Release Statement

Gridded maps of residential/non-residential building classifications and associated building patterns for Kenya (KEN), version 1.0

10 January 2023

This work is part of the GRID3 project with funding from the Bill and Melinda Gates Foundation (BMGF) and the United Kingdom's Foreign, Commonwealth & Development Office (INV-009579, formerly OPP1182425). Method, coding, modelling, and data production have been carried out by Christopher Lloyd. Oversight was provided by Andrew J. Tatem. Contact release@worldpop.org for more information.

The whole WorldPop group are acknowledged for overall project support. The author acknowledges the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. Thanks go to Maksym Bondarenko (WorldPop) for support with online publication, and to Attila Lazar and Edith Darin for internal review of data and the release statement. The author acknowledges the efforts of WorldPop's partners (United Nations Population Fund (UNFPA), Center for International Earth Science Information Network (CIESIN) in the Earth Institute at Columbia University, and the Flowminder Foundation) in supporting access to the building footprints. The author thanks Io Blair-Freese (formerly of BMGF) and Heather Chamberlain (WorldPop) for providing coordination between WorldPop and BMGF and Ecopia.

CITATION

Lloyd, C.T and Tatem, A.J. 2023. Gridded maps of residential/non-residential building classifications and associated building patterns for Kenya (KEN), version 1.0. University of Southampton: Southampton, UK. doi:10.5258/SOTON/WP00756

RELEASE CONTENT

1. KEN_building_class_metrics_v1_0_classification_binary.tif
2. KEN_building_class_metrics_v1_0_residential_mean_score.tif
3. KEN_building_class_metrics_v1_0_residential_count.tif
4. KEN_building_class_metrics_v1_0_nonresidential_count.tif
5. KEN_building_class_metrics_v1_0_residential_density.tif
6. KEN_building_class_metrics_v1_0_nonresidential_density.tif

Rasters are provided for countries identified using ISO3 country codes.

FILES DESCRIPTION

The geotiff rasters have a spatial resolution of approximately 100m (3 arc seconds). Their coordinate reference system is WGS84. Each building has been considered in the grid cell that contained the centroid⁺ of its building footprint. NAs represent grid cells that contain no building footprint centroid⁺.

KEN_building_class_metrics_v1_0_classification_binary.tif

This raster contains the binary classification for all buildings per each grid cell (as a binary integer. 1=residential, 0=non-residential) derived from modelling.

KEN_building_class_metrics_v1_0_ residential_mean_score.tif

This raster contains the mean percentage score (as a decimal number) for all buildings per each grid cell. A value of 1 indicates highest likelihood that all buildings within the grid cell are residential. A value of 0

indicates highest likelihood that all buildings within the grid cell are non-residential. This layer enables the user to pick a threshold for the residential/non-residential cut-off threshold other than that identified as optimal during modelling.

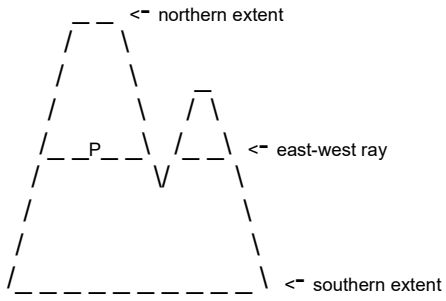
KEN_building_class_metrics_v1_0_residential_count.tif
This raster contains a count of residential buildings per each grid cell.

KEN_building_class_metrics_v1_0_nonresidential_count.tif
This raster contains a count of non-residential buildings per each grid cell.

KEN_building_class_metrics_v1_0_residential _density.tif
This raster contains a measure of the number of residential buildings per grid cell area in square kilometres (km^2), i.e. building count divided by the area in square kilometres of the grid cell. If needed, the grid cell area can be retrieved by dividing the count raster by the density raster.

