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

WorldPop, University of Southampton

28 March 2021

Contents

1   Introduction

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. Here we present an approach that integrates several different data sources available at differing spatial scales to estimate South Sudan's population at a high spatial resolution.

In this report we describe the methods used to produce the 'South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0' (Dooley et al 2021).

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

The key elements of our approach are:

a) disaggregating projected census population estimates for 2020 to a high spatial resolution using a number of ancillary geospatial datasets, including building footprints, which depict factors known to relate to human population presence (*baseline population*)
b) using geocoded internally displaced populations and building footprints to demarcate destination spatial extents of displaced populations (*in-displacement*)
c) modelling where people have been displaced from at a high spatial resolution using geospatial data including conflict locations (*out-displacement*)
d) combining the projected population estimates (in a) and displaced population estimates (in b and c) to produce the final adjusted population estimates that account for displacement (*final population*)

In this report we present model assessments, provide example areas from the results and discuss future work needed to improve mapping efforts in South Sudan.

Note that the final dataset is most likely to represent South Sudan's population distribution as of September 2020 given the age of the input data.

2   Methods

All data preparation and analysis was carried out using R version 4.0.2 (R Core Team, 2020). We used the World Geodetic System 1984 (WGS84) projection system for all geospatial data. If input data was not in this projection, the data was converted accordingly. We used the WorldPop South Sudan mastergrid for the grid positioning of all rasters throughout the analysis. All output data is provided in raster format with WGS84 projection and approximately 100m x 100m resolution (0.0008333 decimal degrees grid).

2.1   Boundaries, settled area and unadjusted (baseline) county population totals

To define our study area, we used the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) operational administrative level 2 boundaries (counties) dataset (OCHA, 2020). Grid cells within this area were defined as 'settled' if they contained any building centroids in the Digitise Africa building footprints dataset (Ecopia AI and Maxar Technologies, Inc., 2019; 2020) for South Sudan and its neighbouring countries (Central African Republic, Democratic Republic of Congo, Ethiopia, Kenya, Sudan, Uganda). For the unadjusted (baseline) population estimates, we used county level population projections for 2020 from the 2008 census produced by the South Sudan National Bureau of Statistics (National Bureau of Statistics, 2015). These projections are derived using fertility and mortality rates, and do not account for displacement.

2.2 Mapping of internally displaced persons

The International Organisation for Migration (IOM) collect extensive information about internally displaced persons (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. The data is collected from key informants at each location and verified by the IOM team who report on the consistency of the data and whether it is in line with observations. IOM's Baseline Assessment Round 9 covers almost every area of the country and provides the highest resolution of locations of IDP populations compared to other datasets where estimates are given for large areas only. However, we acknowledge that there is likely to be large and varying uncertainty in the IDP counts. In this study we do not consider the uncertainty in IDP estimates as it would be incredibly difficult to quantify given the data collection methodology. All IOM data mentioned in this report refers to their Baseline Assessment Round 9. Note that we refer to a single location of IDPs as an IDP population.

As of September 2020, IOM reported 1,615,765 IDPs residing inside South Sudan. Of the total number of IDPs, an estimated 554,081 people were displaced from a different county and 1,061,684 people were displaced from within their same county. All reported IDP populations were 28,000 people or less, with the exception of UNMISS Bentiu Protection of Civilians camp which had 97,321 IDPs.

In this study we aim to identify the likely area in which IDP populations may be living, which then allows us to redistribute those populations over the most likely settled areas. This is necessary because defined boundaries of the areas in which IDPs reside are not available. IOM classifies locations of IDP populations as either a 'displacement site' or within a 'host community'. We assume that 'displacement sites' have been purposefully constructed or converted to accommodate IDPs and, therefore, are occupied by IDPs only. 'Displacement sites' include UN Protection of Civilians sites and converted public buildings such as schools. Conversely, we assume that host communities can have both IDPs and non-IDPs. Because of these assumptions, we mapped the area occupied by 'displacement sites' first and then excluded these areas for the mapping of all other populations (IDP host communities, unadjusted (baseline) population and areas people have been displaced from).

There were a total of 99 open 'displacements sites' recorded by IOM (a further 23 closed sites were also reported to have no IDPs present as of September 2020). There was one open site recorded for Korijo IDP Camp Zone 1, 2 and 3, and we split this into three separate sites based on the proportion of IDPs across the three camps and their locations in July 2017 (IOM, 2017). We merged the sites recorded for 'Panyiduay Hospital' and 'School Panyiduay' as they had the same location coordinates. This meant that IDPs across the two sites occupy the same location in our final dataset.

After these data edits, we mapped the resulting 100 'displacement sites' by identifying the nearest settled grid cell to the site's recorded location coordinates and then buffering an area around the focal cell. The size of the buffered area varied per site and was big enough such that the number of IDPs per settled cell did not exceed 500 people. We capped the buffer size to a maximum of 3.5km, and therefore allowed more than 500 people per settled cell where the maximum buffer was reached. The distance restriction was implemented so that 'displacement sites' weren't incorrectly mapped as sparse and sprawling. We used a people per cell restriction rather than people per building because 'displacement sites' are likely to have non-permanent structures such as tents and so we didn't want to rely on the count of building footprints in these areas reflecting the situation at the time of data collection. The high maximum value of 500 people per grid cell was selected to allow for potentially high population densities in low resource situations and to restrict the spatial areas covered by these IDP only sites to realistic spaces in urban areas.

