Release Statement

Modelled gridded population estimates for Maniema Province in the Democratic Republic of Congo version 4.1.

30 August 2024
Original Release: 30 August 2024

Abstract

This data release provides gridded population estimates (spatial resolution of 3 arc-seconds, approximately 100 m grid cells) for Maniema province in the Democratic Republic of Congo (DRC), along with estimates of the number of people belonging to various age-sex groups. The project team used the Pre-Distribution Registration Survey (PDRS) data from the National Malaria Control Programme (PNLP) collected as part of anti-malarial campaigns in the Democratic Republic of the Congo for 2022, settlement footprint and geospatial covariates to model and estimate population numbers at grid cell level using a Bayesian statistical hierarchical modelling framework. The approach facilitated simultaneous accounting for the multiple levels of variability within the data. It also allowed the quantification of uncertainties in parameter estimates. These model-based population estimates can be considered as most accurately representing the year 2022. This time period corresponds to the PDRS survey date for Maniema. Although the methods were robust enough to explicitly account for key random biases within the datasets, it is noted that systematic biases, which may arise from sources other than random errors within the observed data collection process, are most likely to remain.

These data were produced by the WorldPop Research Group at the University of Southampton. This work was part of the GRID3 – Phase 2 Scaling project, with funding from the Bill and Melinda Gates Foundation (INV-044979). Project partners included the GRID3 Inc., the Center for International Earth Science Information Network (CIESIN) in the Earth Institute at Columbia University and WorldPop at the University of Southampton. The final statistical modelling was designed, developed, and implemented by Chris Nnanatu. Data processing was done by Ortis Yankey and Amy Bonnie with additional support from Tom Abbott and Heather Chamberlain. Project oversight was done by Attila Lazar and Andy Tatem. The PDRS data from the malaria insecticide treated net (ITN) distribution campaigns was collected, processed, anonymised and shared by the PNLP and its implementing partners. The settlement footprint data was prepared and shared by CIESIN

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.

RELEASE CONTENT

1. COD_Maniema_province_population_v4.1_gridded.zip
2. COD_Maniema_province_population_v4.1_agesex.zip

LICENSE

SUGGESTED CITATIONS

Nnanatu C., Yankey O., Bonnie A., Abbott T. J., Chamberlain H., Lazar A. N., Tatem A. J. 2024. Bottom-up gridded population estimates for Maniema province in the Democratic Republic of Congo (2022), version 4.1. https://dx.doi.org/10.5258/SOTON/WP00773

FILE DESCRIPTIONS

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

COD_Maniema_province_population_v4_1_gridded.tif

This geotiff raster contains estimates of total population size for each approximately 100m grid cell (0.0008333 decimal degrees grid) across Maniema province. The values are the mean of the posterior probability distribution for the predicted population size in each grid cell. Gridcells with values of 0 represent areas that were mapped as unsettled according to building footprints data.

COD_Maniema_province_population_v4_1_lower.tif

This geotiff raster contains estimates of the lower bound credible interval (2.5% CI) for each grid cell across Maniema. The values are the 2.5% posterior probability distribution for the predicted population size in each grid cell. The lower bound estimates cannot be summed across grid cells to produce a lower credible interval measure for a multi-cell area. Gridcells with values of 0 represent areas that were mapped as unsettled according to building footprints data.

COD_Maniema_province_population_v4_1_upper.tif
This geotiff raster contains estimates of the upper bound credible interval (97.5% CI) for each grid cell across Maniema. The values are the 97.5% posterior probability distribution for the predicted population size in each grid cell. The upper bound estimates cannot be summed across grid cells to produce an upper bound credible interval measure for a multi-cell area. Gridcells with values of 0 represent areas that were mapped as unsettled according to building footprints data.

COD_Maniema_province_population_v4_1_agesex.zip
This zip file contains 40 geotiff rasters at a spatial resolution of 3 arc-seconds (approximately 100 m). Each raster provides gridded population estimates for an age-sex group per grid cell across Maniema. 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 by 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 “f15_49” (all females between the ages of 15 and 49, inclusive). These data were produced using age-sex proportions from the Malaria Indicator Cluster Survey (MICS) for the DRC for the year 2017. The age-sex proportions were applied to the gridded population estimates (COD_Maniema_province_population_v4_1_gridded.tif) to allocate the population to the different age-sex classes. 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.

