. Three core metrics were calculated: two related to building counts and one for total building area (Table 2). Metrics of building counts were calculated based on both building footprint centroids and any part of a building polygon intersecting a grid cell. The primary metric related to building area was the total area of all footprints in a grid cell. Three additional metrics related to building footprint area were also calculated: the minimum, maximum and mean of all building footprints in a grid cell. All metrics were calculated using open-source Python code, adapted from Foks et al. (2020) and Heris et al. (2020). The input building footprint data products did not utilise a common set of national boundaries, resulting in the processed gridded metrics having different spatial extents for the same country. These gridded building metrics for each data source and country/tile were then mosaiced together to produce Africa-wide raster datasets. Further details of the data processing and mosaicing process are provided in Appendix C. Similar gridded outputs for the Ecopia year 1 building footprints are openly available to download (Dooley et al., 2020). Our subsequent comparative analysis focusses on the three core metrics (Table 2).

3.2 Calculation of summary statistics

For each of the sources of building footprint data, summary statistics were calculated for a range of spatial scales, including national and subnational administrative units and with stratification into rural and urban classes. For national and administrative level 1 (AL1) units, the count and total area of building footprints for each source of building footprints were calculated. The counts of building footprints (Figure 1 and Supplementary Table A.1) were calculated from the centroid-based count of building footprints rasters, while the total areas of building footprints (Supplementary Figure A.6 and Supplementary Table A.1) were calculated from the total area of building footprints rasters. In addition, the count of grid cells with one or more building footprints was calculated (Figure 2). The national and AL1 boundaries used in all comparative analyses, were sourced from GADM (Global Administrative Areas) database (GADM, 2022).

At national level also, the mean count of building footprints (Figure 3) and the mean value of total area of building footprints per grid cell (Supplementary Figure A.10) were calculated, with stratification by rural and urban settings. Urban and rural stratification was based on the GHS-SMOD dataset (Pesaresi et al., 2019). The SMOD L1 classification (3 classes: “urban centre”, “urban cluster”, “rural”) was predominantly used, with the more detailed L2 classification (7 classes) utilised when a more granular stratification of urban/rural types was beneficial (e.g. Supplementary Figures A.8 and A.11). For analyses at the national level, results were also grouped geographically based on the five UN Regions of Africa (Northern Africa, Eastern Africa, Western Africa, Middle Africa and Southern Africa).

The remaining analyses focussed still on count and total area, but with consideration also for the spatial similarity of building footprint datasets. The Jaccard Coefficient was calculated to assess spatial similarity of datasets on a pairwise basis. For each dataset a binary raster was produced, where grid cells with 1 or more building footprints had a value of 1, and all grid cells without building footprints had a value of 0. Identification of grid cells with 1 or more building footprints was based on any part of a building footprint intersecting with any part of the grid cell (not centroid-based). The Jaccard Coefficient ($JC = |A \cap B| / |A \cup B|$) is calculated as the area of intersection between two datasets (A and B), divided by the area of union in the two datasets. When applied in a pairwise fashion to our binary rasters, this is calculated as the area of common grid cells with a value of 1 in both raster A and raster B, divided by the area of all grid cells with a value of 1 in raster A and/or raster B.

Jaccard coefficient values were calculated for each pairwise combination of datasets, and summarised at national level (Figure 4), with stratification by GHS-SMOD classes at regional level (Figure 5 and Supplementary Figure A.11). Based on the same binary classification, the various combinations of building footprint datasets were mapped (Figure 6) and summarised on a national scale (Figure 7). Finally for the subset of grid cells with 1 or more building footprints from all four source datasets, the values for building footprint count and total area in each grid cell were extracted and plotted for each pair of datasets (Supplementary Figure A.12).

4. Results

The four building footprint data products are available for most countries in Africa, although several countries, particularly in north Africa, have no building footprint data in one or more sources (Figure 1). Subnational mapping of building footprint counts per administrative level 1 (AL1) unit shows considerable heterogeneity between the four different data products.

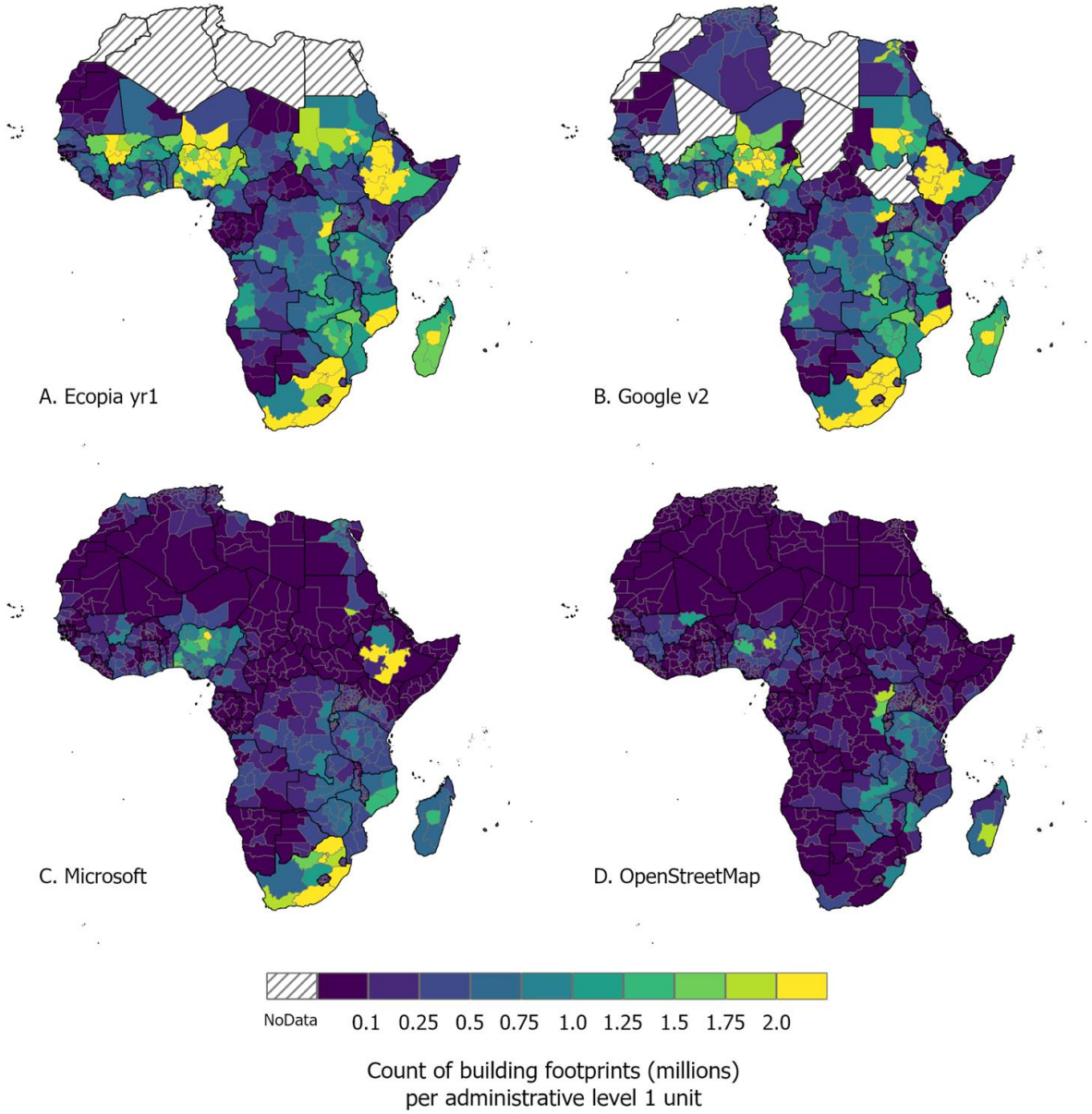

Figure 1: Sub-national variation in building footprint counts between data products (Ecopia year 1, Google v2, Microsoft and OSM as available in January 2023), shown for administrative level 1 units (n=879) for all countries in Africa

