Release Statement

Modelled gridded population estimates for Cameroon 2022, version 1.0

16 December 2022
Original Release: 16 December 2022

Background

This data release statement provides a description of the various modelled population and number of households datasets for Cameroon along with some key methodological insights. The datasets include gridded population estimates, gridded estimates of number of households (with spatial resolution of 3 arc-seconds, approximately 100 m grid cells), aggregated administrative units level population estimates, aggregated administrative units level estimates of number of households, and the estimates of the number of people belonging to various age-sex groups.

The data were produced by the WorldPop Research Group (www.worldpop.org) at the University of Southampton in collaboration with the Cameroon National Institute of Statistics (NIS) who provided most of the model input datasets, as 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.

Population modelling and estimation experts at WorldPop combined multiple nationally representative household listing datasets received from the Cameroon NIS with satellite-based settlement data and geospatial covariates to train geospatial statistical model parameters which were used to estimate population numbers and number of households at high-resolution grid cells using advanced Bayesian hierarchical statistical modelling frameworks. The datasets were anonymised due to confidentiality issues. The use of Bayesian hierarchical inference approach meant that we were able to simultaneously account for multiple levels of variability within the data and quantify uncertainties in the parameter estimates. The design and implementation of the statistical modelling processes was led by Chris Nnanatu, supported by Ortis Yankey. Demographic data and Geospatial covariates processing were led by Ortis Yankey and Thomas Abbott, respectively. Coordination and stakeholder engagement was led by Chris Nnanatu, Attila N. Lazar and Ortis Yankey, with oversight provided by Andrew J. Tatem, supported by Edith Darin.

These model-based population and number of households estimates can be considered as most accurately representing the year 2022. This period corresponds to the reference year of the settlement extent datasets as well as the input household listing

datasets which were collected between 2021 and 2022. We have followed very rigorous statistical modelling approaches to account for various sources of random and systematic biases. However, while our model was robust enough to explicitly account for random biases within the datasets, there could still be some sources of (most likely insignificant) systematic biases that remained unaccounted for.

The authors followed global best practices 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 estimates of population and number of households, not official government statistics.

RELEASE CONTENT

1. CMR_household_v1_0_gridded.zip
2. CMR_household_v1_0_admin.zip
3. CMR_population_v1_0_gridded.zip
4. CMR_population_v1_0_agesex.zip
5. CMR_population_v1_0_admin.zip

LICENSE

SUGGESTED CITATION

Nnanatu C.C., Yankey O., Abbott T. J., Assane, G., 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

FILE DESCRIPTIONS

The projection for all GIS files is the geographic coordinate system WGS84 (World Geodetic System 1984). All the geotiff raster files of the model outputs have spatial resolutions of approximately 100m grid cell (0.0008333 decimal degrees grid cell).

1. **CMR_household_v1_0_gridded.zip**

   This folder contains five geotiff raster files:

   CMR_ household_v1_0_gridded.tif

   This geotiff raster contains estimates of total number of households for each grid cell across Cameroon. The values are the mean of the posterior probability distribution for the predicted number of households in each grid cell. NA values represent areas that were mapped as unsettled according to building footprints data [1].

   CMR_ household_v1_0_lower.tif

   This geotiff raster contains the lower bound of the estimates of number of households for each grid cell across Cameroon. The values are the 2.5% posterior probability distribution (credible interval, CI) for the predicted number of households 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. NA values represent areas that were mapped as unsettled according to building footprints data [1].

   CMR_ household_v1_0_upper.tif

   This geotiff raster contains the upper bound of the estimates of number of households for each grid cell across Cameroon. The values are the 97.5% posterior probability distribution (credible interval) for the predicted number of households 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. NA values represent areas that were mapped as unsettled according to building footprints data [1].

   CMR_ household_v1_0_median.tif

   This geotiff raster contains the estimates of the median number of households for each grid cell across Cameroon. The values are the 50% posterior probability distribution for the predicted number of households in each grid cell. NA values represent areas that were mapped as unsettled according to building footprints data [1].

