Technical Guidance Note

Date: August 2020 (updated version 2)
Keywords: Population and Housing Census, Modelled Population Estimates, Census Under Coverage

The Value of Modelled Population Estimates for Census Planning and Preparation

A population and housing census is among the most complex and massive peacetime exercises a nation undertakes. In most low-and-middle-income countries, the population and housing census is the primary source of data on the number of people, their spatial distribution, age and sex structure, living conditions and other key socioeconomic characteristics. It further provides a comprehensive picture of housing and its conditions in the country. Census data are critical for good governance, development, risk reduction and crisis response, social welfare programmes and business market analyses. However, a census remains a logistically expensive and challenging exercise and is ideally conducted once a decade.

National Statistical Offices (NSOs) often face obstacles that can impact the successful implementation, data quality and overall completion of a national population census. Innovative statistical approaches are being developed and evaluated to support, and potentially improve, census planning and implementation. Modelled population estimates, with their small area estimates of population density, hold the potential to support census planning and improve quality.

Statistical modelling approaches to estimating populations are developed based on the correlation between population density and geospatial covariate data layers. These correlative relationships can be leveraged into a statistical model to predict population density, with certain levels of uncertainty, depending on the initial input data. These models offer the advantage of relatively low-cost and may provide alternative ways of deriving more recent population estimates to inform census planning and implementation. Yet, by virtue of being probabilistic, model-based estimates should not be treated in the same way as true population counts, but more as an opportunity to refine planning when the most recent census is largely outdated. Concretely, estimates rely on existing or specifically implemented sample household surveys, such as those produced in undertaking household enumerations for surveys. These samples are then combined with geospatial covariate data layers to form the basis of the statistical model. Once the model has been produced, other independent data sources, such as household surveys, can be used to verify newly modelled population estimates.

Population is modelled as a function of satellite imagery-based or geospatially derived covariate layers, such as distance to roads, night-time lights intensity or land cover classification. Often, the best predictors of population density are the presence, densities and patterns of built structures. Assuming that populations are mostly located in settled areas, we can estimate the population of a country by combining the population densities with maps of settled areas. This requires that settlement mapping is undertaken, most often using recent, very high-resolution satellite imagery. The model can then predict population densities or people per unit area for any areal units, such as administrative areas or census enumeration areas, or even grids of 100 meters by 100 meters.

Regarding reliability, modelled estimates contain uncertainty that can result in order-of-magnitude confidence intervals around the predicted population density. To date, there is no mechanism for deciding what level of uncertainty is acceptable, or whether it should be included in decision-making. For general indicators based on survey data, the recommendation is to consider that an indicator is not robust when the relative standard deviation (standard deviation over the mean) is less than 33 percent. Yet, even if population estimates are subject to imprecisions, they may serve as:

- input for census planning and census cartography
- substitute to enumeration in inaccessible areas
- input data to update the master sampling frame
- evaluating census coverage
- anonymizing census results
- combining population and housing censuses with agricultural censuses

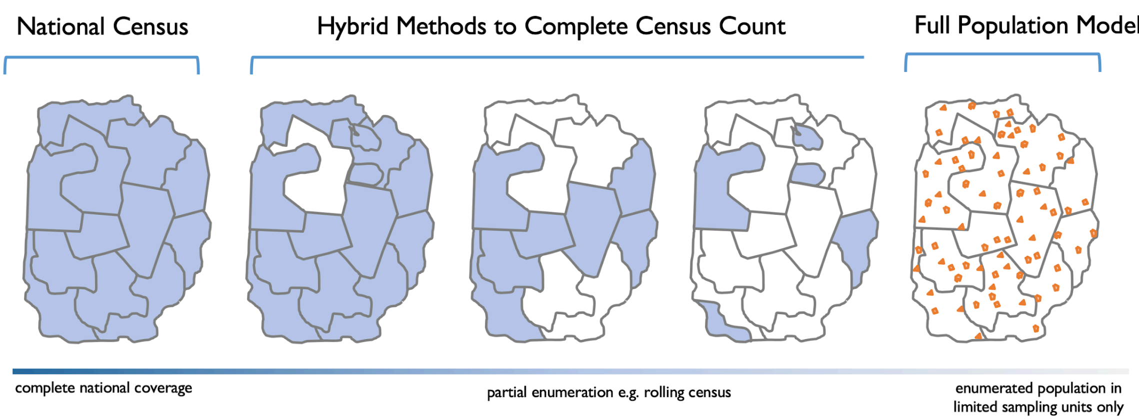
Input for Census Planning and Census Cartography

The planning process for census and cartographic fieldwork fundamentally relies upon population projections (or forecasts) by region, district municipality, village and Enumeration Area (EA). These population counts inform resource allocation for the implementation of the census, including funds, allocated time for individual census stages, personnel and operations. However, official population projections produced right after the census may become increasingly outdated over time, if not continuously updated with more recent intercensal survey data. Further, while population projections can be accurate at the national and regional levels, they may be inaccurate at lower levels, especially the EA level. As modelled population estimates are based on the most recent geospatial covariates available, they often provide a more accurate estimate at district, village or EA levels.

Modelled population estimates, in combination with road/river network geospatial data can be used to plan the most efficient routes of cartographers. More efficient fieldwork can reduce overall costs and the amount of time required to complete cartography. Due to the immense financial cost of a census, it is critical that potential opportunities for savings are considered. These cost savings can be derived from the amount of time that is needed for a cartographer or enumerator to complete their fieldwork by more efficient routing, more accurate planning in how many cartographers are needed for each EA, etc. Further, satellite imagery-based settlement mapping improves the planning and updating of cartography as it can identify small settlements that were previously not shown on maps. Also, modelled population estimates together with data on river and road networks can be used to derive new enumeration areas.

