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

Edith Darin, **Mathias Kuépié**, **Hervé Bassinga**, **Gianluca Boo**, **Andrew J. Tatem**, **Translated by Paul Reeve**

In Population **Volume 77, Issue 3, October 2022**, pages 437 to 464

ISSN 0032-4663
ISBN 9782733280294
DOI 10.3917/popu.2203.0467

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

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).

II. Producing a gridded estimate of the population of Burkina Faso in two steps

1. Population data in Burkina Faso

The main source of population data in Burkina Faso is the population census. Due to the incomplete updating of vital records, Burkina Faso's public statistics use demographic projections to maintain updated population numbers between censuses. After the fourth population census was conducted in 2006, the country produced three documents presenting demographic projections (Guengant et al., 2009; INSD, 2009, 2017), which described a scenario where its population reached 21 million in 2020. However, when evaluating and using such a figure, it is important to examine the underlying hypotheses critically and to consider the variables used to produce it, its time horizon, and the sociopolitical context. In particular, the deterioration of the security situation since 2015 has led to significant changes in the occupation of different areas within the country.

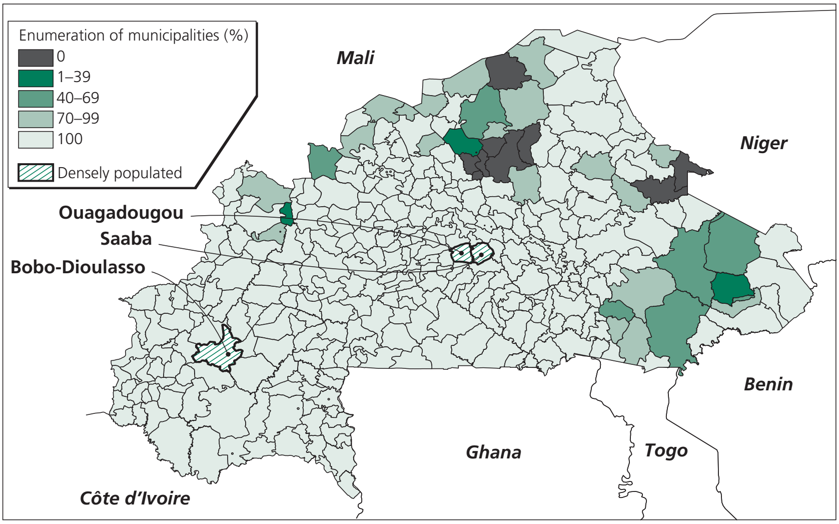
Given the urgent need for up-to-date demographic data and despite a challenging security situation, Burkina Faso carried out its fifth PHC in 2019. However, despite the use of strategies intended to ensure coverage of the entire country (recruitment of interviewers through local referrals, adaptation of communication in these areas, involvement of security authorities, etc.), some localities considered too dangerous were not covered. Out of the 351 municipalities in the country, 52 were only partially covered, and nine were not covered at all. Out of 25,023 enumeration areas, 1,206 (or 4.8%) could not be enumerated. These enumeration areas are located mainly in the north and the east of the country, particularly at the borders with Mali and Niger (Figure 1).

2. Geographical indicators of population sizes

To fill in missing population information, one needs fine-grained geographical data available for the country as a whole and able to act as indicators of inhabitation.

One set of variables relates to the environmental context of inhabited areas. For the present study, we chose gridded climate data produced by the Climatic Research Unit (Harris et al., 2020) and the land cover classification of the European Space Agency (Buchhorn et al., 2020). In addition, we also used the map of the hydrological network produced by the Geographical Institute of Burkina Faso (2015). A second set of variables seeks to describe the country's infrastructure based on the geolocalization of localities grouped into three administrative classes (city, village, and hamlet) and the road network. These are drawn from the national topographical database (Institut géographique du Burkina Faso, 2015). The Malaria Atlas Project also modelled access to cities (Weiss et al., 2018), as did Tusting et al. (2019) for housing conditions. A third dataset, which provides information on the current distribution of the population, is its past distribution as modelled by the WorldPop research unit (2018). This spatial projection of the figures from the previous census offers a good reference for predicting the current population.

