A bottom-up population modelling approach to complement the population and housing census

Edith Darin[1], Gianluca Boo[1], and Andrew J Tatem[1]

[1]WorldPop, Department of Geography and Environment, University of Southampton, Southampton, UK

Introduction

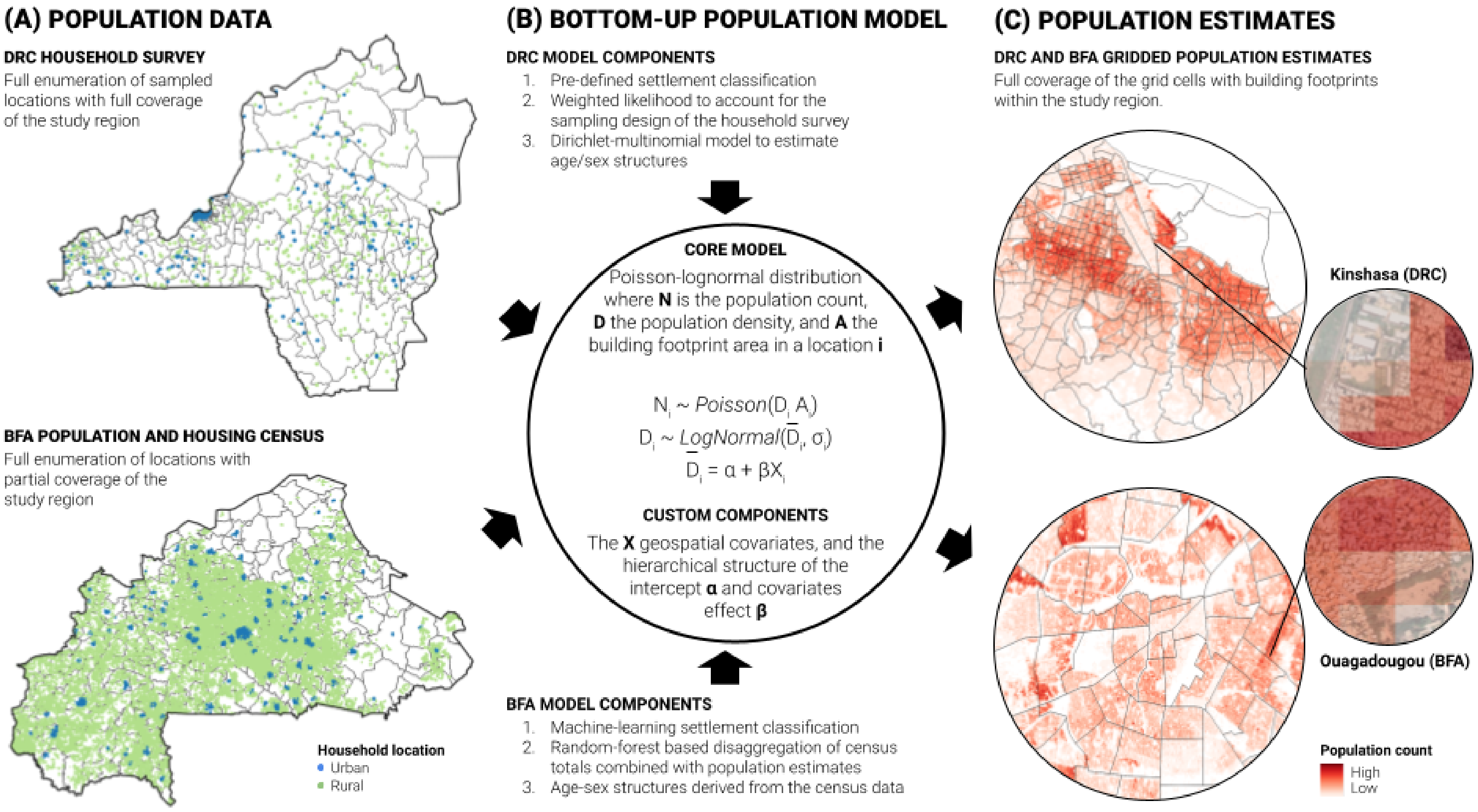
The population and housing census provides essential information for local, national and international decision-making and intervention. Granular data on population counts and age/sex structures can support the planning of critical infrastructures, such as schools, health facilities and transportation networks. The census exercise involves the complete and simultaneous enumeration of the entire population of a country. This effort requires massive logistical and financial resources, which are generally mobilised on a decennial basis (Juran et al. 2020). However, logistical challenges may become insurmountable in countries where political instability, conflicts and natural disasters impede the work of the enumerators. As a consequence, the most vulnerable countries often have either outdated or partial census data.

