Release Statement

Modelled gridded population estimates for Zambia 2019, version 1.0

9 April 2020

Original Release: 7 April 2020

These data were produced by the WorldPop Research Group at the University of Southampton. This work is part of the GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) project funded by the Bill and Melinda Gates Foundation (BMGF) and the United Kingdom's Department for International Development (OPP1182408). Project partners include WorldPop at the University of Southampton, the United Nations Population Fund (UNFPA), Center for International Earth Science Information Network (CIESIN) in the Earth Institute at Columbia University, and the Flowminder Foundation. The Zambia Statistics Agency supported and facilitated this work, and provided the household survey datasets. The modelling work was led by Claire A. Dooley with support from Douglas R. Leasure. Geospatial data processing was carried out by Heather R. Chamberlain, Claire A. Dooley and Oliver Pannell. Coordination and stakeholder engagement was led by Heather R. Chamberlain and Claire A. Dooley with support from Polly Marshall. Oversight was provided by Andrew J. Tatem and Attila N. Lazar.

Note, these data are operational population estimates and are not official government statistics.

CITATION

WorldPop (School of Geography and Environmental Science, University of Southampton). 2020. Bottom-up gridded population estimates for Zambia, version 1.0. https://dx.doi.org/10.5258/SOTON/WP00662

RELEASE CONTENT

1. ZMB_population_v1_0_gridded.zip
2. ZMB_population_v1_0_admin.zip
3. ZMB_population_v1_0_agesex.zip
4. ZMB_population_v1_0_mastergrid.tif
5. ZMB_population_v1_0_sql.sql
6. ZMB_population_v1_0_tiles.zip

FILE DESCRIPTIONS

ZMB_population_v1_0_gridded.zip

This zip file contains two files:

ZMB_population_v1_0_gridded.tif

This geotiff raster contains estimates of total population size for each approximately 100m grid cell (0.0008333 decimal degrees grid) across Zambia. The values are the mean of the posterior probability distribution for the predicted population size in each grid cell. NA values represent areas that were mapped as unsettled according to building footprints data [8]. Note: This raster is accompanied by one ancillary file that contains metadata (ZMB_population_v1_0_gridded.tif.xml).

ZMB_population_v1_0_uncertainty.tif

This geotiff raster contains estimates of uncertainty in the population estimates within each approximately 100m grid cell across Zambia. 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 uncertainty in population estimates across the country. Uncertainty estimates cannot be summed across grid cells to produce an uncertainty measure for a multi-cell area. Uncertainty for multiple cells can be calculated using the cells' posterior predictions. Note: This raster is accompanied by one ancillary file that contains metadata (ZMB_population_v1_0_uncertainty.tif.xml).

ZMB_population_v1_0_admin.zip

This zip file contains two shapefiles.

ZMB_population_v1_0_admin_level1.shp

This shapefile contains the most up-to-date province (administrative level 1) boundaries. The production of these boundaries data was facilitated by GRID3. These are not official government boundaries. This file contains the mean, median, and upper and lower 95% credible intervals of the summed posterior probability distributions of people for all cells within a province. Note: the shapefile consists of a .shp file with the necessary accompanying files of the same name (with extensions .prj, .shx, .dbf).

ZMB_population_v1_0_admin_level2.shp

This shapefile contains the most up-to-date district (administrative level 2) boundaries. The production of these boundaries data was facilitated by GRID3. These are not official government boundaries. This file contains the mean, median, and upper and lower 95% credible intervals of the summed posterior probability distributions of people for all cells within a district. Note: the shapefile consists of a .shp file with the necessary accompanying files of the same name (with extensions .prj, .shx, .dbf).

Note that the upper and lower credible intervals cannot be summed to produce credible intervals for their province, nor can the credible intervals for the provinces be summed to produce credible intervals at the

national level. Credible intervals for a given area are calculated by summing the posterior probability distributions of people for each cell with the area of focus.

