**Release Statement**

**Modelled gridded population estimates for Kwango 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-metre grid cells) for Kwango Province in the Democratic Republic of Congo (DRC). 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 DRC between the periods 2021 to 2023, settlement extents 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. 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 Gates Foundation (INV-044979). Project partners included GRID3 Inc, 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 Chris Nnanatu. Data processing was done by Ortis Yankey with additional support from Heather Chamberlain, Assane Gadiaga and Krishnaveni KS. Project oversight was done by Attila Lazar, Chris Nnanatu 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 extent 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](mailto:release@worldpop.org).*



**RELEASE CONTENT**

1. COD_Kwango_province_population_v4.4_gridded.zip
2. COD_Kwango_province_population_v4.4_agesex.zip

**LICENSE**



**SUGGESTED CITATION**

Nnanatu C., Yankey O., Chamberlain H., Krishnaveni K S, Lazar A. N., Tatem A. J. 2025. Bottom-up gridded population estimates for Kwango Province in the Democratic Republic of Congo (2023), version 4.4. WorldPop, University of Southampton. doi: https://dx.doi.org/10.5258/SOTON/WP00863

**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 for each settled pixels, 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. Grid cells within the national boundary with values of NA represent areas that were mapped as unsettled according to building footprints data, while any other NA values represent areas mapped as being outside national boundary.

**COD_Kwango_province_population_v4_4_gridded.tif**

This geotiff raster contains estimates of total population size for each approximately 100-metre grid cell (0.0008333 decimal degrees grid) across Kwango. The values are the mean of the posterior probability distribution for the predicted population size in each grid cell.

**COD_Kwango_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 the across Kwango. 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.

**COD_Kwango_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 Kwango. 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.

**COD_Kwango_province_population_v4_4_agesex.zip**
This zip file contains 40 geotiff rasters at a spatial resolution of 3 arc-seconds (approximately 100-metre grid cells). Each raster provides gridded population estimates for an age-sex group per grid cell across Kwango. 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 2024 WorldPop Global subnational population pyramids for the DRC. The age-sex proportions are available per a given province. The age-sex proportions were applied to the gridded population estimates (COD_Kwango_province_population_v4_4_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 4.4 (22 December 2025)

- This is the original release of population for Kwango Province in the DRC [doi: 10.5258/SOTON/WP00863] (as described in this release statement).
- 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 year 2023, but because of the different ages of the input data used to build the model, a more precise time point cannot be assigned. The PDRS data that was used as the response variable was collected between 2001 and 2003, 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.

Grid cell alignment is based on a mastergrid. Note that this version's (v4.4) mastergrid aligns with versions 4.1, 4.2 and 4.3 but does not align with previous DRC gridded population layers, namely versions v1.0, v2.0, v3.0. We updated the mastergrid in 2024 to ensure grid cell alignment across all new 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 Democratic Republic of Congo (DRC) population modelling was derived from the PDRS malaria bednet campaign data, collected between 2021 and 2023 across the country's provinces. The PDRS dataset included GPS coordinates of all households within each province. For the modelling, we constructed small-area spatial units by aggregating 100 m grid cells into larger blocks considering the building density differences across the DRC. In urban areas, the minimum aggregation was 4 × 4 cells, while in rural areas, where building density was lower, larger aggregations were applied to ensure meaningful population estimates.

### Settlement Data

We used the The GRID3 COD - Settlement Extents v3.1 (CIESIN, 2024) as the input settlement layer for the population modelling. We used the spatial point shapefile depicting the centroids of settled grid cells at 3-arc seconds (or ~100 meters). This data contained the building count and the building area. We converted the building count per point by rasterizing it to WorldPop mastergrid. The building count raster was then used in population modelling.

### 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 85 geospatial covariates initially tested, 21 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 modelling.

| Description | Source | Link/Reference |
|---|---|---|
| Coefficient of variation building area (3 X 3 moving window) | WorldPop | Woods et al (2024) |
| Coefficient of variation building area (9 X 9 moving window) | WorldPop | Woods et al (2024) |