We cross referenced a subset of the resulting 'displacement sites' areas with publicly available spatial data relating to 22 UN IDP camps (https://unitar.org/maps/countries/94) and found that the areas matched well. However, there is no available spatial data for non-UN camps and cross referencing was therefore not possible for non-UN 'displacement sites'. In the absence of this data, we highlight that there are potential uncertainties in our results. In particular, the maximum of 500 people per grid cell may have been too high for rural 'displacement sites' and may have led to spatial extents that are too small (if the maximum was 250, say, the spatial extent would have been bigger because the buffer size would have needed to be bigger to allow for the lower people per cell). With limited data available for non-UN sites, it is outside the scope of this study to assess our mapped 'displacement sites' thoroughly, and we recommend that further evaluation be done in collaboration with NBS, IOM and other data collectors.

There were 2,097 'host community' sites with recorded IDPs. To map the spatial extents of these populations from their georeferenced point locations, we incrementally increased a buffer around the location until there was no more than 1 IDP per building within the buffer. Given the possible national population totals, reasonable household sizes of 5.5-6.5 persons (average household size in the 2008 census was 5.9; Minnesota Population Center, 2020) and the number of building footprints, we estimate that approximately 50% of the building footprints are residential. Based on this assumption, along with a reasonable average limit of 2 IDPs per residential building in host community settings, we considered 1 IDP per building to be a sensible maximum, while acknowledging that this could vary significantly between IDP populations across the country. If buffer areas of different IDP populations overlapped, we allowed the limit per building within the overlapping area to increase (with the limit being 1 IDP per building per IDP population). We applied a maximum buffer of 50km in order to prevent IDPs being mapped in potentially incorrect locations. We chose 50km because there was always at least one building within that distance of the recorded IDP point location.

2.3 Mapping of unadjusted population (population distribution in the absence of displacement)

We disaggregated the 2020 county level population projection estimates (NBS, 2015) to a high spatial resolution grid using a random forest machine learning-based dasymetric approach (Stevens et al, 2015). This approach allowed us to predict grid cell level population estimates based on modelled relationships with geospatial data (covariates) while preserving the projected county-level population totals. Here we used a bespoke set of covariates that included data potentially important for predicting the distribution of South Sudan's unadjusted population projection. We started with a total of 36 covariates relating to land use types, physical attributes, climate and the built environment (Tables A1-A4). The steps taken in our approach were:

- Prepare covariate layers so that their spatial extent matches the OCHA boundaries
- Mask out the areas considered to be IDP 'displacement sites' (e.g. UN Protection of Civilians sites, i.e. not 'host communities') across all covariates
- Calculate mean covariate values across the settled grid cells per county
- Calculate population density per county (population per settled area defined as the sum of areas of grid cell containing building centroids)
- Run iterations of the random forest model, with log mean population density as the response variable, dropping covariates that have low importance scores in each
- Run the final model that included important covariates only
- Predict log population density weighting layer using the final model and covariate values for each settled grid cell (excluding those classified as a displacement site)
- Distribute the aggregate (county-level) population counts to grid cells using the weighting layer

See Figure A1 for the distribution of unadjusted population counts and log unadjusted population density (model response variable) across counties.

2.4 Mapping locations where people have been displaced from

Data on where people have been displaced from is notoriously difficult to collect due to multiple settlement and administrative unit names used for given locations and challenges in recording multiple origins across large displaced populations. The IOM data summarises where the majority of current IDPs have been displaced from. This is reported for each IDP population (i.e. at each destination location) and by the following arrival time periods: a) 2014-2015; b) 2016-2017; c) 2018 pre- Revitalised Agreement for the Resolution of Conflict in South Sudan (R-ARCSS); d) 2018 post- R-ARCSS; e) 2019; and, f) 2020. While we know that different IDPs may have arrived at a destination location from different places of origin during a given time period, this data is the best resource compared to other available data. From this county origins data, we calculated the total estimated number of IDPs that had been displaced from each county.

In addition to 1,615,765 IDPs, there was an estimated 2,185,117 refugees living outside of South Sudan (as of September 2020; UNHCR, 2020). Very little data exists for the place of origin of these refugees. Here we combined

information from two datasets to estimate the number of refugees displaced from each county but emphasise the unquantified and potentially large uncertainty in these estimates. The first dataset was a household survey conducted by the UN Refugee Agency (UNHCR) in each of the countries surrounding South Sudan (UNHCR, 2019). This survey covered a total of 6,964 households across 15 refugee camps: Ethiopia (3 camps), Uganda (3 camps), Sudan (2 camps), Kenya (3 camps), the Democratic Republic of Congo (3 camps) and the Central African Republic (1 camps). While data on county of origin was not available, the survey reported the proportion of respondents from each state in South Sudan. We applied these proportions to the total number of refugees to estimate the numbers displaced from each state, and then split these state level estimates across counties based on the reported 'not yet returned' estimates provided at each location surveyed in the IOM Baseline Assessment. 'Not yet returned' is an estimate of how many people have been displaced from the survey location and not returned, reported by the key informants. The 'not yet returned' numbers have a very large level of uncertainty due to recall biases and gaps in coverage for areas with consistently high conflict (as the Baseline Assessment targets locations of displaced persons and/or returnees, not areas where there are neither). Therefore, we did not use the exact numbers but instead used the proportion of 'not yet returned' across counties within a given state. The 'not yet returned' estimates were used for this step in estimating the origin of refugees, rather than using the origin data of IDPs because the origins of those displaced internally to neighbouring counties can be very different to the origins of those displaced into neighbouring countries. For example, Yei county has experienced some of the highest levels of conflict in the country yet there are very few IDPs reportedly displaced from Yei. Instead, people were forcibly displaced into Uganda, and this is evident in the relatively high estimates of 'not yet returned' for the county.