KEN_building_class_metrics_v1_0_nonresidential_density.tif
This raster contains a measure of the number of non-residential buildings per grid cell area in square kilometres (km^2), i.e. building count divided by the total number of square kilometres in the grid cell. If needed, the grid cell area can be retrieved by dividing the count raster by the density raster.

+ For simplicity, the term centroid is used throughout this document to indicate the centre point within each building. However, true building centroids are often located outside of the building polygon, perhaps because the building is L-shaped or because several polygons represent the building and so these are grouped into one geometry (i.e. multi-polygon geometry). It is undesirable for building centroids to be located outside of building polygons because allocation of building to a raster grid cell is less accurate. Hence, rather than use the st_centroid function to define the centre of a building we instead use the st_point_on_surface function (Pebesma, 2018). Building centre is accordingly defined as the point half-way along the longest segment of the east-west ray that intersects the building and lies half-way between the northern and southern extents of the building geometry (see illustration below).



$_P$ = Centre of building as defined by st_point_on_surface function

RELEASE HISTORY

Version 1.0 (10 January 2023)

- This is the original release of the data

SOURCE DATA

The classified building data have been modelled using building footprints (polygon features) provided by the Digitize Africa project (Ecopia.AI and Maxar Technologies, 2020/2021), and building footprint and highway data provided by OpenStreetMap (© 2020-2022 OpenStreetMap contributors; geofabrik.de). Digitize Africa is a two-year project funded by the Bill and Melinda Gates Foundation to map buildings and roads in 51 countries across sub-Saharan Africa using satellite imagery and artificial intelligence (AI) to support humanitarian assistance and sustainable development. Maxar provided their Vivid satellite imagery mosaics (50 cm resolution) and Ecopia.AI generated the building footprints using their artificial intelligence-assisted feature extraction techniques (Ecopia, 2021).

The building classification model utilizes impervious surface data provided by DLR (World Settlement Footprint impervious surface data, 2019; pending public release). These data have been combined with settlement extents provided by CIESIN, Columbia University and Novel-T (GRID3 Built-up Areas and Small Settlement Areas, Versions 01 to 01.02. 2021-2).

The model has been trained using building label data provided by OpenStreetMap (OSM).

METHODS OVERVIEW

The classified building layers are generated in two steps: (1) The modelling of individual buildings into residential and non-residential classes, and then (2) the aggregation of these results at the grid cell-level.

1. Model input data are preprocessed using GIS, geospatial, database, and statistical software (OSGeo - GDAL, GRASS GIS, Spatialite, and R), discussed in detail in Lloyd et al. (2020). The modelling follows the methodology of Sturrock, et al. (2018), but with countrywide coverage. To handle potential building heterogeneity, the classification model is run separately for urban and rural areas, and outputs combined. World Settlement Footprint 2019 impervious surface data (DLR, pending public release) and GRID3 Built-up Areas (BUA) and Small Settlement Areas (SSA) Settlement Extents (CIESIN & Novel-T, 2021-2) data are combined to represent urban areas for input to the model. Elsewhere, impervious surface data are considered representative of rural areas (i.e. including small hamlets and isolated buildings). Building classification predictions are combined with known building labels (used to train and test the model), with the latter taking precedence, to produce model outputs.

2. The raster datasets have been produced using a modified version of the code used to produce 'Gridded maps of building patterns throughout sub-Saharan Africa', version 2.0 (Dooley, Leasure, Boo, and Tatem. 2021). All data processing has been carried out using R (R Core Team, 2013). The classified building footprint polygons (model output) have been converted into centroid$^+$ points in UTM projection using the *st_point_on_surface* function in the *sf* R package (Pebesma, 2018). The points have then been re-projected to WGS84 using *sf's st_transform* function so that their corresponding cell IDs could be identified in the WGS84 projected raster master grid. The WorldPop master grid national boundary (WorldPop et al. 2018) defines the extent, resolution, and coordinate reference system of the output building footprint metric layers. Building centroid$^+$ cell IDs have been obtained using the *cellFromXY* function in the *raster* R package (Hijmans & van Etten, 2012). The 'settled' grid cells in the rasters have at least one building footprint centroid$^+$ within the pixel area. No areas have been masked out.