RELEASE HISTORY

Version 4.1 (30 August 2024)

- This is the original release of the data for Haut-Katanga province [doi: 10.5258/SOTON/WP00778] and Haut-Lomami province [doi: 10.5258/SOTON/ WP00777].
- This is a major update of the data for Maniema province [doi: 10.5258/SOTON/WP00773] (as described in this release statement) and Tanganyinka province [doi: 10.5258/SOTON/ WP00774].
- This data release utilizes operational National Malaria Control Programme data, composite, openly accessible building footprint datasets and a new mastergrid.

Version 3.0 (4 January 2022) [doi:10.5258/SOTON/WP00720]

- Original release of the population dataset for the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, KasaïOriental, Lomami and Sud-Kivu provinces.

Version 2.0 (27 May 2020) [doi:10.5258/SOTON/WP00669]

- Major revision of the population dataset for the Kinshasa, Kongo Central, Kwango, Kwilu and MaiNdombe provinces based on finer resolution input data.
- The settled extent is no longer derived from settlement data but from building footprints data.
- Population estimates for the different age and sex groups are no longer derived from existing agesex proportions but the original microcensus data.
- Gridded population estimates were added for individual age-sex groups (COD_population_v2_0_agesex.zip).
- Uncertainty tiles "COD_population_v1_0_tiles_uncertainty.zip" were removed because they were discontinued for use in WorldPop web applications.

Version 1.0 (20 May 2019) [doi:10.5258/SOTON/WP00658]

- Original release of the population dataset for the Kinshasa, Kongo Central, Kwango, Kwilu and Mai-Ndombe provinces.

ASSUMPTIONS AND LIMITATIONS

These population estimates most likely represent the 2022 time, but because of the different ages of the input data used to build the model, a precise time point cannot be allocated. The PDRS data that was used as the response variable was collected in May 2022, while geospatial covariates data were collected from different time periods between 2021 and 2023. Similarly, the CIESIN settlement layers were produced in 2024. The inherent heterogeneity in the temporal alignment of these covariates may introduce uncertainties and potential inaccuracies in the modelling process.

Data on population per household (household size), collected during ITN distribution campaigns, was aggregated to calculate total population count for a given spatial unit. Given that the number of ITNs received by a household is proportional to the household size, there is an incentive for respondents to potentially inflate counts of population per household. The presence of inflated household sizes in the input population data would likely introduce systematic biases in the modelled estimates.

The model does not directly account for external factors such as migration, displacement, or sudden demographic changes, which could significantly influence population dynamics. Consequently, the estimates may not fully reflect dynamic population shifts occurring beyond the scope of the input data.

Grid cell alignment is based on a mastergrid. Please note that the mastergrid used for this version (v4.1), differs from previous versions of gridded population estimates for DRC (v1.0, 2.0 and 3.0) and other existing WorldPop data products. The mastergrid used for this version has been updated so as to ensure grid cell alignment with future WorldPop data products.

SOURCE DATA

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

PDRS Data

The input population dataset used for the population modelling for Maniema province was the PDRS malaria bednet campaign data. The PDRS dataset, which was collected in 2022, provided detailed information on a given household for which a bednet was issued, such as the household size, the number of bednets issued, the number of children in the household, the number of males, and the number of females, among others. Household sizes ranged from 1 to 50 people, with a median of 6 individuals per household.
Although the malaria bednet campaign was designed to distribute bednet to every household within the province, a preliminary exploratory data analysis carried out on the PDRS data indicated that some households were not visited during the campaign. Specifically, we found that some health areas (Kibeleketa, Kyolo, Oku, Kitamuna, Kasese, Kabakaba, Elimu, and Kasesa) were not visited, while others were not completely covered. Health area boundary was obtained from CIESIN (CIESIN, 2023).
The GPS points of all households within the Maniema province were provided in the PDRS data. We implemented population modelling for small spatial units, utilising

unofficial boundaries similar to census Enumeration Areas ("pre-EAs"; Qader et al., 2024). The household-level data on population counts was spatially aggregated to these spatial units, by summing the household size data for all GPS points within each pre-EA boundary.

Settlement Data

Settlement data was provided by CIESIN in the form of raster files (CIESIN, 2024). We obtained two different settlement products, namely (i) settlement area, which indicates the area of a grid cell that is settled; and (ii) building count, which is the number of buildings within a given cell. Each of these settlement layers was used in separate analyses together with the observed population count and ancillary geospatial data in robust statistical modeling. After using each settlement layer in the analysis, we compared model metrics and the gridded population layer from both layers. Settlement building count provided more realistic population numbers at the gridcell level and hence was retained for the final population predictions.