After applying this procedure we found that four counties (Ibba, Western Equatoria; Morobo, Central Equatoria; Nagero, Western Equatoria; Panyikang, Upper Nile) had less than 10% of their census projection population remaining when the total number of people displaced from them was subtracted. Because of the high uncertainty in these displacement origin numbers, we re-estimated the total number of refugees from these four counties such that they did not exceed 90% of the census projection. In reality, it is possible that county populations are depleted to a larger extent, but we erred on the side of caution to avoid under-representing current populations of these counties in the final dataset. Reducing the number of refugees displaced from these four counties resulted in a total of 2,062,059 refugees being accounted for instead of the reported 2,185,117.

We summed the estimated number of IDPs and refugees displaced from each county, and then applied the random forest disaggregation methodology to map estimates of people displaced from each settled grid cell across the country. For this we used the same steps as outlined in section 2.3. See Figure A1 for the distribution of estimated counts of people displaced from each county as well as the model response variable (log estimated density of people displaced from each county). Again, these settled grid cells did not include those classified as purposefully constructed or converted 'displacement sites'. In addition to the 36 covariates used for predicting unadjusted population counts (Tables A1-A4), we generated covariates relating to conflicts (Table A5) as variables relating to conflicts are critical for predicting where people have been displaced from. We used the reported conflict events between January 2014 and September 2020 in the Armed Conflict Location Events Database (ACLED; Raleigh, 2010). After initial comparisons of different potential measures, we created covariates for distance to: a) all 'Battles' and 'Violence against civilians' events; b) all events that resulted in 5 or more fatalities; c) all events that resulted in 20 or more fatalities; and, d) all events that resulted in 50 or more fatalities. Each layer was created separately for each year between 2014 and 2020.

2.5 Final population distribution accounting for displacement

To produce the final, high spatial resolution gridded population distribution, we simply applied the following equation using the grid cell level datasets:

Population count = unadjusted (baseline) pop. count + IDP count - pop. count displaced from location

This corresponds to the equation in the introduction section of this report:

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

3 Results

3.1 Settled area

Figure 1 shows the dates of the satellite imagery used to extract the building footprints. Most buildings were extracted from imagery between 2018 and the first quarter of 2020. In this study we did not consider seasonal differences in population distribution but note that the distribution of structures is largely representative of patterns corresponding to October to March rather than April to September. Given that seasonal migration and seasonal flooding are significant events in South Sudan, we strongly recommend that further work be carried out to understand changes in the presence of shelters/structures across seasons, and whether these changes fully reflect changes in population distributions, or whether additional data unrelated to buildings is needed to map seasonal differences.
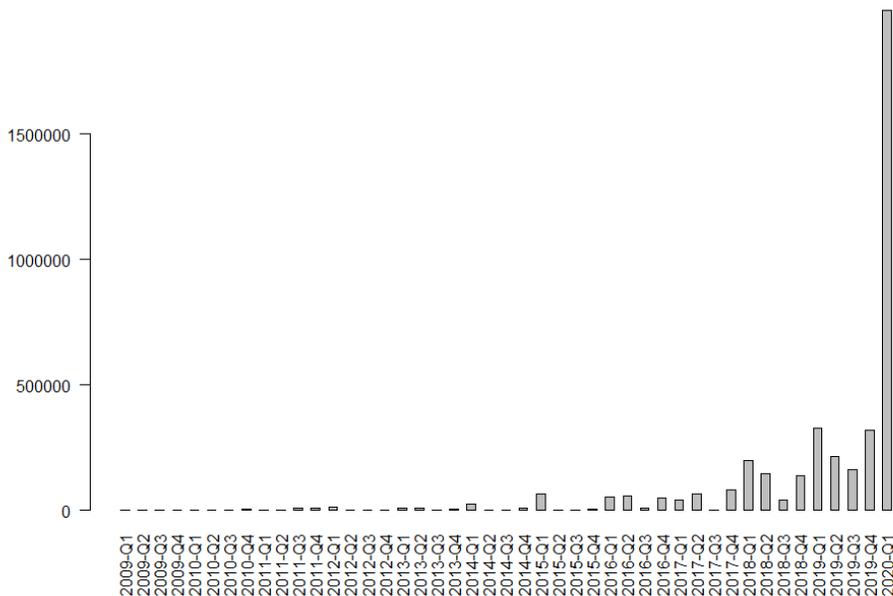

Figure 1: Histogram for date of satellite imagery used to extract footprint per structure by Ecopia.AI and Maxar Technologies, Inc., 2019; 2020.

3.2 Mapping of internally displaced persons

The mean and median number of IDPs per grid cell was: 344 and 224 for 'displacements sites' and 4 and 2 for 'host community' sites (Figure 2). Our method of incrementally increasing the buffer around georeferenced locations was able to capture a reasonable area of buildings that could house IDPs in urban and many rural areas. Figure 3 shows an example of the areas mapped as IDP locations across South Sudan's capital, Juba. In some sparse rural areas, particularly Jonglei in the east of the country, IDP coordinate locations were very far away from their nearest building footprints. Further work is needed to understand the overlap of point locations and building footprints in the context of individual IDP population, e.g. what type of community do the IDPs join? What is the primary housing type? Is the spatial extent in which IDPs live fluctuating significantly in short periods of time? With limited data available on the spatial extent of non-UN camps, it is outside the scope of this study to address these questions and thoroughly evaluate the mapped IDPs. Here we provide a starting point to facilitate future work on mapping IDPs from community level point locations, and recommend further evaluation in collaboration with NBS, IOM and other data collectors.

A total of 1,431 grid cells out of 1,054,006 settled grid cells were classified as 'displacement sites' across the country. These cells were considered to contain IDPs only and were not included in the mapping of the unadjusted census projections and where people have been displaced from.