The highest subnational counts (greater than 2 million building footprints per AL1 unit) are observed for the Google and Ecopia data products (Figure 1). The highest count of building footprints for Google and Ecopia datasets are found for the state of Oromia, Ethiopia (n=14,838,851 and n=13,218,707 respectively). Counts of Microsoft building footprints in excess of 2 million occur in just a few AL1 units in South Africa and Nigeria; with only 2,203,691 building footprints in Oromia state,

Ethiopia in comparison. OSM building footprint counts per AL1 unit are generally lower than for other data products. The highest counts of OSM building footprints are found in Fianarantsoa, Madagascar (n=1,879,313), Ituri, DRC (n=1,873,858) and Bauchi, Nigeria (n=1,772,727). Counts of building footprints for OSM exceeded counts for Ecopia, Google and Microsoft in 33, 88 and 264 AL1 units respectively (excluding countries with no building footprints published). Most AL1 units have relatively low counts of OSM building footprints with 622 units (71%) having fewer than 100,000 building footprints. The proportion of units with fewer than 100,000 Microsoft building footprints is lower at 43% (380/879), with even lower proportions for Ecopia (24%) and Google (32% of AL1 units). The mean count of building footprints per AL1 unit for Ecopia, Google, Microsoft and OSM was 556,615, 534,676, 241,978 and 110,616 respectively. Similar sub-national patterns are observed in terms of total area of building footprints (Supplementary Figure A.6).

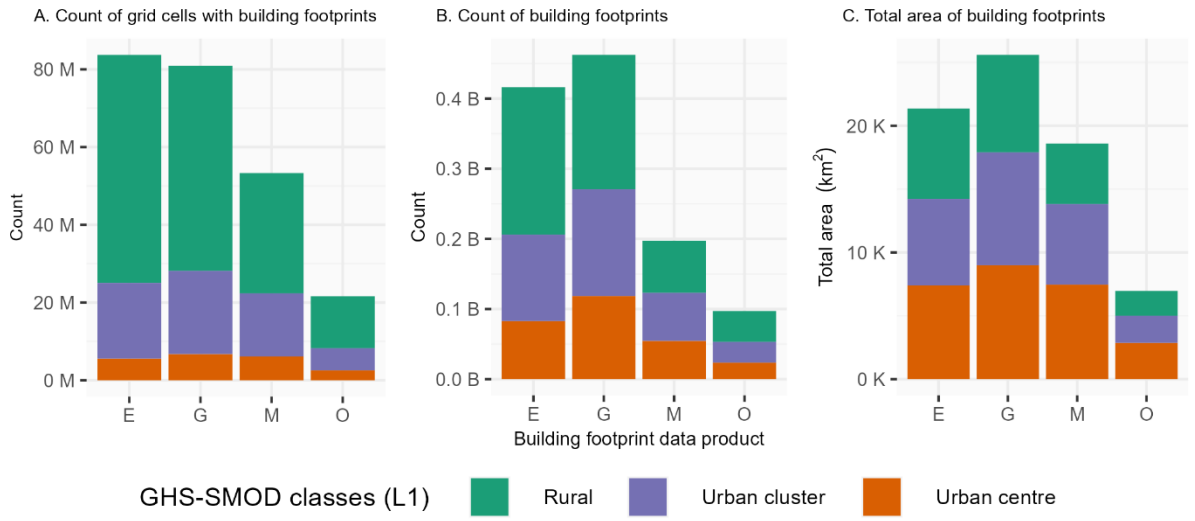

Figure 2: The count of grid cells with building footprints (a), the count of building footprints (b) and the total area of building footprints (c) across all countries in the African region, for each building footprint data product (E=Ecopia year 1, G=Google v2, M=Microsoft and O=OSM), with stratification by GHS-SMOD L1 classes. Counts of grid cells are shown in millions (M), counts of building footprints are shown in billions (B), and total area in thousands (K) of km^2.

The comparison of building counts and total building area is complicated by gaps in coverage in the building footprint data. For Ecopia building footprints, the geographic extent is clearly defined as all countries in sub-Saharan Africa. The coverage of the Google building footprints is not documented, but their interactive data viewer shows clear gaps in coverage for some countries (e.g. Chad, Libya, Mali, Morocco and South Sudan), and for some subnational units. In Figure 1 this can be clearly seen in terms of differences in counts between Google v2 and Ecopia in the provinces of Cabo Delgado in northern Mozambique and North Kivu in eastern DRC (Supplementary Figures A.2-A.5). Since publishing their “global” building footprints, Microsoft has provided a coverage file, which indicates quite extensive gaps in coverage. This can be clearly seen in Figure 2a, as the total number of grid cells with building footprints for Microsoft is considerably lower than total number for Ecopia and Google, although still more than double the count of grid cells with OSM building footprints. When stratified by Region and GHS-SMOD classes (Supplementary Figures A.7-A.9), the disparity in count of grid cells with building footprints generally increases for rural settings and is reduced in urban classes. For OSM buildings, for which incomplete coverage is expected, the number of grid cells with building footprints is the lowest of any data product (Figure 2a), accompanied by the lowest counts (Figure 2b) and total area of building footprints (Figure 2c) across the whole of the UN Africa region.

As Figure 1 shows, there is considerable national and subnational variation in the count and total area of building footprints between the four data products, in part at least due to differences in

completeness of coverage. To help understand what else may be driving such differences, the mean count of building footprints and the mean summed area of building footprints per grid cell (3 arc seconds resolution) are summarised for each country in Figure 3 and Supplementary Figure A.10 respectively. Grid cells with no building footprints were excluded in the calculation of mean values. In these figures, mean building footprint count and summed area are stratified by rural/urban types, as defined by the GHS-SMOD L1 classification (Pesaresi et al., 2019). In addition, the total count or area of building footprints for each dataset, country and SMOD class are represented by the point size.

Figure 3 shows that for urban centres in countries for which both Google and Ecopia building footprints are available, generally Google has the highest mean grid cell count, followed most often by Ecopia. For locations which are similar in terms of their rural/urban classification, there is much variation between countries in the mean value of building footprint counts per grid cell. This variation is most pronounced in urban centres with the mean count per grid cell varying by over 20 building footprints per grid cell between datasets in some countries, such as Sudan and Burkina Faso. Potentially this may reflect differences between data products in the definition and representation of individual buildings. The proportion of building footprints that are in rural or urban settings also varies between datasets and countries (Figure 3), reflecting differences in the distribution of settlements and possible spatial biases in the building footprint datasets. Considering the proportion of building footprints per SMOD class in each dataset and country, Ecopia had the greatest proportion of building footprints in rural settings (median national value of 50.8%). In comparison, the median value was 42.6% for Google, 37.6% for Microsoft and 34.3% for OSM. Conversely, for urban centres, the greatest proportion of building footprints in such settings was for OSM, with a median national value of 27.1%. Median national values for Ecopia, Google and Microsoft were 17.4%, 24.4% and 26.0% respectively.

