Release Statement

South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0

29 March 2021

These data include gridded estimates of population sizes at approximately 100 m resolution with national coverage across South Sudan. This includes estimates of total population sizes and population counts in 40 different age-sex groups. It also includes a breakdown of the total population sizes into internally displaced persons (IDPs) and non-IDPs. These results were produced using publicly available census projections from the South Sudan National Bureau of Statistics and displacement data from the International Organisation for Migration (IOM) and the United Nations Refugee Agency (UNHCR), as well as building footprints from Maxar/Ecopia that were derived from recent satellite imagery. Note that this dataset is most likely to represent South Sudan's population distribution as of September 2020 given the age of the input data.

These data were produced by the WorldPop Research Group at the University of Southampton. This work is 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 Foreign, Commonwealth & Development Office (INV 009579, formerly OPP 1182425). 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. We acknowledge the whole WorldPop Research Group for overall project support. This work was led by Claire A Dooley with support from Chris Jochem, Doug R Leasure, Alessandro Sorichetta, Attila Lazar and Andy Tatem. The authors acknowledge associated support services at the University of Southampton, in the completion of this work. Analyses undertaken were approved by the University of Southampton Faculty Ethics Committee (ERGO II 49112).

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.

RELEASE CONTENT

1. SSD_population_v2_0_gridded.zip
2. SSD_population_v2_0_agesex.zip
3. SSD_population_v2_0_intermed.zip
4. SSD_population_v2_0_mastergrid.tif
5. SSD_population_v2_0_sql.sql
6. SSD_population_v2_0_tiles.zip
7. SSD_population_v2_0_methods.pdf

LICENSE

SUGGESTED CITATION

For files 1-6, i.e. the dataset:

Dooley CA, Jochem WC, Leasure, DR, Sorichetta A, Lazar AN and Tatem AJ. 2021. South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00709

For file 7, i.e. the methods report:

Dooley CA, Jochem WC, Sorichetta A, Lazar AN and Tatem AJ. 2021. Description of methods for South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00710

RELEASE HISTORY

Version 2.0 (this release)

- For this dataset we used more recent input data such that version 2.0 represents the population of South Sudan for approximately September 2020, whereas version 1.0 provides population estimates for 2019.
- For version 2.0, the input displacement data provided by IOM (Baseline Assessment Round 9) covered a larger area of South Sudan compared to the IOM data (Baseline Assessment Round 5) used for version 1.0. This means that the two datasets are not directly comparable, i.e. the difference for a given area between v1.0 and v2.0 may be due to the input data used rather than a reflection of estimated change from 2019 to 2020.
- The method used for version 2.0 differs to that used for version 1.0. This change in method was so that we could produce stand alone data layers for internally displaced persons (IDPs) and non-IDPs separately.
- For version 2.0, we used the national boundary used by IOM whereas version 1.0 we used the WorldPop mastergrid.
- We used building footprints that had been updated based on more recent satellite imagery in certain areas, as well as building footprints that were previously not available (close to the national borders) for version 1.0.

Version 1.0 (6 December 2019)

- Original release of the 2019 South Sudan adjusted population projections dataset.

FILE DESCRIPTIONS

All spatial data files are in geographic coordinate system WGS84 (World Geodetic System 1984: EPSG 4326).

1. SSD_population_v2_0_gridded.zip

This zip file contains three files:

SSD_population_v2_0_gridded_population.tif
This geotiff raster contains estimates of total population size for each approximately 100 m grid cell (0.0008333 decimal degrees grid) across South Sudan. NA values represent grid cells where no building footprints were present. Zero values represent grid cells that contain building footprints but are estimated to contain no people due to displacement of people away from those grid cells. These population estimates include decimals (e.g. 10.3 people). This provides more accurate population totals when grid cells are summed. A population estimate of 0.5 people in each of two neighboring grid cells would indicate an expectation that one person lives somewhere within those two grid cells.

SSD_population_v2_0_gridded_nonidps.tif
This geotiff raster contains estimates of non-internally displaced persons (non-IDPs) for each approximately 100 m grid cell (0.0008333 decimal degrees grid) across South Sudan, i.e. the number of people who have not been displaced from another area. This raster plus the *SSD_population_v2_0_gridded_idps.tif* raster equal the values given in the *SSD_population_v2_0_gridded_population.tif* raster.