| | | |
|---|---|---|
| Coefficient of variation building count (9 X 9 moving window) | WorldPop | Woods et al (2024) |
| Euclidean distance to cropland natural vegetation 2020 | WorldPop | Woods et al (2024) |
| Euclidean distance to sparse vegetation 2020 | WorldPop | Woods et al (2024) |
| Euclidean distance to Tree/Herbaceous cover 2020 | WorldPop | Woods et al (2024) |
| Euclidean distance to Urban areas 2020 | WorldPop | Woods et al (2024) |
| Euclidean distance to bare areas 2020 | WorldPop | Woods et al (2024) |
| GHS Built Surface Non-Residential | GHS | https://human-settlement.emergency.copernicus.eu/download.php |
| GHS Building volume | GHS | https://human-settlement.emergency.copernicus.eu/download.php |
| Dry matter productivity 2022 | Copernicus | https://land.copernicus.eu/global/products/ba |
| Precipitation 2022 | Copernicus | https://land.copernicus.eu/global/products/ba |
| Euclidean distance to ACLED Strategic development locations 2022 | ACLED | https://acleddata.com/ |
| Euclidean distance to OSM education facilities 2023 | OSM | https://www.openstreetmap.org |
| Euclidean distance to OSM place of worship 2023 | OSM | https://www.openstreetmap.org |
| GHS – residential layer 2020 | GHS | https://human-settlement.emergency.copernicus.eu/download.php |
| NDVI values 2021 | Copernicus | https://land.copernicus.eu/global/products/ba |
| Slope | SRTM | https://www.viewfinderpanoramas.org/dem3.html |
| Nighttime light | Earth Observation Group | https://eogdata.mines.edu/products/vnl/ |
| Euclidean distance to GRID3 health facilities | GRID3 | https://data.grid3.org/search?tags=COD |
| Euclidean distance to Grid3 all roads | GRID3 | https://data.grid3.org/search?tags=COD |

### Age-Sex Proportions

We used the 2024 WorldPop Global subnational population pyramids (Bondarenko et al 2025) to calculate the age-sex proportions for Bas-Uele. We multiplied our gridded population estimates (COD_Kwango_province_population_v4_4_gridded.tif) by the gridded age-sex proportions to produce COD_Kwango_province_population_v4.4_agesex.zip.

## METHODS OVERVIEW

The key steps of our approach were as follows:

- Cleaning household dataset from the PDRS by removing extreme outliers from the data.
- Summarizing the household sizes from the PDRS dataset to get the total population at the modelling unit
- 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.

### Data cleaning

The data cleaning process followed a series of steps to ensure that our population data was reliable and consistent:

- We used microcensus data collected in selected provinces of the DRC in 2018 and 2021 as a benchmark for cleaning the PDRS dataset. These provinces included Kinshasa, Kongo-Central, Kwango, Kwilu, Mai-Ndombe, Haut-Katanga, Haut-Lomami, Ituri, Kasai, Kasai-Oriental, Lomami, and Sud-Kivu.
- For each province with microcensus data, and for each modelling unit within the province, we calculated the mean population density (i.e. people per building) by dividing the observed total (microcensus) population with the number of buildings extracted from the CIESIN dataset. We then summarized these densities for each province, identifying key statistics such as the maximum observed density within the province.

- We repeated the same process using the PDRS data by calculating the mean population density (i.e. people per building) for each modelling unit from the total observed population and the extracted (CIESIN) building counts.
- To find and address the likely extreme values, we removed any modelling units, where the mean PDRS population density value was higher than the maximum population density observed in the microcensus for that province. For provinces without microcensus data, we applied a global threshold of 29.98 people per building, that was calculated as the maximum population density of all the microcensus observations.
- Furthermore, we removed modelling units with no CIESIN building count and also those without observed PDRS population.

After implementing the above data cleaning steps, 4% of the modelling unit were dropped from the data. This process helped us ensure that observed spatial demographic characteristics were considered, and thus the dataset was free from unrealistic high values and suitable for use in the modelling work that followed.

**Statistical Modelling**

All data processing, statistical modelling, and analyses were carried out using R version 4.4.2 (R Core Team, 2023), 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). Bayesian hierarchical modelling was implemented using the R-INLA package version 24.12.11 (Rue et al. 2009). Modelled estimates of the population were produced using a bottom-up population modelling framework (Wardop et al., 2018), which utilises a Bayesian statistical inference framework that can be implemented using either a Markov chain Monte Carlo (MCMC)-based strategy (Leasure et al., 2020; Boo et al. 2022; Darin et al., 2022) or the integrated nested Laplace approximation in conjunction with the stochastic partial differential equation (INLA-SPDE; Rue et al., 2009; Lindgren et al., 2011) techniques recently developed by Nnanatu et al. (2022) in the context of Cameroon (Nnanatu et al., 2022; Nnanatu et al., 2025a), and applied in Papua New Guinea (Nnanatu et al., 2024) and the Democratic Republic of Congo (e.g., Nnanatu et al., 2025b).

**Model Specification**

In general, the population count $N_i$ at a given (ideally geolocated) area unit is assumed to be Poisson-distributed, such that $N_i \sim Poisson(\lambda_i)$. However, in the context of small area population modelling (Leasure et al., 2020; Boo et al. 2022 ; Darin et al.,2022; Nnanatu et al., 2022; Nnanatu et al., 2024a; Nnanatu et al., 2024b; Nnanatu et al., 2025), a key assumption of the Poisson model which requires both mean and variance to be equal is often violated due to overdispersion in which case $mean(N_i) \neq var(N_i)$.