Our comparative analysis has so far been limited to relatively coarse geographic scales (national and AL1 units, with stratification by GHS-SMOD classes), and has not assessed spatial similarity in data products at higher spatial resolutions. Figure 4 shows Jaccard coefficient values, per dataset and country in a pairwise-fashion, considering all grid cells with one or more building footprints. Jaccard coefficient values range from 0.0 (no similarity) to 1.0 (spatially identical), with higher values indicating greater spatial similarity in terms of grid cells (3 arc second spatial resolution) with building footprints.

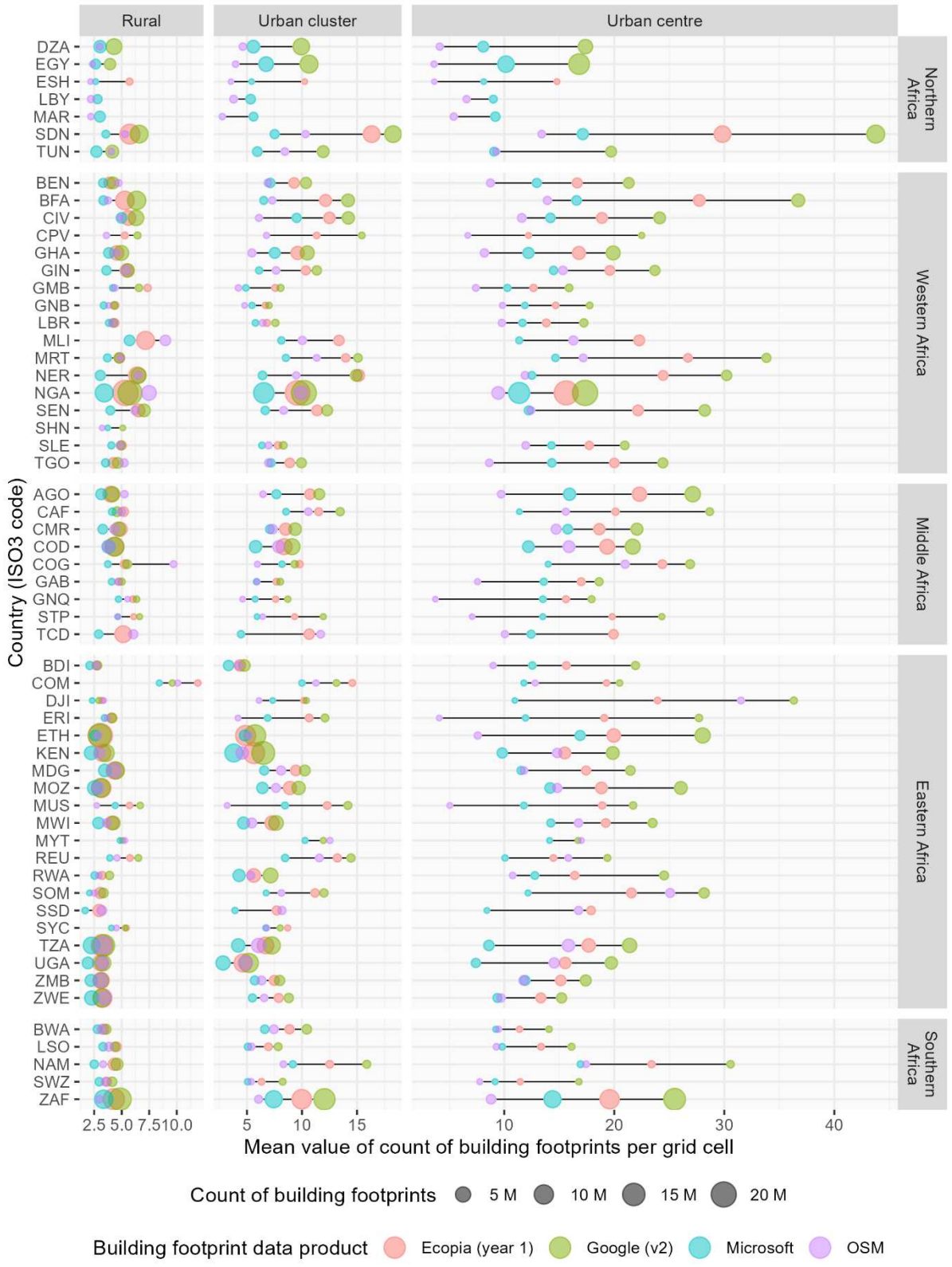

Figure 3: Mean count of building footprint per grid cell, shown for each country and dataset, stratified by rural/urban settings based on the GHS-SMOD L1 classification. The total count of building footprints per strata, dataset and country is represented by the point size (M = million).

Figure 4: Jaccard coefficient values for pairwise comparison of building footprint dataset spatial similarity (E=Ecopia year 1, G=Google v2, M=Microsoft and O=OSM), comparing grid cells with 1 or more building footprints, summarised for each country. Higher Jaccard coefficient values indicate greater spatial similarity. For countries with no building footprints in a particular data product (see Supplementary Tables A.1 and B.1), coefficient values will be 0.00.

For most countries, Jaccard coefficient values are generally highest for the comparison of Ecopia and Google building footprints (median=0.60), excluding countries where Ecopia building footprints are not available (Northern Africa region and Mayotte). The lowest Jaccard coefficient values for most countries are typically found in pairwise comparisons with OSM (median=0.208), as would be expected given the limited coverage of OSM building footprints (Figures 1, 2 and Supplementary Figure A.1 and Supplementary Table A.1). A limited number of countries have higher Jaccard coefficient values for pairwise comparisons with OSM building footprints, with values exceeding 0.65 for Reunion, Lesotho, Seychelles and the Gambia. Pairwise comparisons with Microsoft building footprints show considerable variation between countries (range=0.00-0.89, median=0.33). For the four countries with national Microsoft datasets (Kenya, Nigeria, Tanzania and Uganda), Jaccard

coefficient values for pairwise comparisons with Microsoft building footprints tend to be higher (median=0.54).

In Figure 5, Jaccard coefficient values are calculated across each region with stratification by GHS-SMOD L1 classes. The highest coefficient values are found for urban centres in Southern Africa. Coefficient values in this region are particularly high for the pairwise comparison of Google-Ecopia, Microsoft-Ecopia and Google-Microsoft, with coefficient values increasing in the move from rural (median=0.65), through urban clusters (median=0.91) to urban centres (median=0.94). In contrast, pairwise comparisons involving OSM data in this region have low Jaccard coefficient values (median=0.18), although the highest values are associated with urban centres (median=0.25). The trend of increasing Jaccard coefficient values with increasing urbanicity is maintained when the more granular GHS-SMOD L2 classification is applied (Supplementary Figure A.11). The lowest Jaccard coefficient values in each GHS-SMOD class and region, are found for pairwise comparisons involving OSM buildings. For each SMOD class, the highest Jaccard coefficient values for pairwise comparisons with OSM are for the Eastern Africa region, which is logical given less disparity in OSM building counts in this region (Figure 1).