Figure 1. Coverage of the 351 municipalities (communes) in the 2019 census of Burkina Faso



Source: Authors' construction based on census data collected by INSD in 2019.

The final, fundamental source of data is the map of buildings. Previous attempts to estimate the population from survey data have been based on the manual delineation of rooftops (Checchi et al., 2013; Hillson et al., 2014). Today, the automatic identification of building outlines through artificial-intelligence satellite-image analysis algorithms (Ecopia.AI and Maxar Technologies, 2019) is expanding potential opportunities for population estimation. Not only does this provide extremely fine-grained information on built areas and their surfaces (down to 50 cm resolution), it can also be used to produce variables that characterize the arrangement of buildings, such as their average perimeter or the distance between them. For the present study, we use the following characteristics: the number, perimeter, and area of buildings, as well as the distance between them.

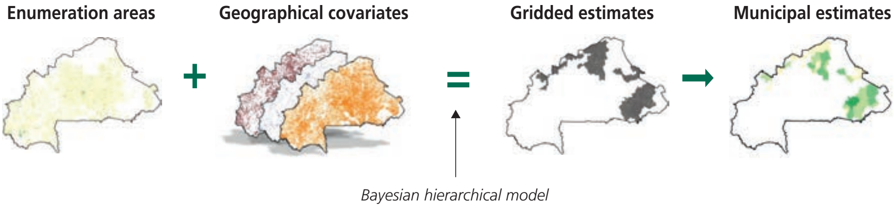
Combining all these different information sources brings out the strengths of a gridded approach. By dividing the territory into cells of identical size, these data can be assembled at the level of a single analytical unit. For vector data that take the form of points, lines, or polygons, such as lines in the road network or points for sites, the Euclidean distance to the nearest grid cell is calculated. For polygons representing buildings, the characteristics of the buildings within a given grid cell are summarized using various statistical indicators (mean, standard deviation, maximum, minimum, median, and coefficient of variation). For the variables deduced from buildings, a focal mean is also calculated, i.e. the mean of the grid cells located within 100 m, 1 km and 5 km from each square, to describe its residential context.

3. First model: estimation of non- or partially enumerated municipalities

The first model (Figure 2) predicts the population sizes of municipalities not enumerated in the census, or only partially. To do this, we use the sizes and GPS coordinates of the enumerated households. Population sizes are calculated for fully covered enumeration areas, and their relation to the geographical covariates is modelled. The complete coverage of the spatial covariates then enables us to predict the population for each cell in non-enumerated areas. By aggregating the estimates according to the geographical boundaries of administrative units, we obtain an estimate of the population sizes of the non-enumerated municipalities and the associated levels of uncertainty.

Bayesian modelling is used to deal with the heterogeneity of population data, and in particular with variations unexplained by the spatial covariates (Appendix A). By assuming that parameters are random variables, Bayesian estimation can be used to quantify the uncertainty associated with the input data (Ferreira et al., 2020). The absence of entire municipalities from the census data means that spatial modelling in the strict sense, which operates on the basis of geographical proximity, is impossible. The spatial logic of the distribution of the population is then approached in terms of a nested hierarchy of administrative structures at different geographical levels (municipality, province, and region). Finer variations can finally be characterized using high-resolution spatial covariates (Leasure et al., 2020a).

Figure 2. Bottom-up modelling of population sizes in non- or partially enumerated municipalities

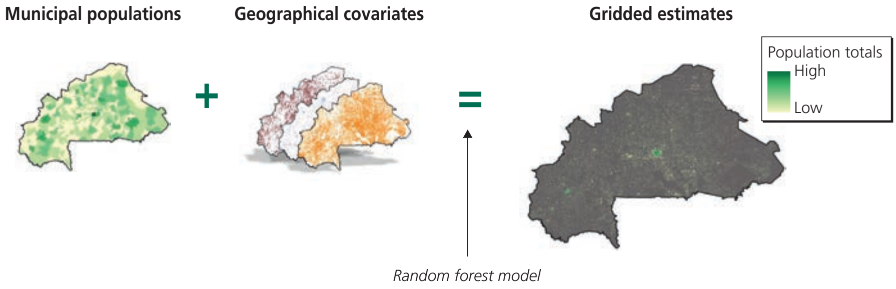


Source: Authors’ construction, 2022.

