GRID3

# Core Spatial Data for Sub-Saharan Africa:

## A report on key spatial data available for development practitioners

September 2021

# Table of Contents

# Introduction

This report was drafted by the Geo-referenced Infrastructure and Demographic Data for Development (GRID3) programme.[1] GRID3 is founded on the premise that there needs to be a set of core spatial data that:

- are capable of being affordably generated at high quality in any country in the world, regardless of baseline capabilities
- enable dramatic increases in effectiveness of development and humanitarian interventions at a suitable level of quality
- achieve their maximum impact when integrated and made interoperable with the rest of a country's data ecosystem

This is part of a broader effort to help users make informed choices about which spatial datasets to incorporate into their planning, assessment, and delivery tools. While information about core spatial data layers options is already available, it is often not very helpful to development practitioners (and the technicians who support them) in choosing among the data layer options. Default metadata fields are often not well understood, are often not highly relevant, and often omit some of the most important information needed for decision making. As geodata are increasingly integrated into planning and management tools, it is absolutely critical that the data quality and format fit the intended use. GRID3's efforts are aimed at helping people make effective choices along these lines.

## Summary

GRID3 created a curated list of datasets within each of five core data layer themes: population, settlements, subnational administrative boundaries, health facilities, and roads. To be included, each dataset had to meet the following criteria:

- publicly accessible, without restriction on use
- recent, not more than three years old or projected to present day unless it is official government data
- complete coverage of the country in question
- comes with complete documentation detailing the input source and applied method/edits

The evaluation of whether a dataset met these criteria varied according to the core spatial data theme.

This report provides an overview of selected core spatial data that are available and accessible in multiple sub-Saharan African countries. The report lists these datasets and their metadata. This report will not recommend a single dataset as being superior to others; such a recommendation is not possible, as the needs of development practicioners differ from case to case. The report will, however, allow users to compare data descriptions and metrics so that they can make an informed decision about which dataset best suits their purposes.

1. The initiative is funded by the Bill & Melinda Gates Foundation and the United Kingdom's Foreign Commonwealth & Development Office. It is implemented by WorldPop at the University of Southampton, Flowminder Foundation, the United Nations Population Fund, and the Center for International Earth Science Information Network (CIESIN) at Columbia University.

## Intended Stakeholders

This report is expected to be of use to many geodata users in the development field, such as urban planners, disaster managers, health system strengtheners (e.g. COVID-19 response planners), and many others. An expected user is an individual collecting and evaluating data to be used to plan service delivery, and make decisions about target populations/access to care/overall resource planning and prioritisation. The assumption is that the user has at least some knowledge around the importance and use of spatial data for decision making.

In order to choose the most appropriate datasets, comparisons can be conducted within metadata to assist the user in balancing considerations (such as recency and resolution) and apply their own weights to these metrics, based on the application. Datasets listed in each of the five categories can be combined to create secondary datasets (e.g. administrative units with population data).

It is envisioned that data would benefit from a general overview of the data via a short explanation and comparison tables (this report) and a more detailed Data Explorer (to be created).

# The Core Data Themes

This report organises core spatial data into five themes:

- Population
- Settlements
- Subnational Administrative Boundaries
- Health Facilities
- Roads

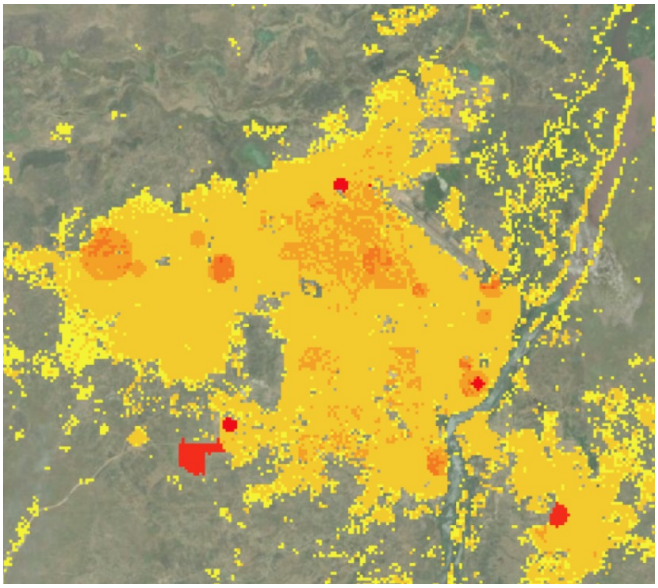
A key value of this collection is that it allows a side-by-side comparison of multiple datasets within a theme. This can help data users find the most appropriate dataset for their specifc (e.g. COVID-19-related) use case. The dataset overview table can be found here, but the next sections describe and discuss each.