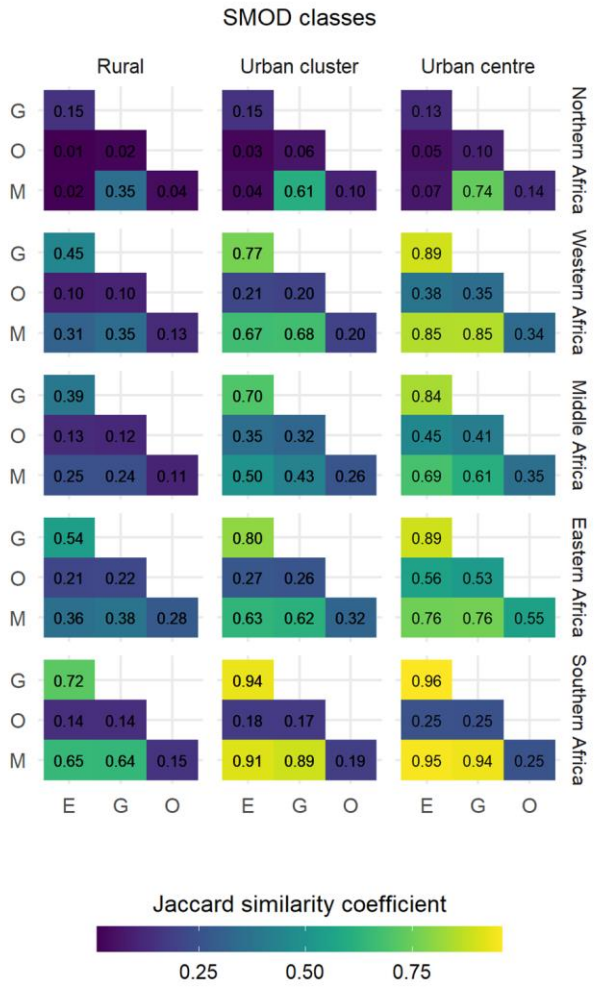

Figure 5: For each region and GHS-SMOD L1 class, Jaccard coefficient values are shown for pairwise combinations of building footprint datasets (E=Ecopia yr1, G=Google v2, M=Microsoft and O=OSM). Jaccard coefficient values were calculated based on grid cells with one or more building footprints.

The Jaccard coefficients shown in Figure 4 and Figure 5 provide insights into the spatial similarity of building footprint datasets on a pairwise basis. Using the same binary classification of grid cells (a grid cell with one or more building footprints is classified as settled), this comparison is further expanded in Figures 6 and 7. For each grid cell, the number of data products (n=1-4) with one or more building footprints is calculated, and the building footprint product(s) identified. Examples of these counts and dataset combinations are mapped for three of the largest cities in Africa (Cairo, Kinshasa and Lagos) in Figure 6, and summarised for all countries in Figure 7.

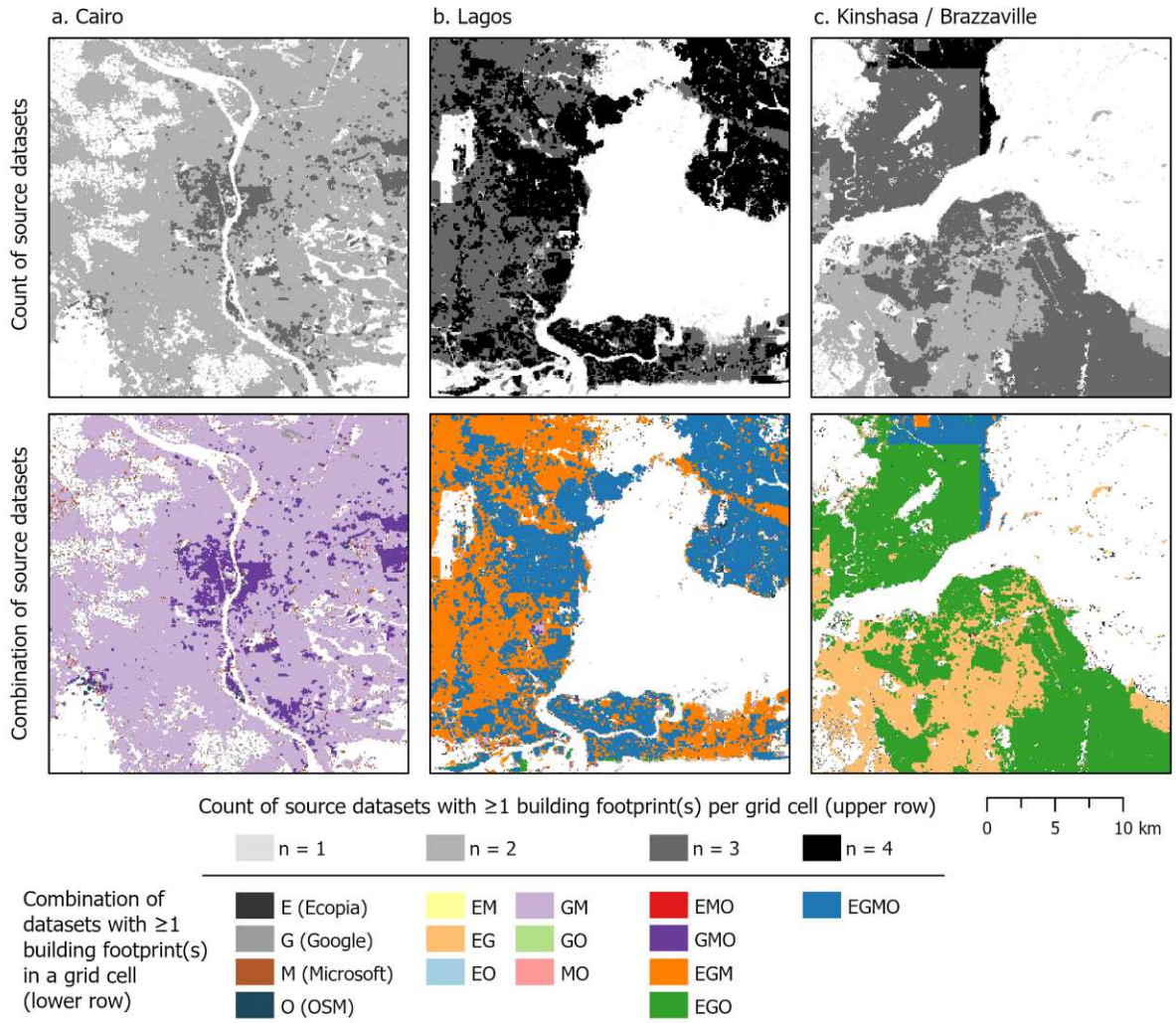

Figure 6: The count of building footprint datasets (upper row) and combination of datasets (lower row) for each grid cell with 1 or more building footprints (E=Ecopia yr1, G=Google v2, M=Microsoft and O=OSM), mapped for (a) Cairo, (b) Lagos and (c) Kinshasa/Brazzaville.

In Figure 6a, the upper map shows that the majority of Cairo has building footprints from 2 data products (n=2) and the lower map shows a combination of Google and Microsoft ("GM"). In central Cairo, the upper map shows a dataset count of 3 and the lower map shows the combination of GMO, indicating the addition of OSM building footprints. For Lagos (Figure 6b), the count of data products is 3 or 4 for most areas, predominantly corresponding to dataset combinations of EGM (Ecopia, Google and Microsoft) or EGMO (all four data products). Figure 6c shows a different picture for Kinshasa and Brazzaville, with building footprints from all data products limited to the north and east of Brazzaville, and a clear "edge" to this area suggesting a data tiling effect. The remainder of Brazzaville has a data product count of 3, primarily corresponding to Ecopia, Google and OSM (EGO). Similarly for Kinshasa, we find 2 (EG) or 3 (EGO) data products for most grid cells, indicating that building footprints for Kinshasa are missing from the Microsoft building footprint data product. In Brazzaville, Microsoft building footprints (EGMO) are limited to the north and far east of the city, along the Congo River.