4. Second model: countrywide gridded estimate

The second model (Figure 3) seeks to disaggregate, at the grid-cell level, all municipal population sizes, i.e. census data adjusted by the PES for fully covered municipalities, and the estimates produced in the previous step for partially enumerated and totally inaccessible municipalities. We begin by estimating the relationship between municipal population density and spatial data aggregated at the level of the administrative unit.

Figure 3. Top-down modelling of the disaggregation of countrywide population totals at the grid-cell level



Source: Authors' construction, 2022.

Research conducted by the WorldPop research unit (Tatem, 2017) shows that the statistical learning approach, more specifically using the random forest algorithm, is the best way to produce such models, due to its predictive power, flexibility, and robustness to multicollinearity (Sorichetta et al., 2015; Stevens et al., 2015, 2020). This approach is based on the concept of a decision tree, which models a dependent variable by dividing the input data into subgroups via thresholds on the explanatory covariates and by calculating a prediction for each subgroup, i.e. the mean of the dependent variable. The aim of decision trees is to identify the partition that, for each observation, minimizes the difference between the observed dependent variable and the predicted value, while avoiding overfitting (e.g. creating a partition of one observation per subgroup). The random forest method extends this approach by sampling the input data and estimating a decision tree for each sample to make the result more robust to statistical noise (Breiman, 2001).

Once the model has been estimated, given the availability of fine-grained spatial data, it can be used to predict population density at the grid-cell level. This prediction is then used as a weighting in disaggregating municipal populations drawn from the hybrid census, to capture spatial dynamics of populations at the submunicipal level.

III. Putting the models into practice

1. Bottom-up modelling and the quality of georeferencing

The complex preparation of input data

To produce a population dataset that can be used in modelling, we need units where population counts and their connection to the associated geographical area are reliable. In the absence of digitized geographical boundaries, the only georeferencing available is household GPS data (87% of observations). The first challenge is to reconstruct the edges of these areas and extract the built area to be used in estimating population density. To do this, circles of different radii were drawn around household GPS points, depending on the settlement type (see Figure 4 for an example).

The second challenge is to select reliable enumeration areas. Here we apply the following four criteria: (a) complete enumeration of the area; (b) reliable information on area size; (c) homogeneity of residential settlement type within the area; and (d) non-overlap between areas. Various indicators (rate of missing GPS points per enumeration area; distance between GPS points and the barycentre of the enumeration area; standard deviation of the GPS coordinates belonging to the same area; number of enumeration areas per 100 m by 100 m grid cell; number of GPS points per grid cell; number of people per building; and population density) were defined in order to construct the most geographically accurate possible database. In the end, 15,817 enumeration areas were selected, or 69% of the initial dataset. Since the selection procedures and choice of radius for encircling GPS points have an impact on the predicted data, a sensitivity analysis is presented in Table 1.

To avoid overestimating the population of areas with security problems due to population movements towards more secure areas with census coverage, migrants who reported recently moving from the missing areas were removed from the basis for the estimation using the question ‘What municipality did you live in last year?’ A post-estimate adjustment was then carried out at the municipal level by adding the migrants to their census municipality and subtracting them from their municipality of origin.

Finally, to integrate the urban/rural distinction, whose mapping has not been digitized, a binary classification model of built areas at the grid-cell level was constructed based on the typology of the enumeration areas. We extrapolated this urban/rural typology using a gradient boosting machine algorithm. This consists in the sequential estimation of a series of decision trees, giving greater weight at each step to observations predicted less well during the previous iteration (Friedman, 2001). Two basic geographical variables were chosen: distance to the main urban centres—the 45 provincial capitals (*chefs-lieux*) and the four medium-sized towns of Bitou, Niangokolo, Garango, and Pouytenga—and the number of buildings within a 500 m radius, since the goal is to estimate the contours of the provincial capitals based on the built density observed in enumeration areas classified as urban. The estimates were calculated in R (R Core Team, 2020) using the *caret* package (Kuhn, 2008). The model was chosen among a set of possible classification algorithms—Adaboost, random forests, support vector machines, and generalized linear models—based on the largest area under the ROC curve. This measures the model’s ability to distinguish between urban and rural areas (0.98 on a scale of 0 to 1 for the gradient boosting machine algorithm).