SSD_population_v2_0_gridded_idps.tif
This geotiff raster contains estimates of internally displaced persons (IDPs) for each approximately 100 m grid cell (0.0008333 decimal degrees grid) across South Sudan, i.e. the number of people who have been displaced from another area. This raster plus the *SSD_population_v2_0_gridded_nonidps.tif* raster equal the values given in the *SSD_population_v2_0_gridded_population.tif* raster.

2. SSD_population_v2_0_agesex.zip

This zip file contains 40 rasters in geotiff format. Each raster provides gridded population estimates for an age-sex group. These were derived from the *SSD_population_v2_0_gridded_population.tif* raster. Note that, in this dataset, we do not provide age-sex group estimates for non-IDPs and IDPs separately. 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 in the groups “f80” and “m80”. We provide an addition four 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 by multiplying regional age-sex proportions (given in *SSD_population_v2_0_agesex_table.csv*) and the gridded total population estimates (*SSD_population_v2_0_gridded_population.tif*). The geographic areas covered by each region in *SSD_population_v2_0_agesex_table.csv* are provided in the raster *SSD_population_v2_0_agesex_regions.tif.*

3. SSD_population_v2_0_intermed.zip

This zip file contains two intermediary files that were used to calculate the non-IDP raster (*SSD_population_v2_0_gridded_nonidps.tif*). For more details about the relationship between each of the files, please see Figure 1. The two files included in *SSD_population_v2_0_intermed.zip* are:

SSD_population_v2_0_intermed_censusproj.tif

This geotiff raster contains estimates of 2020 census projections for each approximately 100 m grid cell (0.0008333 decimal degrees grid) across South Sudan. Zero values in this census projection raster represent grid cells that contain only IDPs, e.g. UN IDP camps. This 2020 census projection raster represents the population distribution if there had been no displacement since the 2008 census.

SSD_population_v2_0_intermed_displacedfrom.tif

This geotiff raster contains estimates of people displaced from each approximately 100 m grid cell (0.0008333 decimal degrees grid) across South Sudan.

4. SSD_population_v2_0_mastergrid.tif

This raster contains 1s for each approximately 100m grid cell in South Sudan which contain buildings according to the Maxar/Ecopia building footprints. NA values indicate grid cells that were considered unsettled or outside of South Sudan.

5. SSD_population_v2_0_sql.sql

This SQLite database contains estimates of total population size in each grid cell. This database is source data for the woprVision web application (https://apps.worldpop.org/woprVision/?data=SSDv2.0) and it can be queried using the wopr R package (Leasure et al 2020).

This contains a table with the following columns:

- "cell" contains a cell ID to identify the location. Cell IDs correspond to those the cell IDs of SSD_population_v2_0_mastergrid.tif.
- "x" and "y" columns contain WGS84 coordinates for the centroid of the grid cell.
- "Pop" column contains column contains a population estimate for each grid cell.
- "agesexid" column contains the region ID for the age-sex proportions that are provided in *SSD_population_v2_0_agesex_table.csv* and *SSD_population_v2_0_agesex_regions.tif*.
- "area" contains the grid cell area in hectares.

6. SSD_population_v2_0_tiles.zip

This tiled web map allows for rapid display of the approximatively 100 m gridded total population estimates across the study area (i.e. *SSD_population_v2_0_gridded_population.tif*). These can be used to develop web applications for the model results. The tiles were created in XYZ format (i.e. compatible with Leaflet) with full coverage of the study area for the zoom levels 1 to 14. These tiles are source data for the woprVision web application (https://apps.worldpop.org/woprVision/?data=SSDv2.0).

7. SSD_population_v2_0_methods.pdf

This document provides a detailed description of the methodology developed and implemented to produce this dataset.