   CMR_household_v1_0_uncertainty.tif

   This geotiff raster contains estimates of uncertainty in the estimated number of households within each grid cell across Cameroon. The uncertainty values are calculated as the ratio of the differences between the upper (97.5% CI) and the lower

(2.5% CI) credible intervals of the posterior prediction and the mean of the posterior prediction (i.e., uncertainty = (upper bound – lower bound)/mean). These numbers provide a comparable measure of uncertainty in the estimates of number of households per grid cell across the country. Note that the uncertainty estimates cannot be summed across grid cells to produce an uncertainty measure for a multi-cell area. However, uncertainty for multiple cells can be calculated using the cells' posterior predictions.

2. CMR_household_v1_0_admin.zip

The folder contains three .csv files.

CMR_ household_v1_0_admin_level0.csv

This csv file contains the estimated total number of households for the entire country (administrative level 0). The .csv file contains the estimates of the total number of households, along with the lower, median, upper and uncertainty of the estimates of the number of households.

CMR_ household_v1_0_admin_level1.csv

This csv file contains estimates of the total number of households for the 10 regions of Cameroon (administrative level 1). The .csv file contains the ID, the names of the regions, the estimated total number of households per region, the lower, median, upper and uncertainty estimates of the estimated number of households per region. The .csv file can be joined to regional shapefile boundary of Cameroon and visualized using any mapping software.

CMR_ household_v1_0_admin_level2.csv

This csv file contains estimates of the total number of households for the departments (divisions) in Cameroon (administrative level 2). The .csv file contains the ID, names of the departments, the estimated total number of households per department, lower, median, upper and uncertainty estimates of the estimated number of households per department. The .csv file can be joined to department shapefile boundary of Cameroon and visualized using any mapping software.

3. CMR_population_v1_0_gridded.zip

This zip file contains four files:

CMR_population_v1_0_gridded.tif

This geotiff raster contains estimates of total population size for each grid cell across Cameroon. 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 [1].
Note: This raster is accompanied by two ancillary files that contain metadata (CMR_population_v1_0_gridded.tif.aux.xml; CMR_population_v1_0_gridded.tif.ovr).

CMR_population_v1_0_lower.tif
This geotiff raster contains estimates of the lower bound credible interval (2.5% CI) for each grid cell across Cameroon. The values are the 2.5% posterior probability distribution for the predicted population size in each grid cell. Note that the lower bound estimates cannot be summed across grid cells to produce a lower credible interval measure for a multi-cell area. However, lower bound estimates for multiple cells can be calculated using the cells' posterior predictions. NA values represent areas that were mapped as unsettled according to building footprints data [1].
Note: This raster is accompanied by two ancillary files that contain metadata (CMR_population_v1_0_lower.tif.aux.xml; CMR_population_v1_0_lower.tif.ovr).

CMR_population_v1_0_upper.tif
This geotiff raster contains estimates of the upper bound credible interval (97.5% CI) for each grid cell across Cameroun. The values are the 97.5% posterior probability distribution for the predicted population size in each grid cell. Note that the upper bound estimates cannot be summed across grid cells to produce an upper bound credible interval measure for a multi-cell area. However, upper bound estimates for multiple cells can be calculated using the cells' posterior predictions. NA values represent areas that were mapped as unsettled according to building footprints data [1].
Note: This raster is accompanied by two ancillary files that contain metadata (CMR_population_v1_0_upper.tif.aux.xml; CMR_population_v1_0_upper.tif.ovr).

CMR_population_v1_0_uncertainty.tif
This geotiff raster contains estimates of uncertainty in the population estimates within each grid cell across Cameroon. The uncertainty values are the difference between the upper (97.5% CI) and the lower (2.5% CI) credible intervals of the posterior prediction divided by the mean of the posterior prediction: uncertainty = (upper – lower)/mean. These numbers provide a comparable measure of uncertainty in population estimates across the country. Note that the uncertainty estimates cannot be summed across grid cells to produce an uncertainty measure for a multi-cell area. However, uncertainty for multiple cells can be calculated using the cells' posterior predictions.

Note: This raster is accompanied by two ancillary files that contain metadata (CMR_population_v1_0_uncertainty.tif.aux.xml; CMR_population_v1_0_uncertainty.tif.ovr).