6
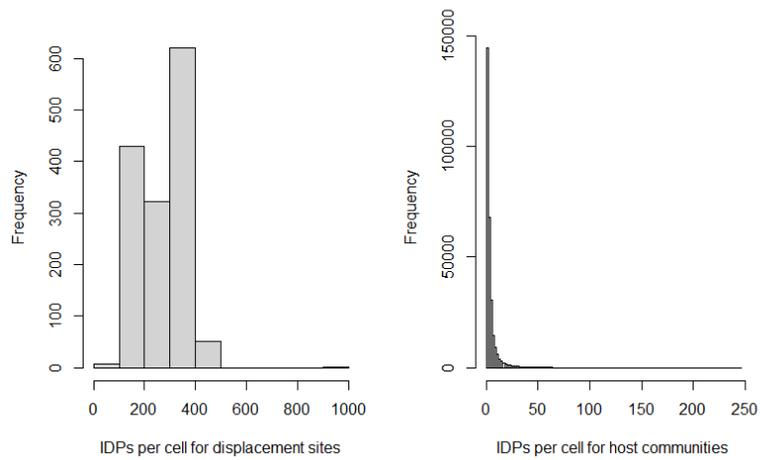
Figure 2: Histograms of IDPs per cell for 'displacement sites' (left) and 'host communities' (right).

IDP location type

- Displacement site
- Host community
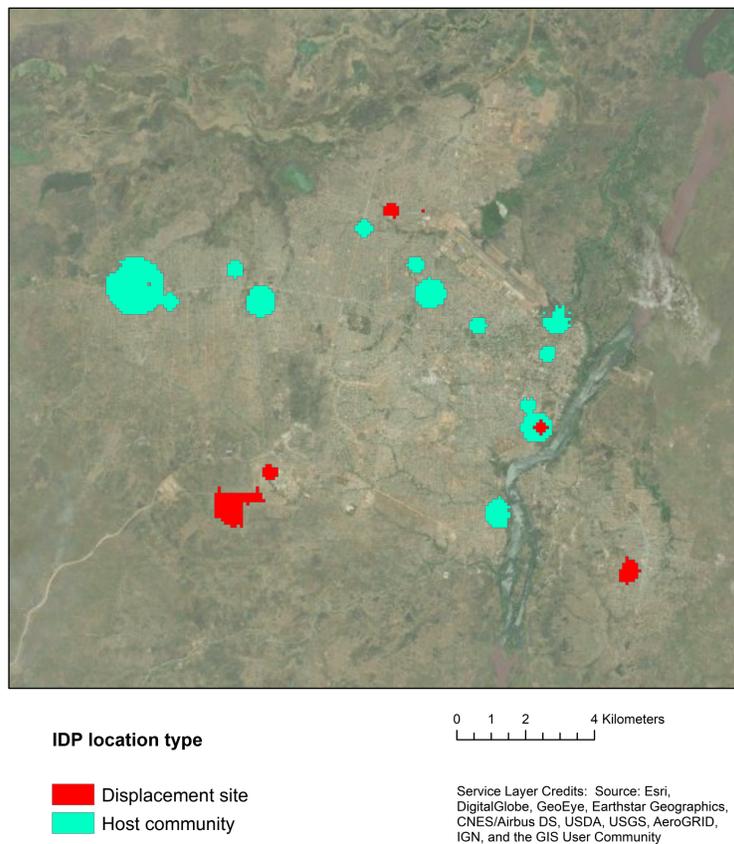
Figure 3: Mapped spatial extents of IDP populations across Juba derived using reported point locations and estimates IDP counts as of September 2020 (IOM, 2021) and building footprint (Ecopia.AI and Maxar Technologies, Inc., 2019; 2020).

3.3 Mapping of unadjusted population (population distribution in the absence of displacement)

The variance explained was 50.02% and the mean of squared residuals was 0.28, for the final random forest model used to predict the unadjusted census population counts at a high spatial resolution. Given the relatively low number of administrative unit (79 counties) to fit the model and the high uncertainties in projections from a census conducted 12 years ago, we consider these results to be reasonable. Figure 4 shows covariate importance for the final model and highlights that specific monthly precipitation means and building footprint metrics were among the best variables from the covariate set for estimating the spatial distribution of the unadjusted population.
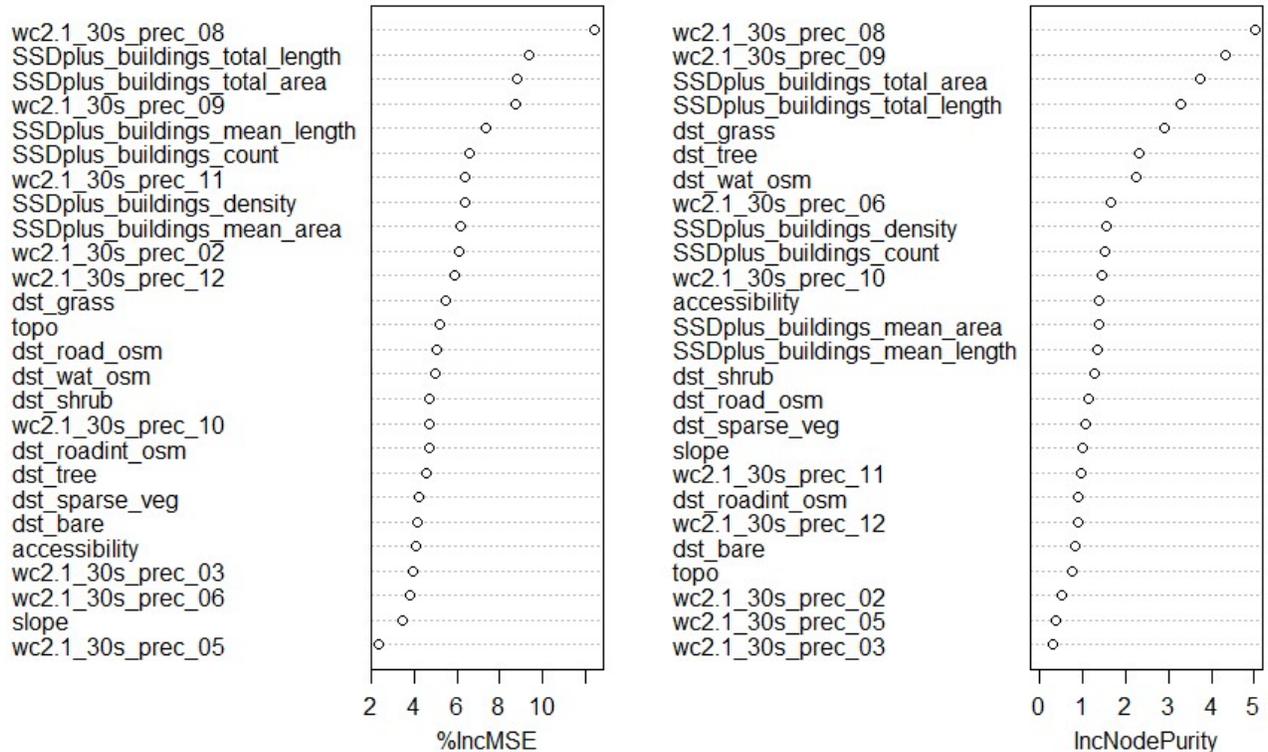