The binary classification raster (*_classification_binary.tif) represents the mode binary building classification per grid cell, derived from the model. Where, for a given grid cell, there are equal numbers of residential and non-residential buildings, the grid cell is classified as residential.

The mean score raster (*_residential_mean_score.tif) represents the mean percentage likelihood of a grid cell containing residential buildings. The modelling technique invariably uses unequal amounts of residential and non-residential data in model training and testing. Hence, the model uses a country specific cut-off threshold at which equally good performance is achieved when classifying residential and non-residential

structures. This threshold is applied by the model to produce spatially comparable binary building classifications (Sturrock et al. 2018), which are subsequently averaged per grid cell to produce the binary classification raster (see above). The mean score raster allows the user to pick a different threshold for the residential/non-residential cut-off threshold than the one identified as optimal during modelling.

Building counts per grid cell have been calculated by simply summing the number of residential or non-residential building centroids[+] (i.e. binary building classification) for each cell ID. Building density per grid cell have been calculated by dividing the building count in a cell by the area of the cell. Grid cell area have been calculated in square kilometres using the *area* function in the *raster* R package.

KEN MODEL METRICS

24,125,591 buildings total
20,949,136 residential
3,176,455 non-residential
60,407 labelled buildings used in training and testing

Urban modelling
0.9266501 AUC
0.793 res/non-res cut-off threshold
85.3% residential correctly classified in the testing dataset
85.3% non-residential correctly classified in the testing dataset

Rural modelling
0.9752602 AUC
0.935 res/non-res cut-off threshold
91.7% residential correctly classified in the testing dataset
91.7% non-residential correctly classified in the testing dataset

LIMITATIONS

In urban locations where building density is greatest, there will typically be greater numbers of residential than non-residential buildings per ~100m pixel in raster output. This is because buildings with mixed use are included in the residential classification for the purpose of modelling, and because residential buildings tend to be smaller than non-residential (e.g warehouses, universities, factories, malls) buildings. Hence, the binary classification raster will be more likely to reflect a residential classification in urban pixels, and in count rasters the count of residential buildings is likely to be greater than non-residential in urban pixels. In urban locations where non-residential buildings predominate, there will typically be fewer buildings per pixel and the binary raster will be more likely to reflect a nonresidential classification.

Modelling is most successful where data on structure type are available for a reasonable subset of buildings with a reasonable spatial distribution within a given country. In some modelling instances, this information is available for only a relatively small number of structures. In such instances, predictions are extrapolated in neighbouring regions. To handle potential building heterogeneity, the classification model is run separately for urban and rural areas per country, and outputs combined.

There may be some error in OSM or other label dataset attributes or bias could be introduced if building type information is only available for certain categories of structures. Ground truth data (i.e. household surveys) provide valuable additional building label data in some countries modelled.

This modelling is only really useful where every structure has been mapped or where the completeness can be quantified. If gaps exist of unknown size, as is often the case, estimates of numbers of residential structures will be low. While OSM does not currently have a mechanism to estimate the completeness of the building enumeration data, efforts are underway to provide this information (Sturrock et al. 2018). Building footprint accuracy and completeness are discussed in Lloyd et al (2020).

There are instances where OSM building polygons are duplicated in source data due to field mapping/data management error. Where this occurs, geometrically identical duplicates are systematically removed from the building dataset in preprocessing to ensure that pixel values in building rasters are as accurate as possible. Due to the size of the source dataset, rare non-identical duplicates cannot be systematically removed without also removing many other building polygons for which delineation tolerance has led to modest overlap between two adjacent structures. Non-identical duplicate building polygons therefore remain in the dataset, and respective isolated pixel values in the building rasters are thus less accurate for these buildings.