Figure 7 summarises the count and combination of building footprint data products for all countries, showing the proportion of grid cells per country with each combination. A blank (white) value indicates that no building footprints are available for a particular data product (e.g. no Microsoft building footprints are available for Cabo Verde). Where 4 data products are available, grid cells with building

footprints can have 1 of 15 possible combinations of data products (x-axis). Grid cells with building footprints from all 4 (EGMO), are the greatest proportion of any combination only for a minority of countries (The Gambia, Comoros, Reunion, Tanzania, Seychelles, Zambia, Botswana, Lesotho and Eswatini). For all other countries, the most common combination of 3 data products per grid cell is Ecopia, Google and Microsoft (EGM), followed by Ecopia, Google and OpenStreetMap (EGO). The most common combination of 2 data products per grid cell is Ecopia and Google (EG). For grid cells with building footprints from only 1 data product, most commonly these are either from Ecopia (E) or Google (G).

As a result of the limited coverage of building footprint data products in Northern Africa, combinations of data products for grid cells in this region are limited. For Algeria, Egypt and Tunisia, the greatest proportion of grid cells with building footprints is for the combination of Google and Microsoft. For Western Sahara, Libya and Morocco, the greatest proportion of grid cells are associated with building footprints from a single source, which for Western Sahara is Ecopia and for Libya and Morocco is Microsoft. In contrast, countries in Southern African have high proportions of grid cells with building footprints from 3 or 4 data products (median value of 69.8% for all countries in the region). There is also a very low proportion of grid cells in countries in the Southern African region with a single building footprint dataset (median value of 18.9%). For the subset of grid cells with building footprints from all 4 data products, the relationships between count and total area of building footprints are further explored in Supplementary Figure A.12.

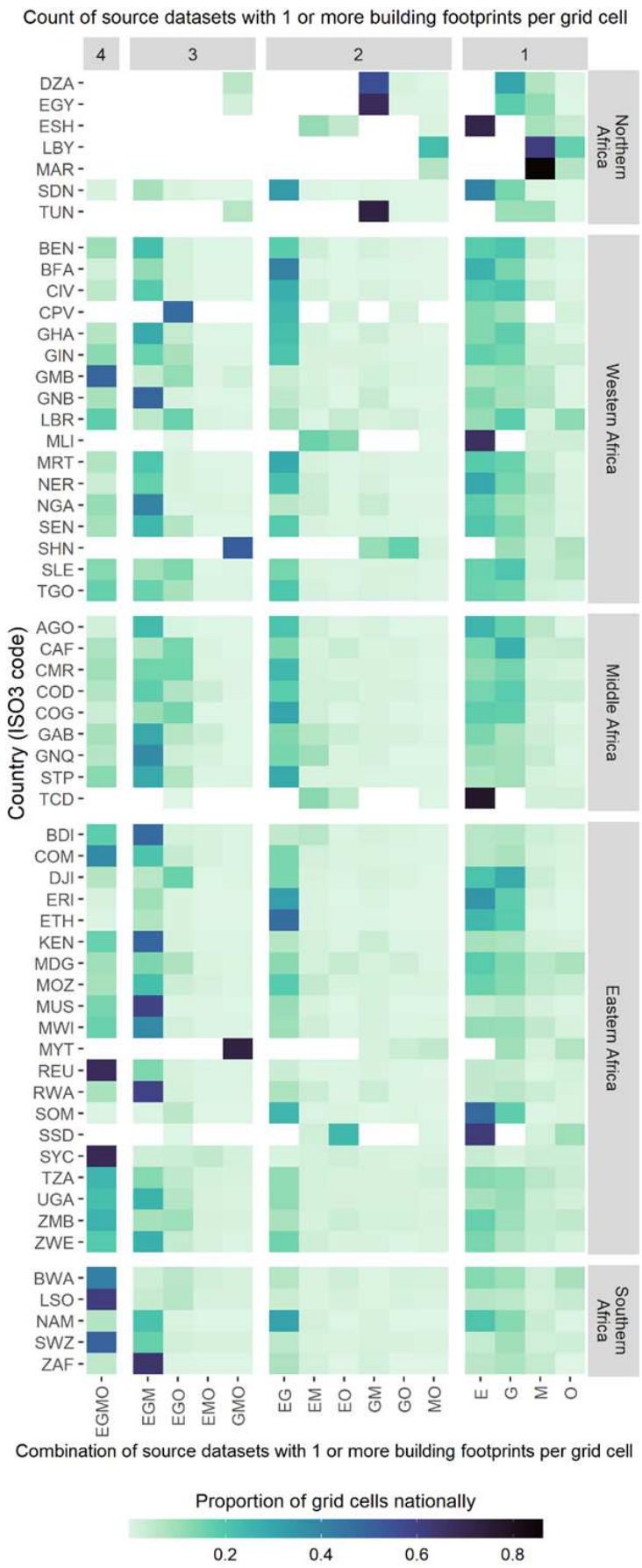

Figure 7: For each country, every grid cell with one or more building footprints (from any dataset) is classified in terms of the number of datasets with one or more building footprints (n=1-4). These are further classified in terms of the combinations of source building footprint datasets (E=Ecopia year 1, G=Google v2, M=Microsoft and O=OSM), where EG represents grid cells with 1 or more building footprints from both Ecopia and Google and no building footprints from Microsoft or OSM. For each country, the proportion of all grid cells with one of more building footprints in each class is shown.

5. Discussion

The growth in the availability of building footprint datasets has provided new opportunities for research and data-informed planning and decision making related to settlements and the built environment (e.g. Buchanan et al. (2020), Huang and Wang (2020), Jochem et al., (2021), Domingo et al. (2023)). The effectiveness of such work however is largely dependent on the accurate mapping of buildings and built infrastructure. Assessing the accuracy of building footprints for countries in Africa though is difficult, given the lack of definitive reference dataset(s). However, understanding how currently available building footprint datasets are similar, how they differ and the extent to which they are comparable is the first step in understanding how such datasets can be used effectively in research and decision-making. In our comparative analysis of building footprint data from four sources (Ecopia, Google, Microsoft and OSM), we have utilised a range of approaches to compare data products in the absence of definitive reference datasets. Our results show that substantial differences exist between data products, and therefore consideration is needed by data users in terms of the suitability of a building footprint dataset for their intended application and geographic context.

Our results show that considerable variations exist between available building footprint data products across countries in Africa. In general, for most locations, Ecopia and Google building footprints consistently have the greatest counts and total area of building footprints when assessed at the national (Table A.1) and subnational level (Figure 1 and Supplementary Figure A.6). There are however multiple African countries for which Ecopia and/or Google building footprints are not available (Figure 1). If locations are stratified in terms of urban/rural settings (Figure 3), then for urban centres and urban clusters, the mean count of building footprints per grid cell is generally highest for Google, followed most often by Ecopia. In rural clusters, there is less variability between data products in terms of mean count per grid cell. As well as consistently having the highest counts and total area of building footprints, Google and Ecopia have the greatest spatial similarity, both in terms of grid cells with at least one building footprint (Figures 3-7) and the correlation between counts and total area of building footprints per grid cell (Supplementary Figure A.12).

