Abstract

Sub Theme: Data and Methods

Session: Hybrid census opportunities in Africa

Topic: South Sudan Population Estimation Survey, 2021: Up-to-date Modelled Population Estimates

Submitted by
1. Dr Ademola Olajide (olajide@unfpa.org)
2. Francis Tukwasibwe (tukwasibwe@unfpa.org)

Key words: Micro census, satellite imagery, geospatial data, population estimation, South Sudan

1. Introduction/background

Population numbers at local levels are fundamental data for many development applications, including the delivery and planning of services, election preparation and response to disasters. In resource-poor settings recent and reliable demographic data at subnational scales can often be lacking or incomplete. National population and housing census data can be outdated.

The population distribution of South Sudan has undergone substantial changes in recent years, including internal and international displacement and migration due to conflicts and recent periods of severe flooding. Moreover, there have been only limited, large-scale data collection activities in the country, and the previous population and housing census was completed in 2008, prior to South Sudan's independence. The result of these conditions is significant uncertainty in the current population of the country and the future population distribution and dynamics. Yet in order to support South Sudan's growth and development, up-to-date population data at subnational levels are needed.

Within the 2030 agenda, the growing requirement for spatially disaggregated population data has triggered the exploration of new data sources at different geographical scales and time periods, especially in highly stressed countries and countries without a recent population census.

Recent advances in satellite imagery and geospatial data, along with statistical methods are presenting opportunities for model-based approaches to estimating population (UNFPA, 2020; Wardrop et al., 2018). These model-based approaches generally rely on observations of population in sample sites collected from across a study region (micro census). These population data are combined with geospatial data layers in a statistical model to predict and map population for the unsampled areas in a study region. These types of model-based predictions of the population have been referred to as a "bottom-up" approach (Wardrop et al., 2018). While modelled estimates do not replace the need to conduct a full, national census, they can provide important information for policy, planning, service delivery, and other operational needs that require up-to-date estimates of the population (UNFPA, 2020).

2. The Goal and objectives of the South Sudan PES

The goal of the 2021 Population Estimation Survey was to contribute to the improvement of quality of life of South Sudanese through the provision of current and reliable population estimates for development planning, policy formulation and service delivery, as well as for monitoring and evaluating population programmes.

The specific objectives were to:

- generate reliable modelled population estimates on population density and basic demographic characteristics for all levels of administration;

- generate indicators for monitoring and evaluation of programs prior to conduct of the population and housing census and other population-based surveys.
- provide information to be used for development of advocacy materials for policy- makers.

3. Methodology

Unlike in a Population and Housing Census which is a complete count of people in the country, the Population Estimation Survey was based on enumeration of all people from sampled sites (micro census). A total 1,536 sample sites were selected in areas across South Sudan based satellite-derived information and a system of equally sized grid squares. The target population was all households, household members and residents within the bounds of the sampled sites.

The population estimation approach used in South Sudan is a two-step method to first predict a baseline population and then adjust that distribution to reflect likely internal displacement. The result is a high spatial resolution estimate of the total population for every location in South Sudan, reflecting the current population at the time the PES was conducted in summer 2021. The baseline population is first estimated using a statistical model to understand the relationship between observed demographic characteristics in the PES data and mapped socio-environmental covariate layers describing the local context. The statistical model is then used to predict the baseline population into unsurveyed locations based on these modelled relationships. In the second step, the baseline population is adjusted for recent displacements and migrations. A subsequent modelling component is used to generate age-sex disaggregated population estimates.

Schematic of a Population Estimation Survey: Micro Census (Population Counts) ➕Geospatial Covariates ➡ Population Estimation (Prediction of population in the un surveyed areas based on covariates, using statistical modelling)

The covariates used for for the Population Estimation Survey are those that are i) strongly correlated to population density and ii) available consistently across all areas where the population estimation is required. The covariates used are Enhanced vegetation index, Flood events, Distance to floods, Distance to main roads, Building area (mean), Distance to conflicts (2015 – 2020), Travel time to major cities, Distance to water, Slope and Elevation. When mapped as a geospatial layer, the covariates aid the prediction of population into nonsurveyed areas. To identify the covariate values associated with each analysis cluster, the GPS points of the enumerated households were buffered by 20 meters and the mean, minimum, maximum, and range of covariate values were extracted. A multi-step process was used to select the covariates used in the model. Descriptive statistics and exploratory data visualisations were first used to compare the distribution of each covariate to the observed populations in the PES. Highly correlated covariates ($\rho > 0.8$) were excluded to reduce the effects of multicollinearity in the model. Each covariate was then separately regressed against population density and a backwards stepwise selection process was used to select the best subset of covariates based on Akaike Information Criterion (AIC).