ZMB_population_v1_0_agesex.zip
This zip file contains 36 rasters in geotiff format. Each raster provides gridded population estimates for an age-sex group. Files are labelled with either an “m” (male) or an “f” (female) followed by the number of the first year of the age within the age group 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”. These data were produced using the gridded age-sex proportions from Carioli et al. [1] by multiplying the gridded population estimates (ZMB_population_v1_0_gridded.tif) by the gridded age-sex proportions which differ by region. While this data represents population counts, values contain decimals, i.e. fractions of people. This is because we do not estimate which grid cell each individual in a given age group occupies. 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.

ZMB_population_v1_0_mastergrid.tif
This raster contains 1s for each approximately 100m grid cell (0.0008333 decimal degrees) in Zambia which are considered to contain people, 0 values indicate grid cells within Zambia that were considered unsettled and thus not containing people. NAs show grid cells considered as outside of Zambia. The binary classification was determined based on the presence of at least one building centroid [8] in a grid cell.

ZMB_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 the source data for WorldPop tools used to display and analyze these model results. Note that these 10,000 samples do not necessarily produce a fully converged posterior distribution. The fully converged Bayesian model contained three MCMC chains. We limited the SQL database to 10,000 samples due to file size considerations (the SQL database is approximately 50 GB).

ZMB_population_v1_0_tiles.zip
This tiled web map allows for rapid display of the 100 m gridded population estimates across Zambia. These can be used to develop web applications for these model results. The tiles were created in XYZ format (i.e. compatible with Leaflet) with full coverage of Zambia for zoom levels 1 to 14.

RELEASE HISTORY
Version 1.0 (7 April 2020)
- This is the original release of the data

ASSUMPTIONS AND LIMITATIONS

The assumptions and limitations are as follows:
- Administrative boundary datasets include all areas of the country. Any settled pixels outside of the boundaries shown in ZMB_population_v1_0_admin_level1_population.shp were not included.
- Date of dataset. We believe the year the data represent is early 2019, however, we cannot pinpoint an exact time because the input data was collected at different time points. We also cannot assign a specific month to the dataset for the same reason.

- Some urban areas outside of Lusaka have the highest measures of uncertainty. This is because most of the input data representing urban areas came from Lusaka, which may not be representative of other urban settings.
- The modelled population counts in areas that primarily have non-residential buildings, may be over-estimated. These areas have significantly fewer estimated people than other settled areas of the same size, however, when compared to limited data for these primarily non-residential areas, they appear to be too high. Caution should be taken when using the population data for industrial (and other primarily non-residential) areas.
- We assume that the building footprints data is accurate and that each building polygon corresponds to a building structure.
- There are some small areas where the extent of the building footprints data does not cover the full extent of the district and province boundaries. Because of this, there may be a relatively small number of missing settled grid cells in the districts adjoining the national boundary.

SOURCE DATA

The key datasets used to produce the modelled population estimates are:

- Pre-census pilot mapping exercise (2019) – household totals and GPS locations were used.
- Livestock and aquaculture census survey (2018) [10] – household totals, GPS locations and cluster sample weights were used.
- Saving Mothers, Giving Life survey (2017) [9] – household totals and GPS locations were used.
- Building footprints for Zambia [8] – vector polygons.

METHODS OVERVIEW

The key steps of our approach were as follows:

- Cleaning of survey household GPS location data
- Creating cluster boundaries that incorporated all settled areas of survey clusters using GPS location data
- Developing a Bayesian statistical model that included:
    - An observation sub-model to take advantage of the data on reporting of the missing households present in two of the three survey datasets
    - A process sub-model that describes the relationship between population counts per building area and three geospatial covariates – mean building area, building density and the coefficient of variation in mean building area. The process model included a hierarchical random intercept with a hierarchy of districts nested within provinces and provinces nested with settlement types. Settlement types used were urban, peri-urban and rural
    - A weighted likelihood to account for the different sampling procedures across the three datasets
- Fitting the model to the data, checking model convergence
- Carrying out 10-fold out-of-sample cross-validation to check model fit. This involves checking model statistics when the model is fit to subsets of the data
- Predicting the population size for each settled pixel across Zambia using the model

All data processing and analysis was carried out using R (v.3.6.0) [6] and JAGS (v4.3.0) [5] with the exception of the rasterisation of the district and province boundaries which was carried out in ArcGIS Pro [3].