## Population Data

The population data provides information on how people are distributed in space. There are several population datasets available and each of them uses different methodologies to model numbers and distributions. The building block of this information is typically national census data although there are exceptions. A national census seeks to count the total population of a country at a spatial resolution limited to pre-defined administrative boundaries, i.e. enumeration areas. A census is onerous and expensive to conduct, and therefore undertaken only every 10 years. The various population datasets selected for inclusion in the collection use different methods and data sources (e.g. census, survey data, registers). They estimate population distributions across both space and time. Ancillary data (such as roads, elevation, water bodies, and building footprints from high-resolution satellite imagery) may be used to inform the population distribution model. Thus, population estimates vary across datasets depending on input data and methodology; these variations are important considerations when deciding which dataset is the most appropriate for a particular application. It is important to highlight that these population estimation methods and datasets do not replace the population and housing census.

*Overview of which datasets were chosen and why*

The datasets included in the population core spatial data theme meet criteria considered to be important in responding to development challenges, including COVID-19 response. All are publicly available, use census information (if available), have nearly complete coverage of sub-Saharan Africa (except the bespoke GRID3 population models), have clear and complete metadata and documentation, and represent a population estimate from within the last three years (although some official estimates may not). Since censuses are the main source of population counts used to produce these datasets, the accuracy of the estimates depends heavily on the recency of each country's census. They differ mainly in input datasets used, creation methodology, and spatial resolution/ format of the final product.



Gridded population estimates for South Sudan

### Unadjusted and adjusted population estimates

The producers of gridded population datasets may provide "unadjusted" or "adjusted" population estimates. Unadjusted gridded population estimates are based on raw subnational census data or census-based projected estimates. To produce adjusted estimates, the unadjusted subnational data or projected estimates are multiplied by a country-specific adjustment factor to force the resulting adjusted subnational estimates to sum to the national estimate provided in the United Nations WorldPopulation Prospects (WPP). The population estimates provided in the WPP often correct for over- or under-reporting in the nationally-reported figures and are broadly accepted as accurate among the international community; thus, adjusting to the UN estimates may improve the reliability of the gridded datasets. This adjustment may however not be in line with government preference. To meet the needs of users who may prefer working with unadjusted estimates, most producers of gridded population data release both adjusted and unadjusted datasets.

The next sections introduce the selected population datasets and their methodologies. It is worth highlighting that most of these belong to the so-called 'top-down' population estimation method, where trusted subnational administrative totals (official or projected) are disaggregated to individual grid cells. The only exception is the GRID3/WorldPop country-specific models, that use bespoke statistical models to extrapolate census and survey data and thus create population estimates for the entire country with high resolution and quantified uncertainty (i.e. bottom-up models).[2]

It is worth noting that the Gridded Population of the World (GPWv4), WorldPop, and Facebook datasets described below use the same subnational census inputs pre-processed by CIESIN for GPWv4 revision 11, but they differ in the methodologies used to allocate the population estimates to grid cells.

The datasets selected for the population theme are listed below and an overall description is added to each dataset.

2. Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman and A. J. Tatem (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. DOI: 10.1073/pnas.1715305115. http://www.pnas.org/content/pnas/115/14/3529.full.pdf

- **Official National Population Statistics**

  This is the authoritative national population dataset (whether census, register, or survey). These datasets will always serve as relevant points of comparison and for many uses will be a required layer, despite their limited temporal and spatial resolution. The data may be available from national agencies as tables (e.g. Excel, CSV, PDF) or joined to boundaries at a matching administrative level (e.g. shapefile, geodatabase). In some cases, the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) matches the tabular data to boundaries, performs some quality checks and validation, and releases the dataset as a Population Statistics Common Operational Dataset (COD-PS) on the Humanitarian Data Exchange (HDX) platform.

- **UN OCHA Population Statistics Common Operational Datasets (COD-PS)**

  Population Statistics Common Operational Datasets (COD-PS) include: (1) demographic tables in spreadsheet (XLSX and CSV) formats, and (2) gazetteers of feature names and P-codes (place codes). COD-PS datasets can be linked by database or GIS to COD-AB (administrative boundary) datasets, when available, using the P-codes as a key. The UN OCHA COD Portal provides information on the status of the datasets for each country and links to the data on the HDX platform (https://data.humdata.org/dataset).