4. Limitations

The model results show some uncertainty in the population estimate for South Sudan. This uncertainty originates from the combination of different sources of data as well as potential errors in the input sources. It reflects the observed variation in the population density from the PES field work which could not be fully explained by the statistical model. Uncertainty also arises from the data processing steps used to create the analysis units to correct for the lack of site boundaries. There is also uncertainty in the number and location of displaced persons from the IOM data; however, this uncertainty remains unquantified. This work has several assumptions and limitations. Because the sample frame and the predictions are based on the presence of structures in satellite imagery, populations in temporary shelters may be missed.

Similarly, there is also no specific accounting for nomadic populations or seasonal migration in this estimate. Most of the satellite imagery was collected in 2020; however, in areas where buildings have recently changed the

population may be over- or under-predicted. Moreover, the extracted building footprints do not distinguish residential from non-residential building and as a result population may be predicted into non-residential areas. The age-sex structure model only uses information from 20 the PES sample and applies the modelled proportions to the total population (baseline plus IDPs). This assumes that IDPs have the same age-sex structure which may not necessarily be true. Future work could improve on the population predictions in several ways.

This approach can never replace the rich production of data on the individual, family, household or community generated by a traditional population and housing census. However, where a traditional census cannot be fully executed in all locations of a given country due to insecurity or other concerns, then this hybrid approach could produce population estimates for small areas or uniform, detailed grids in the absence of traditional census data.

5. Results

Overall, the results of the population model suggest that there have been considerable changes in the population distribution in South Sudan since the last census in 2008. The overall population of South Sudan has likely grown, and the distribution of the population has shifted geographically. The model of age-sex structures suggests a potential gap in males aged 20-49. These prime working age adults may be outside the country for work or schooling, or they may have experienced higher mortality during the civil war and recent conflicts. The base of the age pyramids is also smaller than might be expected. There is insufficient evidence to tell if this reflects a true change in fertility or infant mortality, or possibly a reluctance of respondents to list infants in the household roster. Comparing the model-based estimates using the PES data with existing population projections for the country highlights these differences. The largest populations are now likely in Warrap and Northern Bahr el Gazal states while areas in the eastern and northeastern parts of the country (especially Upper Nile and Jonglei states) have relatively lower populations. These changes likely reflect impacts of conflicts, floods, and return migrations all leading to internal displacements. Outside of these four states, the different sources of population estimates are generally closer. Note that the current population projections for the country produced by the National Bureau of Statistics (SSNBS, 2014; 2015) are based on the 2008 census and assume a smooth, uninterrupted growth rate of between 3% and 4% per year. The method used to produce the Common Operational Dataset (UN OCHA, 2021) is not well documented; however, it is described as a population model that does consider recent internal displacements using data from IOM DTM.

Table 1. Population Size by State and Sex

State and Administrative Area	Total Female	Total Male	Total
Upper Nile	401,830	388,317	790,147
Jonglei	428,547	362,558	791,105
Unity	481,593	411,187	892,780
Greater Pibor Admin Area	129,631	110,471	240,102
Ruweng Admin Area	126,407	108,009	234,416
Western Bahr el Ghazal	290,425	272,130	562,555
Northern Bahr el Ghazal	993,310	931,032	1,924,342
Warrap	1,401,801	1,237,683	2,639,484
Lakes	680,910	584,563	1,265,473
Abyei Admin Area	71,262	62,696	133,958
Western Equatoria	346,886	316,347	663,233
Central Equatoria	693,689	630,832	1,324,521
Eastern Equatoria	536,638	445,264	981,902
Total	**6,582,929**	**5,861,089**	**12,444,018**