ASSUMPTIONS AND LIMITATIONS

We link county level (administrative level 2) census projections to publicly available administrative boundaries provided by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) dataset (OCHA, 2020). These boundaries will likely differ to the exact boundaries used for the 2008 census and so the disaggregated census numbers (*SSD_population_v2_0_intermed_censusproj.tif)* and, consequently, the total population sizes per grid cells (*SSD_population_v2_0_gridded_population.tif)* may not be accurate in counties where the mismatch in boundaries is significant. Unfortunately, it is difficult to identify where these mismatches may be due to the limited availability of data on the boundaries used for the 2008 census.

We assume that the building footprints dataset is a true representation of 'settled' areas, i.e. areas containing people. Further work is needed to quantify error rates in the building footprints data as well as how these errors vary across settlement types.

The data available for place of origin of IDPs is given as crude estimates and that available for refugees is only for a small subsample. We have used the available data to estimate the total number of displaced persons per county of origin. These estimates could include large inaccuracies, and we recommend that future work focuses on collecting data on the number of people displaced from each location across South Sudan, as this is likely to be the biggest source of error in our population estimates.

We assume that the number of IDPs recorded in the International Organisation for Migration data (IOM 2020; 2021) is correct. We use simple rules to disaggregate the total number of IDPs at each destination site across nearby buildings. The real spatial extents covered by IDPs may differ from that shown in the IDPs raster (*SSD_population_v2_0_gridded_idps.tif*) and suggest that this dataset be used only as a guide for household surveys.

UNHCR reports that there are 300,000 Sudanese refugees residing in South Sudan. We have not included these people in our estimates due to the lack of information about where in the country they may live.

Generally, we recommend that these population estimates be used as a guide and that household listings be carried out prior to any household surveys in South Sudan due to the uncertainties associated with all input population data used to generate this dataset.

The population counts for specific age-sex groups should be used with caution. These are calculated using regional proportions and do not consider differences between settlements within regions nor differences between IDP and non-IDP populations.

SOURCE DATA

Boundaries. We used publicly available administrative level 2 boundaries provided by the OCHA dataset (OCHA, 2020).

Building footprints. We used rasterised building footprint centroids to define settled grid cells. We used the methodology outlined in Dooley et al 2020, applied to the latest building footprints provided by Ecopia.AI and Maxar Technologies, Inc. (2019 and 2020) for South Sudan and its surrounding countries so that the full extent of the boundaries dataset was covered by the building footprints data.

Census projections. We used the South Sudan National Bureau of Statistics county level census projections for 2020 (National Bureau of Statistics, 2015).

Displacement data. IOM collects extensive information about IDPs in South Sudan. From 1st July 2020 to 30th September 2020, IOM conducted Round 9 of their 'Baseline Assessment' of IDPs (IOM, 2020; 2021) which includes estimated numbers of IDPs at more than 2,000 locations across South Sudan. We used the two excel spreadsheets released by IOM for Round 9 (IOM, 2020; 2021).

Refugee data. We used the United Nations Refugee Agency (UNHCR) estimates of South Sudanese refugees living outside of South Sudan as of September 2020. We also used the UNHCR's 'Regional intention survey of South Sudanese refugees' to estimates the number of refugees displaced from each county (UNHCR, 2019).

Model covariates. Please see Method Report (Dooley et al 2021) for more details.

- WorldPop Global covariates which includes slope, topography, nighttime lights and distance to different land use types (Lloyd et al, 2019)
- WorldClim covariates (Fick and Hijmans, 2017)
- HydroSHEDS covariates (Lehner et al, 2008)
- Building footprint metrics covariates (calculated from Ecopia.AI and Maxar Technologies, Inc. 2019; 2020) ; following method of Dooley et al, 2020
- Armed Conflict Location Events Database (Raleigh et al, 2010)

METHODS OVERVIEW

For a detailed description of our approach, please see Dooley et al 2021. Here we provide an overview.

South Sudan's last population and housing census was conducted in 2008 prior to its independence from Sudan in 2011. Estimating the population of South Sudan and mapping its spatial distribution is incredibly challenging due to ongoing conflict, flooding and famine that continues to drive large scale movement within the country as well as across national borders into neighbouring countries. We generated this dataset using an approach that integrates several different data sources to estimate South Sudan's population at a high spatial resolution.

