**Release Statement**

# Census-cartography-based gridded population estimates for Mali (2020), version 1.0.

30 December 2022

These data consist of modelled gridded population estimates produced at a spatial resolution of approximately 100m across Mali. The estimates comprise a combination of total population counts at enumeration area level collected by the census cartography team of the Mali Statistics Office and modelled population counts created using a Bayesian statistical model for areas that could not be covered by the cartography team because of security issues. The main input data for the model are the cartography data collected in the safe part of the country in 2019-2020 (628 out of 714 *Communes* –administrative level 3–, that is 87% of the country territory). Other essential input data include metrics derived from building footprints, which were automatically delineated by Ecopia.AI in 2021 using satellite imagery collected by Maxar Technologies between 2010 and 2021. The modelled population estimates represent the period of the census cartography, but their consistency may be impacted by the accuracy of the building footprints.

These data were produced by the WorldPop Research Group at the University of Southampton as part of the GRID3 Project, GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) programme funded by the Bill and Melinda Gates Foundation (BMGF) and the United Kingdom's Foreign, Commonwealth & Development Office (INV 009579, formerly OPP 1182425). The study was approved by the Faculty Ethics Committee of the University of Southampton (ERGO II 64957). The project was led by the Center for International Earth Science Information Network (CIESIN) at Columbia University, in collaboration with the WorldPop Research Group at the University of Southampton, the United Nations Fund for Population (UNFPA) and the Malian Institut National de la Statistique (INSTAT).

The production of these data was led by Edith Darin (WorldPop) with support from Matthias Kuépié and Jean Wakam (UNFPA), Abdoul Karim Diawara, Assa Gakou and Siaka Cissé (Institut National de la Statistique), and Attila N Lazar (WorldPop) and Andrew J Tatem (WorldPop). The authors acknowledge the support of their respective institutions in the completion of this work.

## CITATION

WorldPop and Institut National de la Statistique du Mali. 2022. Census-cartography-based gridded population estimates for Mali (2020), version 1.0. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00745

## LICENSE



## RELEASE CONTENT

1. MLI_population_v1_0_gridded.tif
2. MLI_population_v1_0_agesex.zip
3. MLI_population_v1_0_mastergrid.tif
4. MLI_population_v1_0_sql.sql

## FILE DESCRIPTIONS

All the GeoTIFF rasters presented below were georeferenced using the WGS84 datum (World Geodetic System 1984: EPSG 4326) with a consistent spatial resolution of 0.0008333 decimal degrees (i.e. approximately 100m).

### 1. MLI_population_v1_0_gridded.tif

This GeoTIFF raster represents estimates of total population counts within grid cells of approximately 100m across Mali The raster values are the mean of a posterior distribution (calibrated by the census cartography totals when available) and therefore include decimals (e.g. 0.5 people). An estimate of 0.5 people in two neighbouring cells indicates that one person lives somewhere within those two cells. NA values represent cells where no building footprint is present.

### 2. MLI_population_v1_0_agesex.zip

This zip file contains 38 GeoTIFF rasters representing estimated population counts for specific age and sex groups within grid cells of approximately 100m. The rasters were created *post-hoc* by multiplying the total population counts provided in the *MLI_population_v1_0_gridded.tif* raster and national age and sex proportions from the United Nations demographic yearbook [7], reproduced in *MLI_population_v1_0_agesex_table.csv. The location of the regions of the subnational age-sex pyramids are in MLI_population_v1_0_agesex_regions.tif.*

34 rasters represent commonly reported age and sex groups labelled with either an “f” (female) or an “m” (male) followed by the number of the first year of the corresponding age class. The age groups consist of five-year bins labelled with a “0”, “5”, “10”, etc. Eighty-year-olds and over are represented in the groups “f80” and “m80”. Three additional rasters represent demographic groups frequently targeted in public health campaigns. These groups are labelled as “under5” (all females and males under the age of five), “under15” (all females and males under the age of 15) and “f15_49” (all females between the ages of 15 and 49, inclusive).

### 3. MLI_population_v1_0_mastergrid.tif

This binary GeoTIFF raster has a value of one if a grid cell of approximately 100m contains at least one building footprint and zero if no building footprint is present. NA values indicate grid cells outside the boundaries of Mali.

### 4. MLI_population_v1_0_sql.sql

This SQLite database contains estimates of the total population size in each grid cell of approximately 100m. This database is source data for the wopr R package and the associated woprVision web interface [1]. The database contains a table with the following columns:

- **cell** — contains a unique identifier for each grid cell of approximately 100m corresponding in the MLI_population_v1_0_mastergrid.tif raster.
- **x** and **y** — contain the coordinates of the centroid of each grid cell.
- **pop** — contains the full posterior distribution of the estimated total population counts for each grid cell.
- **agesexid** — contains the unique identifier linking each grid cell to the relative province-level age and sex proportions.