4. CMR_population_v1_0_admin.zip

This zip file contains three csv files:

CMR_population_v1_0_admin_level0.csv

This csv file contains the estimated population totals for the entire country (administrative level 0). The .csv file contains the total population estimates, the lower, median, upper and uncertainty of the population estimates.

CMR_population_v1_0_admin_level1.csv

This csv file contains the estimated population totals for each of the 10 regions of Cameroon (administrative level 1). The .csv file contains the ID, the names of the regions, the total population estimates, the lower, median, upper and uncertainty of the population estimates for each region. The .csv file can be joined to regional shapefile boundary of Cameroon and visualized using any mapping software.

CMR_population_v1_0_admin_level2.csv

This csv file contains the estimated population totals for each of the departments (divisions) in Cameroon (administrative level 2). The .csv file contains the ID, the names of the departments, the total population estimates, the lower, median, upper and uncertainty of the population estimates for each department. The .csv file can be joined to department shapefile boundary of Cameroon and visualized using any mapping software.

5. CMR_population_v3_0_agesex.zip

This zip file contains two sub-folders namely, ‘Age_sex gridded raster’ and ‘National age_sex pyramid’. The ‘Age_sex gridded raster’ folder contains 40 geotiiff raster files at a spatial resolution of 3 arc-seconds (approximately 100 m), while the National age_sex pyramid’ folder contains only one raster file. Each raster provides gridded population estimates for an age-sex group per grid cell across Cameroon. For the Age_sex gridded raster’ folder, the raster 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 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 projected age-sex proportions [2] to allocate the gridded population to the different age-sex classes in each grid cell. 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

Note that the upper and lower credible intervals cannot be summed to produce credible intervals for their region, nor can the credible intervals for the regions be summed to produce credible intervals at the national level. Credible intervals for a given area are calculated from the posterior probability distributions of people for each cell within the area of focus.

NOTE: *Aggregated totals for administrative unit 3 (sub-division or arrondissements) were not included in the folders following advice by the Cameroon NIS due to boundary alignment challenges within the level 3 shapefile.*

RELEASE HISTORY

Version 1.0 (16 December 2022)
This is the original release of the data.

SOURCE DATA

We provide the sources of the input datasets used for the population modelling below:

1) Demographic Data

The demographic datasets used for the population modelling comprises nationally representative household listing datasets and the respective shapefiles of the surveyed Enumeration Areas as well the shapefiles for administrative units 1 (regions), 2 (departments or divisions) & 3(sub-divisions or *arrondissements*).

Household listing datasets

Of the seven household listing datasets received from the Cameroon National Institute of Statistics (NIS), five datasets were eventually selected and used for the population modelling after series of rigorous exploratory data analyses. The selected datasets are:

- Cameroon Malaria Indicator Survey (CMIS 2022)
- Employment and Informal Sector Survey (EESI, 2021)

- The 2021/2022 fifth Cameroon Households Survey phases 1 (ECAM5 Phase 1, 2021/2022)
- The 2021/2022 fifth Cameroon Households Survey phases 2 (ECAM5 Phase 2, 2021/2022)
- The 2021/2022 fifth Cameroon Households Survey phases 3 (ECAM5 Phase 3, 2021/2022)

Administrative boundaries

In addition to the shapefiles for all the household listing datasets listed above, the Cameroon NIS also shared the following administrative units shapefiles:

- Regional shapefile (received in 2022)
- Departmental shapefile (received in 2022)
- Sub-divisional shapefile (received in 2022)

2) Settlement Data

The settlement data used for the statistical modelling was provided by the building footprints for Cameroon obtained from Maxar/Ecopia [1]. These satellite imagery-derived datasets are provided as vector polygons from which information regarding number of buildings and other building characteristics/metrics are obtained.

3) Geospatial Covariates

We collated from various sources, a wide range of geospatial covariates which were eventually tested for their relationship with population density and number of households. The geospatial covariates include land uses and land cover data, and other remote sensing data such as climate variables such as temperature and rainfall, physical features and infrastructure such as roads and schools, and conflict data. From the collated geospatial covariates, we used a generalized linear model (GLM)-based stepwise selection method to select the best sets of covariates for each of the population density and number of household models. The selected covariates were then further accessed for multi-collinearity and statistical significance. Then all statistically significant geospatial covariates with variance inflation factor (VIF) of less than 5 were retained for the final models. In all, eight (8) covariates were selected for the population density model while nine (9) covariates were selected for the model for estimating the number of households. The descriptions, sources and spatial resolutions of these final geospatial covariates are presented in Table 1 below.