Our approach models the population distribution to likely settled locations and adjusts for displacement. Conceptually, the population in a given location is expected to follow:

Final population = baseline population + in-displacement - out-displacement

Figure 1 shows a schematic of our approach, and indicates the output raster included in this data release corresponding to each element of the approach. The main methodological elements are:

a) We disaggregated projected county level census population estimates for 2020 to a high spatial resolution using a number of ancillary geospatial datasets which depict factors known to relate to human population presence. For this disaggregation, we implemented a random forest machine

learning-based dasymetric approach outlined in Stevens et al, 2015. (*baseline population*; *SSD_population_v2_0_intermed_censusproj.tif*).

b) We used geocoded internally displaced populations and building footprints to demarcate destination spatial extents of IDP populations. We then disaggregated the IDPs across their population's extents (*in-displacement*, *SSD_population_v2_0_gridded_idps.tif*)

c) We disaggregated county level estimates of the number of people displaced from any area within a county. Again, we implemented the random forest machine learning-based dasymetric approach (Stevens et al, 2015) using ancillary geospatial datasets that included those relating to conflict and flooding (*out-displacement*, *SSD_population_v2_0_intermed_displacedfrom.tif*)

d) We subtracted the layer produced in c) from the layer produced in a) to generate the non-IDPs raster *(SSD_population_v2_0_gridded_nonidps.tif*), *i.e. baseline population - out-displacement*. We then added the non-IDPs and IDPs (in b) layers to produce the final adjusted population estimates that account for displacement (*final population*, *SSD_population_v2_0_gridded_population.tif*).

The age-sex disaggregated population data was produce post-hoc, by multiplying regional age-sex proportions and the gridded total population estimates together.

All data preparation and analysis was carried out using R version 4.0.2 (R Core Team, 2020). The code used to generate this dataset is available here: https://github.com/cadooley/SSD_pop_v2.0