Substitute to Enumeration in Inaccessible Areas

While conducting a census, some countries face security or access challenges in certain parts of their territory, with no possibility to send enumerators there. In that case, it is possible to model population numbers and densities of inaccessible areas, using either the census data that was collected or specially designed survey data (Figure 1). The accuracy of the estimates will depend upon the quality of the enumeration in accessible areas, the intrinsic performance of the model and the likelihood of the modelling assumptions. The developed population model is always tailored to the situation at hand (i.e. data availability and population context) with the aim of producing the best possible population estimates given the input data and time that are available.



Common Operational Datasets on Population Statistics

Common Operational Datasets (CODs) are authoritative reference datasets needed to support operations and decision-making for all actors in humanitarian response. COD-PS (Common Operational Datasets on Population Statistics) are updated sex-and age disaggregated population dataset at the lowest geographical level available. They are to be used in humanitarian settings only to respond to data requirements for crisis response and intervention. As such, their guiding principle is the more flexible and pragmatic criteria of 'best-available' instead of official statistical standards. In that case, it is possible to estimate COD-PS and densities of inaccessible areas using the census data that was partially collected to build a model. As such, where modelled population data are the best-available source for updated sex, age and geographically disaggregated data, they may be used as COD-PS to support operations and decision-making in humanitarian settings.

Master Sampling Frames and Population Estimates

Once census data are validated, a master sample of EAs is generally drawn from the census database. This sample is used during the intercensal period to conduct the major socioeconomic and demographic surveys of the country. The parameters of the master sample tend to deviate from that of the population over time, due to population growth and changes in its spatial distribution. Depending on the type of census cartography from which the initial master sample was drawn, gridded population estimates can be used to update the master sampling frame in two ways. If the census cartography was digital, one can update the master sampling frame by updating populations of the Eas to gridded estimates. If the cartography was not digital, updating populations of the Eas to gridded estimates cannot be applied since Eas boundaries would not be in numerical format and it would not be possible to overlay the gridded populations layers to the Eas layers. In this situation, we would advise drawing a new master sample from the gridded population, considering each grid or group of contiguous groups of grids as primary sampling units. This second strategy is an improvement of the well-known area sampling, since in this case, we have a more accurate estimate of the population of the “area” (which in this case is a grid or a group of contiguous grids). As for other use, this is only possible if the modelled estimates are robust enough to reflect changes in population parameters.

Assessment of Census Coverage

Generally, after the enumeration phase, census coverage is usually assessed through different methods and data sources (demographic projections, civil registration data, health data, education data, etc.), which might include the implementation of a post-enumeration survey.
With geospatial population estimates, we have at our disposal an additional source of comparison. For geospatial population estimates to serve as a source for census coverage assessment, the estimation should be conducted using data from enumeration areas with perfect coverage and good data quality. Once again, the geospatial estimates should be considered only if the model performance is high and the underlying assumptions realistic.

Anonymization of Census Results

Statistical methods can be used to anonymize the census dataset (Figure 1, left panel). Individual observations can be aggregated to gridded high-resolution population totals using GIS software, and statistical methods. With such methods, the observations (e.g. administrative totals) can be disaggregated to 100x100m resolution and thus create an anonymous data layer usable by the government for decision making and survey planning and the general public, if publicly released.

Combining with Agricultural Censuses

Combining the population and housing census with the agricultural census within the national statistical system is consistent with the Global Strategy to Improve Agricultural and Rural Statistics[1]. At the operational level this combination can be achieved by 1) establishing an integrated survey framework, 2) developing a master sample frame for agriculture, and 3) integrating the data management systems. Experience shows that collecting limited and well-defined agricultural data during the population and housing census can substantially contribute to building an efficient master frame for agricultural censuses and surveys in many developing countries. Further, modelled population estimates provide an opportunity to refine the sampling frame for agricultural censuses as population densities reflect soil occupation, which provides information on potential agricultural areas with low population density and potential non-agricultural areas with high population density. This information could be used to then further refine agricultural sampling frame, through:

- combining modelled estimates with geographic information from the agricultural census and
- classifying urban or populated areas that are to be excluded from the construction of the sampling frame for agricultural surveys, as this would support avoiding the use of resources for field operations in areas without agricultural use.

Conclusion

Modelled population estimates are potentially useful at different steps of a census, from planning to implementation, from quality assessment to data usage. However, it is essential to note that these estimates should be used with caution: different models will produce different population estimates. Fortunately, mechanisms exist to objectively 'choose' the most statistically accurate model and validate it. Despite the ability to choose the intrinsically best model, it is imperative to recall that a geospatial population model cannot correctly reflect the real world. As Box said, "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

This technical brief (Version2) has been prepared by Sabrina Juran, Mathias Kuepie, Maureen Jones, Catalina Valle, Kathrin Weny, Mathias Nathan from the United Nations Population Fund in collaboration with Andy Tatem, Attila Lazar, Douglas Leasure and Heather Chamberlain from WorldPop at the University of Southampton.

[1] Food and Agriculture Organization of the United Nations. Global strategy to improve agricultural and rural statistics. 2010 http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/ICAS5/Ag_Statistics_Strategy_Final.pdf