## Release Statement

### Modelled gridded population estimates for the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces in the Democratic Republic of the Congo (2021), version 3.0.

04 January 2022

These data consist of modelled gridded population estimates produced at a spatial resolution of approximately 100m across the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces in the Democratic Republic of the Congo (DRC). The estimates comprise total population counts created using a Bayesian statistical model and *post-hoc* breakdowns in 40 age and sex groups. The main input data were derived from a dedicated microcensus survey carried out in the targeted provinces throughout March and April 2021. The microcensus was led by the Flowminder Foundation, the École de Santé Publique de Kinshasa, the WorldPop Research Group at the University of Southampton and the Bureau Central du Recensement, which is part of the Institut National de la Statistique of the DRC. 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 microcensus but their consistency may be impacted by the accuracy of the building footprints, particularly in the areas where the satellite imagery used for automatic delineation was outdated.

These data were produced by the WorldPop Research Group at the University of Southampton as part of the GRID3 Mapping for Health Project. This project was delivered under the leadership of the Ministry of Public Health, Hygiene and Prevention of the DRC and funded by Gavi, the Vaccine Alliance (RM 867204 20A2). The project was led by the Flowminder Foundation and the Center for International Earth Science Information Network (CIESIN) at Columbia University, in collaboration with the WorldPop Research Group at the University of Southampton and national partners including, but not limited to, the École de Santé Publique de Kinshasa and both the Bureau Central du Recensement and the Institut National de la Statistique. This work was a continuation of the 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 62716).

The production of these data was led by Gianluca Boo (WorldPop) with support from Roland Hosner (Flowminder Foundation), Pierre Z Akilimali (École de Santé Publique de Kinshasa), Edith Darin (WorldPop), Heather R Chamberlain (WorldPop), Warren C Jochem (WorldPop), Patricia Jones (WorldPop), Roger Shulungu Runika (Institut National de la Statistique), Henri Marie Kazadi Mutombo (Bureau Central du Recensement), Attila N Lazar (WorldPop) and Andrew J Tatem (WorldPop). The authors acknowledge the support of their respective institutions in the completion of this work.

*The WorldPop Research Group at the University of Southampton, their partners and funders offer these data on a "where is, as is" basis. They do not offer explicit or implicit warranties 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. The authors followed rigorous procedures designed to ensure that the input data, the applied methods and the results are of reasonable quality. If users encounter apparent errors or misstatements, they should contact release@worldpop.org.*

## CITATION

G Boo, R Hosner, PZ Akilimali, E Darin, HR Chamberlain, WC Jochem, P Jones, R Shulungu Runika, HM Kazadi Mutombo, AN Lazar and AJ Tatem. 2021. Modelled gridded population estimates for the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces in the Democratic Republic of the Congo (2021), version 3.0. WorldPop, University of Southampton, Flowminder Foundation, École de Santé Publique de Kinshasa, Bureau Central du Recensement and Institut National de la Statistique. doi:10.5258/SOTON/WP00720

## LICENSE

## RELEASE CONTENT

1. COD_population_v3_0_gridded.tif
2. COD_population_v3_0_agesex.zip
3. COD_population_v3_0_mastergrid.tif
4. COD_population_v3_0_sql.sql
5. COD_population_v3_0_tiles.zip
6. COD_population_v3_0_methods.zip

## 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. COD_population_v3_0_gridded.tif

This GeoTIFF raster represents estimates of total population counts within grid cells of approximately 100m across the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces. The raster values are the mean of a posterior distribution 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. A detailed description of the methods used to produce this data is provided in the *COD_population_v3_0_methods.zip* file.

### 2. COD_population_v3_0_agesex.zip

This zip file contains 40 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 *COD_population_v3_0_gridded.tif* raster and age and sex proportions derived from the microcensus data for each province. A detailed description of the methods used to produce this data is provided in the *COD_population_v3_0_methods.zip* file.

36 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. “f0” and “m0” are population counts of under one-year-olds for females and males. “f1” and “m1” are population counts of one- to four-year-olds for females and males. Over four years of age, the age groups consist of five-year bins labelled with a “5”, “10”, etc. Eighty-year-olds and over are represented in the groups “f80” and “m80”. Four additional rasters represent demographic groups frequently targeted in public health campaigns. These groups are labelled as “under1” (all females and males under the age of one), “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. COD_population_v3_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 the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces.

### 4. COD_population_v3_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 COD_population_v3_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.
- **count** — contains the count of building footprints in each grid cell.

### 5. COD_population_v3_0_tiles.zip

This zip file contains a tiled web map allowing for the rapid display of the estimated total population counts presented in the *COD_population_v3_0_gridded_population.tif* raster in dedicated web interfaces. The tiles were created in an XYZ format compatible with Leaflet for the zoom levels 1 to 14. These tiles are source data for the woprVision web interface.

### 6. COD_population_v3_0_methods.zip

This zip file contains three html documents providing a detailed description of the methodology developed and implemented to produce these data. The file *COD_population_v3_0_methods_sampling.html* describes the sampling design developed for the microcensus. The file *COD_population_v3_0_methods_microcensus.html* describes the microcensus data and associated processing steps. The file *COD_population_v3_0_methods_statistical_model.html* describes the statistical model used to estimate total population counts and age and sex breakdowns within grid cells of approximately 100m across the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces.

## DATA SOURCES

**Building footprints** — we accessed the latest building footprints produced by Ecopia.AI using Maxar Technologies satellite imagery [1] collected between 2010 and 2021 for the DRC and rasterized them with a resolution of approximately 100m. In doing so, we computed the number of centroids within each grid cell, similar to the work carried out by Dooley et al. [8]. Building footprints were also used to derive some of the model covariates presented below. The building footprint data are not openly available.

