Release statement

Gridded disaggregated population estimates for Kenya (2021), version 1.0.

22 July 2022

These data were produced by the WorldPop Research Group at the University of Southampton. This work was part of the GRID3 project with funding from the United Nations Children's Fund (UNICEF) - Population Modelling for use in Routine Health Planning and Monitoring project (contract no. 43335861). Projects partners included the Kenya Unicef Regional and Country Offices, WorldPop research group at the University of Southampton and the Center for International Earth Science Information Network in the Columbia Climate School at Columbia University. Assane Gadiaga (WorldPop) led the input processing and the modelling work following the Random Forest (RF)-based dasymetric mapping approach developed by Stevens et al. (2015). Thomas Abbott supported the covariates processing work, as well as Christopher Lloyd, particularly for the processing of residential/non-residential building footprints. In-country engagements were done by Benard Mitto, Justine Dowden (CIESIN) and Maria Muniz (Unicef). Using the 2009 and 2019 census data from the Kenya's National Bureau of Statistics (KNBS), the US Census Bureau released the census-based total population projections, population by age and sex and digital sub-counties boundaries. Duygu Cihan helped in the preparation of these input population data. Attila N Lazar, Edith Darin and Heather Chamberlain advised on the modelling procedure. The work was overseen by Attila N Lazar and Andy J Tatem.

The authors followed rigorous procedures designed to ensure that the used data, the applied method and thus the results are appropriate and of reasonable quality. If users encounter apparent errors or misstatements, they should contact WorldPop at release@worldpop.org.

WorldPop, University of Southampton, and their sponsors offer these data on a "where is, as is" basis; do not offer an express or implied warranty of any kind; do not guarantee the quality, applicability, accuracy, reliability or completeness of any data provided; and shall not be liable for incidental, consequential, or special damages arising out of the use of any data that they offer.

RELEASE CONTENT

KEN_population_v1_0_gridded.tif

KEN_population_v1_0_agesex.zip

KEN_population_v1_0_mastergrid.tif

LICENSE

SUGGESTED CITATIONS

Gadiaga A. N., Abbott T. J., Chamberlain H., Lloyd C. T., Lazar A. N., Darin E., Tatem A. J. 2022. Census disaggregated gridded population estimates for Kenya (2021), version 1.0. University of Southampton. doi:10.5258/SOTON/WP00747

FILES DESCRIPTIONS

The projection for all GIS files is the geographic coordinate system WGS84 (World Geodetic System 1984).

KEN_population_v1_0_gridded.tif

This geotiff raster, at a spatial resolution of 3 arc-seconds (approximately 100m at the equator), contains estimates of the total population size per grid cell across Kenya. NA values represent areas that were mapped as unsettled based on gridded building patterns derived from building footprints (Dooley and Tatem, 2020). These data are stored as floating-point numbers rather than integers to avoid rounding errors in aggregated population totals for larger areas.

KEN_population_v1_0_agesex.zip

This zip file contains 40 GeoTIFF rasters representing estimated population counts for specific age and sex groups within grid cells of approximately 100m. We provide 36 rasters for the commonly reported age-sex groupings of sequential age classes for males and females separately. These are labelled with either an "m" (male) or an "f" (female) followed by the number of the first year of the age class represented by the data. "f0" and "m0" are population counts of under 1-year olds for females and males, respectively. "f1" and "m1" are population counts of 1 to 4 year olds for females and males, respectively. Over 4 years old, the age groups are in five year bins labelled with a "5", "10", etc. Eighty year olds and over are represented in the groups "f80" and "m80". We provide four additional rasters that represent demographic groups often targeted by programmes and interventions. These are "under1" (all females and males under the age of 1), "under5" (all females and males under the age of 5), "under15" (all females and males under the age of 15) and "f1549" (all females between the ages of 15 and 49, inclusive).

These data were produced *post-hoc* by multiplying the total population counts provided in the *KEN_population_v1_0_gridded.tif* raster and age and sex proportions derived the US Census bureau age-sex projections for each sub-county. While this data represents population counts, values contain decimals, i.e. fractions of people. This is because both the input population data and age-sex proportions contain decimals. For this reason, it is advised to aggregate the rasters at a coarser scale. For example, if four grid cells next to each other have values of 0.25 this indicates that there is 1 person of that age group somewhere in those four grid cells.

KEN_population_v1_0_mastergrid.tif

This geotiff raster contains the rasterised administrative units used to perform the population disaggregation, with a spatial resolution of approximately 100m grid cell (0.0008333 decimal degrees). The pixel values are IDs referring to the administrative boundary polygons (sub-counties) that match the corresponding units in the input population data.

RELEASE HISTORY

Version 1.0 (22 July 2022) doi:10.5258/SOTON/WP00747
Original release of this data set.

SOURCE DATA

- Digital Sub-county boundary and their projected population totals and age/sex group totals for 2021 based on the 2019 Population and Housing census were provided by the US Census bureau in a shapefile format (population projection).
- Gridded building patterns (building count, building total area, building mean area, building area variance, building density, building length, building mean length and building length variance) were derived from building footprints by Dooley and Tatem (2020).
- Additional geospatial covariates (Lloyd et al., 2019), representing factors related to population distribution (distance to land cover maps, mean precipitation and temperature, slope and elevation, motorized friction surface, walking friction surface, travel time to city, distance to coastline, protected areas, schools, health facility, market place, place of worship, local roads, main roads, railway station, road intersection, and built settlement, and night-time lights), were obtained from the "Global High-Resolution Population Denominators Project" (OPP1134076) and the Copernicus Climate Data Service (Muñoz Sabater, 2021). Because building footprints contain different structure types (e.g. industrial areas, warehouses), information on the residential status of the buildings, such as building footprints predicted residential/non-residential classification and probability of residential was also used. The classification method used for building footprints-derived residential/non-residential status is described by Lloyd et al. (2020).