## DATA SOURCES

**Building footprints** — we accessed the latest building footprints produced by Ecopia.AI using Maxar Technologies satellite imagery [1] collected between 2010 and 2020 for Mali and rasterized them with a resolution of approximately 100m, based on the centroid of the building footprint. They were also used to derive some of the model covariates presented below. The building footprint data are not openly available.

**Population totals** — we accessed the total population counts for 25466 enumeration areas from the census cartography [2].

**Administrative boundaries** — we accessed vector boundaries for Mali produced by INSTAT [2] and rasterized them with a resolution of approximately 100m.

**Settlement classes** — we derived four settlement classes (i.e. urban, periurban, village and hamlet) by applying de Bellefon et al method [8] on the building footprints [1].

**Model covariates** — we created 34 model covariates and selected six of them to be implemented in the model:

- **Number of building footprints** (in a focal window of 100m) within grid cells of approximately 100m resolutions following the methodology developed by Jochem et al. [9].
- **Coefficient of variation of building footprint area** (in a focal window of 200m) within grid cells of approximately 100m resolutions following the methodology developed by Jochem et al. [9].
- **Distance to conflicts** (Euclidean distance) between January 2015 and December 2020 within grid cells of approximately 100m resolutions using the ACLED data [5].
- **Annual biomass production** (re-sampled from an original resolution of approximately 1km) in 2019 within grid cells of approximately 100m resolutions using the Action contre la faim data [6].
- **Disaggregated projected population count** in 2020 within grid cells of approximately 100m resolutions using the WorldPop data [4].
- **Distance to urban centers** (Euclidean distance) within grid cells of approximately 100m resolutions using the INSTAT data [2].

## METHOD OVERVIEW

We developed a Bayesian hierarchical model to estimate total population counts within grid cells of approximately 100m, similarly to Leasure et al. [10] and Boo et al. [11].

The model was fitted using the population count of the 25466 enumerations areas that were fully covered by the census cartography. See Figure 1 for their locations.