Figure 4: List of covariates used to predict the unadjusted census population counts, ordered by importance in the model. See Tables A1-A4 for descriptions of the covariates. The %IncMSE indicates the increase of the mean squared error and the IncNodePurity is a measure of the total increase in node purity, when the given variable is randomly permuted.

3.4 Mapping of where people have been displaced from

For the final random forest model estimating the origin of displacement, the variance explained was 46.34 % and the mean of squared residuals was 0.81. We found that the conflict covariates were the most important of the covariates set (Figure 5). These results show that there is huge potential for using geospatial layers derived from conflict data to map and understand where people have been displaced from. Again, the level of variance explained by the model is reasonable given that the model is based on only 79 administrative units and the high uncertainties in county level counts. The county level estimates were based on summaries of place of origin for whole IDP populations as well as very limited origin data of refugees. In order to improve the accuracy of mapping where people have been displaced from, finer scale origin data is needed.
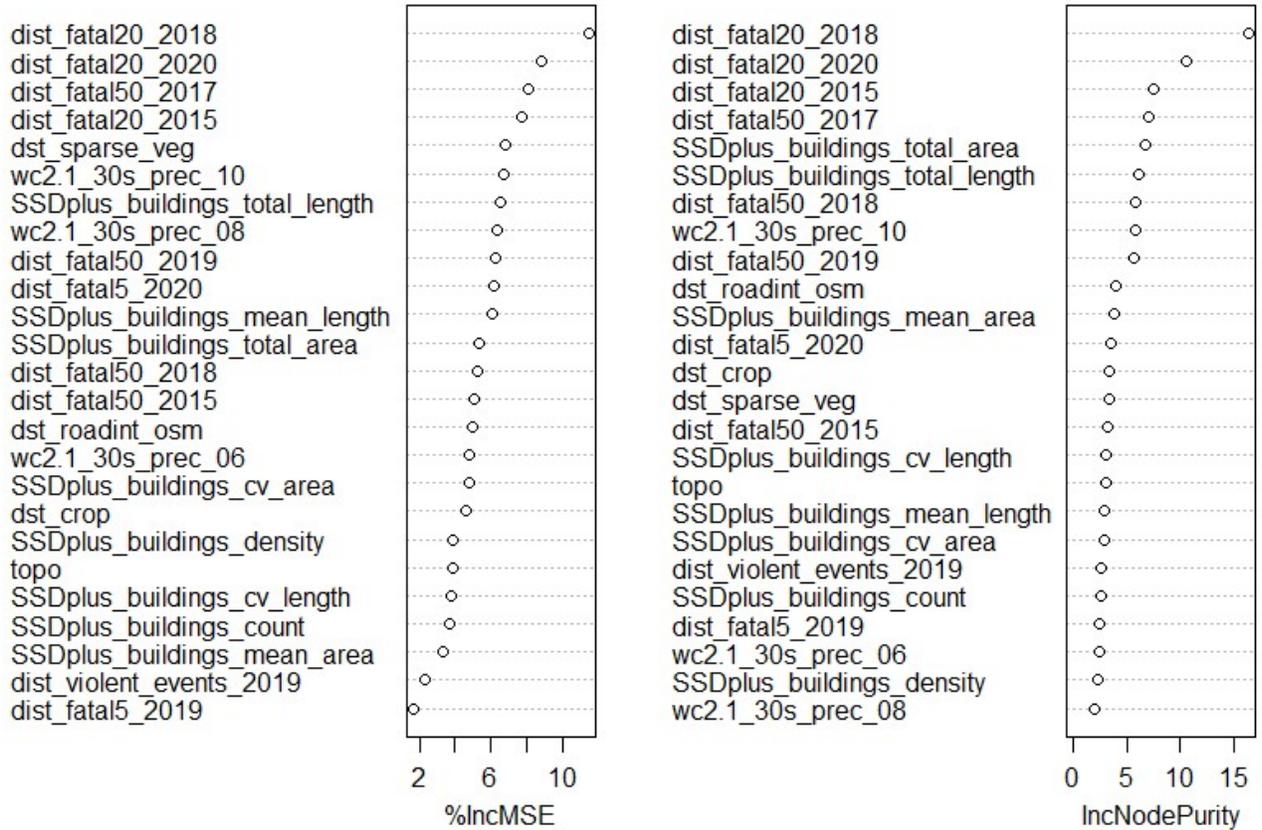
Figure 5: List of covariates used to predict the locations where people have been displaced from, ordered by importance in the model. See Tables A1-A5 for descriptions of the covariates. The %IncMSE indicates the increase of the mean squared error and the IncNodePurity is a measure of the total increase in node purity, when the given variable is randomly permuted.

