Release Statement

Bayesian gridded population estimates for Ghana 2018, version 1.0

28 August 2020

These data include gridded estimates of population sizes at approximately 100 m resolution with national coverage across Ghana. This includes estimates of total population sizes, population counts in 36 different age-sex groups, people per household, people per building, households per building, and statistical measures of uncertainty. These results were produced using publicly available census microdata (IPUMS) from the 2010 Ghana census and building footprints from Maxar/Ecopia that were derived from recent satellite imagery.

This data release was produced by the WorldPop Research Group at the University of Southampton with funding from the Bill and Melinda Gates Foundation to provide critical spatial data and population estimates to support polio surveillance and eradication (INV-002697). Statistical modelling was led by Doug Leasure with oversight from Andy Tatem and coordination with the Gates Foundation Polio Team from Vince Seaman. Chris Jochem, Edith Darin, and Attila Lazar provided internal WorldPop peer reviews that helped to improve the results and documentation. We acknowledge the whole WorldPop Research Group for overall project support. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. Analyses undertaken were approved by the University of Southampton Faculty Ethics Committee (ERGO II 61033).

SUGGESTED CITATION

Leasure DR, Tatem AJ. 2020. Bayesian gridded population estimates for Ghana 2018, version 1.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00680

RELEASE HISTORY

Version 1.0 (28 August 2020)

- Original release of this population dataset.

RELEASE CONTENT

1. GHA_population_v1_0_gridded.zip
2. GHA_population_v1_0_agesex.zip
3. GHA_population_v1_0_mastergrid.tif
4. GHA_population_v1_0_sql.sql
5. GHA_population_v1_0_tiles.zip

FILE DESCRIPTIONS

The map projection for all GIS files is WGS84.

1. GHA_population_v1_0_gridded.zip

This zip file contains six files:

GHA_population_v1_0_gridded_population.tif

This geotiff raster contains estimates of total population size for each approximately 100m grid cell (0.0008333 decimal degrees grid or 3 arc seconds) across the study area. The values are the mean of the Bayesian posterior probability distribution for the predicted population size in each grid cell. NA values represent grid cells where no building footprints occurred. These population estimates includes decimals (e.g. 10.3 people). This provides more accurate population totals when pixels are summed. A population estimate of 0.5 people in each of two neighboring pixels would indicate an expectation that 1 person lives somewhere within those two pixels.

GHA_population_v1_0_gridded_uncertainty.tif

This geotiff raster contains estimates of uncertainty in the population estimates for each approximately 100m grid cell across the study area. The uncertainty values are the difference between the upper and lower 95% credible intervals of the posterior prediction divided by the mean of the posterior prediction: (upper – lower)/mean. These numbers provide a comparable measure of relative uncertainty in population estimates across the country.

Note: We did not include a raster of upper and lower credible intervals for population estimates because they cannot be summed across grid cells to produce accurate credible intervals for population estimates within a multi-cell area, and most end-users need this information for multi-cell areas. Credible intervals for multiple cells can be calculated using the woprVision web application (https://apps.worldpop.org/woprVision), the wopr R package (Leasure et al 2020a), or manually calculated from the cells' posterior predictions in GHA_population_v1_0_sql.sql (described below).

GHA_population_v1_0_gridded_hh.tif
This geotiff raster contains estimates of the number of households in each approximately 100m grid cell across the study area.

GHA_population_v1_0_gridded_pph.tif
This geotiff raster contains estimates of the average number of people per household in each approximately 100m grid cell across the study area.

GHA_population_v1_0_gridded_hpb.tif
This geotiff raster contains estimates of the average number of households per building footprint in each approximately 100m grid cell across the study area.

GHA_population_v1_0_gridded_ppb.tif
This geotiff raster contains estimates of the average number of people per building footprint in each approximately 100m grid cell across the study area.

2. GHA_population_v1_0_agesex.zip

This zip file contains 40 geotiff rasters that contain counts of people for each age-sex group in each approximately 100m grid cell across the study area.

We provide 36 rasters for the commonly reported age-sex groupings of sequential age classes for males and females separately. These files 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 needed 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 "f15_49" (all females between the ages of 15 and 49, inclusive).