Building footprints from OSM and Microsoft generally have lower values in terms of total area and counts (Figure 1, Figure 2, Supplementary Table A.1 and Figure A.6). Comparing counts and total areas nationally, or for AL1 units, shows clear differences compared to Ecopia and Google, however these spatially aggregated values may hide further spatial variation, particularly associated with heterogeneity in the coverage and completeness of Microsoft and OSM building footprint datasets (Figure 2 and Supplementary Figures A.7-A.9). This spatial heterogeneity in coverage means that for some locations, it may be that OSM or Microsoft building footprints are the most accurate and comprehensive data. However, this will depend on the location, context and spatial extent of the area of interest, and for an end user to identify this necessitates a time-intensive manual review of the available building footprint datasets, potentially including comparisons against recent satellite imagery. In general, Microsoft building footprints appear to have better coverage in urban areas than in rural areas (Supplementary Figure A.8), but this is by no means universal, with some major cities and urban areas having no Microsoft building footprints (e.g. Kinshasa, DRC – Figure 6c).

Differences between the four data products could be due to multiple factors, including differences in feature extraction algorithms, differences in what is considered as a building and differences in post-processing. There can also be differences related to the satellite imagery used in feature extraction, such as differences in the age of satellite imagery, imagery resolution, presence of cloud cover or haze, and the spatial extents for which imagery were available/included. The spatial heterogeneity in coverage of Microsoft and OSM building footprint is influenced by several of these factors associated with the methods used in producing the datasets. For example, the tiling effect (and missing tiles) of mosaiced satellite imagery used by Microsoft, and the limited spatial extent of feature digitisation activity by OSM contributors (often through spatially targeted mapping campaigns), are both visible in Figure 6, most clearly for Kinshasa.

For building footprint data products that are created through automated feature extraction, if an area has incomplete coverage with suitable satellite imagery, then the resulting extracted building footprints will also have incomplete coverage. For data users that are interested in using building footprint datasets across large areas, any incompleteness is a potential problem, particularly if it is widespread. This issue is particularly prevalent for Microsoft building footprints where there are seemingly large gaps in satellite imagery coverage, including for some major cities (Figure 6c). For the Microsoft "global" dataset, users can consult the accompanying coverage extent file to identify if a lack of building footprints is due to (i) no processing of satellite imagery in an area (as presumably occurred in the Microsoft building footprints in Kinshasa) or (ii) no buildings detected. However even with this knowledge, the widespread nature of the coverage gaps in the Microsoft data may render the data unusable in some applications. Neither Ecopia nor Google provide coverage extent information but doing so would be beneficial to users.

In contrast to feature-extracted building footprints, the nature of OSM as a VGI data product results in inevitable heterogeneity in coverage and completeness – a topic of ongoing research. Our analysis shows that for all regions in Africa, the count of grid cells with OSM building footprints, along with the count and total area of OSM building footprints is the lowest of any data product (Supplementary Figure A.7). Calculated across all urban centres in sub-Saharan Africa, the count/total area of OSM buildings as a percentage of Ecopia (27%/24%) and Google (22%/26%) is somewhat similar to recent estimates of OSM building completeness for this region (Herfort et al., 2023). In major cities, where it might be assumed that OSM building footprints have reasonable coverage even if not all buildings are digitised, coverage is still very patchy even when all grid cells with one or more building footprints are considered (Figure 6).

Somewhat confusingly, efforts to integrate OSM data with other data sources are not always clearly distinguished and labelled as such. For example, the Daylight Map Distribution (https://daylightmap.org/) integrates Microsoft building footprints and is provided as a basemap in ESRI's ArcGIS software. Arguably, as this basemap is labelled as OpenStreetMap, ESRI users could gain a false impression of OSM building completeness. This is also relevant when considering sources of training data for use with feature extraction algorithms, particularly given limited existing building training datasets for African countries (Wang et al., 2022). If OSM buildings (in locations with comprehensive coverage) are used as training data, considerations around circularity may be needed, especially with a growing number of feature-extracted building footprint datasets that potentially may be integrated into OSM in some way. In addition, recent initiatives to develop open-source algorithms and training datasets are lowering barriers to entry for new building footprint datasets to be developed, particularly in previously data-sparse settings. For example, the Replicable AI for Microplanning (Ramp) open-source deep learning model (Hallas et al., 2022) has been developed specifically for use in low- and middle-income countries. Similarly, BEAM (Building and Establishment Automated Mapper) is an open-source tool for mapping building footprints from drone imagery (UNITAC, 2022).

When building footprints are extracted or digitised from satellite imagery, the building footprint features inherently represent building locations, shapes and extents at the time that the imagery was captured. Without information on imagery dates, data users are unable to assess the time point that the building footprints represent, which for some uses renders the data unusable (e.g. change detection). For some building footprint data products, the satellite imagery time period spans a few years, but for others this can extend to over a decade (Table 1). Ecopia is the only data producer to provide detailed information on the temporal coverage of satellite imagery, with imagery date included as an attribute (Supplementary Figure A.1). In the context of our comparative analysis of data products, the lack of information on imagery dates is therefore a caveat in the interpretation of results.

The inclusion of satellite imagery dates in all building footprint datasets would considerably expand their utility.

Our comparative analysis of four data products has assessed counts and total areas of building footprints at a range of spatial scales, but inevitably has several limitations. Firstly, there are likely differences between the building footprint datasets in terms of what is considered as a building and how individual buildings are delineated. For datasets produced by automated feature extraction, this is determined by the feature extraction and post-processing algorithms, but limited information on this is released by data producers. Some building footprint products permit overlaps between polygons (Google, Microsoft and OSM) while Ecopia doesn't. We have summarised overlaps in the input data files (Supplementary Tables B.3-B.5) but have not accounted for these in calculating gridded metrics of total building area, since the overall impact is most likely small. Secondly, the temporal coverage of satellite imagery likely also varies between datasets and the lack of information on imagery dates means there is potentially large temporal mismatches between datasets, but the extent to which this influences our analysis results is not possible to identify without knowing the imagery dates.

Additionally, confidence scores are only provided for Google building footprints. Given the subjective nature of selecting a confidence threshold value and the fact that confidence scores were only available for a single data product, we opted to not use the confidence score and included all Google v2 building footprints. If a confidence threshold had been applied, then building footprint counts and total area would be reduced. In assessing spatial similarity of building footprint datasets, we utilised a single threshold of building footprint counts and considered all grid cells with 1 or more building footprints, but further work should explore a range of threshold values. Our comparisons between data products focussed on building footprint count and area metrics. Further work should explore calculating geometry and morphology measures, such as those relevant to characterising urban morphology (Jochem and Tatem (2021), Fleischmann et al. (2022)). Furthermore, to understand the impact of utilising different building footprint datasets in a variety of applications, and the extent to which the choice of datasets affects outcomes and results, further work is needed to explore the use of existing building footprint datasets in a range of contexts.

Finally, the lack of definitive reference datasets for buildings across the African continent means that our analysis is limited to comparisons solely between available building footprint data products, all of which have largely been digitised or extracted from satellite imagery. With the growing geographic coverage of building footprint data products, further work should expand this comparison to include building footprints from authoritative sources, with care taken in maximising temporal alignment between datasets. Such comparisons are needed across a diverse range of settlement typologies, as previous comparative work has tended to focus on specific locations (single city or single country and US/Europe-centric) with a typically narrow typology of settlements and urban form.

The currently available building footprint data products provide a snapshot of the built environment, not for a single timepoint, but mosaiced together from many single timepoints, often spanning multiple years. Looking to the future, with the launch of new satellites and sensors, the frequency with which sufficiently high-resolution imagery to map buildings is collected, should only improve. The establishment of a high-quality baseline dataset should enable coarser-resolution imagery to be regularly utilised for change detection, with focussed use of higher-resolution imagery. This provides potential opportunities to move away from a single snapshot and towards regular and dynamic mapping, such is already available for landcover data products (e.g. Google's Dynamic World). Integration with new and additional data sources should also enable enhanced attribution of building footprints, with characteristics such as building use, heights, building materials and addressing.