  Population statistics delivered as core Common Operational Datasets have been evaluated according to UN OCHA Field Information Services (FIS) criteria to be "fully usable," "partly usable," or "requires improvement" to support humanitarian response. As of 2018, the United Nations Population Fund (UNFPA) and UN OCHA formally agreed that UNFPA Regional Offices (together with HQ and CO colleagues) will identify the "best-available" sex-and-age-disaggregated dataset for each country and discuss their findings with UN OCHA's Information Management (IM) at the regional (and country) level. The agreed dataset is then presented to the Information Management Working Group (IMWG) and IM network for further validation and adoption. If adopted, the dataset becomes the official population statistics COD (COD-PS) for a particular country, and in most cases the COD-PS will be publicly available at HDX. The process and standards for COD-PS are outlined on the COD Confluence Wiki. For some countries, the COD population data will align with WPP estimates.

  The preferred candidate dataset for a COD-PS is the most recent subnational population projections prepared by the National Statistics Office (or a similar authority). If government data are available but outdated, or unavailable (or too outdated), then UNFPA projections or mode- based estimates are produced using the Bayesian population projection framework. This ensures the methodological approach used to construct COD-PS at some subnational level (i.e. ADM-1 and below) is consistent with the approach of the UN's official national population projections (known as the WorldPopulation Prospects).

  "As it is a humanitarian tool, the COD-PS is not required to be an official statistical output constructed according to the international standards of official statistics. Rather it is intended to be updated annually according to the best-available humanitarian data standard, or as humanitarian needs and priorities change, and allows for the input data and estimation/ projection methods to be of a lower standard than official statistical standards."[3]

3. See UN OCHA. IM Toolbox. Available at https://humanitarian.atlassian.net/wiki/spaces/imtoolbox/pages/2493349951/Population+Statistics+COD+COD-PS.

## Gridded Population of the World (GPW) UN WPP-Adjusted Population Count, v4.11, Gridded Population of the World (GPW) Population Count, v4.11, and Basic Demographic Characteristics, v4.11

The GPWv4.11 population count datasets provide gridded versions of national census data for the years 2000, 2005, 2010, 2015, and 2020 at a resolution of 30 arc-second (~ 1km). They are made available by the Center for International Earth Science Information Network (CIESIN) Columbia University, and the NASA Socioeconomic Data and Applications Center (SEDAC). These datasets can be downloaded and used for historic data investigations. Counts for the different years are arrived at through use of official census data or official projections where they are available, and careful extrapolation at the smallest geographic unit available where official projections are insufficient. Two versions are provided, one that matches each country's official population total, and one that matches totals published in the 2015 Revision of the United Nations WorldPopulation Prospects (UN WPP).[4] No other manipulation or modelling of the official data takes place.

Though there are more recently published population datasets available, GPWv4 was selected because it is a simple spatial representation of population distribution. The only inputs are census population counts and census geographies. No covariates are used to model the distribution of the population, a factor that allows users to integrate the dataset with other core spatial datasets without risk of endogeneity. This reflects the original motivation for creating the dataset, which was to provide a spatial population layer for use in analysis with data from the physical sciences.

To produce the Population Count dataset, census inputs at the most detailed administrative level possible were collected from results of the 2010 round of censuses (2005-2014). Subnational growth rates were calculated from the most recent census and a previous census year and used to extrapolate the subnational census data to produce estimates for the years 2000, 2005, 2010, 2015, and 2020. A proportional allocation gridding algorithm, utilising approximately 13.5 million national and subnational administrative units, was used to allocate the population estimates to 30 arc-second (~ 1km) grid cells, essentially distributing the subnational estimates evenly within the administrative units.

To produce the UN WPP-Adjusted data set, the population count estimates were adjusted such that the dataset's country totals match the national estimates reported in the 2015 UN WPP, and then the same gridding algorithm was used to disaggregate the adjusted estimates.

The GPWv4.11 Basic Demographic Characteristics dataset provides gridded estimates of male and female populations by age for the year 2010. These estimates are derived by applying the calculated proportions of males and females in each 5-year age group (from ages 0-4 to 85+) for each country's census year to the 2010 estimates of total population (the unadjusted Population Count dataset). The estimates are gridded as described above.

The main disadvantage of the GPW dataset is that the algorithm may allocate population to grid cells that do not actually have persons living there, especially if the input data are at a low resolution, i.e. at a low administrative level. The simplicity of the model means that people may be allocated to a grid cell regardless of the presence of buildings (indicating settled areas) in the grid cell. Thus, the accuracy of the GPW model for a given country depends on the resolution of the input data for that country.