In addition to the rasterized population estimates for specific age-sex groups, we provide a spreadsheet of age-sex proportions (GHA_population_v1_0_agesex_table.csv) estimated from IPUMS data for different regions. The regions are defined in the accompanying raster (GHA_population_v1_0_agesex_regions.tif).

3. GHA_population_v1_0_mastergrid.tif

This geotiff raster contains 1s for each settled approximately 100m grid cell (0.0008333 decimal degrees) across the study area. 0 values indicate grid cells that were considered unsettled and thus not containing people. NAs show grid cells considered as outside the study area.

4. GHA_population_v1_0_sql.sql

This SQLite database contains samples (n=10,000) from the Bayesian posterior predictions of population size in each grid cell. These can be used to derive the posterior distribution for population totals for larger areas that contain more than one grid cell. This database is source data for the woprVision web application (https://wopr.worldpop.org/woprVision) and it can be queried using the wopr R package (Leasure et al 2020a).

The SQL database contains a single table (Nhat) that includes the population predictions. This table contains the following columns:

- “cell” contains a cell ID to identify the location. Cell IDs correspond to those the cell IDs of GHA_population_v1_0_mastergrid.tif.
- “x” and “y” columns contain WGS84 coordinates for the centroid of the grid cell.
- “Pop” column contains a comma-separated string of population estimates which are the MCMC samples from the predicted posterior distribution for the population estimate in that grid cell.
- “agesexid” column contains the region ID for the age-sex proportions that are provided in GHA_population_v1_0_agesex_table.csv and GHA_population_v1_0_agesex_regions.tif.
- “area” contains the total settled area in hectares. This corresponds to the total building area from Dooley et al (2020) in their raster GHA_buildings_v1_1_total_area.tif.

5. GHA_population_v1_0_tiles.zip

This tiled web map allows for rapid display of the approximatively 100 m gridded population estimates across the study area (i.e. GHA_population_v1_0_gridded.tif). These can be used to develop web applications for the model results. The tiles were created in XYZ format (i.e. compatible with Leaflet) with full coverage of the study area for the zoom levels 1 to 14. These tiles are source data for the woprVision web application (https://wopr.worldpop.org/woprVision).

ASSUMPTIONS AND LIMITATIONS

The files included in this data release have undergone an internal WorldPop peer review in which the data were assessed by two researchers not involved in the work. The internal review was intended to ensure that methods were appropriate, documentation was adequate, the results were fit for purpose, and to identify assumptions and limitations that end-users need to be aware of. Specific locations where potential issues were identified are included in Appendix A of this release statement.

We labeled the population estimates as 2018 because this was the most common satellite image acquisition year (Fig. 1) for imagery used by Maxar Technologies and Ecopia.AI (2020) to derive the building footprints. The year of satellite images varied across the country from 2009 to 2019, depending on the exact location. Refer to the raster *GHA_buildings_v1_1_imagery_year.tif* (Dooley et al 2020) for a map of the imagery years for building footprints.

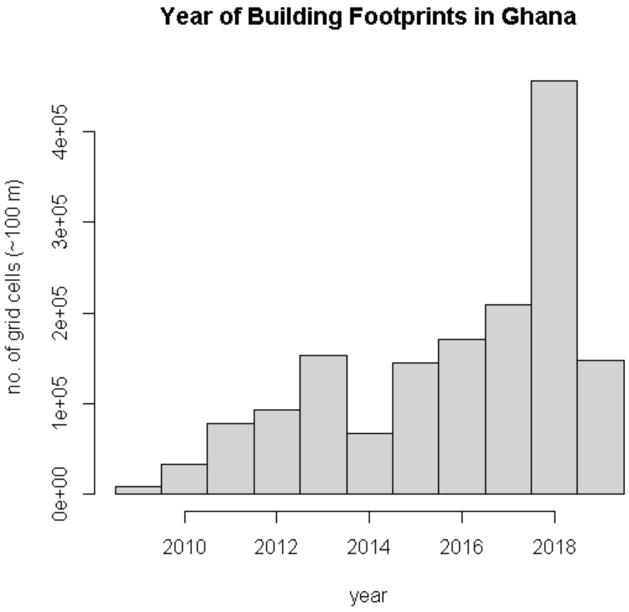

Figure 1. Year of satellite imagery used to create building footprints.