For this reason, the mean parameter $\lambda_i$ is usually expressed in terms of population density to account for spatial aggregation error (e.g., Leasure et al 2020, Nnanatu et al 2022). Typically, the mean parameter is given as $\lambda_i = \mu_i B_i$, where $B_i$ is the total number of buildings within a pre- enumeration area (pre-EA) $i$ and

$$D_i = \frac{pop_i}{B_i} \qquad (1)$$

is the population density defined as the number of people per building which follows a Gamma distribution given by

$$D_i \sim Gamma(\alpha_1, \alpha_2) \qquad (2)$$

where $\alpha_1$ and $\alpha_2$ are the shape and rate parameters with mean $\mu_i = \alpha_1/\alpha_2$ and variance $\phi = \alpha_1/\alpha_2^2$, respectively. The predicted population density $\widehat{D}_i$ for pre-EA $i$ is given by

$$\widehat{D}_i = \exp\left(X_i^{\boldsymbol{T}}\boldsymbol{\beta} + Z_i^T\boldsymbol{\gamma} + \xi(s_i) + \zeta_i\right) \qquad (3)$$

where $X$ and $Z$ are the design matrices of fixed effect covariates (e.g., average annual precipitation, average annual temperature, distance to crop land) and random effects (e.g., settlement type), respectively. Also, the terms $\boldsymbol{\beta} \in \mathbb{R}^{(K\times 1)}$ and $\boldsymbol{\gamma}$ are the vectors of fixed effects regression parameters and random effects variances, respectively. While the terms $\xi(s_i)$ and $\zeta_i$ are the spatially varying and spatially independent random effects accounting for spatial autocorrelations and dissimilarities between observations, respectively. We have that the term $\xi(s_i)$ is a Gaussian Random Field (GRF) such that

$$\xi(s_i) \sim GRF(\mathbf{0}, \Sigma) \qquad (4)$$

where $\Sigma$ is a dense covariance matrix. The INLA-SPDE approach allows us to approximate the GRF using a computationally efficient Gaussian Markov Random Field (GMRF) by discretising the continuous spatial domain using mesh (Lindgren et al., 2011). The random term $\zeta_i$ is assumed to follow a zero-mean Gaussian distribution specified by

$$\zeta_i \sim Normal\left(0, \sigma_\zeta^2\right) \qquad (5)$$

where $\sigma_\zeta^2 > 0$ is a variance parameter. Then, finally, the predicted population counts at grid cell $g$ is obtained as

$$\widehat{N}_g = \widehat{D}_g \times B_g \qquad (6)$$

where $\widehat{D}_g$ is the predicted population density in grid cell $g$ using the corresponding grid cell covariate values and the model parameter values based on equation (3); $B_g$ is the corresponding building count for grid cell $g$ $(g = 1, \dots, G)$. The prediction covariates included $G$ grid cells at 100m-by-100m resolution, and population counts were predicted in each grid cell that contains values of building counts.

All models were implemented within the integrated nested Laplace approximation (INLA; Rue et al, 2009) in conjunction with the stochastic partial differential equation

(SPDE Lindgren et al, 2011) frameworks. It allowed us to gain more computational advantage by discretizing the entire study location continuous space into a Gaussian Markov random fields (GMRF) process. To ensure flexibility and better capture local variabilities within the data, we used the Penalized Complexity (PC) (Simpson et al., 2017) on the standard deviation parameters throughout, such that a small probability of 0.01 is assigned to the standard deviation $\sigma$ being greater than 1, that is, $P(\sigma > 1) = 0.01$.

**Model fit checks**

Model fit checks was conducted using the Bias, mean absolute error (MAE), the root mean square error (RMSE) and the correlation coefficient (CC). Smaller values of for the Bias, MAE and RMSE indicate better fit and predictive ability while larger values of CC indicate better predictive ability.  Table 2 below shows the model fit metrics

Table 2. Goodness of fit results

| | |
|---|---|
| Bias | 1.57 |
| MAE | 163.99 |
| RMSE | 286.92 |
| CC | 0.66 |

Further posterior inference and grid cell predictions were also carried out. The prediction at the grid cell uses the model parameters of the best fit training set model to predict population counts at 100m-by-100m grid cells across the study location using the corresponding grid cell values of the geospatial covariates and the building counts.

**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 and Heather Chamberlain for reviewing the data and providing thoughtful suggestions prior to this release.

**WORKS CITED**


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 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 v3.1. New York: GRID3. https://doi.org/10.7916/d6gy-yh28. Accessed [22nd August 2025].

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.

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.

Nnanatu, C. C., Yankey, O., Dzossa, A. D., Abbott, T., Gadiaga, A., Lazar, A., & Tatem, A. (2025a). Efficient Bayesian hierarchical small area population estimation using INLA-SPDE: integrating multiple data sources and spatial-autocorrelation.

Nnanatu, C., Yankey, O., Abbott, T., Bonnie, A., Chamberlain, H., Lazar, A., & Tatem, A. (2025b). Modelled gridded population estimates for Haut-Lomami Province in the Democratic Republic of Congo version 4.2.

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

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

Simpson, D. P., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. (2017). “Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors.” *Statistical Science* 32 (1): 1–28.

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. https://www.pnas.org/doi/10.1073/pnas.1715305115

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