The German Aerospace Center (DLR) is currently working on the development and validation of the World Settlement Footprint 2019 Imperviousness (WSF2019-Imp) layer that underpins the building classification modelling upon which this raster dataset is based. Hence, limitations of the data should be published soon. The WSF2019-Imp is the beta version of a thematic layer estimating the percent impervious surface (PIS) of the pixels marked as settlements in the WSF2019 binary layer (Marconcini et al. 2021), at ~10m spatial resolution for the year 2019. The WSF2019 layer is produced using Sentinel 1 (S1) radar data and Sentinel 2 (S2) optical imagery (both with ~10 m-spatial resolution). The increased spatial resolution of the S2 data has allowed for a better identification of building structures compared to previously available data (Esch et al. 2022), improving the built-up coverage, especially in suburban and rural settings (Palacios-Lopez, 2021).

REFERENCES

Center for International Earth Science Information Network (CIESIN), Columbia University, & Novel-T. 2021-2. GRID3 Settlement Extents, Versions 01, 01.01, and 01.02. https://academiccommons.columbia.edu/search?f%5Bauthor_ssim%5D%5B%5D=Novel-T&f%5Bsubject_ssim%5D%5B%5D=Human+settlements

Dooley C. A, Leasure D.R, Boo G, and Tatem A.J. 2021. Gridded maps of building patterns throughout sub-Saharan Africa, version 2.0. University of Southampton: Southampton, UK. Source of building footprints "Ecopia Vector Maps Powered by Maxar Satellite Imagery"© 2020/2021. doi:10.5258/SOTON/WP00712.

Ecopia.AI and Maxar Technologies. 2020/2021. Digitize Africa building footprint data. https://www.ecopiatech.com/commerical-landing-africa

Ecopia.AI and Maxar Technologies. 2021. The Most Comprehensive Map of Buildings in the USA. https://www.ecopiatech.com/post/the-most-comprehensive-map-of-buildings-in-the-usa

Esch T, Brzoska E, Dech S, Leutner B, Palacios-Lopez D, Metz-Marconcini A, Marconcini M, Roth A, Zeidler J. 2022. World Settlement Footprint 3D - A first three-dimensional survey of the global building stock. Remote Sensing of Environment 270. 112877. https://doi.org/10.1016/j.rse.2021.112877

Hijmans R.J. & van Etten J. 2012. raster: Geographic analysis and modeling with raster data. R package version 2.0-12. http://CRAN.R-project.org/package=raster

Lloyd C.T, Sturrock H.J.W, Leasure D.R, Jochem W.C, Lázár A.N, Tatem A.J. 2020. Using GIS and Machine Learning to Classify Residential Status of Urban Buildings in Low and Middle Income Settings. Rem Sens 12 (23). 3847. https://doi.org/10.3390/rs12233847

Marconcini M, Metz-Marconcini A, Esch T, Gorelick N. 2021. Understanding Current Trends in Global Urbanisation - The World Settlement Footprint Suite. GI_Forum 2021, Issue 1, 33-38. https://austriaca.at/0xc1aa5576%200x003c9b4c.pdf

Palacios-Lopez D, Bachofer F, Esch T, Marconcini M, MacManus K, Sorichetta A, Zeidler J, Dech S, Tatem A.J, Reinartz P. 2021. High-Resolution Gridded Population Datasets: Exploring the Capabilities of the World Settlement Footprint 2019 Imperviousness Layer for the African Continent. Remote Sens. 13, 1142. https://doi.org/10.3390/rs13061142

Pebesma, E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 10(1), 439–446. doi: 10.32614/RJ-2018-009, https://doi.org/10.32614/RJ-2018-009 .

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sturrock H.J.W, Woolheater K, Bennett A.F, Andrade-Pacheco R, Midekisa A. 2018. Predicting residential structures from open source remotely enumerated data using machine learning. PLoS ONE, 13, e0204399. https://doi.org/10.1371/journal.pone.0204399

WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University. 2018. Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00651