4. UN WPP estimates are based on all available sources of data on population size and levels of fertility, mortality, and international migration for 235 distinct countries or areas comprising the total population of the world.

- **WorldPop Constrained Individual Countries UN Adjusted Population Counts, Constrained Individual Countries Population Counts, and Age and Sex Structures (Constrained)**

  The WorldPop Constrained Population Count datasets provide adjusted and unadjusted gridded population estimates for all SSA countries for the year 2020. The adjusted estimates are based on WorldPop's own population projections using the latest census dataset, whereas the "UN Adjusted" dataset presents estimates adjusted to match the UN WPP national population estimates for that year. The Random Forest method is used to disaggregate the projected administrative population totals to 100m x 100m resolution (3 arc-second) using relevant global geospatial datasets to inform the spatial distribution of the population. In addition to the administrative level census data and geospatial covariate datasets, WorldPop uses a very high-resolution, imagery-based building footprint layer to only allocate population to settled pixels and to refine the population distribution within settlements.

  The WorldPop Age and Sex Structures dataset provides 2020 estimates of the number of males and females per grid cell broken down by age groups (0-1 and 5-year groups up to 80+). Estimates are based on data from national censuses and household survey data conducted as close to 2020 as possible. A table of proportions of the total population by gender and age groups was produced for each country, for the smallest possible geographic unit, and used to produce a raster layer of subnational age/sex structure.[5] The WorldPop UN Adjusted Population Count rasters were multiplied by the age/sex structure rasters to produce rasters of population counts by age and sex. The data are constrained to 3 arc seconds (~100 m) grid cells mapped as settled.

  The WorldPop Constrained 2020 datasets were selected because allocating population based on relevant covariates to known built-up areas improves the spatial representation of population distribution. It avoids the problem of allocating population across what are actually uninhabited areas, as happens if one evenly allocates population across each census unit's area.

- **Facebook High Resolution Population Density Maps + Demographic Estimates (AKA HRSL)**

  The Facebook dataset offers gridded population estimates (as counts) for 2020 at 1 arc-second (~ 30m) for 46 of the 48 SSA countries (Somalia and South Sudan are currently missing). The dataset is the result of a collaboration between Facebook Connectivity Lab and CIESIN. The methodology employs a machine learning algorithm to extract buildings from commercially available high-resolution satellite imagery and classifies 30m grid cells with buildings as settled. Proportional allocation is then used to evenly distribute census-based population estimates (adjusted to the UN WPP) from CIESIN's GPWv4r11 to these settled grid cells, regardless of the number or size of the building(s). No other covariates are used. Version 1.5 (forthcoming) will include updates for all countries based on more-recent imagery and a refined algorithm that improves the detection of buildings.

5. See Pezzulo, C. et al. (2017) Available at: https://www.nature.com/articles/sdata201789

- **WorldPop/GRID3 bespoke Population Estimates**

  Country-specific gridded population estimates are special products where the inputs and methods are tailored to the specific country in question. One might need the more precise disaggregation of a trusted admin total, maybe matching it to the official statistics (top-down method), or new, reliable estimates are desired due to the lack of suitable census data (bottom-up method). In both cases, the best available, highest resolution population data acquired from non-publicly accessible government data sources and country specific covariates are used. Typically, due to the country specific inputs, these custom models are more accurate than the top-down models listed above even when the census data are recent. A bottom-up approach uses detailed survey data and geospatial datasets to create a bespoke statistical model that predicts the high resolution population distribution in areas with no observations and with quantified estimates of uncertainty. This approach has the potential to achieve higher levels of accuracy than the top-down methods, but accuracy depends on the input data quality and representativeness. The cost of production of the bottom-up models (in terms of time and effort) is however higher. The WorldPop bespoke top-down and bottom-up country methods use similar geospatial covariates for modelling, although the bottom-up models typically use a considerably larger set of covariates. WorldPop currently offers bespoke gridded population data products for eleven SSA countries, most of which were produced by the GRID3 programme (many of those in collaboration with national statistical offices). These bespoke products are especially useful if there are geographic gaps in census enumeration or if the census was not conducted recently. The WorldPop/GRID3 bespoke population estimates can be downloaded from the WorldPop Open Population Repository and the GRID3 Data Hub.