**Microcensus data** — we derived total population counts and age and sex breakdowns within the 1,397 survey clusters targeted in a microcensus led by the Flowminder Foundation [2] between March and April 2021. Inclusion criteria, imputation and other data processing steps are described in the *COD_population_v3_0_methods_microcensus.html* report contained in the file *COD_population_v3_0_methods.zip*. The microcensus data are not publicly available.

**Administrative boundaries** — we accessed vector boundaries for the Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu provinces in the DRC produced by the Bureau Central du Recensement [3] and rasterized them with a resolution of approximately 100m. The original administrative boundaries are in the process of being consolidated and are currently not publicly available.

**Settlement classes** — we derived four settlement classes (i.e. urban, periurban, village and hamlet) by reclassifying GHS-SMOD data [4] and rasterized them with a resolution of approximately 100m. The original classes 10 and 11 were reclassified as hamlet, the classes 12 and 13 as village, the classes 21, 22 and 23 as periurban and the class 30 as urban.

**Model covariates** — we created 1,178 model covariates and selected four of them to be implemented in the model:

- **Average perimeter of the building footprints** (log-transformed) within grid cells of approximately 100m resolutions following the methodology developed by Jochem et al. [9].
- **Compactness of the building footprints** (log-transformed) within grid cells of approximately 100m resolutions following the methodology developed by Jochem et al. [9].
- **Monthly variability of dry matter productivity** (re-sampled from an original resolution of approximately 300m) between April 2000 and March 2021 within grid cells of approximately 100m resolutions using the Copernicus Dry matter Productivity (DMP) [5] data.
- **Monthly variability of surface air temperature** (re-sampled from an original resolution of approximately 300m) between April 2000 and March 2021 within grid cells of approximately 100m resolutions using the Copernicus ERA5 [6] data.

## 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]. We then computed *post-hoc* breakdowns in 40 age and sex groups by multiplying the total population counts by relative age and sex proportions at the province level. These proportions were derived from the microcensus data, weighted using to account for the different probabilities of selection of the clusters implied by the stratification defined in the sampling design [12]. A full report with a detailed description of the sampling design is available in the *COD_population_v3_0_methods_sampling*.html report and of the statistical model in the *COD_population_v3_0_statistical_model.html* report, both contained in the file *COD_population_v3_0_methods.zip*.

## ASSUMPTIONS AND LIMITATIONS

We assume that the population counts and age and sex characteristics derived from the microcensus data are accurate. However, throughout the microcensus, the surveyors faced considerable security and logistical challenges that may have impacted the quality of the data collection. We also assume that the building footprints dataset are an accurate representation of potential residential locations at the time when the microcensus was conducted, although inaccuracies have been observed in relation to the satellite imagery used for the automatic delineation and the methods resulting both in false positives and negatives. We also consider that the boundaries of the provinces 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 each grid cell containing the centroid of a building footprint polygon is potentially residential. For this reason, the total population counts 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 *de-jure* population [13] based on the place of residence between March and April 2021, when the microcensus was carried out. The population with no physical address (e.g. homeless) at the time of the survey were not included in the estimates. We consider that age and sex proportions are similar within the same province, although variations across rural and urban areas are likely to occur.

We assume that the processes observed within the areas targeted in the microcensus 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) [14]. Whether this potential issue was considered both in the sampling design and statistical modelling, some mismatches between total population counts estimated at the microcensus-cluster and grid-cell level have been observed.

## DATA CITED

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

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

**[3]** Bureau Central du Recensement (BCR). 2018. Report des Limites Administratives — République Démocratique du Congo. [Dataset].

**[4]** Florczyk A, Corban C, Ehrlich D, Carneiro Freire S, Kemper T, Maffenini L, Melchiorri M, Pesaresi M, Politis P, Schiavina M, Sabo F, Zanchetta L. 2019. GHSL Data Package 2019, Publications Office of the European Union, Luxembourg. doi:10.2760/0726. [Dataset]. https://ghsl.jrc.ec.europa.eu/ghs_smod2019.php

**[5]** Copernicus Global Land Service. 2021. Dry Matter Productivity. [Dataset]. https://land.copernicus.eu/global/products/dmp

**[6]** Copernicus Global Land Service. 2021. Copernicus ERA5 — Monthly Temperature. [Dataset]. https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation

## WORK CITED

**[7]** 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.

**[8]** Dooley CA, Boo G, Leasure DR, Tatem AJ. 2021. Gridded maps of building patterns throughout sub-Saharan Africa, version 2.0. WorldPop, University of Southampton. Source of building footprints: Ecopia Vector Maps Powered by Maxar Satellite Imagery (C) 2019-2021. doi:10.5258/SOTON/WP00712.

**[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, Darin E, Leasure DR, Dooley CA, Chamberlain HR, Lazar AN, Tschirhart K, Sinai C, Hoff NA, Fuller T, Musene K, Batumbo A, Rimoin AW, Tatem AJ. 2021, High-resolution population estimation using household survey data and building footprints. [Preprint]. doi:arXiv:2106.07461 [stat.AP].

**[12]** Cochran WG. 1977. Sampling techniques. 3d ed. New York: Wiley (Wiley series in probability and mathematical statistics).

**[13]** United Nations. 1991. Handbook of Vital Statistics Systems and Methods, Volume 1: Legal, Organisational and Technical Aspects United Nations Studies in Methods, Glossary, Series F, No. 35, United Nations, New York.

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

## RELEASE HISTORY

**Version 3.0** (this release) [doi:10.5258/SOTON/WP00720]

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

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

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

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

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