In a small proportion of grid cells (12,638 out of 1,054,006) we found that the estimated number of people displaced from them was larger than their unadjusted population count. This was not unexpected given the high levels of uncertainty in both the census projections and the counts of people displaced from counties, and shows that the county level uncertainty propagated to the grid cell level. The affected grid cells were concentrated to a small number of counties in Upper Nile State and along the southern national border, and correspond to the areas of highest conflict (see top-right map of Figure A1 for variation in estimates of people displaced across counties with lighter colours indicating counties where large numbers have been displaced from). For these grid cells where the estimated number of people displaced from them exceeded their unadjusted population count, we replaced the value in the origin of displacement layer with the unadjusted population count, such that the final population estimates are set to zero.

3.5 Final population distribution accounting for displacement

For the final population data, the mean, median and maximum grid cell level population counts were 10.6, 7.1 and 921.0, respectively. In Figure 6 we present the final population distribution across Juba, and in Table 1. we show the final state level totals.

Table 1. 2020 state level population counts from census projections (unadjusted to account for displacement) and counts for the South Sudan V2.0 dataset that are adjusted to account for displacement. *This is the number of refugees we deducted from the census counts given the available county level population data and place of origin information. The official number of refugees as of October 2021 was 2,185,117 (UNHCR, 2020).

State	NBS population projection estimate	High resolution SSD v2.0 population estimates accounting for displacement
Upper Nile	1,550,846	1,230,934
Jonglei	2,064,566	1,656,252
Unity	990,898	838,610
Warrap	1,532,386	1,528,778
Northern Bahr el Ghazal	1,133,147	1,063,200
Western Bahr el Ghazal	526,652	447,759
Lakes	1,196,067	1,143,392
Western Equatoria	870,055	475,330
Central Equatoria	1,750,020	1,505,782
Eastern Equatoria	1,583,049	1,254,524
Total inside South Sudan	**13,206,622**	**11,144,563**
Outside country		2,062,059*
Total	**13,206,622**	**13,206,622**

population count adjusted for displacement

- 0 - 15
- 15 - 25
- 25 - 50
- 50 - 75
- 75 - 100
- 100 - 250
- 250 - 921
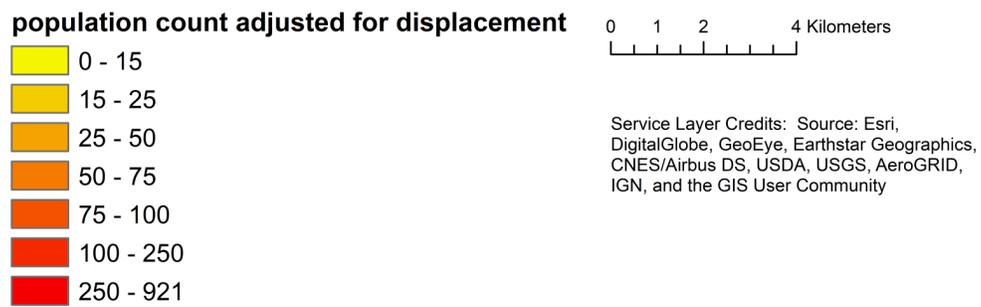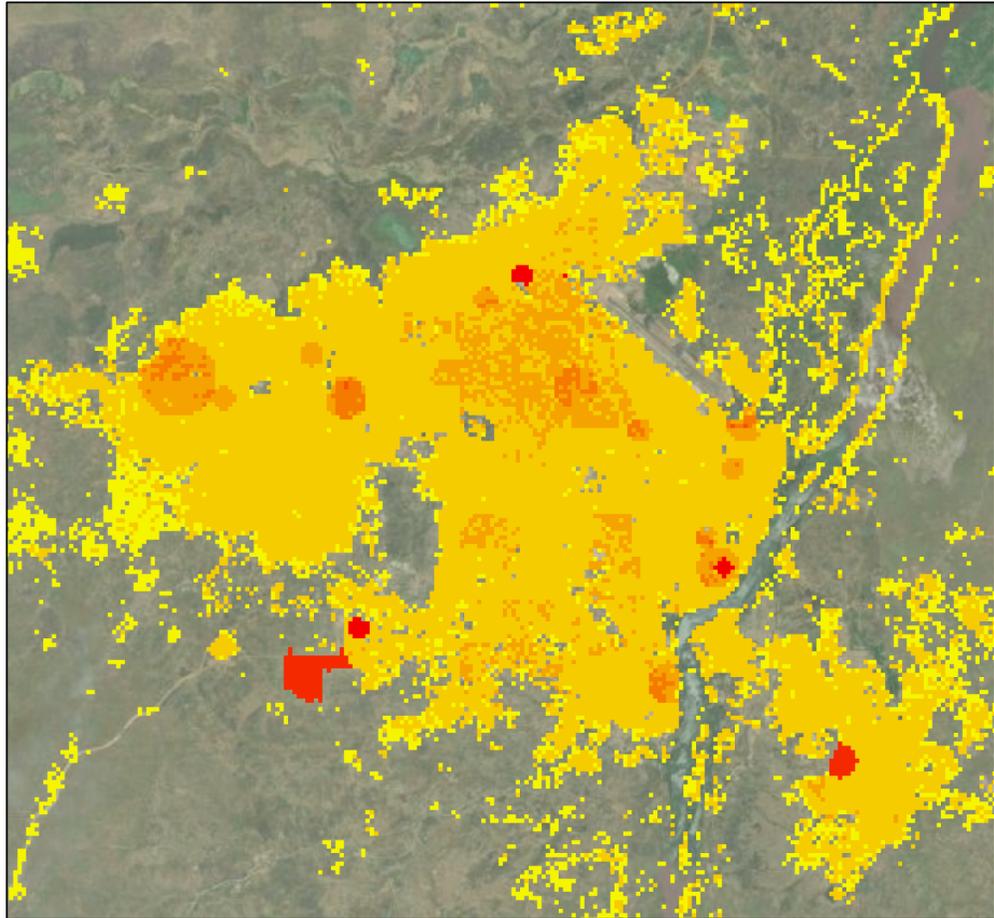
Figure 6: Final grid cell level population counts accounting for displacement across Juba. The highest population counts (red) can been seen in areas where 'displacement sites' are located and relatively high population counts (dark orange) in areas where IDPs are present in host communities.