Fig 1. Population Pyramid

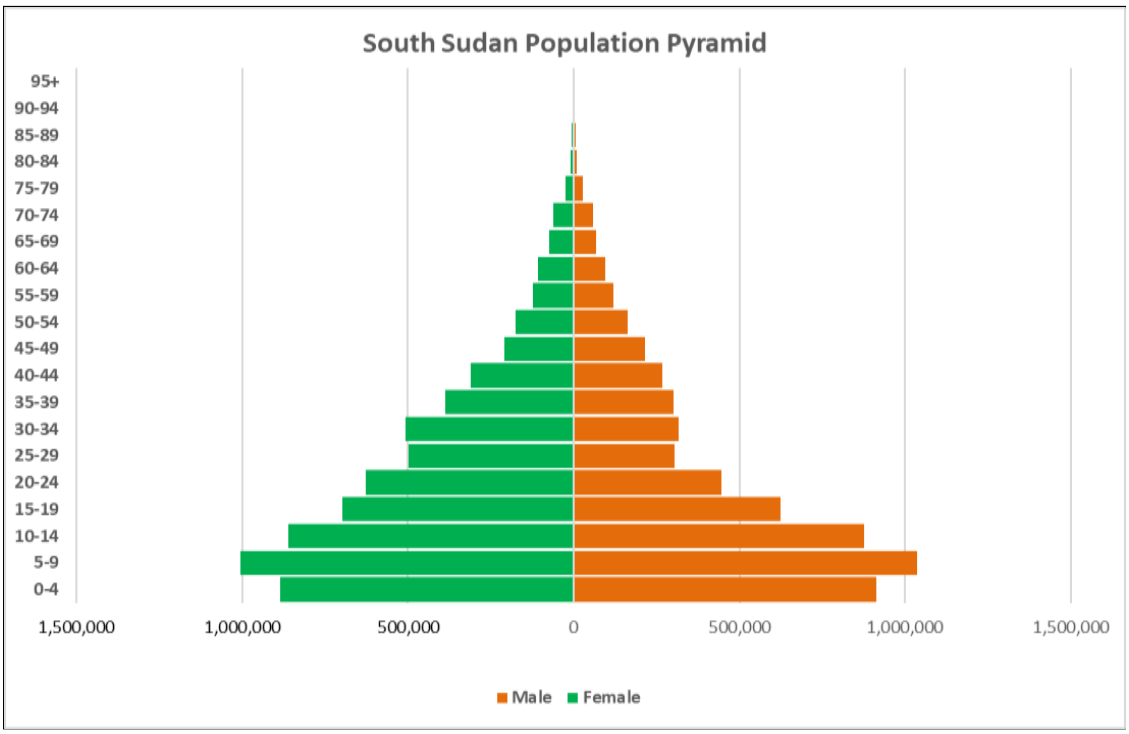
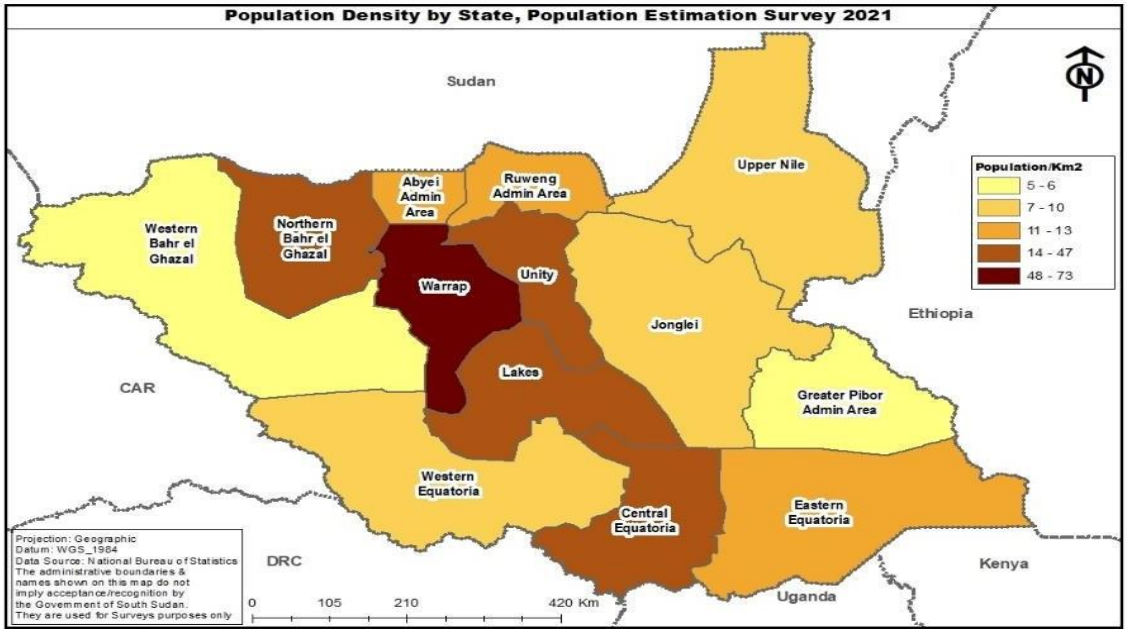