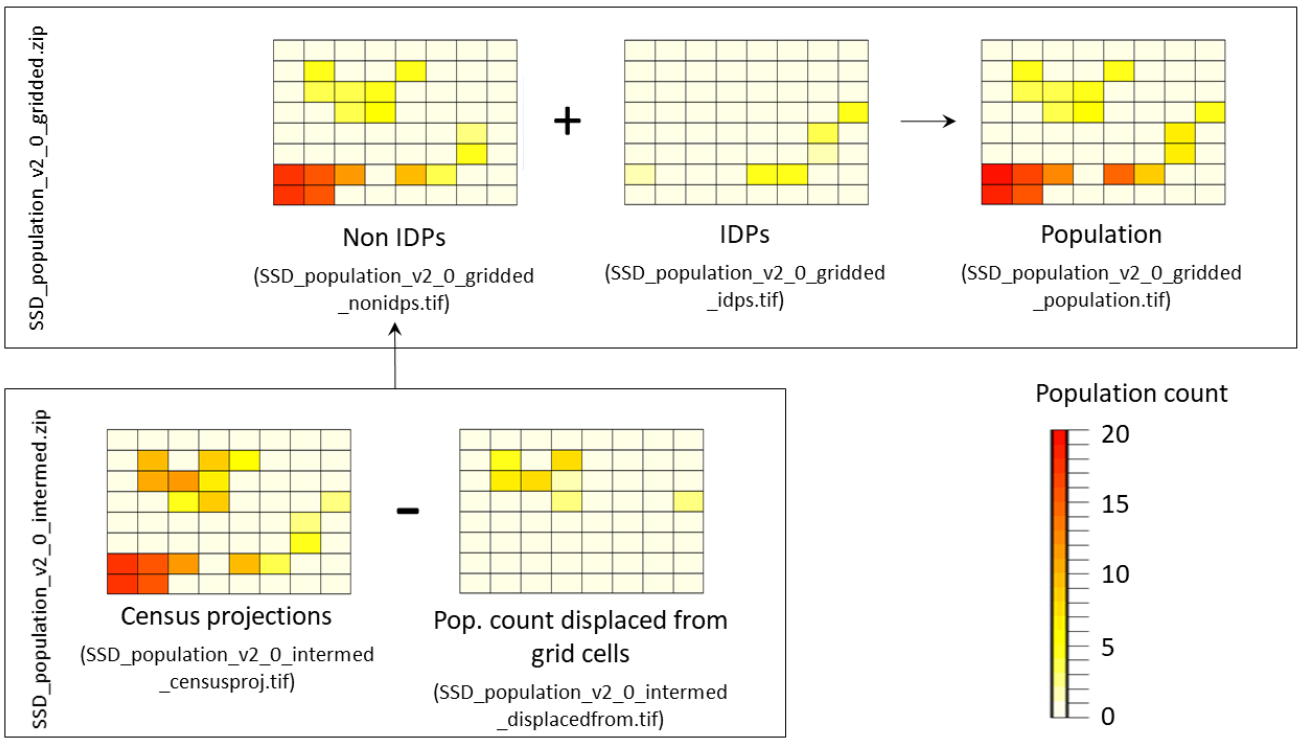

Figure 1. Schematic of the approach used to generate high spatial resolution population estimates, adjusted for displacement, for South Sudan 2020. Note, this is for illustrative purposes and does not correspond to a specific location in the dataset.

REFERENCES

Dooley CA, Boo G, Leasure DR, Tatem AJ. 2020. Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. doi:10.5258/SOTON/WP00677.

Dooley CA, Jochem WC, Sorichetta A, Lazar AN and Tatem AJ. 2021. Description of methods for South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00710

Ecopia.AI and Maxar Technologies, Inc. 2019. Digitize Africa Data. Ecopia.AI and Maxar Technologies, Inc. Data downloaded on 1st December 2020 for COG, UGA, KEN, ETH, SDN, CAF.

Ecopia.AI and Maxar Technologies, Inc. 2020. Digitize Africa Data. Ecopia.AI and Maxar Technologies, Inc. Data downloaded on 1st December 2020 for SSD.

Fick, S.E. and R.J. Hijmans, 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37 (12): 4302-4315.

IOM Displacement Tracking Matrix. 2020. South Sudan - Baseline Locations Round 9. Data released 1st December 2020 https://displacement.iom.int

IOM Displacement Tracking Matrix. 2021. South Sudan - Baseline Assessment Round 9 - IDP And Returnee. Data released 31st January 2021 https://displacement.iom.int

Leasure DR, Bondarenko M, Darin E, Tatem AJ. 2020. wopr: An R package to query the WorldPop Open Population Repository, version 0.4.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00688. https://github.com/wpgp/wopr

Lehner, B., Verdin, K., Jarvis, A. 2008. New global hydrography derived from spaceborne elevation data. Eos, Transactions, AGU, 89(10): 93-94. Data is available at www.hydrosheds.org

Lloyd CT, Chamberlain H, Kerr D, Yetman G, Pistolesi L, Stevens FR, Gaughan AE, Nieves JJ, Hornby G, MacManus K, Sinha P, Bondarenko M, Sorichetta A, Tatem AJ. 2019. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big Earth Data, 3(2), 108-139. https://dx.doi.org/10.1080/20964471.2019.1625151

National Bureau of Statistics. 2015. "Population Projections, South Sudan. From 2015 - 2020." https://ssnbs.org/home/document/census/population-projections-for-south-sudan-by-county-from-2015-to-2020

OCHA. 2020. South Sudan (SSD) Administrative Boundary Common Operational Database (COD-AB). Accessed via: https://data.humdata.org/dataset/south-sudan-administrative-boundaries

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raleigh, C., Linke, A., Hegre, H. and Karlsen, J. 2010. "Introducing ACLED-Armed Conflict Location and Event Data." Journal of Peace Research 47(5) 651660. Data downloaded from https://www.acleddata.com/data/ on January 2021

Stevens F.R., Gaughan A.E., Linard C., Tatem A.J. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." PLOS ONE 10 (2). Public Library of Science: 1-22. https://doi.org/10.1371/journal.pone.0107042

UNHCR. 2019. Regional intention survey of South Sudanese refugees. https://microdata.unhcr.org/index.php/catalog/224

UNHCR. 2020. Regional overview of the South Sudanese refugee population: 2020 SOUTH SUDAN REGIONAL RRRP. Downloaded from: https://data2.unhcr.org/en/documents/details/79631