Release Statement

Modelled gridded population estimates for Tshuapa Province in the Democratic Republic of Congo, version 4.4.

22 December 2025

Abstract

This data release provides gridded population estimates (spatial resolution of 3 arc-seconds, approximately 100 m grid cells) for Tshuapa 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 2023, settlement footprints 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 2023. This time period corresponds to the PDRS survey date for Tshuapa. 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 & Melinda Gates Foundation (INV-044979). Project partners included GRID3, the Center for Integrated Earth System Information (CIESIN) within the Columbia Climate School at Columbia University, and WorldPop at the University of Southampton. The final statistical modelling was designed, developed, and implemented by Ortis Yankey. Data processing was done by Ortis Yankey with additional support from Heather Chamberlain. Project oversight was done by Chris Nnanatu, Attila Lazar, and Andy Tatem. The PDRS data from the malaria insecticide treated net (ITN) distribution campaigns were collected, processed, anonymised, and shared by the PNLP and its implementing partners. The settlement footprint data was prepared and shared by CIESIN (2024). The data has been clipped to GRID3-CIESIN health area extent (version 8.0) (CIESIN, 2025).

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. These data are operational population estimates and are not official government statistics.

RELEASE CONTENT

1. COD_Tshuapa_province_population_v4.4_gridded.zip
2. COD_Tshuapa_province_population_v4.4_agesex.zip

LICENSE

SUGGESTED CITATIONS

Yankey O., Nnanatu C., Chamberlain H., Lazar A. N., Tatem A. J. 2025. Bottom-up gridded population estimates for Tshuapa Province in the Democratic Republic of Congo (2022), version 4.4. WorldPop, University of Southampton. doi: https://dx.doi.org/10.5258/SOTON/WP00877

FILE DESCRIPTIONS

The projection for all GIS files is the geographic coordinate system WGS84 (World Geodetic System 1984). Kindly note that 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 somewhere in those four grid cells.

COD_Tshuapa_province_population_v4_4_gridded.tif

This geotiff raster contains estimates of total population size for each approximately 100m grid cell (0.0008333 decimal degrees grid) across Tshuapa 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_Tshuapa_province_population_v4_4_lower.tif
This geotiff raster contains estimates of the lower bound credible interval (2.5% CI) for each grid cell across Tshuapa. 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_Tshuapa_province_population_v4_4_upper.tif
This geotiff raster contains estimates of the upper bound credible interval (97.5% CI) for each grid cell across Tshuapa. 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_Tshuapa_province_population_v4_4_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 Tshuapa. 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 older 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 2024 WorldPop Global subnational population pyramids for the DRC. The age-sex proportions are available per a given province. Hence, we applied the age-sex proportions for Tshuapa to the gridded population estimates (COD_Tshuapa_province_population_v4_4_gridded.tif) to allocate the population to the different age-sex classes.

RELEASE HISTORY

Version 4.3 (22 December 2025)

- This is the original release of the data for Tshuapa Province [doi: 10.5258/SOTON/WP00877] (as described in this release statement).
- This data release utilizes operational National Malaria Control Programme data, composite, openly accessible building footprint datasets and a new mastergrid.
- This data is released as part of a collection of population estimates for all 26 DRC provinces: https://wopr.worldpop.org/?COD/Population/v4.4

ASSUMPTIONS AND LIMITATIONS

These population estimates most likely represent the 2023 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 2023, while geospatial covariates data were collected from different time periods between 2020 and 2023. Similarly, the CIESIN settlement layers were produced in 2024. The inherent heterogeneity in the temporal alignment of these datasets used in the modelling 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 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 (v4.2), is the same as in version 4.1, but 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 Tshuapa Province was the PDRS malaria bednet campaign data. The PDRS dataset, which was collected in 2023, 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.

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, while others were not completely covered.

The GPS points of all households within the 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 all buildings whose centroid falls within a given cell, and (ii) building count, which is the number of building centroids 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 use 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 assessed for multi-collinearity and statistical significance. Eventually, of the 80 geospatial covariates initially tested, 7 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
Coefficient of variation – building count 2024	WorldPop	Woods et al (2024)
Euclidean distance to Herbaceous/ Grassland landcover type 2020	WorldPop	Woods et al (2024)
Euclidean distance to water bodies 2021	WorldPop	Woods et al (2024)
Standard deviation temperature 2022	Copernicus	https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-agrometeorological-indicators?tab=form
Euclidean distance to OSM places of worship 2023	OSM	https://www.openstreetmap.org/#map=3/68.59/70.05
Mean burnt area per 100m pixel in 2021	Copernicus	https://land.copernicus.eu/global/products/ba
Euclidean distance to ACLED riot locations 2022	ACLED	https://acleddata.com/

Age-Sex Proportions (MICS Data)

We used the 2024 WorldPop Global subnational population pyramids (Bondarenko et al 2025) to calculate the age-sex proportions for Tshuapa. The MICS data had age-sex proportions per a given province. Hence, we multiplied our gridded population estimates (COD_Tshuapa_province_population_v4_4_gridded.tif) by the age-sex proportions(grouping) to produce COD_Tshuapa_province_population_v4.4_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).
- Household sizes from the PDRS data point ranged between 0 and 20. Out of 421677 PDRS data points, 589 data points had a household size of 0. Points with a household size of 0 were non-residential buildings such as hotels, marketplaces, military training camps, etc. These observations were hence removed from the dataset.