- **Other Population Datasets**

  A few additional datasets were considered and ultimately not selected for this collection for varying reasons. The GHS Population Grid (GHS POP), part of the Global Human Settlement Layer (GHSL) collection produced by the Joint Research Council (JRC) allocates GPWv4r11 to built areas, as HRSL does, but uses much coarser remote sensing data to identify built-up areas and therefore delineates settlements with less precision (as well as omits entirely many smaller settlements). Landscan, produced by Oak Ridge National Laboratory, is freely available only to educational/research institutions. All other users must pay for the data. Similarly, Esri's 2016 WorldPopulation Estimate dataset is available only to ArcGIS users.

***Which metrics/metadata were chosen and why***

The complexity involved in the creation of these datasets can be difficult to digest. Nonetheless, a user will often find themselves faced with the choice of several datasets; there is a need to compare them and make an informed choice that can be articulated and justified. While the POPGRID tool attempts to summarise the differences and provide information and links to download documentation and data, this report and the upcoming GRID3 tool seeks to provide comparison information at a more granular level. With this in mind, key metadata selected for inclusion in the comparison (such as resolution, time period of the data, spatial coverage, update cycle, license, and methodology) are listed for comparison.

***Detailed overview of metadata comparison***

The table linked below shows the above-mentioned datasets compared to one another according to their metadata and metrics. This table gives a general overview only. Please see the link to the table here.

## Settlements Data

COVID-19 response and mitigation measures need to take place at all levels of government. In most countries, the spatial unit below the admin unit 4 or 5 is the village/settlement level. Often, settlements are grouped to form health facility catchment areas, but catchment area maps or settlement maps are unavailable at the country level. For this reason, an overview of existing settlement data is included here.

***Overview of which datasets were chosen and why***

Both polygon and point layers were chosen as the settlement layers. The polygons can show the size of the settlements but may not include names. The points typically show an approximate location of a settlement (not necessarily the center) and can include at least one of the village's names (if there is more than one). The following layers represent settlement data that are available for multiple countries in SSA. This section also includes building footprints that were released by Google in July 2021.

The datasets included are listed below and described in detail.

- **Geographic Names for Geopolitical Areas from GNS (National Geospatial-Intelligence Agency, NGA)**

  GNS (The GEOnet Names Server) of NGA is the official repository of standard spellings of all foreign geographic names, approved by the United States Board on Geographic Names (US BGN). These may not be in line with local spellings. The NGA prepares country-based policies on geographic names standardisation for all countries. These country policies describe the methods that are used to achieve standard spellings of geographic names in a given country. The standard spellings are approved for use in U.S. government publications. The individual country policies, the most recent updates, and the next estimated updates are all provided on the NGA homepage.

  Datasets are available from satellites (remote-sensed data), scientific sampling campaigns, intergovernmental agreements, and from Central Intelligence Agency (CIA) maps available to the general public. Higher-resolution maps for areas with limited GIS capabilities (e.g. SSA) are not available. Additional data sources include administrative and economic records, newspapers, and social media, or model-based investigations carried out by NGA (regarding disease outbreaks, and food/water scarcity). The database contains information about location, administrative division, and quality of the geographic feature, such as feature effective date and termination date, date of name modification, and feature modification (if a new feature has been added to the dataset).

  Historical features are those that have been disestablished (e.g. a former first-order administrative division), those that are destroyed by human or natural means and are no longer locatable (e.g. a town permanently flooded by a river impoundment), or whose existence has been otherwise invalidated. The date in the “Terminate Date” field in the Geographic Names Database indicates historical features. This date may be an effective date of a documented change or the date of the database update.

- **GRID3 Settlement Extents**

GRID3 settlement extents constitute a comprehensive set of settlement polygons per country. This work has been undertaken as part of the GRID3 programme.

The settlement extents and classification are derived solely from the Maxar building footprints, and no ancillary datasets are used. The center points of building footprint features are converted to a 3 arc-second raster grid of building densities. Shell-up contours are then generated using the building density grid to delineate settled vs. unsettled areas. The shell-up method includes contours that start at the lower bounds, but includes all grid cells with building densities to the upper bounds of the grid. For example, a shell up contour of 10 would include all grid cells with a building density of 10 or more. Contours with a building density of one or more are used to create the settlement extent polygons.