Geospatial Covariates

A wide variety of geospatial covariates, which are related to population distribution, were considered in the modelling. These geospatial covariates include land uses and land cover data, climate variables such as temperature and rainfall, physical features and infrastructure such as roads and schools, and conflict data. Population model covariates were selected using a generalized linear model (GLM) – based stepwise selection method. The selected covariates were further accessed for multi-collinearity and statistical significance. Eventually, of the 85 geospatial covariates initially tested, 5 were retained as the best fit covariates with variance inflation factor (VIF) of less than 5. The descriptions of these final geospatial covariates are presented in Table 1 below.

Table 1. Selected geospatial covariates for the modelling.

Description	**Source**	**Link/Reference**
Euclidean distance to Cropland/Natural vegetation LC type per 100m pixel for 2020.	WorldPop	Woods et al (2024)
Coefficient of Variation – Length of Microsoft Building Footprint	Microsoft	https://github.com/microsoft/GlobalMLBuildingFootprints

Euclidean distance to Educational Facilities 2023	OSM	https://www.openstreetmap.org
Euclidean distance to Main Roads 2023	OSM	https://www.openstreetmap.org
Digital Elevation Model (Slope)	SRTM	https://www.viewfinderpanoramas.org/dem3.html

Age-Sex Proportions (MICS Data)

We used the 2017 MICS dataset (INS, 2017) to calculate the age-sex proportions for Maniema. We multiplied our gridded population estimates (COD_Maniema_province_population_v4_1_gridded.tif) by the gridded age-sex proportions to produce COD_Maniema_province_population_v4.1_agesex.zip.

METHODS OVERVIEW

The key steps of our approach were as follows:

- Cleaning and summarizing the household sizes from the PDRS dataset to get the total population at the pre- Enumeration Area (pre-EA) level (Qader at al. 2024).
- Geospatial covariates were subjected to robust covariate selection for model training and parameter estimation.
- We developed a hierarchical Bayesian statistical model using the INLA-SPDE approach (Lindgren et al. 2011) to fit and predict the population count.
- Population estimates were predicted at grid cell level using the grid cell values of the covariates selected at the model training level.

Statistical Modelling

In general, within the context of bottom-up population modelling (Leasure et al. 2022, Boo et al., 2022; Darin et al., 2022, Nnanatu et al. 2022), the observed population count at area unit k, y_k, is a Poisson distributed random variable with mean parameter $\lambda_k = \bar{d}_k B_k$ where k is the estimation unit (e.g., enumeration area), while $\bar{d}_k$ and B_k are the mean parameter of the corresponding population density and the number of buildings/settled area, respectively. That is,

$$y_k \sim \text{Poisson}\left(\bar{d}_k B_k\right) \qquad (1)$$

Then, the transformed mean population density $\bar{d}_k$ is assumed to be linked to a set of geospatial covariates with log-link function:

$$\log(\bar{d}_k) = \mu + \sum_{j=1}^{J} \beta_j x_{kj} + \sum_{l=1}^{L} f_l(z_{kl}) \qquad (2)$$

where μ is the intercept parameter, $\beta = (\beta_1, \dots \beta_J)$ is a vector of fixed effects coefficients of the $(x_1, \dots, x_J)$ geospatial covariates; $f_l(.)$ is a function of L random effects covariates including those that capture variability in the population estimates due to settlement type, cluster location and spatial autocorrelations. The population density (defined as people per building or people per settled area) is assumed to be a Gamma distributed random variable with parameters α and γ with mean and variance given by $\bar{d}_k = \alpha/\gamma$ and $\sigma_d^2 = \alpha/\gamma^2$, respectively.

The inclusion of spatial autocorrelation requires the use of computationally efficient statistical modelling software. Thus, the integrated nested Laplace approximation (INLA; Rue et al 2009; Lindgren et al., 2011) is used via the R-INLA statistical package. Note that the method described above predicts population count at regular grid cells using the parameter values trained at the cluster/pre-EA level by calculating the predicted grid-cell level population density as

$$\hat{d}_g = \exp\left(\hat{\mu} + \sum_{j=1}^{J} \hat{\beta}_j x_{gj} + \sum_{l=1}^{L} \hat{f}_l\left(z_{g_l}\right)\right) \qquad (3)$$

where $\{x_g\}_{g=1}^{G}$ are the corresponding grid cell level values of the geospatial covariates used in training the model at the cluster level, so that the overall predicted population count across the G 100m by 100m grid cells is given by

$$\widehat{pop} = \sum_{g=1}^{G} B_g \hat{d}_g \qquad (4)$$

where B_g is the corresponding building count or the size of settled area in grid g. We assumed default INLA priors for each of the parameter estimates which have been found to be robust..

