Population-Environment Research Network (PERN) Cyberseminars
https://www.populationenvironmentresearch.org/cyberseminars

Perspective on Applications of Gridded Demographic Datasets in Poor Data Settings

Dr Alessandro Sorichetta,
School of Geography and Environmental Science, University of Southampton
Email: A.Sorichetta@soton.ac.uk

PERN Cyberseminar on Application of Gridded Population and Settlement Products in Geospatial Population-Environment Research
14 – 18 October 2019

Context

Both census-based and estimated population and demographic data are typically made available either in tabular form (e.g., UNDESA, 2109) or associated to irregularly-shaped spatial units representing administrative level boundaries ranging from states/provinces to enumeration areas (US Census Bureau, 2019).

While such data are still useful for reporting and statistical purposes, having accurate population and demographic data at the finest possible spatial level are crucial for other applications such as: assessing the number of people potentially affected by natural and man-made disasters, organizing relief operations, delivering health and educational services, organizing elections, planning vaccination campaigns and delivering bed nets, estimating infectious disease risks, burdens, and dynamics, as well as measuring and monitoring Sustainable Development Goals. Indeed, aggregated population and demographic data, particularly in low-density and relatively larger administrative units, do not accurately represent the true spatial distribution of both the total population and specific demographic groups of interest. Furthermore, aggregated data referring to political or administrative units often presents analytical challenges when used to investigate the reciprocal relationship between population distribution and environmental factors, such as climate and land use change, deforestation, urbanization, and pollution, which may not be representative at the same “arbitrary” spatial level at which the population and demographic data are aggregated.

Nevertheless, for the majority of low- and middle-income countries, either characterized by rapid development and urbanization or most severely and disproportionately affected by both natural disaster and infectious disease morbidity, accurate population and demographic data at the finest spatial level are often either difficult to obtain or simply unavailable.

For these reasons, since the early-1990s, there has been an increasing effort to produce consistent and comparable spatially-explicit population and demographic datasets by using a range of approaches, assumptions, and input data to disaggregate administrative unit-based figures to a regular grid of fixed spatial resolution (Leyk et al., 2019). Such gridded datasets, also enabling flexible integration with other types of geospatial data (including Remote Sensing-based human settlement datasets and natural hazard footprints), are nowadays largely and increasingly used for analysis and modelling in a growing number of fields. Some examples of how they are used to support applications in data poor settings are presented in the next section, along with some important considerations about their fitness for use.

"Real-World" Applications

As highlighted in Balk et al. (2006), "the fewer the assumptions and inputs that are used in the construction of gridded population datasets, the fewer the restrictions that have to be imposed on the appropriateness of use in a wide variety of applications". However, as mentioned by Leyk et al. (2019) in their background paper for this cyberseminar, it is extremely important to consider the "Fitness for use" of a given gridded population dataset for an intended purpose. Thus, especially for specific analyses in which the accuracy of population distribution is critical, the use of highly modeled gridded population and demographic datasets, obtained using ancillary dataset, should be preferred.

For example, in the context of spatial modeling of infectious and non-communicable disease distributions and dynamics, as much accurate as possible population distribution data are required for correctly enumerating disease burdens and populations at risk. Furthermore, with substantial focus and investments placed on (i) better estimating the prevalence (Bhatt et al., 2015), endemicity (Battle et al., 2019) and suitability (Messina et al., 2016) of multiple diseases at the grid cell level, it is key to use reliable denominators for measuring the health metrics associated to them – especially in countries where spatially detailed and up-to-date census data are not available. In this framework, highly modeled WorldPop gridded population datasets are used to provide the basis for health metrics supporting, among others, the IHME Global Burden of Disease, the Malaria Atlas Project, and the Trachoma Atlas.

Alegana et al. (2015) demonstrate also that, in absence of reliable, up-to date and/or detailed census-based data, accurate spatially-explicit age-structure datasets can be produced by integrating geolocated household survey data with geospatial covariates (in this case with uncertainty quantification). The estimated proportions of the population under 5 years are used to guide polio vaccination allocations, plan vaccinator routes and logistics, and estimate coverage rates in Northern Nigeria. Furthermore, the corresponding gridded under-5 dataset forms part of the Nigerian Vaccination Tracking System through its mapping tool.

Similarly, in natural and man-made disaster situations there is a need to estimate, as much accurately as possible, the number of potentially affected people to determine both the scale of the event and relief needed. Considering that it is highly unlikely that such events are going to impact areas aligned with (coarse) census units, in countries and areas where detailed population data are not available, UNITAR-UNOSAT regularly use WorldPop gridded population and demographic datasets to assess the number of people potentially affected by natural disasters, as well as their characteristics such as age and sex (examples include, among others, the 2019 Peru Earthquake and Tropical Cyclone Idai). UNOCHA also uses WorldPop gridded datasets in the Libya Humanitarian Needs Overviews Report to estimate the percentage of population in need within each province in 2015.