4 Discussion

Here we present a unique approach for integrating different datasets to produce high spatial resolution population estimates that account for displacement in a high conflict and data limited setting. We estimate the approximate time point that the data represents as being September 2020, given the mix of input data. We recommend the following areas for future work:

- Validation of IDP population spatial extents - here we applied a set of rules to map IDP populations, but further information about the populations is needed to help understand their densities compared to that of host populations.
- Collating information on place of origin of displaced persons - there is very little coherent and accurate origin data. The UNHCR survey used in this study had a relatively small sample size and only reported data at the state level. Similar representative household surveys of displaced populations could provide valuable data on place of origin at a fine scale. This would open up options to carry out causal analyses as well as improve the county level estimates for the top-down approach we implemented.
- Development of informative covariates for improving the accuracy of top-down models - uncertainty in county level counts needs to be reduced before this step, however, with more accurate estimates further testing of potential covariates could improve the results. Specifically:
    - Other covariates relating to fatalities and event types that could be derived from the conflict location data
    - Alternatives to the flow accumulation covariate that may be indicative of flooding events, e.g. topographic wetness index. Although natural disasters were responsible for only $< 3\%$ of IDP displacement (IOM 2020; 2021)
    - Other climatic variables - here we only used monthly precipitation. We did test alternative models that included 3 month means and sums in place of the individual monthly precipitation covariates, but the latter produced better model fit. WorldClim provides other variables on temperature and additional metrics such as maximum and minimums.
- Consideration of seasonal dynamics including nomadic groups.

5 Appendix

Table A1: List of WorldPop Global covariates (Lloyd et al, 2019)

Covariate	Layer name
Accessibility	accessibility
Slope	slope
Topography	topo
VIIRS Nighttime Lights	viirs_nl
Distance to Bare Area	dst_bare
Distance to Cropland Natural Vegetation	dst_crop
Distance to Herbaceous Cover, Grassland, Mosses	dst_grass
Distance to Tree/ Herbaceous Cover, Flooded, Fresh/Saline/Brackish Water	dst_mix
Distance to Shrubland	dst_shrub
Distance to Sparse Vegetation	dst_sparse_veg
Distance to Tree Cover	dst_tree
Distance to Water Bodies, Permanent Snow and Ice	dst_wat
Distance to Roads in Open Street Map	dst_road_osm
Distance to Road Intersections in Open Street Map	dst_roadint_osm
Distance to Water in Open Street Map	dst_wat_osm

Table A2: List of WorldClim covariates (Fick and Hijmans, 2017)

Covariate	Layer name
Mean monthly precipitation from 1970-2020: January	wc2.1_30s_prec_01
Mean monthly precipitation from 1970-2020: February	wc2.1_30s_prec_02
Mean monthly precipitation from 1970-2020: March	wc2.1_30s_prec_03
Mean monthly precipitation from 1970-2020: April	wc2.1_30s_prec_04
Mean monthly precipitation from 1970-2020: May	wc2.1_30s_prec_05
Mean monthly precipitation from 1970-2020: June	wc2.1_30s_prec_06
Mean monthly precipitation from 1970-2020: July	wc2.1_30s_prec_07
Mean monthly precipitation from 1970-2020: August	wc2.1_30s_prec_08
Mean monthly precipitation from 1970-2020: September	wc2.1_30s_prec_09
Mean monthly precipitation from 1970-2020: October	wc2.1_30s_prec_10
Mean monthly precipitation from 1970-2020: November	wc2.1_30s_prec_11
Mean monthly precipitation from 1970-2020: December	wc2.1_30s_prec_12

Table A3: List of HydroSHEDS covariates (Lehner et al, 2008)

Covariate	Layer name
Flow accumulation	hydrosheds_flow_acc

Table A4: List of Building Footprint Metrics covariates (calculated
from Ecopia.AI and Maxar Technologies, Inc. 2019; 2020)

Covariate	Layer name
Building count	SSDplus_buildings_count
Coefficient of variation of building area	SSDplus_buildings_cv_area
Coefficient of variation of building perimeter	SSDplus_buildings_cv_length
Building density: count divided by grid cell area	SSDplus_buildings_density
Mean building area	SSDplus_buildings_mean_area
Mean building perimeter	SSDplus_buildings_mean_length
Total building area	SSDplus_buildings_total_area
Total building perimeter	SSDplus_buildings_total_length