***Comparison of GRID3 and NGA data quality***

Only 46 percent of the NGA settlement points fall within the GRID3 settlement extent polygons. Most of the discrepancies were observed in rural areas. The settlement structure in rural areas differs from that of large settlements; the houses are very small and not always captured by the methods used to generate building footprints. Sometimes in rural areas a settlement point falls between the house groups that are part of the same settlement. If the distance among those building groups exceeds 500m, the settlement point will not be validated by the GRID3 contour polygons. Also, the discrepancies may be caused by the update methods applied by NGA (see 5.1.3); a settlement will be declared as a historic place if its existence has been invalidated. In countries experiencing unrest, the pace of updates may be slow.

- **OpenStreetmap (OSM)**

Both settlement points with names and settlement extents are available on the OSM website. OSM data are mainly collected and maintained by volunteers. In some countries, public authorities also provide data. OSM data can be more up-to-date than data from public authorities because of the continuous maintenance by volunteers, but may not be as comprehensive or officially endorsed. Updates are sometimes made daily, while pre-processed data are updated at a slower pace. The most recent data can be downloaded immediately; however, older versions of the data are still available. Points of interest (shops, for example) might be included before they are open. In contrast, locations that are not open to the public may be missing from OSM. Density of OSM locations in any given area depends on the number of volunteers.

The OSM settlement points layer (gis_osm_places_free_1) contains the settlement name and settlement class (fclass, such as city, village, town, suburb, hamlet, farm, and population counts) for some of the locations. The gis_osm_places_free_1 polygon layer represents the settlement extents and contains the same attributes as the settlement point layer. It usually contains a very small number of locations. A much larger number of settlement extents listed as residential are contained in the OSM land use layer, but the settlement names are rarely available.

- **Google Building Footprints**

  In July 2021 Google released a building footprint layer for 50 countries in Africa. This layer consists of 516 million building detections, across an area of 19.4 million km2. The building footprints are available for download in the polygon geometry. Its attributes include a confidence score indicating the likelihood that something is a building and a Plus Code corresponding to the centre of the building. There is no information about the type of building, its street address, or any other details. The dataset will be updated as new images become available. The data have a relatively open Creative Commons Attribution (CC BY-4.0) license, so the data can be shared (and even used commercially). This is not the case for the Maxar building footprint; hence the Maxar dataset was not included in this report.

- **Government Settlement Data**

  Each country will have official city and village data. However, in SSA much of the land is under traditional leadership and hence rural settlements have not been "officially" mapped and so are often only available as a list of names without coordinates. Villages in the region may have more than one name. Typically, comprehensive village geodata are unavailable and were hence not included in the overview table (but are mentioned here for the sake of completeness).

***Which metrics/metadata were chosen and why***

Not many comprehensive settlement datasets exist in the open domain. As with the other datasets in this report, the comparison is done using the metrics' accessibility and use (public accessibility, use constraints, relationship to corresponding official dataset), recency (vintage of data product and source data), coverage completeness (i.e. more than 10 countries in SSA), and documentation completeness (input source data, methods/edits).

***Detailed overview of metadata comparison***

The table linked to below shows the above-mentioned datasets and compares the corresponding metadata and metrics. Please see the table here.

## Boundary Data

In most countries, planning and budgeting is done at various administrative levels. In some countries, these align with health facility catchment areas. Boundary data for most countries exist in the open domain. They mostly align with official government boundaries (some more than others).

***Overview of which datasets were chosen and why***

For the purpose of providing users with access to the available open boundaries data, the following datasets are included: authoritative data from United Nations Second Administrative Level Boundaries (UN SALB), COD from (UN OCHA), operational or non-authoritative boundary data from the Global Administrative Area Database (GADM), and OSM.

- **UN SALB**

  The SALB programme, in close collaboration with national geospatial information authority of each member state of the United Nations, aims to make available a global repository of authoritative information and geospatial data about the administrative unit structure of countries, down to the second subnational level and through time. These data have been validated and endorsed by the national napping authorities and thus can be used with confidence for official representation of boundary delineation up to the second administrative level. At this time, UN SALB holds boundary data for 29 SSA countries with the goal of publishing data for 15 more countries by December 2021.

- **UN OCHA Administrative Boundary COD (COD-AB)**

  The UN OCHA COD boundaries are widely used by the humanitarian community to provide assistance during crises. The COD boundaries are accessible for download from the HDX repository. The data providers of CODs are usually government authorities; however, they might not be the most recent boundaries or have license restrictions such as "for humanitarian use only," which might restrict data use.

- **GADM**

  GADM provides open access to subnational administrative boundaries for every country in the world. It also has a wide use within the geospatial community and has global and regional coverage available. Despite its wide use, GADM has limited documentation about the source of the input data, contributors, year the data represents, and also requires written permission for redistribution and commercial use.