METHODS OVERVIEW

Modelling: Following the Random Forest (RF)-based dasymetric mapping approach (Stevens et al., 2015), the popRF 'R' package (Bondarenko et al., 2021) based on Breiman (2001) algorithm was used to model Sub-county total population density as a combination of the geospatial covariates and then to estimate the total population density in each approximately 100 m grid cell (0.0008333 decimal degrees grid or 3 arc seconds). The model could explain 95% of the total population input variance. The list of used covariates is listed in the Appendix.

The gridded population estimates were then combined with the age/sex proportions calculated from the projections for Kenya (population projection) to produce gridded population estimates for each sex group (female and male) at regular age intervals.

ASSUMPTIONS AND LIMITATIONS

This dataset was produced based on the projected 2021 population totals for Sub-county derived from the 2019 Population and Housing Census. Although the enumerated population totals have been projected to 2021, the estimate of population in each Sub-county may not reflect the current population, given the time elapsed since the last census and the necessary assumptions made in projecting the population estimates.

The gridded population estimates are constrained within the settled area derived from gridded building metrics (Dooley and Tatem, 2020). We assumed that the building footprint data (Ecopia.AI and Maxar Technologies, 2020), from which the gridded building metrics were derived, is accurate and that each building polygon corresponds to a building structure. In addition, the distribution of buildings might not represent the current building landscape because of the necessity to use satellite imagery from different years in extraction of the building footprints (e.g. due to cloud coverage) (Dooley and Tatem, 2020). In locations which have recently experienced rapid settlement changes, for example, establishment of new settlements, rapid urban growth or abandonment of settlements, the population estimates are likely to be less accurate.

Finally, very high (>1000 people per pixel) population totals were estimated for the Mathare Sub-County. The Mathare Sub-County population projection totals are consistent with the publicly accessible 2019 census results, therefore, these were not removed from the disaggregation. However, without on-the-ground knowledge of the area, uncertainty remains if these high population densities within the Sub-County (764 people per hectare) and the resulting very high pixel values are realistic.

WORKS CITED

Bondarenko M., Nieves J.J., Forrest R.S., Andrea E.G., Jochem C., Kerr D., and Sorichetta A.

(2021): popRF: Random Forest-informed Population Disaggregation R package, _Comprehensive R Archive Network (CRAN)_, url:https://cran.rproject.org/package=popRF.

Breiman, L. Random forests. Mach. Learn. 45, 5–32 (2001).

Carioli A, Pezzulo C, Hanspal S, Hilber T, Hornby G, Kerr D, Tejedor-Garavito N, Nielsen K, Pistolesi L, Adamo S, Mills J, Nieves JJ, Chamberlain H, Bondarenko M, Lloyd C, Yetman G, Gaughan A, Stevens F, Linard C, James W, Sorichetta A, Tatem AJ. *In prep*. Population structure by age and sex: a multi-temporal subnational perspective.

Dooley, C. A. and Tatem, A.J. 2020. Gridded maps of building patterns throughout sub-Saharan Africa, version 1.0. University of Southampton: Southampton, UK. Source of building Footprints “Ecopia Vector Maps Powered by Maxar Satellite Imagery”© 2020. https://dx.doi.org/10.5258/SOTON/WP00666.

Ecopia.AI and Maxar Technologies. 2020. Digitize Africa data. http://digitizeafrica.ai

Lloyd, C.T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F.R., Gaughan, A.E., Nieves, J.J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., and Tatem A.J., 2019.Global spatio-temporally harmonised datasets for producing high resolution gridded population distribution datasets. Big Earth Data, 3(2), 108-139.

https://dx.doi.org/10.1080/20964471.2019.1625151

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

Muñoz Sabater, J., (2021): ERA5-Land monthly averaged data from 1950 to 1980. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.68d2bb30

Pezzulo C, Hornby GM, Sorichetta A, Gaughan AE, Linard C, Bird TJ, Kerr D, Lloyd CT, Tatem AJ. 2017. Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. *Sci. Data* 4:170089 https://dx.doi.org/10.1038/sdata.2017.89.

Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data.

PLoS ONE 10, e0107042 (2015). https://doi.org/10.1371/journal.pone.0107042

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/WP00646

APPENDIX

List of covariates
Motorized friction surface
Walking friction surface
Distance to cities
Distance to places of education
Distance to marketplaces
Distance to places of worship
Distance to local roads
Distance to main roads
Distance to Water bodies
Distance to railway stations
Distance to cultivated areas 2015
Distance to woody areas 2015
Distance to shrub area edges 2015 (130)
Distance to sparse vegetation areas 2015
Distance to aquatic vegetation areas 2015
Current average total annual precipitation
Current average annual temperature
Slope
Elevation
Distance to coastline
Nighttime lights 2020 VIIRS
Distance to CAT1 protected areas 2017
Buildings area (coefficient of variation)
Buildings mean area
Buildings mean length
Buildings total area
Buildings total length
Buildings count
Buildings density
Residential count
Residential density
Residential mean area
Residential mean length
Residential total area
Residential total length