The concept of bottom-up population modelling for estimating population in the absence of recent census data was described by [7]. Approaches similar to the one used here for Zambia have been carried out for Afghanistan [2], Nigeria [4,11] and DRC [12].

ACKNOWLEDGEMENTS

We thank the Zambia Statistics Agency for anonymising the survey data and providing access to the data in accordance with the data sharing agreement between the University of Southampton and the Zambian government. We thank all the Zambia Statistics Agency staff who participated in the June 2019 survey data workshop which initiated the population modelling work.

WORKS CITED

[1] Carioli, A., Carla Pezzulo, Sophie Hanspal, Theo Hilber, Graeme Hornby, David Kerr, Natalia Tejedor-Garavito, Kristine Nielsen, Linda Pistolesi, Susanna Adamo, Jane Mills, Jeremiah J. Nieves, Heather Chamberlain, Maksym Bondarenko, Chris Lloyd, Greg Yetman, Andrea Gaughan, Forrest Stevens, Cathrine Linard, William James, Alessandro Sorichetta, Andrew J. Tatem. 2020. Population structure by age and sex: a multi-temporal subnational perspective. [draft]

[2] Chamberlain, H.R.*., Clarke D.J.*, Jochem W.C., Wardrop N., Kerr D., Hellali B., Rasuli J., Nabi M., Rahimi S., Ndyanabangi B., Halimi A., Bashir N., Bird T.J., Bengtsson L., Juran S., and Tatem A.J., 2020. National population mapping from spatially incomplete enumeration data [draft].

[3] ESRI. 2018. ArcGIS Pro 2.1 Redlands, CA: Environmental Systems Research Institute

[4] Leasure, D.R., Jochem, W.C., Harper, M., Weber, E.M., Seaman, V., Tatem, A.J. 2020. High resolution population mapping with limited survey data: a hierarchical Bayesian modeling framework to account for uncertainty [draft].

[5] Plummer M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing 124(125):10.

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

[7] Wardrop N.A., Jochem W.C., Bird T.J., Chamberlain H.R., Clarke D., Kerr D., Bengtsson L., Juran S., Seaman V., Tatem A.J. 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences 115, 3529–3537.

DATA CITED

[8] Building Footprints Zambia, Digitize Africa data © 2020 Maxar Technologies, Ecopia.AI

[9] Central Statistical Office (now Zambia Statistics Agency). 2017. Zambia Saving Mothers, Giving Life Maternal Mortality Endline Census in Selected Districts. http://www.savingmothersgivinglife.org/docs/2017-SMGL-Endline-Census-Final-Report.pdf

[10] Ministry of Fisheries and Livestock and Central Statistical Office (now Zambia Statistics Agency). 2019. The 2017/2018 Livestock and Aquaculture Census. https://www.zamstats.gov.zm/phocadownload/Agriculture/The%202017-18%20Livestock%20&%20Aquaculture%20Census%20Summary%20Report.pdf

[11] WorldPop (School of Geography and Environmental Science, University of Southampton). 2019. Bottom-up gridded population estimates for Nigeria, version 1.2. https://dx.doi.org/10.5258/SOTON/WP00655

[12] WorldPop (School of Geography and Environmental Science, University of Southampton). 2020. Bottom-up gridded population estimates for the Kinshasa, Kongo-Central, Kwango, Kwilu, and Mai-Ndombe provinces in the Democratic Republic of the Congo, version 1.0. https://dx.doi.org/10.5258/SOTON/WP00658