- **OSM**

  Subnational boundaries data available on OSM also enjoy wide usage within the geospatial community, due to providing one of the most recent datasets collected through crowdsourcing methods. Despite providing openly available and continually updated boundary data, OSM is limited by its lack of validation (due to being crowdsourced data).

- **geoBoundaries Repository**

  In addition to providing seamless access to the UN SALB and UN OCHA CODs datasets through API URLS, geoBoundaries also includes data from open community contributions that might not have been included anywhere else. These additional data are undergoing automated as well as manual checks to ensure clean topology and the presence of comprehensive metadata parameters (such as year, source, and license of the dataset).

- **National Boundaries**

  National boundaries are typically available at the respective ministry. These may or may not align with the various abovementioned datasets. Nationally maintained boundaries often have topology errors and do not harmonise with higher or lower administrative level boundaries. Because these boundaries are not openly accessible in most countries, they are mentioned in this report for completeness but not included in the comparison table.

***Which metrics/metadata were chosen and why***

The quantitative metrics and metadata parameters used for assessing boundaries were determined based on a consultative process with the partners from the Coalition to Advance Progress of Administrative Boundaries in Africa project. The quantitative comparative metrics include: number of administrative units, the mean, maximum, and minimum for an area, perimeter length, and number of vertices for each administrative level. The metadata parameters will include the year a boundary represents, date of publication, source, license, terms of use, and whether the boundary dataset is authoritative.

The identified boundary datasets are currently being hosted in the geoBoundaries repository. geoBoundaries[6] is an open license resource of boundaries for every country in the world that has been produced and maintained by the College of William & Mary's geoLab since 2017, providing access to open license data available in the three categories of highest precision, simplified country boundaries, or as a global dataset.

The data provided by geoBoundaries are clean of topology errors, include standardised attributes and metadata (including the above-listed comparison metrics), and undergo various levels of quality checks by the William & Mary team. Besides providing API and download access to OCHA CODs, SALB, GADM, and OSM boundaries data from one central database, geoBoundaries also includes unique datasets not available elsewhere by incorporating community contributions from the open data community.

***Detailed overview of metadata comparison***

The quantitative comparison metrics were chosen as an approximation for precision and accuracy of the boundary delineation. Given that a scale at which the boundaries are digitised is often unknown, such calculations as minimum, maximum, and mean number of vertices, as well as area and length of the perimeter per admin level, are used as an approximation of precision and accuracy. The metadata parameters chosen for comparison provide essential information the user would need to evaluate the fitness for use (dependent on the specific application of the data). The metadata parameters are included in the table linked here.

## Health Facilities Data

The mapping of healthcare facilities is an essential component of health service delivery. This layer, combined with population data, necessarily indicates the accessibility of healthcare and provides foundational information for the development of vaccine campaigns, the planning of further healthcare facilities, and planned outreach via community volunteers.

***Overview of which datasets were chosen and why***

Knowing health facility locations, as well as attributes such as name and services available, is crucial for many development, emergency, and health-related activities. These data are especially relevant for vaccine and drug distribution. In the open domain, no complete and up-to-date regional dataset exists for facilities. The datasets described below are the best available.

6. See https://www.geoboundaries.org/geoContrast.html

- **OSM Health Facility Data**

  OSM data are collected by volunteers and are freely available for use. The use of volunteer-collected data, however, can lead to inconsistencies in collection and other quality issues. The Heidelberg Institute for Geoinformation Technology has a project that analyses OSM completeness of health facilities in SSA. It compares the number of facilities in OSM for a given location using the dataset on public health facilities in SSA, developed by Maina et al. (2019) as a reference dataset. Healthsites.io makes the health facility data easily downloadable and also serves as a contributor.

- **SSA facility database from World Health Organization and the Wellcome Trust Research Programme (WHO-KWTRP) Maina et al (2019)**

  Over the last 15 years, a variety of geocoding methods were used in compiling a database of health facilities in SSA. Master health facility lists from governmental and non-governmental sources in 50 countries were collected, cleaned, and merged to create a database of public health facilities managed by governments, local authorities, faith-based organisations (FBOs), and NGOs. These data are available and accessible. A publication details the methods used to create the dataset, which is housed in two venues: Figshare and WHO. The WHO version of the dataset will continue to be updated.

