Technical Guidance Note

Date: May 2020
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. By definition, a population and housing census is an enumeration of the total population of a country and provides data on number of people, their spatial distribution, age and sex structure, living conditions and other key socioeconomic characteristics. These data are critical for good governance, development, risk reduction and crisis response, social welfare programmes and business market analyses. In most developing countries, censuses are the primary source of data on the size and spatial distribution of a population and its key characteristics, and this is likely to remain the case for the foreseeable future. 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. Modeled population estimates, with their predictive surface of population density, hold the potential to support census planning and improve quality.

Model-based 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way as true population counts, but more as an opportunity to refine planning when the most recent census is largely outdated.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 population surveys, such as those derived in the household enumeration stage of a 2-stage cluster sample. 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 modeled population predictions.

Population counts are then converted to population densities, 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, one of the best predictors of population density is the presence or absence of built structures. Assuming that populations are mostly in settled areas, we can estimate the population of a country by combining the densities within settled areas. This requires that settlement mapping is undertaken, most often using recent, very high-resolution satellite imagery. The model can then predict population density because it is based on the entire landscape of a country. These predictions are typically derived in terms of population density, or people per unit area, such as grids of 100 meters by 100 meters, which can be applied to any aerial units, such as administrative areas or census enumeration area delineations.

Regarding reliability, modelled estimates contain uncertainty that can result in order-of-magnitude differences in 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 more 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
- and potentially for evaluating census coverage, under certain circumstances.

Modelled Population Estimates as an Input for Census Planning and Census Cartography

The planning process for census and cartographic fieldwork is fundamentally reliant upon population counts of regions, district municipalities, villages and Enumeration Areas (EA). These population counts inform resource allocation for the implementation of the census, including funds, allocated time for individual census stages, laboratory and field personnel and operations. Generally, prior to any census pre-test, planning is based upon projections from the previous census of the population of different geographic units. In ideal circumstances, these data are at most eight years old. 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 most recent geospatial covariates available, they often provide a more accurate estimate at district, village or EA levels. For instance, in the case of Nigeria, projections are currently based on the 2006 Census, while the recent modelled population estimates produced by GRID3 are based on 2018 survey data and recent geospatial covariates. Therefore, these models could be considered instead of classical demographic projections when starting the planning process of a census.

Modelled population estimates, in combination with road/river network geospatial data can be used to plan most efficient routes of cartographers. More efficient fieldwork can reduce overall costs and 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 field work 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.

Population Estimates as Substitute to Enumeration in Inaccessible Areas

While conducting census, some countries are facing 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 the census data that was collected. The accuracy of the estimates will depend upon the quality of the enumeration in the accessible parts, the intrinsic performance of the model and the likelihood of the modelling assumptions.

Updating Master sampling Frame 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. Obviously, with change in the population over time, the parameters of the master sample deviate from that of the population. A further advantage of gridded population estimates is that they can provide an update of the master sampling frame. As for other use, this is only possible if the model is robust enough to reflect changes in population parameters.

Census Coverage Assessment and Population Estimates

Generally, after the enumeration phase, census data are compared with other reliable sources (demographic projections, civil registration data, health data, education data, etc.) and unexplained gaps serve as signals of coverage problems. With geospatial population estimates, we have to 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.

Conclusion

Modelled population estimates are potentially useful at different steps of a census, from planning to implementation. However, it is important 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. 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 (Version1) has been prepared by Sabrina Juran, Mathias Kuepie and Maureen Jones, Population and Development Branch, Technical Division, United Nations Population Fund in collaboration with Andy Tatem and Heather Chamberlain, WorldPop Project, University of Southampton