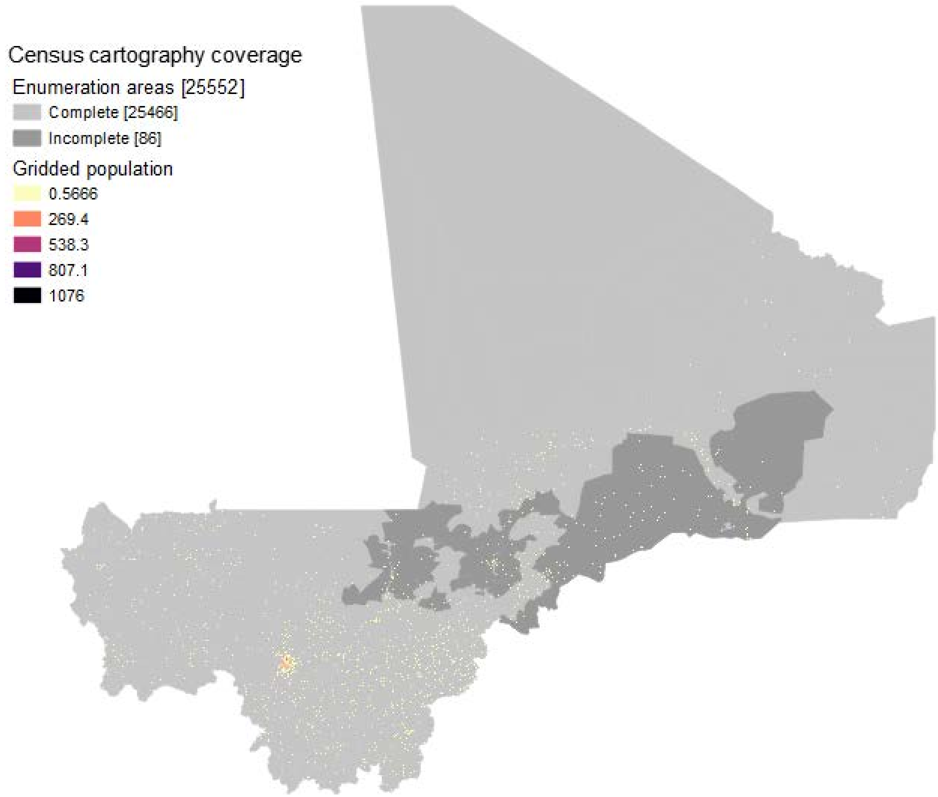


*Figure 1 Census cartography coverage*

The model is based on assuming that population count is drawn from a Poisson distribution whose parameter is decomposed in a latent population density described by a lognormal distribution, multiplied by the observed building footprint area. We define the hierarchy for estimating the intercept of the latent mean population density based on the *settlement classes* (see Data source section), the *Région* (administrative level 1) and the *Cercle* (administrative level 2). We estimated a random variance of the mean population density based on the cartography data collection type (paper-based vs tablet-based). As geospatial covariates, we used the one listed in Data Sources section, following a covariates selection process based on running the full model with one covariate at a time, and keeping covariates with the best prediction accuracy and the lowest multicollinearity. We add a random covariate slope based on the *settlement classes* for estimating the effect of *Distance to conflict*, *Distance to urban centers*, *Number of building footprints* and *Coefficient of variation of building footprint area*.

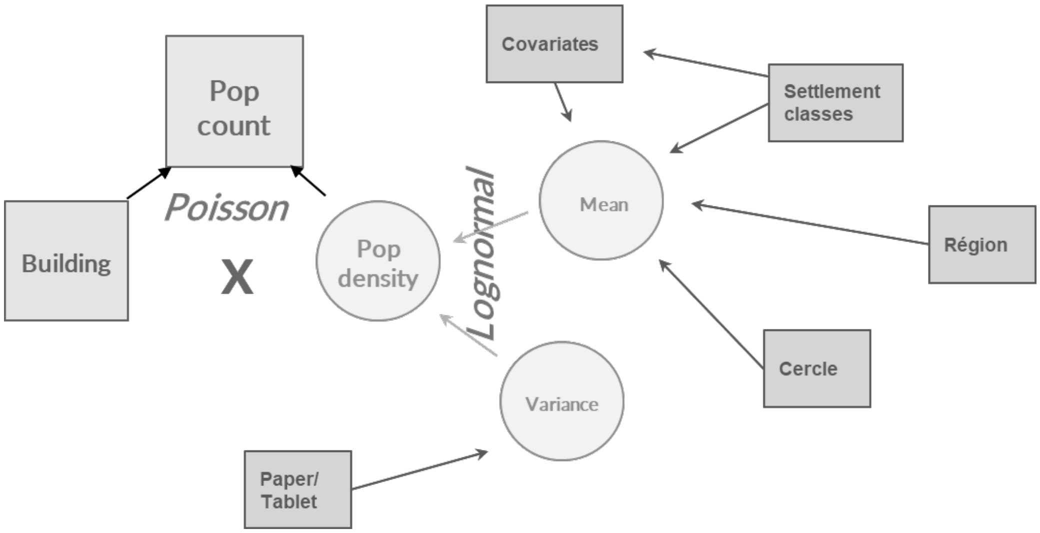


*Figure 2 Schema of population model*

For the enumeration areas that were fully covered by the census cartography, we can compare the predicted population count with the observed population count. In-sample, the average residual (bias) was 16.7%, the average of the absolute residuals (inaccuracy) was 34.7%, the standard deviation of the residual (imprecision) was 60.3% and 94% of the observed population count fall into the 95% confidence interval of their predictions.

We then predicted posterior distribution of population count for all grid cells of approximately 100m that were within Mali boundaries [2] and contained at least one building footprint centroid [1]. For grid cells falling within enumeration areas that were fully covered by the cartography ground data collection, we used the posterior population prediction as weight to disaggregate observed enumeration area totals. For grid cells falling within area not covered or partially coverage by the cartography ground data collection, we keep the posterior distribution of the above population model as they are. We then computed *post-hoc* breakdowns in 37 age and sex groups by multiplying the mean population count prediction at grid cell level by the national age and sex proportions [7]. Final gridded mean population estimates can be visualised through the *woprVision* interface [13].

## ASSUMPTIONS AND LIMITATIONS

We assume that population counts and age and sex characteristics derived from the UN demographic yearbook are accurate. However, they are based on projections and are available only at national level. We also assume that the building footprints dataset are an accurate representation of potential residential locations at the time when the cartography was conducted, although inaccuracies have been observed in relation to the satellite imagery used for the Ecopia automatic feature extraction and thus the method is likely resulting both in false positives and negatives. We also consider that the boundaries of the Cercle are accurate, even though they may differ from boundaries produced by national, international, and other relevant bodies. For this reason, the geographic extent of the gridded population estimates may not be aligned with other data sources.