Finally, Thomson et al. (2017) describe how highly modelled gridded population datasets might be used as a sample frame to select primary sampling units for complex household surveys in countries were data are outdated or inaccurate. In particular, they demonstrate the possibility of replicating the 2010 Rwanda Demographic and Health Survey (DHS) using an R-based "GridSample" algorithm to sample the 2010 UN adjusted WorldPop gridded population dataset for Rwanda (stratifying by 30 districts and oversampling in urban areas).

Regarding correlation analyses, it is important to note that to avoid issues relating to endogeneity, highly modeled gridded population datasets should not be used to make

predictions about any of the geospatial covariates used to model the population distribution or, similarly, to explore (spatial) relationships between the latter and the former. In such cases, it would be highly recommended to either use unmodeled datasets (such as in Cohen and Small, 1998) or re-modelling the population distribution without using the ancillary dataset of interest.

For example, in The State of the Pacific's RMNCAH Workforce 2019 Report, UNFPA used WorldPop gridded population datasets, produced without using health facility locations as an ancillary dataset, to derive country-based gridded pregnancy datasets (James, et al. 2018) and combine them with travel time to the nearest facility providing emergency obstetric and newborn care (EmONC) services. This is done for 15 Pacific countries to identify areas underserved by RMNCAH (Maternal, Newborn, Child and Adolescent Health) workers and estimate the number of pregnancies potentially not having access to EmONC facilities.

Similarly, Gaughan et al. (2019) examine the spatial relationship between gridded CO2 emissions disaggregated using remote sensing-based night-time light data (ODIAC) and gridded population datasets, for Vietnam, Cambodia and Laos, driven by a set of ancillary datasets not including night-time lights. This is done to characterize potential errors and uncertainties associated with only using night-time light data to disaggregate "residential" CO_2 emissions.

Final Considerations

In order to select the most appropriate gridded population and demographic datasets, end-users should carefully consider the objective(s) of their analysis, the assumptions and modelling approach used to produce the gridded datasets, as well as the covariates used to produce them. Furthermore, it is also important to consider that, even if the use of geospatial covariates and advanced statistical modelling techniques produces a more accurate representation of population distribution (Stevens et al., 2015; Sorichetta et al., 2015), each ancillary dataset used are often model outputs themselves and thus they have a degree of uncertainty that will carry over into the gridded datasets.

To this regard, the POPGRID Data Collaborative platform provides extended documentation/metadata and visual comparison tool for better understanding the source of uncertainties associated to both the input data and the model approach used to produce the various gridded population datasets. Nevertheless, as stated by Bai et al. (2018) "quantifying the accuracy of population distribution maps is recognized as a critical and challenging task" mostly due to the lack of available ground-truth compatible population data. To this end, the POPGRID Validation & Intercomparison Working Group has recently started a working document for collecting ideas, data, and metadata for extending and advancing previous validation efforts performed by the World Bank for Malawi, Engstrom et al. (2019) for SriLanka, Bai et al. (2018) for China, and Hall et al. (2012) for Sweden.

References

Alegana V.A. et al., 2015. Fine resolution mapping of population age-structures for health and development applications. Journal of The Royal Society Interface, 12(105), 20150073.

Bai Z. et al., 2018. Accuracy assessment of multi-source gridded population distribution datasets in China. Sustainability, 10(5), 1363.

Balk D. et al., 2016. Determining Global Population Distribution: Methods, Applications and Data. Adv. Parasit 62, 119–156.

Battle K.E. et al., 2019). Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. The Lancet.

Bhatt S. et al., 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature, 526(7572), 207.

Cohen, J.E., & Small, C., 1998. Hypsographic demography: the distribution of human population by altitude. Proceedings of the National Academy of Sciences, 95(24), 14009-14014.

Gaughan A.E. et al., 2019. Evaluating nighttime lights and population distribution as proxies for mapping anthropogenic CO2 emission in Vietnam, Cambodia and Laos. Environmental Research Communications, 1(9), 091006.

James W.H.M. et al., 2018. Gridded birth and pregnancy datasets for Africa, Latin America and the Caribbean. Sci. Data 5:180090.

Leyk S. et al., 2019. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth Syst. Sci. Data, 11, 1385–1409.

Messina J.P. et al., 2016. Mapping global environmental suitability for Zika virus. Elife, 5, e15272.

Stevens F.R. et al., 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one, 10(2), e0107042.

Sorichetta, A. et al. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Sci. Data 2:150045 doi: 10.1038/sdata.2015.45 (2015).

Thomson D. et al., 2017. GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data. International journal of health geographics, 16(1), 25.

Engstrom R. et al., 2019. Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka. The World Bank.

Hall, O. et al., 2012. From census to grids: comparing gridded population of the world with Swedish census records. The Open Geogr J, 5, 1-5.