We used data from the 2010 Ghana census to estimate the number of people per household and the proportion of the population in each age-sex group for urban and rural settlements in various regions. This assumes that the patterns of people per household and age-sex structure have not changed in these areas since 2010. This was a necessary assumption in the absence of more recent survey data.

We cannot currently identify building footprints that represent non-residential structures in Ghana. This results in the model predicting populations for buildings that are not residential. This can strongly impact pixel-level predictions in areas that area mostly non-residential buildings. However, population estimates for larger areas should be more robust to this problem because our estimates of the average number of housing units per building footprint includes these non-residential building footprints and should therefore account for the proportion of non-residential buildings across larger areas fairly accurately.

This method cannot map variation in people per household or households per building with precision at the 100 m spatial scale. It estimates these parameters for urban and rural settlements in over 100 geographic units across the country (corresponding to the spatial units from the IPUMS data and the projected census totals used as input data), but it does not attempt to map variation within those units. It accounts for that variation as statistical uncertainty which results in the population estimates having fairly wide credible intervals.

SOURCE DATA

Building footprints. We used rasterized counts of buildings within approximately 100 m grid cells across Ghana (Dooley et al 2020). These were derived from building footprints for sub-saharan Africa that were created from recent high-resolution satellite imagery (Ecopia.AI and Maxar Technologies 2020).

Population counts. We used microdata samples from the 2010 Ghana census that were publicly available from IPUMS International (Minnesota Population Center 2020), whose original source was the Ghana Statistical Services. These data provide a count of the number of people and their age and sex from thousands of housing units with representative national coverage. Each household record is geo-tagged to geographic areas nested within larger regions. These level 1 and level 2 geographic boundaries were available for download from IPUMS International. These publicly available household-level census microdata do not contain GPS coordinates for privacy reasons.

Projected census totals. We used population totals for level 3 administrative units (WorldPop et al. 2018a) that were projected to the year 2018 (WorldPop et al. 2018b) from the 2010 Ghana census.

Settlement types. We used the urban and rural classifications of enumeration areas from the Ghana Statistical Service (Ghana Statistical Services 2010).

METHODS OVERVIEW

The method is a hybrid between commonly used top-down and bottom-up approaches (Wardrop et al 2018) that are often used to map populations when complete census results are not available. It is like bottom-up approaches because it uses household-level survey data that do not have full coverage of the country. It is like top-down approaches because it uses projected census totals to constrain population estimates. Unlike other top-down approaches (Stevens et al 2018), this model enforces a “soft constraint” in which the population totals can deviate significantly from the census projections provided as input data if the weight of evidence from the rest of the model suggests the projections are inaccurate. This approach is unlike top-down and bottom-up approaches because it cannot use high-resolution geospatial covariates other than the count of buildings in each 100 m pixel.

We developed a hierarchical Bayesian model to estimate the following parameters:

1. People per household for urban and rural areas from 102 geographic areas (i.e. level 2 IPUMS boundaries),
2. Households per building footprint for urban and rural areas from 170 geographic areas (i.e. level 3 admin units of the projected census totals),
3. Proportion of the population in each of 36 age-sex groups for urban and rural areas from 10 regions (i.e. level 1 IPUMS boundaries), and
4. Statistical uncertainty for all parameter estimates.

The statistical model used the following components:

1. A Poisson Hurdle model to represent one-inflated distributions of people per household with log-normal overdispersion,
2. A Beta-Binomial model to represent probabilities of single-person households,
3. A Dirichlet-Multinomial model to represent age-sex structure,
4. Hierarchical random effects to share information among settlement types and geographic areas for some parameters (Leasure et al 2020b),
5. A log-normal measurement error model to account for uncertainty in the projected census totals used as input data.

The model was fit using RStan software (Stan Development Team 2020) and the R statistical programming language (R Core Team 2020). Model predictions were computed using the Iridis 5 High Performance Computing Cluster at the University of Southampton.