- **Country Data**

  Single country datasets from a given ministry of health and donor agencies typically exist in each country. An overview study by South et al. (2020) sought to document publicly available official health site data. They reported that only 7 of 52 African countries have openly available official master facility lists (MFL). These MFL's are made publicly available as PDFs, mapping software, and Excel-downloadable databases. In addition, there are 20 single-country datasets (including the 7 official MFLs) from ministries and donor agencies that are publically available.

***Which metrics/metadata were chosen and why***

Metrics to compare the different datasets include the number of facilities in each country, whether the datasets contained official data, if they were "stamped" as official data from the government agency, the methodology used to create the dataset, the date the dataset represents, and the frequency of updates. There is a lot of variability in the health sites data due to the crowd sources and each country's specific needs or choices.

***Detailed overview of metadata comparison***

The overview table can be found here.

## Roads Data

Roads data also exist in the public domain. This is especially critical for logistics planning and for creating travel time analyses. Roads data are also a helpful element for orientation in maps. The subsections below describe the two main roads datasets that are available and accessible.

- **OSM Roads**

  OSM country level road datasets were included because they are a convenient package of roads by country. OSM is crowdsourced and as such is not exhaustive and may not always be authoritative; but few road databases exist and so this source is included. (Country datasets from such sources as WFP available on HDX were not included, as these datasets pull from OSM and were considered to be indistinguishable.)

- **Facebook Roads**

  The Facebook Artificial Intelligence (AI) road datasets contain only the AI-predicted roads that are missing from OSM. As such, this dataset should be used with the OSM roads dataset for complete road coverage. The Facebook-developed tool, RapiD, edits and adds features to OSM by auto-detecting roads from satellite imagery with data integrity checks. The tool uses deep learning and weakly supervised training to predict road networks from commercially available high-resolution imagery. The roads predicted by AI are added only after human validation.

***Which metrics/metadata were chosen and why***

The roads datasets, although they should be used together, are compared using the same metrics and metadata as the datasets mentioned above. The comparison is done using the metrics of accessibility and use (public accessibility, use constraints, relationship to corresponding official dataset), recency (vintage of data product and source data), coverage completeness (i.e. more than 10 countries in SSA) and documentation completeness (input source data, methods/edits).

***Detailed overview of metadata comparison***

The table linked to below shows both datasets and their metadata and metrics side by side. Please see the table here.

## Tools for Data Selection

Development practitioners are encouraged to read the input data and method descriptions thoroughly and explore the core spatial datasets before deciding on which dataset meets their needs best.

There are multiple tools that can be used to explore these datasets available online, such as:

- GRID3 Data Hub (GRID3 population, settlement, boundaries and infrastructure datasets): https://data.grid3.org
- woprVision (WorldPop's bespoke population models): https://apps.worldpop.org/woprVision/
- PopGrid (comparison tool for top-down population estimates): https://sedac.ciesin.columbia.edu/mapping/popgrid/
- WorldPop Demograpic data portal (subnational age/sex structures): https://www.portal.worldpop.org/demographics/

While the tools listed above are each focused on data produced by a specific organisation or programme, GRID3 is also creating a Data Explorer tool which will allow users to compare and choose core spatial data for practical applications in sub-Saharan Africa, and bring open source data from different producers under one tool. While this report references only datasets that are available for at least 10 countries in sub-Saharan Africa, the Data Explorer tool will also draw on data that may not be available for all countries. This will allow users to choose the best possible dataset for their needs.

## Contributors

Olena Borkovska, Eniko Kelly, Attila Lazar, Marc Levy, Chisimdi Onwuteaka, Linda Pistolesi, Silvia Renn, Johanna Snell, Corey Sobel, Antoinette Wannebo

GRID3

GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) works with countries to generate, validate and use geospatial data on population, settlements, infrastructure, and subnational boundaries. GRID3 combines the expertise of partners in government, United Nations, academia, and the private sector to design adaptable and relevant geospatial solutions based on capacity and development needs of each country.

grid3.org
info@grid3.org
@GRID3Global

**Funders & Partners:**

The GRID3 programme is funded by a grant from the Bill & Melinda Gates Foundation and the United Kingdom's Foreign, Commonwealth & Development Office. It is managed by Columbia University's Center for International Earth Science Information Network (CIESIN) and implemented by the United Nations Population Fund (UNFPA), WorldPop at the University of Southampton and the Flowminder Foundation.

Bill & Melinda Gates foundation · UKaid · Columbia Climate School Center for International Earth Science Information Network · UNFPA · WorldPop · FLOWMINDER.ORG