6. Conclusion

Whilst all building footprint datasets are conceptually trying to represent the same thing, our analysis shows that for African countries, the available building footprints differ considerably in their representation of buildings, with big differences in data products between countries and across urban-rural contexts. This creates a situation where end users need to beware. Given the lack of a definitive reference source against which to assess each dataset, our comparative analysis of these four data products has focussed on highlighting similarities and differences. Ultimately, there is no conclusive best dataset. The four building footprint data products are not interchangeable, and users of these data need to assess the data for suitability in terms of the context, time period, spatial scale and use case that is of interest. There are several concrete steps that producers of building footprints can take to enhance the usability of their products. Firstly, data producers should provide information on the spatial extent of each dataset, including which (if any) areas have been excluded from processing to distinguish between true gaps in coverage and possible errors/omissions. Secondly, information on the date of satellite imagery used for extraction or digitisation should be included as an attribute in each building footprint feature to assess data currency. Finally, the documentation available for all datasets should be enhanced, including details of the data processing and methods. This documentation should include details on building footprint geometry (such as whether overlaps are permitted), details of rates of omission/commission and how that varies across countries and urban-rural contexts.

7. References

1. Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, *8*(5), 232, https://doi.org/10.3390/ijgi8050232.

2. Arribas-Bel, D., & Fleischmann, M. (2022). Understanding (urban) spaces through form and function. *Habitat International*, *128*, 102641, https://doi.org/10.1016/j.habitatint.2022.102641.

3. Barnsley, M. J., & Barr, S. L. (1996). Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. *Photogrammetric engineering and remote sensing*, *62*(8), 949-958.

4. Biljecki, F., & Chow, Y. S. (2022). Global building morphology indicators. *Computers, Environment and Urban Systems*, *95*, 101809, https://doi.org/10.1016/j.compenvurbsys.2022.101809.

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

6. 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.

7. Boo, G., Darin, E., Thomson, D. R., & Tatem, A. J. (2020). A grid-based sample design framework for household surveys [version 1]. *Gates Open Re*s, 4, https://doi.org/10.12688/gatesopenres.13107.1

8. Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., ... & Tatem, A. J. (2022). High-resolution population estimation using household survey data and

building footprints. *Nature communications*, *13*(1), 1330, https://doi.org/10.1038/s41467-022-29094-x.

9. Buchanan, M. K., Kulp, S., Cushing, L., Morello-Frosch, R., Nedwick, T., & Strauss, B. (2020). Sea level rise and coastal flooding threaten affordable housing. *Environmental Research Letters*, *15*(12), 124020, https://doi.org/10.1088/1748-9326/abb266.

10. Cao, Q., Huang, H., Wang, W., & Wang, L. (2023). Characterizing urban densification in the city of Wuhan using time-series building information. *Landscape Ecology*, 1-21. https://doi.org/10.1007/s10980-023-01718-7

11. Chamberlain, H. R., Lazar, A. N., & Tatem, A. J. (2022). High-resolution estimates of social distancing feasibility, mapped for urban areas in sub-Saharan Africa. *Scientific Data*, *9*(1), 711, https://doi.org/10.1038/s41597-022-01799-0.

12. Checchi, F., Stewart, B. T., Palmer, J. J., & Grundy, C. (2013). Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations. International journal of health geographics, 12(1), 1-12, https://doi.org/10.1186/1476-072X-12-4.

13. Chiwele, D., Lamson-Hall, P. & Wani, S. (2022) Informal settlements in Lusaka. *International Growth Centre and UN-Habitat*, https://www.theigc.org/wp-content/uploads/2022/02/Informal-settlements-in-Lusaka-web.pdf.

14. Corbane, C., Pesaresi, M., Kemper, T., Politis, P., Florczyk, A. J., Syrris, V., ... & Soille, P. (2019). Automated global delineation of human settlements from 40 years of Landsat satellite data archives. *Big Earth Data*, *3*(2), 140-169, https://doi.org/10.1080/20964471.2019.1625528.

15. Darin, E., Kuépié, M., Bassinga, H., Boo, G., Tatem, A. J., & Reeve, P. (2022). The Population Seen from Space: When Satellite Images Come to the Rescue of the Census. *Population*, 77(3), 437-464, https://doi.org/10.3917/popu.2203.0467.

16. Domingo, D., Van Vliet, J., & Hersperger, A. M. (2023). Long-term changes in 3D urban form in four Spanish cities. *Landscape and Urban Planning*, *230*, 104624, https://doi.org/10.1016/j.landurbplan.2022.104624.

17. Donnay, J. P., Barnsley, M. J., & Longley, P. A. (Eds.). (2000). *Remote sensing and urban analysis: GISDATA 9*. CRC Press.

18. Dooley, C. A., Tatem, A. J., & Bondarenko, M. (2020). Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. *University of Southampton: Southampton, UK*, https://doi.org/10.5258/SOTON/WP00677 [DATASET]

19. Dukai, B., Ledoux, H., & Stoter, J. E. (2019). A multi-height LoD1 model of all buildings in the Netherlands. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 51-57, https://doi.org/10.5194/isprs-annals-IV-4-W8-51-2019.

20. Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., ... & Strano, E. (2017). Breaking new ground in mapping human settlements from space–The Global Urban Footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, *134*, 30-42, https://doi.org/10.1016/j.isprsjprs.2017.10.012.

21. Facebook, 2019. RapiD – The OpenStreetMap editor driven by open data, AI, and supercharged features (GitHub Repository), https://github.com/facebook/Rapid.

22. Fleischmann, M., & Arribas-Bel, D. (2022). Geographical characterisation of British urban form and function using the spatial signatures framework. *Scientific Data*, *9*(1), 546, https://doi.org/10.1038/s41597-022-01640-8.

23. Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2022). Methodological foundation of a numerical taxonomy of urban form. *Environment and Planning B: Urban Analytics and City Science*, *49*(4), 1283-1299, https://doi.org/10.1177/23998083211059835.

24. Foks, N., Heris, M. P., Bagstad, K. J. & Troy, A. A (2020), A Code for Rasterizing Microsoft's Building Footprint Dataset. U.S. Geological Survey https://doi.org/10.5066/P9XZCPMT.

25. Fugate, D., Tarnavsky, E., & Stow, D. (2010). A survey of the evolution of remote sensing imaging systems and urban remote sensing applications. In Rashed, T., Jürgens, C. (Eds.)*, Remote sensing of urban and suburban areas* (pp. 119-139), *Remote Sensing and Digital Image Processing*, vol 10. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4385-7_7.

26. GADM (2022) Global Administrative Areas Database, version 4.1, https://gadm.org/download_world.html [DATASET]

27. Gibson, L., Cicione, A., Stevens, S., & Rush, D. (2022). The influence of wind and the spatial layout of dwellings on fire spread in informal settlements in Cape Town. *Computers, Environment and Urban Systems*, *91*, 101734, https://doi.org/10.1016/j.compenvurbsys.2021.101734

28. Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*, 211-221, https://doi.org/10.1007/s10708-007-9111-y.

29. Hallas, M., Price, R., & Haithcoat, J. (2022, December). Replicable AI for Microplanning (ramp): Democratizing Geospatial Data Science for Global Health. In *AGU Fall Meeting Abstracts* (Vol. 2022, pp. GC42E-0755).

30. Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS International Journal of Geo-Information*, *2*(4), 1066-1091, https://doi.org/10.3390/ijgi2041066.

31. Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific reports*, *11*(1), 3037, https://doi.org/10.1038/s41598-021-82404-z.

32. 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.

33. Heris, M. P., Foks, N. L., Bagstad, K. J., Troy, A., & Ancona, Z. H. (2020). A rasterized building footprint dataset for the United States. *Scientific data*, *7*(1), 207, https://doi.org/10.1038/s41597-020-0542-3.

34. Huang, X., & Wang, C. (2020). Estimates of exposure to the 100-year floods in the conterminous United States using national building footprints. *International Journal of Disaster Risk Reduction*, *50*, 101731, https://doi.org/10.1016/j.ijdrr.2020.101731.

35. Jensen, J. R., & Cowen, D. C. (1999). Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric engineering and remote sensing*, *65*, 611-622.

36. Jochem, W. C., & Tatem, A. J. (2021). Tools for mapping multi-scale settlement patterns of building footprints: An introduction to the R package foot. *PLoS One*, *16*(2), e0247535, https://doi.org/10.1371/journal.pone.0247535.

37. Jochem, W. C., Leasure, D. R., Pannell, O., Chamberlain, H. R., Jones, P., & Tatem, A. J. (2021). Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science*, *48*(5), 1161-1179, https://doi.org/10.1177/2399808320921208.

38. Lazarus, E. D., Limber, P. W., Goldstein, E. B., Dodd, R., & Armstrong, S. B. (2018). Building back bigger in hurricane strike zones. *Nature Sustainability*, *1*(12), 759-762, https://doi.org/10.1038/s41893-018-0185-y.

39. Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences*, *117*(39), 24173-24179, https://doi.org/10.1073/pnas.1913050117.

40. Li, J., Huang, X., Tu, L., Zhang, T., & Wang, L. (2022). A review of building detection from very high resolution optical remote sensing images. *GIScience & Remote Sensing*, *59*(1), 1199-1225, https://doi.org/10.1080/15481603.2022.2101727.

41. Microsoft, 2018. Computer generated building footprints for the United States (GitHub Repository), https://github.com/microsoft/USBuildingFootprints. [DATASET]

42. Microsoft, 2019a. Computer generated building footprints for Canada (GitHub Repository), https://github.com/microsoft/CanadianBuildingFootprints. [DATASET]

43. Microsoft, 2019b. Uganda Tanzania Building Footprints (GitHub Repository), https://github.com/microsoft/Uganda-Tanzania-Building-Footprints (accessed 10th January 2023). [DATASET]

44. Microsoft, 2021a. Kenya Nigeria Building Footprints (GitHub Repository), https://github.com/microsoft/KenyaNigeriaBuildingFootprints [DATASET] (accessed 10th January 2023).

45. Microsoft, 2021b. South America Building Footprints (GitHub Repository), https://github.com/microsoft/SouthAmericaBuildingFootprints. [DATASET]

46. Microsoft, 2022. Worldwide building footprints derived from satellite imagery (GitHub Repository), https://github.com/microsoft/GlobalMLBuildingFootprints [DATASET] (accessed 10th January 2023).

47. Nirandjan, S., Koks, E. E., Ward, P. J., & Aerts, J. C. (2022). A spatially-explicit harmonized global dataset of critical infrastructure. *Scientific Data*, *9*(1), 150, https://doi.org/10.1038/s41597-022-01218-4.

48. Park, Y., & Guldmann, J. M. (2019). Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, environment and urban systems*, *75*, 76-89, https://doi.org/10.1016/j.compenvurbsys.2019.01.004.

49. Pesaresi, M., Florczyk, A., Schiavina, M., Melchiorri, M., & Maffenini, L. (2019). GHS settlement grid, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015), R2019A. *European*

Commission, Joint Research Centre (JRC), https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218

50. Price, R., & Hallas, M. (2019, December). Mapping Every Building and Road in sub-Saharan Africa. In *AGU Fall Meeting Abstracts* (Vol. 2019, pp. IN41A-02).

51. Rae, A. (2015). *Urban footprints: some building outline data sources*, Retrieved from http://www.undertheraedar.com/2015/07/urban-footprints-some-building-outline.html (accessed 5th January, 2023).

52. Robinson, C., Nsutezo, S. F., Ortiz, A., Sederholm, T., Dodhia, R., Birge, C., ... & Ferres, J. M. L. (2023). Rapid building damage assessment workflow: An implementation for the 2023 Rolling Fork, Mississippi tornado event. *arXiv preprint,* https://doi.org/10.48550/arXiv.2306.12589.

53. Sanchez-Cespedes, L. M., Leasure, D. R., Tejedor-Garavito, N., Amaya Cruz, G. H., Garcia Velez, G. A., Mendoza, A. E., ... & Ospina Bohórquez, M. (2023). Social cartography and satellite-derived building coverage for post-census population estimates in difficult-to-access regions of Colombia. *Population Studies*, 1-18, https://doi.org/10.1080/00324728.2023.2190151.

54. Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., ... & Quinn, J. (2021). Continental-scale building detection from high resolution satellite imagery. *arXiv preprint,* https://doi.org/10.48550/arXiv.2107.12283.

55. Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., ... & Dang, H. A. H. (2017). Mapping the world population one building at a time. *arXiv preprint,* https://doi.org/10.48550/arXiv.1712.05839.

56. UNITAC. (2022). BEAM (Building & Establishment Automated Mapper) User Manual, Version 11.2022, *UNITAC Hamburg (United Nations Innovation Technology Accelerator for Cities)*, https://unitac.un.org/sites/unitac.un.org/files/beam_user_manual-v2.pdf, (accessed 1st August 2023).

57. Wang, J., Fleischmann, M., Venerandi, A., Romice, O., Kuffer, M., & Porta, S. (2023). EO+ Morphometrics: Understanding cities through urban morphology at large scale. *Landscape and Urban Planning*, *233*, 104691, https://doi.org/10.1016/j.landurbplan.2023.104691.

58. Wang, J., Georganos, S., Kuffer, M., Abascal, A., & Vanhuysse, S. (2022). On the knowledge gain of urban morphology from space. *Computers, environment and urban systems*, *95*, 101831, https://doi.org/10.1016/j.compenvurbsys.2022.101831.

59. Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., ... & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, *115*(14), 3529-3537, https://doi.org/10.1073/pnas.1715305115.

60. Zhang, Y., Zhou, Q., Brovelli, M. A., & Li, W. (2022). Assessing OSM building completeness using population data. *International Journal of Geographical Information Science*, *36*(7), 1443-1466, https://doi.org/10.1080/13658816.2021.2023158.

61. 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.

Acknowledgements

This study is an output of the WorldPop Research Group at the University of Southampton. This work was part of the GRID3 project with funding from the Bill and Melinda Gates Foundation and the United Kingdom's Foreign, Commonwealth & Development Office [grant numbers: INV-009579 and INV-045694]. In addition to WorldPop, GRID3 project partners included the United Nations Population Fund (UNFPA), Center for International Earth Science Information Network in the Columbia Climate School at Columbia University, and the Flowminder Foundation.

Author contributions

Heather Chamberlain: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft**. Edith Darin**: Methodology, Writing - Review & Editing**. Ademola Adewole**: Writing - Review & Editing**. Warren C. Jochem**: Software, Writing - Review & Editing**. Attila Lazar**: Funding acquisition, Writing - Review & Editing**. Andrew Tatem**: Supervision, Funding acquisition, Writing - Review & Editing

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendicesBuildingFootprintcomparisonv20PREPRINT.pdf