In this study, we approached the population modelling using building count settlement layer. Thus, population density was defined as people per building count. The novelty of the modelling approach utilised here is that it allows for the adjustment of potential systematic bias in the input population data within a coherent Bayesian hierarchical population modelling framework while at the same time adjusting for spatial autocorrelation within the observed data.

All data processing and analysis was carried out using R (v.4.2.2) (R Core Team, 2023) and INLA (v 22.05.07) (Rue et al. 2009). The concept of bottom-up population modelling for estimating population in the absence of recent census data was described by Leasure et al. (2020). Approaches similar to the one used here for Maniema have been carried out for Papua New Guinea (WorldPop and NSO PNG, 2022) and Cameroun (Nnanatu et al, 2022).

ACKNOWLEDGEMENTS

We thank the DRC PNLP and its implementing partners for providing access to the anonymised household data collected during malaria ITN distribution campaigns, in accordance with the relevant data sharing agreements. The whole WorldPop group are acknowledged for overall project support. We thank Attila Lazar for reviewing the data and providing thoughtful suggestions prior to this release.

WORKS CITED

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., ... & Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature communications*, *13*(1), 1330.

Center for International Earth Science Information Network (CIESIN), Columbia University and Ministère de la Santé Publique, Hygiène et Prévention, Democratic Republic of the Congo (2023). *GRID3 COD - Health Areas v2.0*. New York: Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). Unpublished.

Center for International Earth Science Information Network (CIESIN), Columbia University. (2024). GRID3 COD - Settlement Extents v3.0 alpha. Unpublished.

Darin, E., Kuépié, M., Bassinga, H., Boo, G., Tatem, A. J., & Reeve, P. (2022). The Population Seen from Space: When Satellite Images Come to the Rescue of the Census. *Population*, *77*(3), 437-464.

Flowminder Foundation, École de Santé Publique de Kinshasa (ESPK), WorldPop (University of Southampton), Bureau Central du Recensement (BCR). 2021. Microcensus survey in the provinces of Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami, and Sud-Kivu (Democratic Republic of the Congo). Version 1.5. [Dataset].

INS (2017)*, Enquête par grappes à indicateurs multiples, 2017-2018, rapport de résultats de l'enquête. Kinshasa, République Démocratique du Congo.*

Leasure, D.R., Jochem, W.C., 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 [in review at Proceedings of the National Academy of Sciences of the United States of America].

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423–498

Nnanatu C.C., Yankey O., Abbott T. J., Lazar A. N., Darin E., Tatem A. J. 2022 Bottom-up gridded population estimates for Cameroon (2022), version 1.0. https://dx.doi.org/10.5258/SOTON/WP00784

Qader S H, Batana Y. M., Kosmidou-Bradley W., Skoufias E., Tatem A. J. 2024. Automatic pre-Enumeration Areas (pre-EAs) delineation and national sampling frame for the Democratic Republic of Congo. Policy Research Working Paper; No. (under review). DRC - Automatic Pre-Enumeration Area Delineation for National Sample Frame Data Report | Data Catalog (worldbank.org)

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Rue, H., Martino, S., & Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society:Series b (statistical methodology), 71(2), 319-392

UCLA-DRC Health Research and Training Program (University of California, Los Angeles) and Kinshasa School of Public Health. 2017 and 2018. Kinshasa, Kongo Central and former Bandundu microcensus survey data.

D. Woods, T. McKeen, A. Cunningham, R. Priyakanto, A. Soricheta , A.J. Tatem and M. Bondarenko. 2024 "WorldPop high resolution, harmonised annual global geospatial covariates. Version 1.0" University of Southampton: Southampton, UK. DOI:10.5258/SOTON/WP00772

WorldPop and National Statistical Office of Papua New Guinea. 2022. Census-independent population estimates for Papua New Guinea (2020-21), version 1.0. WorldPop, University of Southampton. DOI: 10.5258/SOTON/WP00763