Outdated and partial census data can be complemented using different population estimation techniques. In particular, the bottom-up modelling approach leverages population data collected in a set of fully enumerated locations and ancillary geospatial covariates derived, for example, from building footprints or road network data to estimate population counts and age/sex structures throughout a region of interest (Wardrop et al. 2018). As geospatial covariates are available at increasing spatial resolutions, in the order of meters to tens of meters, bottom-up models can estimate population within grid cells of approximately 100m — a finer level of detail than standard census enumeration zones. This level of detail offers the advantage of flexible aggregation of the population estimates within different administrative and functional units, for instance, school catchment areas and health zones.

We present the application of a bottom-up population modelling approach in the Democratic Republic of the Congo (DRC) — the last census was conducted in 1984 — and Burkina Faso (BFA) — the last census was completed in 2020 but only covered 70% of the country. In the DRC, we accessed population data collected in two rounds of household survey carried out in the western part of the country in 2018 and 2019. In BFA, we accessed the census data collected in 2020. We fitted the population data in a Bayesian hierarchical modelling framework and produced gridded population estimates with complete coverage of the two study regions (Leasure et al. 2020). The models showed a good fit for out-of-sample predictions and the resulting bottom-up population estimates are currently used for census support and humanitarian intervention in both countries.

Population data

The bottom-up population modelling approach builds on population counts and age/sex structures collected within a set of representative locations. This input data can be retrieved from different data sources. When the census is outdated, recent household surveys allow us to capture the most recent information on demographic characteristics and geographic distribution of the population. In the DRC, we accessed household survey data involving the full enumeration of approximately 1,000 clusters located in five provinces. When a complete census is not feasible, demographic information can be retrieved from the areas covered with the census. In BFA, we accessed demographic information from approximately 23,000 census enumeration zones. Different population data sources can be flexibly integrated if they include accurate spatial attributes, such as the GPS location of the households or the boundaries of the surveyed locations.



(A) POPULATION DATA

DRC HOUSEHOLD SURVEY

Full enumeration of sampled locations with full coverage of the study region

BFA POPULATION AND HOUSING CENSUS

Full enumeration of locations with partial coverage of the study region

(B) BOTTOM-UP POPULATION MODEL

DRC MODEL COMPONENTS

1. Pre-defined settlement classification
2. Weighted likelihood to account for the sampling design of the household survey
3. Dirichlet-multinomial model to estimate age/sex structures

CORE MODEL

Poisson-lognormal distribution where **N** is the population count, **D** the population density, and **A** the building footprint area in a location **i**

$$N_i \sim \mathit{Poisson}(D_i A_i)$$

$$D_i \sim \mathit{LogNormal}(\bar{D}_i, \sigma_i)$$

$$\bar{D}_i = \alpha + \beta X_i$$

CUSTOM COMPONENTS

The **X** geospatial covariates, and the hierarchical structure of the intercept **α** and covariates effect **β**

BFA MODEL COMPONENTS

1. Machine-learning settlement classification
2. Random-forest based disaggregation of census totals combined with population estimates
3. Age-sex structures derived from the census data

(C) POPULATION ESTIMATES

DRC AND BFA GRIDDED POPULATION ESTIMATES

Full coverage of the grid cells with building footprints within the study region.

Figure 1. The bottom-up population modelling approach developed in the DRC and BFA — **(A)** the input population data, **(B)** the components of the population models and **(C)** the resulting gridded population estimates in the capital cities of Kinshasa (DRC) and Ouagadougou (BFA).

Bottom-up population modelling

We developed a bottom-up modelling approach based on Bayesian hierarchical models to estimate population counts and age/sex structures within grid cells of approximately 100m (Leasure et al. 2020). We modelled population counts using a Poisson-lognormal distribution, including hierarchical slope and variance defined based on settlement type (e.g., urban and rural settlements) and administrative regions (e.g., municipalities). The hierarchical structure was developed specifically for each model to account for local context, data quality and availability. We included bespoke geospatial covariates with full coverage of the region of interest to estimate population counts outside of the surveyed locations. In the DRC model, we also incorporated a weighted-likelihood approach to tackle bias introduced by the sampling design adopted in the household surveys and estimated age/sex structures using a Dirichlet-multinomial distribution process (Gelman et al. 2013). In the BFA model, we combined the population estimates with a top-down disaggregation of the population counts retrieved from the census data to create gridded population estimates for the entire country. Age/sex structures were provided by the national census office.

Population estimates

In both population models, we accessed building footprints extracted from recent satellite imagery and derived morphological and topological attributes both within surveyed locations and grid cells across the study regions. These attributes enabled us to inform bespoke settlement classification techniques developed using a deterministic approach in the DRC model and a machine learning classifier in the BFA model. We also used building footprints attributes as model covariates, for instance, the average building-footprint area, the average distance between building footprints and the average count of building footprints within different focal windows. Lastly, building footprints were also used to constrain the population estimation to grid cells with at least one building centroid. The most relevant attributes derived from building footprints were made openly available throughout sub-Saharan Africa (Dooley et al. 2020).