Table 1. Descriptions and sources of model geospatial covariates

Covariate	Population density model	Number of households model	Description	Source	Year	Format	Resolution
X1	✓	✓	ACLED distance to conflicts	https://acleddata.com/	2021	Raster	100m
X2	✓	✓	ACLED distance to explosions	https://acleddata.com/	2021	Raster	100m
X3	✓	✓	OSM Distance to waterbodies	https://www.geofabrik.de/data/download.html	2022	Raster	100m
X4	✗	✓	Distance to woody-tree area edges	https://www.worldpop.org/project/categories?id=14	2015	Raster	100m
X5	✗	✓	Distance to shrub area edges	https://www.worldpop.org/project/categories?id=14	2015	Raster	100m
X6	✓	✓	Distance to herbaceous area edges	https://www.worldpop.org/project/categories?id=14	2015	Raster	100m
X7	✓	✓	OSM Distance to local roads	https://www.geofabrik.de/data/download.html	2022	Raster	100m

X8	✓	✖	OSM Distance to marketplaces	https://www.geofabrik.de/data/download.html	2022	Raster	100m
X9	✖	✓	OSM Distance to railways	https://www.geofabrik.de/data/download.html	2022	Raster	100m
X10	✓	✖	Slope	https://www.worldpop.org/project/categories?id=14	2000	Raster	100m
X11	✓	✓	Night-time light	https://eogdata.mines.edu/products/vnl	2020	Raster	100m

Note: The ticks (✓) mean used for the modelling while the crosses (✖) indicate that the variable was not used.

Note that most of the covariates are found to be significant predictors of both population density and number of households in Cameroon (e.g., ACLED conflict data, distance to waterbodies, distance to local roads and night-time light), other covariates such as distance to woody areas, distance to shrub area edges and distance to railways only significantly predicted number of households, while Distance to marketplaces and slope only significantly predicted population density (Table 1).

METHODS OVERVIEW

We outline the various statistical analysis and modelling techniques undertaken below:

1) Datasets received from the Cameroon NIS were first catalogued and reviewed.
2) Datasets (household listing data and shapefiles, settlement data, and geospatial covariates) were explored, prepared and cleaned.
    - Datasets were compared with respect to year of collection and spatial coverage.

- Household listing datasets were joined to the respective shapefiles (boundary files) using the common unique IDs, which were either already existing within the data or had to be created.
- Datasets collected within the last two years of the modelling year (2021 - 2022) were retained while others were dropped.
- For the selected datasets, where two or more survey datasets overlapped at the household level (i.e., collected information from the same household), only the household data from the most recent survey were retained. Thus, for all duplicated household geolocation points, data points from the older surveys were dropped.
- Data was screened and checked for anomalous values and any detected anomalies were corrected wherever possible or discarded.
- Missing household sizes within the combined data were imputed using the average (mean) household sizes of each individual dataset.
- In all, there were 2,290 Enumeration Areas (EAs) with 2,587,569 people from 509,628 households across all the combined datasets. Nationally, there were ~5 people per household on average.
- The geospatial covariates and settlement data (building footprint) were extracted from the various sources and prepared for statistical modelling.
- These geospatial covariates were then joined with the combined demographic data for training model.

3) A generalized linear model (GLM)-based stepwise regression approach in combination with the variance inflation factor (VIF) were used to identify and choose the best set of geospatial covariates that significantly predicted population density and number of households whilst minimizing the potential effects of multicollinearity.
4) Bayesian hierarchical geostatistical models were developed using the retained geospatial covariates to train the respective model parameters to predict population density and number of households. These were implemented using the integrated nested Laplace approximation (INLA) in conjunction with the stochastic partial differential equation (INLA-SPDE; [3, 4]). All model implementations were done in R statistical programming software [5] via the R-INLA package according to the following steps:
    - Build the mesh covering the entire Cameroon. The mesh provides the discretized version of the continuously indexed spatial domain.
    - Create the projection matrix, A, calculate the SPDE object, and build the data stack of the covariates and the response variables. The response variable is the population density (people per building) for the population model, and average number of households (number of households per building) for the number of households model.