Table A5: List of conflict covariates based on Armed Conflict Location Events Database (Raleigh et al, 2010); following method of Dooley et al, 2020

| Covariate	Layer name
Distance to events classed as 'Battles' and 'Violence against civilians' in 2014	dist_violent_events_2014
Distance to events classed as 'Battles' and 'Violence against civilians' in 2015	dist_violent_events_2015
Distance to events classed as 'Battles' and 'Violence against civilians' in 2016	dist_violent_events_2016
Distance to events classed as 'Battles' and 'Violence against civilians' in 2017	dist_violent_events_2017
Distance to events classed as 'Battles' and 'Violence against civilians' in 2018	dist_violent_events_2018
Distance to events classed as 'Battles' and 'Violence against civilians' in 2019	dist_violent_events_2019
Distance to events classed as 'Battles' and 'Violence against civilians' in 2020 (Jan-Sept)	dist_violent_events_2020
Distance to any events resulting in 5 or more fatalities in 2014	dist_fatal5_2014
Distance to any events resulting in 5 or more fatalities in 2015	dist_fatal5_2015
Distance to any events resulting in 5 or more fatalities in 2016	dist_fatal5_2016
Distance to any events resulting in 5 or more fatalities in 2017	dist_fatal5_2017
Distance to any events resulting in 5 or more fatalities in 2018	dist_fatal5_2018
Distance to any events resulting in 5 or more fatalities in 2019	dist_fatal5_2019
Distance to any events resulting in 5 or more fatalities in 2020 (Jan-Sept)	dist_fatal5_2020
Distance to any events resulting in 20 or more fatalities in 2014	dist_fatal20_2014
Distance to any events resulting in 20 or more fatalities in 2015	dist_fatal20_2015
Distance to any events resulting in 20 or more fatalities in 2016	dist_fatal20_2016
Distance to any events resulting in 20 or more fatalities in 2017	dist_fatal20_2017
Distance to any events resulting in 20 or more fatalities in 2018	dist_fatal20_2018
Distance to any events resulting in 20 or more fatalities in 2019	dist_fatal20_2019
Distance to any events resulting in 20 or more fatalities in 2020	dist_fatal20_2020
Distance to any events resulting in 50 or more fatalities in 2014	dist_fatal50_2014
Distance to any events resulting in 50 or more fatalities in 2015	dist_fatal50_2015
Distance to any events resulting in 50 or more fatalities in 2016	dist_fatal50_2016
Distance to any events resulting in 50 or more fatalities in 2017	dist_fatal50_2017
Distance to any events resulting in 50 or more fatalities in 2018	dist_fatal50_2018
Distance to any events resulting in 50 or more fatalities in 2019	dist_fatal50_2019
Distance to any events resulting in 50 or more fatalities in 2020 (Jan-Sept)	dist_fatal50_2020
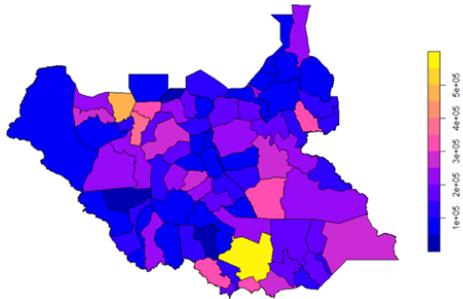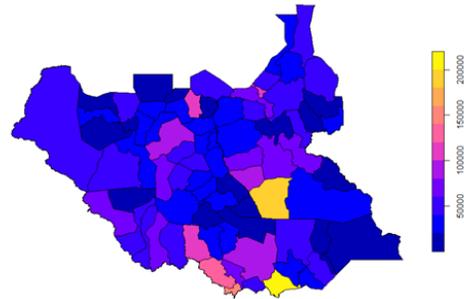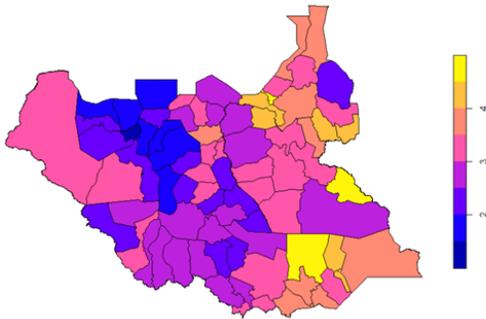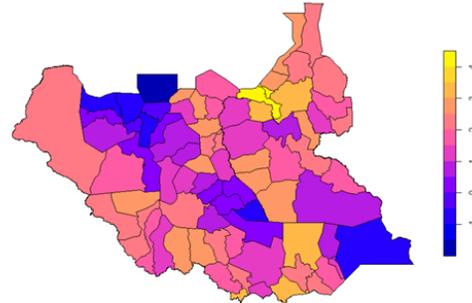
Figure A1: Distribution of county level counts in absolute numbers (top plots) and log densities (bottom plots). In this study we used the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) operational administrative level 2 boundaries (counties) dataset (OCHA, 2020), as shown here.

6 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, 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

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. 2017. "Kajo Keji, Central Equatoria: Paper Registration | Rapid Intentions & Multi-Sectorial Needs Survey | 29 June-12 July 2017".

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

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

Minnesota Population Center. 2020. Integrated Public Use Microdata Series, International: Version 7.3 [South Sudan 2008 Census]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D020.V7.3

National Bureau of Statistics. 2015. "Population Projections, South Sudan. From 2015 - 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.

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.

7 Contributing

The production of this report and its corresponding dataset were led by Claire A. Dooley from the WorldPop Research Group at the University of Southampton. Chris Jochem, Alessandro Sorichetta and Andy Tatem con-

8   Suggested Citation

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

9   License

This report may be redistributed following the terms of a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) License.

10   Code

The code used to generate the dataset (Dooley et al 2021) is available here: https://github.com/cadooley/SSD_pop_v2.0