Fig 2. Population Distribution and density



References

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 .

Center for International Earth Science Information Network (CIESIN), Columbia University and Novel-T. (2021). GRID3 South Sudan Settlement Extents, Version 01. Palisades, NY: Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). https://doi.org/10.7916/d8-30rg-nd63 .

Dooley, C. A., Boo, G., Leasure, D. R., Tatem, A. J. (2020). Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. WorldPop, University of Southampton. https://doi.org/10.5258/SOTON/WP00677

Dooley C. A., Chamberlain H. R., Leasure D. R., Membele G. M., Lazar A. N., Tatem A. J. (2021a). "Description of methods for the Zambia modelled population estimates from multiple routinely collected and geolocated survey data, version 1.0." WorldPop, University of Southampton. doi: 10.5258/SOTON/WP0070.

Dooley, C. A., Jochem W. C., Sorichetta, A., Lazar, A. N., Tatem, A. J. (2021b). "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 and DigitalGlobe (2017) Technical specification: Ecopia building footprints powered by DigitalGlobe. Available at: https://dg-cms- ploads poduction.s3.amazonaws.com/uploads/legal _document/file/109/DigitalGlobe_Ecopia_Building_Footprints_Technical_Specification.pdf (accessed 11 April 2022).

Ecopia.AI and Maxar Technologies. (2020). Digitize Africa data. Ecopia.AI and Maxar Technologies. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. https://dl.acm.org/doi/10.5555/3001460.3001507

Hoffman, M. D., Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." Journal of Machine Learning Research 15, 1593-1623.

IOM Displacement Tracking Matrix. (2022). South Sudan – Baseline Assessment Round 11 –IDP and Returnee. Data released 21 Jan 2022. https://displacement.iom.int .

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., Tatem, A. J. (2020). "National population mapping from sparse survey data: A hierarchical Bayesian modelling framework to account for uncertainty." Proceedings of the National Academy of Sciences, 177(39), 24173-24179. https://www.pnas.org/doi/10.1073/pnas.1913050117

Raleigh, C., Linke A., Hegre, H., Karlsen, J. (2010). "Introducing ACLED-Armed Conflict

Location and Event Data." Journal of Peace Research, 47(5) 651-660. R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/ .

South Sudan National Bureau of Statistics [SSNBS]. (2014). Population Projections, South Sudan from 2008 – 2015. https://www.ssnbss.org/home/document/census/populationprojection-for-south-sudan-by-countyby-sex-from-2008-2015.

South Sudan National Bureau of Statistics [SSNBS]. (2015). Population Projections for South Sudan by County: 2015 – 2020. https://nbs.gov.ss/wpcontent/uploads/2021/05/Population-Projectiosn-of-South-Sudan-2015-2020.pdf.

Stevens, F., 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): e0107042. https://doi.org/10.1371/journal.pone.0107042 .

UN OCHA. (2021). 2022 South Sudan Population Estimates: Endorsed baseline and guidance note. https://data.humdata.org/dataset/cod-ps-ssd .

UNFPA. (2020). "The value of modelled population estimates for census planning and preparation." Technical Guidance Note, August 2020 (updated version 2). https://www.unfpa.org/resources/value-modelled-population-estimates-censusplanning-and-preparation .

Vehtari, A., Gelman, A., Gabry, J. (2017). Practical Bayesian model evaluation using leaveone-out cross-validation and WAIC. Statistics and Computing. 27(5), 1413--1432. doi:10.1007/s11222-016-9696-4 .

Vehtari, Aki, Gelman, Andrew, Simpson, Daniel, Carpenter, Bob, Bürkner, Paul-Christian (2019). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. arXiv preprint arXiv:1903.08008.