- Specify and run multiple nested geostatistical model using the inla() function whilst ensuring that the 'config = TRUE' option is set within the 'control.compute' argument to enable posterior simulation.

  Population density model

  Typically, within the context of population modelling, population count is a discrete random variable which naturally follows a Poisson distribution ($pop_i \sim Poisson(\lambda_i)$) [6-8]. However, to account for spatial aggregation error, the models are often specified in terms of the population density D_i:

  $$pop_i \sim Poisson(\lambda_i = \bar{D}_i B_i) \qquad (1)$$

  where B_i is the number of building within a given EA ($i = 1, 2, \dots, n$), $\bar{D}_i$ is the mean population density with $D_i = pop_i / B_i$ (people per building), and $D_i \sim Gamma(\alpha_1, \beta_1)$; the mean and variance of the population density are $\bar{D}_i = \alpha_1/\beta_1$ and $\sigma_D^2 = \alpha_1^2/\beta_1$, respectively. Then,

  $$\log(\bar{D}_i) = \mu + X_i^T \boldsymbol{\beta} + \zeta + \psi_i \qquad (2)$$

  where,

  - μ is the intercept parameter
  - X_i is a vector of the various geospatial covariates values per EA;
  - $\boldsymbol{\beta} > 0$ is the vector of unknown regression parameters.
  - $\zeta \sim Normal(0, \sigma_\zeta^2)$ is a generic random effect term that accounts for settlement type or the data source differences.
  - $\psi \sim Normal(0, \boldsymbol{Q}^{-1})$ is the Gaussian Random Field (GRF) specified through the INLA-SPDE approach. Where $\boldsymbol{Q}$ is the precision matrix.

  Number of household model

  Similarly, for the number of households model, we assume a Poisson distribution model for the number of households (num, the response variable) such that

  $$num_i \sim Poisson(\bar{h}_i B_i) \qquad (3)$$

  where $\widetilde{H}_i = num_i / B_i$ (households per building) is the is the average number of households per building, and assumed to come from a Gamma distribution, $\widetilde{H}_i \sim Gamma(\alpha_2, \beta_2)$, $E[\widetilde{H}] = \bar{h}$. Then,

  $$\log(\bar{h}_i) = \mu + X_i^T \beta + \zeta + \psi_i \quad (4)$$

  where, μ, X_i, β, ζ, and ψ are as defined above with X_i not necessarily containing the set of covariates.

- Carry out further model fit checks across the nested Bayesian statistical models using the deviance information criterion (DIC; [9]) and/or the widely

applicable information criterion (WAIC; [10]), and the Conditional Predictive Ordinate (CPO; [11]).

- Models with lower WAIC or lower DIC and lower $-\sum \log(CPO)$ are retained as the best fit models.
- Other model fit metrics used for model selection included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute Bias (ABIAS), and correlation coefficient between the observed and predicted response (CORR). Models with smaller MAE, RMSE and ABIAS but with larger CORR are retained as having better predictive abilities.
- The selected best fit models are further tested for predictive ability via cross validation techniques:
  - **In-sample cross validation**: the entire sample was used as the training set and 20% of the sample were randomly selected as the test set. Thus, the test samples were included within the training data hence called 'in-sample'. The model parameters are then used to predict the values of the population density (population count) or the number of households of the test data.
  - **Out-of-sample cross validation**: the entire dataset is divided into 80% training and 20% testing sets. Thus, the test data is not part of the training set hence called 'out-of-sample'. The model parameters based on the training set are used to predict the values of the population density (population count) or the number of households of the test data.
  - For each cross-validation technique implemented, model fit metrics (MAE, RMSE, ABIAS and CORR) are calculated. The estimates of the model fit metrics from the best fit models for the population density model and number of households model are given in Table 2 and Table 3, respectively.