Estimation of the model of population sizes in non-enumerated municipalities

The final model has three hierarchical levels: the type of built area, the region, and the municipality. Five variables were selected because of their correlation with the population density of enumeration areas: the number of buildings within a 5 km radius; distance to rivers classified by INSD as temporary (i.e. which disappear during the dry period); distance to secondary roads; friction surface, which represents the difficulty of crossing a cell and depends on the presence and quality of railways, rivers, and roads, as well as topography; and the gridded projections of WorldPop. The model was estimated using Stan (Carpenter et al., 2017), whose scripts and diagnostics are available on GitHub.[4] The estimation was replicated 9,000 times (3,000 iterations for three Markov chains), thus simulating the entire distribution of parameters and predictions. This distribution is summarized here by the mean prediction and the upper and lower limits of the confidence interval, defined to contain 95% of predictions. To evaluate the model, a cross-validation was applied, by estimating the parameters based on 70% of the selected enumeration areas and predicting the population counts for the remaining 30% (test sample). Figure 4 shows the fit between the distribution of the mean predicted population and that of the observed population for the enumeration areas in the test sample.

Limitations: difficult-to-model input data

Despite the care taken through the procedure for selecting enumeration areas, the data remain very heterogeneous, as reflected by the very large confidence intervals (see limits in Figure 4) and by the standard deviation of prediction errors at the enumeration area level of 263 individuals, representing a coefficient of variation of 5.8. Furthermore, the procedure for selecting and defining the limits of enumeration areas has an impact on the results. To measure that impact, we conducted a sensitivity analysis on two crucial steps:

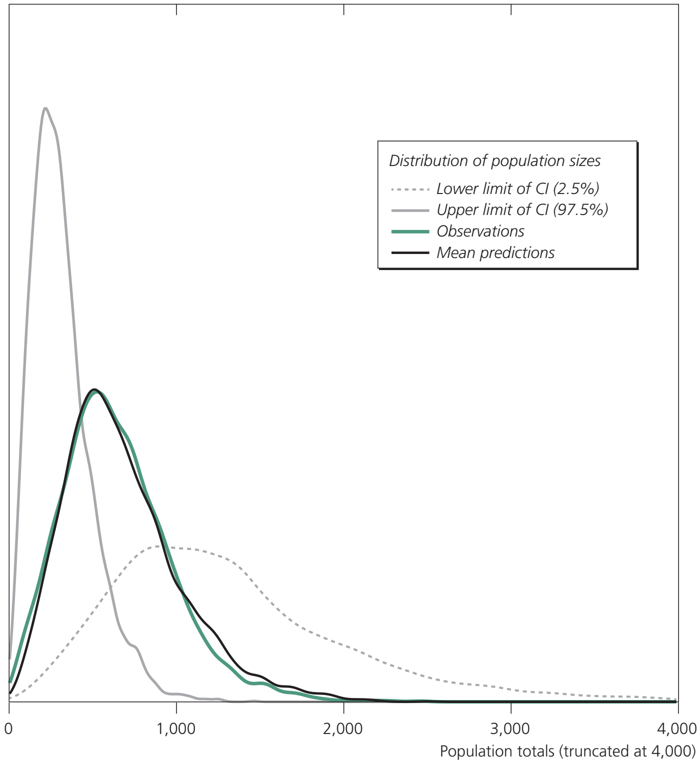
- The choice of radius around the GPS points:
  - Scenario 1: we distinguish between the highly urban municipalities of Ouagadougou, Saaba, and Bobo-Dioulasso and the rest, using a radius of 25 m and 100 m, respectively, in the two cases.
  - Scenario 2: we refine the discrimination between urban areas by separating out the capital Ouagadougou, with a radius of 20 m, the two densely populated urban municipalities of Bobo-Dioulasso and Saaba with a radius of 25 m, the other urban municipalities with a

[4] https://github.com/wpgp/BFA_population_v1_0_methods/tree/main/supplements

radius of 80 m, and rural areas with a radius of 120 m (see Figure 1 for the location of the municipalities).