- 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, $\mathrm{y_k}$, 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,

$$y_k = (\bar{d}_k B_k) \qquad (1)$$

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

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

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 two competing settlement layers, i.e., building count and building area to define population density. Thus, we had two separate models. In the first model, population density was defined as people per building count, and in the second model, population density was defined as people per settled area. These two models were fitted, and the best model based on model metrics was selected for the final predictions.

In this study, we used building count to define population density. Within this framework, we tested three different model re-parameterizations. The first model, Model 1, included fixed effects for the geospatial covariates and a random effect for the Global Human Settlement Layer Degree of Urbanization classes (GHSL-SMOD) (Schiavina et al. 2023). Model 2 extended this by incorporating an additional random effect at the cluster level. The final model, Model 3, further included a spatial random effect component in addition to the specifications in Model 2. These three models were compared, and the model with the best fit was selected for final predictions.

Model fit checks.

Model fit checks and model selection of the three models described above relied primarily on a constellation of model fit metrics, including the absolute bias (BIAS), the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), the Deviance Information Criterion (DIC) and the Pearson correlation coefficient (CORR). A lower value for the absolute bias, MAE and the RMSE and the DIC indicates a better-fit model. A higher value for the Pearson correlation coefficient indicates a better-fit model. Table 2 below provides the model-fit metrics across the three models. Based on the model fit checks, model 3 provided the best fit, and the final population predictions at the grid cell level were based on this model.

Table 2. Model fit metrics.

Models	BIAS	RMSE	MAE	DIC	Corr
Model1	-23.95	519.87	265.52	28795.36	0.75
Model 2	-3.70	453.06	226.26	28686.14	0.81
Model 3	-9.36	432.68	202.83	27944.51	0.83

The novelty of the modelling approach utilised here is that it allows for the adjustment of potential systematic bias in the two settlement layers used as input in defining population density within a coherent Bayesian hierarchical population modelling framework while at the same time adjusting for spatial autocorrelation within the observed data.

Data processing and analysis were conducted in R (v.4.4.1) (R Core Team, 2024), using the packages INLA (v.24.6.27) (Rue et al., 2009), tidyverse (v.2.0.0) (Wickham et al., 2019), sf (v.1.0-17) (Pebesma and Bivand, 2023), and terra (v.1.7-78) (Hijmans et al., 2024). 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 Haut-Katanga have been carried out for Papua New Guinea (Nnanatu et al. 2024) 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 WorldPop group is acknowledged for overall project support particularly Attila Lazar, Heather Chamberlain and Gianluca Boo for reviewing the data and providing thoughtful suggestions prior to this release.

WORKS CITED

Armed Conflict Location and Event Data (ACLED) (2024). Data Export Tool. https://acleddata.com/.

Bondarenko M., Priyatikanto R., Tejedor-Garavito N., Zhang W., McKeen T., Cunningham A., Woods T., Hilton J., Cihan D., Nosatiuk B., Brinkhoff T., Tatem A., Sorichetta A. (2025) Constrained estimates of 2015-2030 total number of people per grid square broken down by gender and age groupings at a resolution of 3 arc (approximately

100m at the equator) R2024B version v1. Global Demographic Data Project - Funded by The Bill and Melinda Gates Foundation (INV-045237). WorldPop - School of Geography and Environmental Science, University of Southampton. DOI:10.5258/SOTON/WP00805

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 Integrated Earth System Information (CIESIN), Columbia University, Ministère de la Santé Publique, Hygiène et Prévention, Democratic Republic of the Congo, and GRID3. 2025. GRID3 COD - Health Zones v8.0 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-Katanga, Ituri, Kasaï, Kasaï-Oriental, Tshuapa, and Tshuapa(Democratic Republic of the Congo). Version 1.5. [Dataset].

Hijmans, R. J., Bivand, R., Dyba, K., Pebesma, E., & Sumner, M. D. (2024). terra: Spatial Data Analysis. https://CRAN.R-project.org/package=terra.

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 modelling framework to account for uncertainty. Proceedings of the National Academy of Sciences of the United States of America, 117(39): 24173–24179.

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

Nnanatu, C., Bonnie, A., Joseph, J., Yankey, O., Cihan, D., Gadiaga, A., ... & Tatem, A. (2024). Small area population estimation from health intervention campaign surveys and partially observed settlement data.

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

Schiavina M., Melchiorri M., & Pesaresi M. (2023). GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030). *European Commission, Joint Research Centre (JRC)*. doi:10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA.

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.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R, … &, Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software,* 4(43), 1686. doi:10.21105/joss.01686.

Woods D, 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