Table 2. Model fit metrics for Population density model

Sample	**MAE**	**RMSE**	**Absolute BIAS**	**CORR**
In-Sample	58.745	82.527	3.997	0.997
Out-of-sample	63.327	93.278	3.212	0.996

Table 3. Model fit metrics for number of households model

Sample	**MAE**	**RMSE**	**Absolute BIAS**	**CORR**
In-Sample	54.491	82.727	4.757	0.816
Out-of-sample	53.492	80.230	3.944	0.820

For the population density model, the model performances at both in-sample and out-of-sample cross validations were similar even though the model did slightly better at in-sample for MAE and RMSE (Table 2). The population density model shows slight better performance for out-of-sample cross-validation when the Absolute biases are compared. However, for the number of households model, the model's predictive ability was consistently higher for out-of-sample prediction than for in-sample prediction (Table 3). Both models produced correlation coefficients of more than 80% at both in-sample and out-of-sample cross-validations. These results indicate that the models performed very well and showed no evidence of overfitting.

5) The sets of geospatial covariates within the best fit model with the best predictive ability identified in step 5 above are then extracted and stacked at the grid cell level for high-resolution estimation and prediction at the 100m grid squares. The respective values of the settlement data (building count) are also extracted for the grid cell prediction.
6) The model parameter values of the best fit models are used for the grid cell prediction according to the following steps:
    - Draw 100 samples of the parameter values from the posterior distribution of the parameters given the data.

    For each set of parameter values at each run,
    - Predict the values of the response variable of interest (population density or number of households) at each grid cell.
    - Back transform the estimates to obtain the original response and save. This is done by taking exponent of the predictions because of the log-link functions used in the modelling.

    For each grid cell g
    - ❖ Obtain the predicted population count ($\widehat{pop}_g$) as the product of the back transformed estimate of the population density ($\widehat{D}_g$) and the corresponding building count (B_g) of the grid cell. That is,

$$\widehat{pop}_g = \widehat{D}_g \times B_g \qquad (5)$$

        - ❖ Calculate the uncertainties in the estimates of the parameters as 95% credible interval of the posterior distribution. This gives the lower and upper bounds of the estimates.
    - Write and save the grid datasets as geotiff raster files.
    - Obtain the disaggregated estimates (for population count only) by age and sex using appropriate population age-sex proportion data.
    - Obtain the aggregated estimates of the response variable of interest at coarser administrative units (e.g., region and department)

Prior Distribution

For both models, we used the priors within R-INLA to estimate the model parameters via Bayesian statistical inference framework:

- $Uniform(0,1)$ for the intercept parameter.
- $Normal(0, 0.01)$ for other fixed effects parameters
- $Gamma(1, 0.00005)$ for all the hyperparameters of the random effects.

All data processing and analyses were carried out using R version 4.0.2 and INLA (v 22.05.07). Thus, users who would like to reproduce these estimates exactly are encouraged to use the same R versions because optimization methods in R and INLA could vary between versions. Finally, the concept of bottom-up population modelling for estimating population in the absence of recent census data was well described in [8].

ASSUMPTIONS, STRENGTHS AND LIMITATIONS

The key assumptions, strengths and limitations of the methodology are listed below:

- Mean imputation was used to impute households with missing household sizes in the household listing dataset using the mean household size of the enumeration area. The imputed household sizes for those missing household may not actually be the true household size; however, it is assumed that the impact of the differences on the model estimates is likely to be insignificant in that data were aggregated to Enumeration Area (EA) level for statistical modelling.
- Data fitting was done using enumeration areas as unit of analysis; however, population predictions were done at the grid cell level (100 X100m). The differences in the unit of analysis may produce a modifiable area unit problem. However, we saw from a simulation study that the impacts of such spatial misalignment on the accuracy of the model estimates waned as the sample size increased.