The fitted statistical model was used to predict the population sizes (and associated parameters) in each 100 m grid cell based on the number of buildings, the settlement classification (urban/rural), and the spatial units to which that location belonged (i.e. level 1, 2, and 3 admin units). Level 1 and 2 units came from IPUMS geographic boundaries and level 3 units came from the projected census totals.

We assessed r-squared, bias, imprecision, inaccuracy, and coverage of credible intervals for parameter estimates using out-of-sample cross-validation. 70% of the IPUMS data were used to fit the model and 30% were held out for cross-validation. For estimating people per household, the model was relatively unbiased (-6.3%; i.e. bias as a percentage of the predicted value) but also imprecise at the household level (76%). This was expected because of the lack of specific location information from the IPUMS data. Our 95% credible intervals contained 96.8% of the out-of-sample data indicating that the credible intervals accurately accounted for this uncertainty in the local-area estimates. For larger areas (i.e. settlement types within IPUMS units and projected census units), the model estimated average household sizes accurately (r-squared = 0.93).

WORKS CITED

Dooley CA, Boo G, Leasure DR, Tatem AJ. 2020. Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. WorldPop, University of Southampton. Source of building footprints: Ecopia Vector Maps Powered by Maxar Satellite Imagery (C) 2020. doi:10.5258/SOTON/WP00677. Data available from https://wopr.worldpop.org/?GHA/Buildings/v1.1

Ghana Statistical Services. 2010. Census enumeration area boundaries and urban/rural classification. Ghana Statistical Services: Accra, Ghana. Non-public dataset obtained via personal communication.

Leasure DR, Bondarenko M, Tatem AJ. 2020a. wopr: An R package to query the WorldPop Open Population Repository, version 0.3.4. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00679. Source code available from https://github.com/wpgp/wopr

Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ. 2020b. National population mapping from sparse survey data: a hierarchical Bayesian modelling framework to account for uncertainty. Proceedings of the National Academy of Sciences. *In press.*

Ecopia.AI and Maxar Technologies. 2020. Building Footprints DRC, Digitize Africa data.

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [Ghana 2010]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D020.V7.2

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Stan Development Team. 2020. RStan: the R interface to Stan. R package version 2.19.3. http://mc-stan.org/.

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. 2018a. Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00651

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. 2018b. Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). Available from ftp://ftp.worldpop.org/GIS/Population/Global_2000_2020/CensusTables/

APPENDIX A

This appendix includes examples of specific locations in which internal reviewers identified potential issues in the results. We provide WGS84 longitude and latitude coordinates to identify locations.

- There is a line or artifact in the grid running N-S on the west side of Accra (around -0.416919, 5.496210). This corresponds to a difference in imagery year used for mapping building footprints. Refer to the raster *GHA_buildings_v1_1_imagery_year.tif* (Dooley et al 2020) for a map of the imagery years for building footprints.

- Building footprints may miss some areas on the edges of cities that did not contain buildings in the satellite images used for building footprints, but buildings have since been built. Potential examples of this are seen in Accra (-0.292867, 5.515089), Sekondi (-1.70742, 4.92462), and New Takarodi (-1.74160, 4.90183).

- We currently do not have a way to identify building footprints that represent non-residential structures. One example is Mallam Market in Accra (-0.279879,5.570480).

- There are some areas missing buildings likely due to cloud cover in the satellite images used to create the building footprints. Potential examples of this are in urban areas on the north side of Accra (-0.254431, 5.695043), Kasoa (-0.43707, 5.50224), and a town in the southwest of the country that is missing population estimates (-2.3708, 5.6142).

- Some pixels contain population estimates where no buildings are visible from Bing and Esri satellite imagery. This may be due to the building footprints erroneously mapping buildings where none occurred. Some examples can be seen along the Mouhoun river (-1.80578,8.62518), in Awudame cemetery (-0.225895,5.566638), and in mines (-2.359, 7.005) and (-2.3426, 7.0265).

- There are over 800 people in a pixel in the center of Kumasi (-1.601718, 6.705048), while neighboring pixels with larger structures have less than half as many people. This may not necessarily be erroneous, but we wanted to highlight the strong effect that differences in building count can have on model results. The building pattern in this pixel is very different than neighboring pixels: lots of small buildings compared to fewer large buildings. A similar example can be seen in the center of Accra (-0.224150 5.545761).