Edith DARIN*, Mathias KUÉPIÉ**, Hervé BASSINGA***,
Gianluca BOO*, Andrew J. TATEM*

The Population Seen from Space: When Satellite Images Come to the Rescue of the Census

Great steps have been made in recent decades in observing the Earth from the sky. Landscapes and infrastructure can now be mapped at an extremely fine spatial scale. These data—particularly useful to geographers—can also benefit demographers. By combining observations of buildings in satellite images with complementary demographic data, population sizes in areas not reached by the census can be estimated. The authors apply this method to the case of Burkina Faso and explain how a hybrid population census can be carried out when data cannot be collected in some areas.

Today, developing public policies requires precise knowledge of the size and characteristics of the population. To respond to this need, national statistical offices must perform counts. National censuses are the foundational data collection operations on the number of inhabitants in each country. The national population is the denominator for many development indicators (Carr-Hill, 2014). Reliably and regularly estimating this denominator is important in all domains (land use planning and development, education, democratic representation, social protection, health, etc.) and at various geographical scales (United Nations, 2017). While traditionally the publication of population sizes is organized by administrative units such as provinces or regions, this format leads to spatial discontinuities that can prove arbitrary and that do not reflect other ways of dividing a territory according to criteria such as employment (employment basin) or health (healthcare districts[1]).

(1) Administrative division of a country based on the organization of the supply of healthcare services.

* WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, United Kingdom; Leverhulme Centre for Demographic Science, Department of Sociology, University of Oxford.

** United Nations Population Fund, Dakar, Senegal.

*** Institut national de la statistique et de la démographie, Ouagadougou, Burkina Faso.

Correspondence: e.c.darin@soton.ac.uk

Population-E, 77 (3), 2022, 437–464 DOI: 10.3917/popu.2203.0467

To remedy this problem at least partially, some countries (including the United States since 1940 and the United Kingdom since 2001) have decided to publish population data at the level of the enumeration area, the smallest operational unit in the census. Others have chosen to publish data based on a division of the territory into grid cells, to provide a standardized unit of analysis that can be aggregated into an effectively unlimited number of spatial combinations. Due to the diversity of grid units (e.g. 200 m square in France, 1 km square in Germany), since 2011 the European Statistical System has been promoting the publication of harmonized gridded data (Backer and Holt Bloch, 2011) to disseminate the results of the 2021 European censuses (INSPIRE, 2014).

Besides increased spatial resolution, gridded population data can play a role when security problems, natural disasters, or political conflicts make it impossible to map or carry out census operations in certain areas. By linking demographic data to their spatial distribution, gridded data allow spatial modelling to be used to estimate the missing population. This has been advocated recently by the United Nations Population Fund (UNFPA) through the notion of *hybrid census*. In a hybrid census, data from accessible areas are combined with high-resolution estimates for inaccessible areas (Jhamba et al., 2020). A pilot study was carried out in Afghanistan in 2017 (UNFPA, 2017).

The spatially uniform units created by using a gridded structure for census data make it theoretically possible to statistically model the population of inaccessible areas. But it is the advent of very high-resolution spatial data that makes such modelling viable. Satellite imagery has long been used to precisely map land cover and night-time light. But software and artificial intelligence are now enabling the extraction of ever-increasing amounts of information, including the nearly perfect tracing of the footprint of all buildings (Ecopia. AI and Maxar Technologies, 2019). These high-resolution footprints of the built environment are very information-rich, and integrating them into the modelling of the population represents a major scientific challenge.

The hybrid census approach is particularly well adapted to the Burkina Faso context. Burkina Faso's National Institute for Statistics and Demography (INSD) carried out its fifth population and housing census (PHC) in late 2019, but security issues in the north and east of the country kept the census from covering nearly 5% of enumeration areas (Institut national de la statistique et de la démographie, 2019). This article begins by proposing a method for estimating populations in these inaccessible areas. This first, 'bottom-up' model is a Bayesian hierarchical model that combines spatial variables with demographic information collected in enumeration areas where counting took place. This estimate is then applied to predict the population of the non-enumerated areas. We then show that a 'top-down' statistical learning model can be used to obtain demographic data at a high geographical resolution (at the grid-cell

level), disaggregating census population counts, for areas with census coverage, and predicted counts, for uncounted areas. The challenge is thus twofold: predicting the population in areas where enumeration could not take place and producing gridded estimates for the full country territory. The underlying challenge is how we can use novel data and innovative statistical methods to cope with recurring problems in counting the population and capturing its spatial distribution.

I. Spatial modelling of the population: what is at stake

1. The challenges of the traditional census

A PHC is a complex operation that must be meticulously organized to ensure the coverage of all residential structures and the entire population. This organization is divided into two major sequential phases: census mapping and enumeration.

The role of census mapping is to survey the full territory of the country, identifying all inhabited places and residential structures, and producing a rapid estimate of the population. Based on this information, each administrative unit in the country is divided into enumeration areas (containing around 1,000 inhabitants in urban areas and 800 in rural areas), finely partitioning the territory. The enumeration phase is kept brief (generally 2–3 weeks) to produce a snapshot of the population while limiting the risk of double counting due to population movements. However, the solutions to the many problems that arise in the field—underestimation of the scale of work required in some areas due to issues with mapping; omission of some areas from the map; multiple complaints and refusal to cooperate by some groups; delayed payment of field personnel; etc.—often come at the cost of the quality and exhaustiveness of the collected information.