- We assume that the building footprints data is accurate, and that each building polygon corresponds to a building structure. However, the building footprint data was collected in 2018. 2018 building footprint was used because we did not have access for current building footprint for the country.
- An assumption is made that the OpenStreetMap data obtained through the geofabrik data portal is accurate at the date of download. These datasets are updated daily and therefore may differ from the time of publication and beyond. This applies to the obtaining of covariates x3, x7, x8 and x9 which occurred on 10th June 2022.
- There are some small areas where the extent of the building footprints data does not cover the full extent of the enumeration area. Because of this, the number of building count for those enumeration areas may be underestimated which may lead to an underestimation of the total population in those areas.
- 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.
- The use of Bayesian statistical inference approach allowed a straightforward approach for uncertainty quantification providing us with the upper and lower bounds estimates of the population counts and number of households at the grid cells.
- The use of INLA approach provided a computational advantage by the use of Laplace approximation techniques to approximate the posterior distribution instead of drawing samples from the stationary distribution, thereby, offering high computational speed while at the same time circumventing the well-known posterior convergence problems that are prevalent among Markov chain Monte Carlo (MCMC)-based solutions.
- In addition, the use of SPDE approach to account for spatial autocorrelation between the modelling units (observation units) provided even more computational efficiency by discretizing the spatially continuous surface of Cameroon using Gaussian Markov Random Fields (GMRF,[12]), thereby avoiding the high computational cost associated with the dense correlation matrix of the Gaussian Process. This allowed us to borrow strength from locations with more observations to estimate population or number of households at locations with little or no observations to improving the accuracy of the estimates at a much greater computational speed.
- Our methodology allowed us to account for the potential variability in population count or number of households due to differences in the data collection strategies

used in the different household listing datasets that were combined. However, the initial models tested showed no significant differences between models with data source random effects and models which did not account for the data source differences. This was probably because the different data collection schemes were very similar in terms of methodology and they were based on the same sampling frame.

- Finally, the hierarchical modelling approach enabled us to account for multiple sources of variabilities in the estimates due to the hierarchical structure within the observed datasets.

ACKNOWLEDGEMENTS

We thank the Cameroon National Institute of Statistics (NIS) for providing access to the anonymized data in accordance with the data sharing agreement between the University of Southampton and the Cameroon government. Specifically, we thank Mr. Anaclet Desire Dzossa, Mrs. Marie Antoinette Fomo and Mr. Wounang Romain for being part of the process throughout the time. The commitment of Mr. Mathias Kuepie the UNFPA Cameroon representative at the time of the project on seeing the success of the project is highly acknowledged. We thank Dr Sarchil Qader for playing a major role throughout the engagement process, and we also thank the entire WorldPop team who supported the project in one way or the other. Finally, we are thankful for the funding provided by the UK Foreign, Commonwealth & Development Office (FCDO) for the training/capacity strengthening workshop on the production of modelled population estimates for Cameroon delivered by the WorldPop experts to the personnel of Cameroon NIS and other participants from other sister agencies in Yaoundé, Cameroon, from 6th to 15th of May 2024.

REFERENCES

[1] Building Footprints Cameroon, Digitize Africa data © 2020 Maxar Technologies, Ecopia.AI

[2] Pezzulo C, Hornby GM, Sorichetta A, Gaughan AE, Linard C, Bird TJ, Kerr D, Lloyd CT, Tatem AJ. 2017. Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. Sci. Data 4:170089 https://dx.doi.org/10.1038/sdata.2017.89.

[3] 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

[4] 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

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

[6] Leasure, D. R., W. C. Jochem, E. M. Weber, V. Seaman and A. J. Tatem (2020). "National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty." Proceedings of the National Academy of Sciences": 201913050. DOI: 10.1073/pnas.1913050117. https://www.pnas.org/doi/pdf/10.1073/pnas.1913050117

[7] Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lazar, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., Tatem, A. J. (2022). "High-resolution population estimation using household survey data and building footprints." Nature Communications, 13, 1330. https://doi.org/10.1038/s41467-022-29094-x

[8] 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

[9] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde, "Measures of Model Complexity and Fit," Journal of the Royal Statistical Society B, vol. 65 , p. 583–639, 2002.

[10] Watanabe, S. (2013). "A Widely Applicable Bayesian Information Criterion." Journal of Machine Learning Research 14: 867–97.

[11] Pettit, L. I. 1990. "The Conditional Predictive Ordinate for the Normal Distribution." Journal of the Royal Statistical Society. Series B (Methodological) 52 (1): pp. 175–84.

[12] Rue H, Held L. (2005). Gaussian Markov random fields. Theory and applications. Chapman & Hall.