While in the DRC model, we only used covariates derived from building footprints (i.e., average building-footprint area, average building-footprint proximity and average building footprint count within a 2km focal window), in the BFA model we also included additional covariates linked to transportation (i.e., distance to secondary roads and friction surface) and water (i.e., distance to water streams) networks. Both models passed standard Bayesian model checks (e.g., convergence and mixing of the MCMC chains and analysis of residuals) and confirmed good predictive performance of population counts, as suggested by an R^2 of 0.79 for the DRC model and 0.63 for the BFA model for out-of-sample model predictions. The resulting bottom-up population estimates are openly available on a dedicated platform (WorldPop 2021) and are currently used to complement the census data in both countries (Boo et al. 2020; WorldPop and INSD 2020).

Discussion and conclusions

We presented the application of a bottom-up population modelling approach in the DRC and BFA. Similar work has been carried out in Nigeria and Zambia and is in progress in other countries of sub-Saharan Africa as part of the Geo-Referenced Infrastructure and Demographic Data for Development (GRID3) programme (GRID3 2021). The role of bottom-up population estimates for census support was recently highlighted by the United Nations Population Fund (Juran et al. 2020). However, the development of bottom-up population models relies upon the availability of a representative set of recent georeferenced population data. In countries with outdated censuses, this data may become available through conventional household surveys and pre-listings or bespoke microcensus surveys.

The estimation of population counts and age/sex structures within grid cells of approximately 100m offers additional advantages compared to conventional census summaries provided by national statistics offices. The regular format and high spatial resolution of gridded population estimates allow for the flexible aggregation within larger functional or administrative units. In addition, the gridded population estimates can be combined with other spatial datasets, such as flood risk areas or infrastructure locations, to better inform decision-making and intervention at the subnational level. Some of these use cases are reported on the website of the GRID3 programme, for instance, the support to routine vaccination campaigns or the definition of census pre-enumeration areas (GRID3 2021).

The population modelling work in the DRC and BFA underscores the flexibility of the bottom-up modelling approach. The input population data can be retrieved for different data sources, such as household surveys or partial censuses, provided that they include accurate spatial attributes. The Bayesian hierarchical modelling can be flexibly customised to the input population data and country context through the definition of the hierarchical set-up and model covariates, but also by including bespoke sub-models, such as the weighted-likelihood component. Bayesian models offer the additional advantage of accounting for uncertainty in the input data and the model parameters. For instance, current developments in bottom-up population modelling are attempting to include measurement error components to account for uncertainty in the building footprints data. Finally, the gridded population estimates can be aggregated within custom spatial units to support specific applications.

References

Boo G, Darin E, Leasure DR, Dooley CA, Chamberlain HR, Lazar AN, Tatem AJ. 2020. Modelled gridded population estimates for the Kinshasa, Kongo-Central, Kwango, Kwilu, and Mai-Ndombe provinces in the Democratic Republic of the Congo (2018), version 2.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00669.

Dooley CA., Boo G, Leasure DR, Tatem AJ. 2020. Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. WorldPop, University of Southampton. Source of building footprints: Ecopia Vector Maps Powered by Maxar Satellite Imagery ©. doi:10.5258/SOTON/WP00677

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 2013. CRC press.

GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) [Internet]. 2021. Available from: https://grid3.org.

Juran S, Kupie M, Jones M, Chamberlain HR, Tatem AJ. 2020 May. The Value of Modeled Population Estimates for Census Planning and Preparation [Internet]. (UNFPA Technical Guidance Note). Available from: https://www.unfpa.org/sites/default/files/resource-pdf/Technical_Guidance_Note-_Value_of_Modeled_Pop_Estimates_in_Census.pdf.

Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ. National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. Proceedings of the National Academy of Sciences. 2020 Sep 29;117(39):24173–9. doi:10/10.1073/pnas.1913050117.

Wardrop NA, Jochem WC, Bird TJ, Chamberlain HR, Clarke D, Kerr D, Bengtsson L, Juran S, Seaman V, Tatem AJ. Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences. 2018 Apr 3;115(14):3529–37. doi:10.1073/pnas.1715305115.

WorldPop, Institut National de la Statistique et de la Démographie (INSD) du Burkina Faso. 2020. Census-based gridded population estimates for Burkina Faso (2019), version 1.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00687.

WorldPop. 2021. WorldPop Open Population Repository [Internet]. Available from https://wopr.worldpop.org.