The new generations of the PHC use satellite imagery, a digital geographical information system and the administration of census questionnaires via tablet. These have drastically improved census mapping, the monitoring of data collection, and thus data quality (Eyinga Dimi, 2019). Nonetheless, given the complexity of enumeration operations and the risks of omission, it is customary, following the enumeration phase, to carry out a representative sampling of enumeration areas by stratum (type of area and/or region) and submit an abbreviated version of the questionnaire. This procedure, known as a post-enumeration survey (PES), is carried out to measure rates of omissions and verify the quality of the collected information. But not all countries perform these surveys. Of the 134 countries that participated in the 2010 round of censuses, only 66% went on to carry out a PES, and of these only three-quarters made use of the results. In Africa, Asia, and South America, the proportion is only one-third (UNFPA, 2019). Moreover, even if the quality of the PES is acceptable,

the size of the population is adjusted homogeneously within strata, masking the dependence of omissions on the quality of the work of particular teams and difficulties in the field in particular areas. Finally, in some cases, significant areas of the country are inaccessible to census teams for physical or security reasons (Buettner and Garland, 2008), so the population there must be estimated in some other way.

2. Spatial data and population estimation

In the context of population censuses, spatial data are mainly treated as an operational tool to facilitate field logistics and ensure the completeness of census mapping. They can also be understood as a vehicle for demographic information in thematic maps where geographical subdivisions are assigned a colour based on their population sizes (Martin, 2011). But these maps do not allow inhabited areas to be distinguished from uninhabited ones (such as lakes or deserts), and make observations strictly dependent on the chosen boundaries, which creates problems when those boundaries are changed. The concept of gridded population was developed to better capture the real spatial distribution of the population. This format was originally developed in the domain of remote sensing, i.e. of the ground level observed from the air or from space. Leyk et al. (2019) dated the first large-scale gridded population to the NASA Goddard Institute for Space Studies' Global Distribution of 1984 Population Density at 1° × 1° Resolution (Fung et al., 1991). However, gridded demographic data first emerged out of Scandinavian statistical institutes in the 1960s (Claeson, 1963).

To understand this 3-decade delay between the Scandinavian initiatives and the first global gridding, it is important to note the difference between gridded data drawn from the aggregation of observations carried out at a finer level of detail than the grid cell, on the one hand, and gridded data derived from a statistical disaggregation model, on the other. In Scandinavian countries, gridded demographic statistics were produced from administrative records that associate individuals in the population with their postal address (Longva et al., 1998). Gridded data, by aggregating individual data, thus serve in this context to address a problem of data confidentiality. In 2010, the European statistical system launched the GEOSTAT project, which promoted the production of harmonized European gridded population data at a scale of 1 km × 1 km. Only 11 countries possess localized data requiring aggregation (Backer and Holt Bloch, 2011). In the other countries, disaggregation models must be used. The idea of refining the spatial representation of the population, excluding uninhabited areas, and thereby producing dasymetric maps,[2] is not a new one (Scrope, 1833). It was originally devised in 1911 by Semionov-Tian-Shansky when designing an atlas of Russia. Interest in this type of

(2) For details, see https://journal.augc.asso.fr/index.php/ajce/article/view/ajce.34.1.147

population map began to grow rapidly beginning in the 1990s, with the development of increasingly high-performance geographical information systems (Petrov, 2012). Geographical data would now help to estimate the precise spatial distribution of the population. With the arrival of new spatial data, various methods have been developed in recent years to spatially disaggregate population data. The increasing availability of remote sensing data on types of land cover (Friedl et al., 2002), night-time light (Elvidge et al., 2017), and climatic data (Harris et al., 2014) has expanded the range of sources that can provide information on local variations in population density. Furthermore, techniques for integrating these different types of data have evolved, from the homogeneous allocation of the population restricted to inhabited areas, to the estimation of local variations using multiple linear regression (Langford, 1991) and more sophisticated statistical learning methods.(3) These approaches to the statistical modelling of the population, which Wardrop et al. (2018) termed 'top-down' population mapping, can be used to keep total population numbers at the original scale of the census data. This assumes that reliable census data covering the entire country are available.

But geographical data can also be used to estimate populations, and thus can be considered predictors of population. In this context, gridding can be used to define a uniform framework for the entire country and thus a common system for enumerated and non-enumerated areas. In geostatistics, this 'bottom-up' approach, which allows a set of observations to be extrapolated to a given area, has been widely used, in particular to estimate the distribution of environmental variables on the basis of surveys (Chilès and Delfiner, 2009). Applying methods from geostatistics to human phenomena, spatial epidemiology then sought to map the incidence of diseases based on the geographical referencing of cases (Lawson, 2013). This approach then spread into other areas of the social sciences. Geolocalized surveys were used to map social phenomena, such as poverty (Alderman et al., 2002), vaccine coverage (Utazi et al., 2019), and housing conditions (Tusting et al., 2019) at the country level. However, these types of studies work with data on prevalence, and not on total population sizes. In spatial ecology, in contrast, the populations of observed species are estimated based on their spatial distribution (Elith and Leathwick, 2009). These approaches have relatively rarely been used to study human populations: two pilot studies have been produced, one for Nigeria (Weber et al., 2018) and the other for Afghanistan (UNFPA, 2017), to respond to the need for recent population data. Working with the Nigerian data, Leasure et al. (2020a) developed a Bayesian model that also allows for the quantification of the uncertainty associated with these estimates. This is the approach we adapt here for estimating the population of areas not covered by the 2019 census of Burkina Faso.

(3) For example, using random forests (Stevens et al., 2015) or maximum entropy (Leyk et al., 2013). The first is the approach taken here (see below).