We assume that the building footprint layer is accurate. Figure 3 shows however the date of imagery that was used, which spans from 2010 to 2020.

We assume that each grid cell containing the centroid of a building footprint polygon is potentially residential. For this reason, population totals and relative age and sex breakdowns may be over-estimated in cells with primarily non-residential buildings (e.g. industrial areas). The estimates represent the

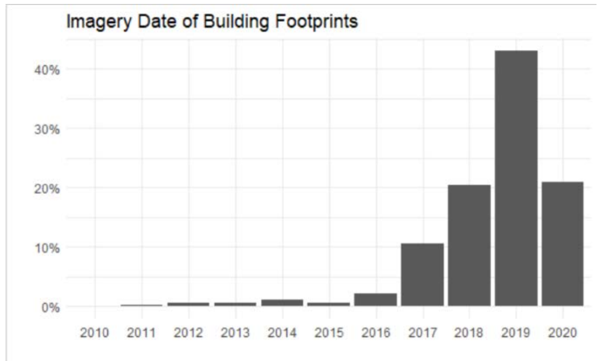


*Figure 3 Imagery date of the building footprint feature extraction*

population based on the place of residence during the census cartography. The population with no physical address (e.g. homeless) at the time of the cartography were not included in the estimates.

We assume that the processes observed within the areas targeted in the enumeration area are reflective of the ones occurring at the grid cell level. However, we expect that for larger areas this relationship is subject to higher degrees of uncertainty because of the modifiable areal unit problem (MAUP) [12]. We assume that the demographic processes, especially the relationship between building footprint count and population count observed within the areas covered by the census cartography are similar to the ones in the areas not covered by the census cartography. Nonetheless areas that could not be covered by the census cartography have experienced a high level of insecurity due to conflicts, such that the relationship between building footprint and population count might be impacted.

## DATA CITED

**[1]** Ecopia.AI and Maxar Technologies. 2021. Digitize Africa data (year 2). [Dataset]. http://digitizeafrica.ai

**[2]** Institut de la Statistique du Mali (INSTAT). 2020. Cartographie du RGPH5 [Dataset].

**[4]** WorldPop, University of Southampton. 2020. Constrained UN-adjusted gridded population estimates. https://hub.worldpop.org/geodata/summary?id=49699.

**[5]** Armed Conflict Location & Event Data Project (ACLED). [Dataset]. www.acleddata.com.

**[6]** Action Contre la Faim Surveillance West Africa. 2019. Annual biomass production (Dry Matter Productivity. [Dataset]. https://data.humdata.org/dataset/acf_biomass_west-africa_raster.

**[7]** United Nations. 2020. Demographic Yearbook 71st Issue. https://unstats.un.org/unsd/demographic-social/products/dyb/dybsets/2020.pdf.

## WORK CITED

**[8]** de Bellefon MP, Combes PP, Duranton G, Gobillon L, Gorin C. 2021. Delineating urban areas using building density. Journal of Urban Economics. 125:103226. https://doi.org/10.1016/j.jue.2019.103226

**[9]** Jochem WC, Leasure DR, Pannell O, Chamberlain HR, Jones P, Tatem AJ. 2021. Classifying settlement types from multi-scale spatial patterns of building footprints. Environment and Planning B: Urban Analytics and City Science. 48(5):1161-1179. doi:10.1177/2399808320921208

**[10]** Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ. 2020. National population mapping from sparse survey data: a hierarchical Bayesian modelling framework to account for uncertainty. Proceedings of the National Academy of Sciences. 117 (39) 24173-24179. doi:10.1073/pnas.1913050117

**[11]** Boo, G., E. Darin, D. R. Leasure, C. A. Dooley, H. R. Chamberlain, A. N. Lázár, K. Tschirhart, C. Sinai, N. A. Hoff, T. Fuller, K. Musene, A. Batumbo, A. W. Rimoin and A. J. Tatem (2022). "High-resolution population estimation using household survey data and building footprints." Nature Communications 13(1): 1330. DOI: 10.1038/s41467-022-29094-x. https://doi.org/10.1038/s41467-022-29094-x.

**[12]** Openshaw S. 1983. The Modifiable Areal Unit Problem — Concepts and Techniques in Modern Geography. Norwich, UK: Geo Books.

**[13]** Leasure DR, Bondarenko M, Darin E, Tatem AJ. 2020. wopr: An R package to query the WorldPop Open Population Repository, version 1.0.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00716. https://apps.worldpop.org/woprVision/?data=MLIv1.0.

## RELEASE HISTORY

**Version 1.0** (this release) [doi:10.5258/SOTON/WP00745]

Original release of the population dataset.