- The maximum accepted residential density threshold:
  - fixed for all enumeration areas; or
  - proportional to the maximum residential density in each province.

Figure 4. Comparison of the distribution of mean predicted and observed population sizes in the test sample of enumeration areas (4,745 areas)



Note: The distributions of the bounds of the 95% confidence interval (CI) associated with each predicted population are shown in grey. The mean error is 45 individuals out of a mean of 660 per enumeration area.
Source: Authors' calculations based on census data collected by INSD in 2019.

To evaluate the predictions derived from the different versions of the input data, they can be compared with the observed populations for fully enumerated municipalities. Table 1 illustrates the dilemma of the choice of metric, with absolute error (root mean square error [RMSE]) lower for Procedures 1 and 3 and relative error (relative RMSE) lower for Procedure 4. Relative error allows the size of municipalities to be factored in, keeping in mind that the areas to be predicted do not include highly populated municipalities. Moreover, note the predictions' strong sensitivity to the required cleansing procedures given the quality of the collected data of the GPS points.

Table 1. Analysis of model sensitivity in relation to data cleansing procedures

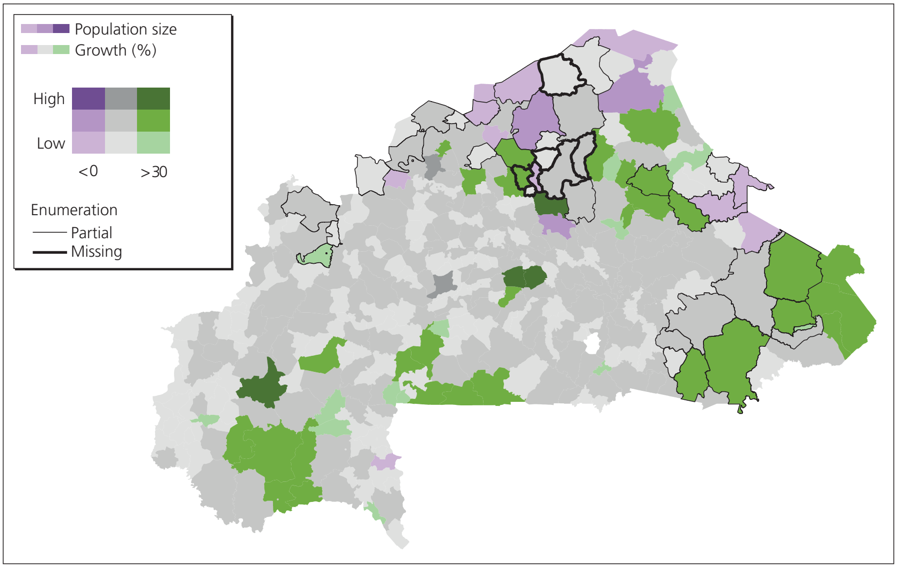
	Data cleansing procedure		% of data discarded	RMSE	Relative RMSE
	GPS radius scenario	Density threshold			
1	Scenario 1	0.16	5	13,862	40%
2	Scenario 2	0.16	4	21,142	33%
3	Scenario 2	0.9 × maximum	7	18,023	27%
4	**Scenario 2**	**0.8 × maximum**	**9**	**19,274**	**24%**

Note: The data cleansing procedures are the scenarios for choice of radius around GPS points and the maximum threshold for acceptance of population densities (number of people per m^2 of built area). The quality indicator—root mean square error (RMSE)—compares the predictions with the INSD counts for fully enumerated municipalities. Relative RMSE was calculated from the relation of the errors to the size of the associated municipalities. The selected model is shown in bold.
Source: Authors' calculations based on census data collected by INSD in 2019.

Analysis of demographic estimates

Figure 5 illustrates the 2019 census of Burkina Faso, whose preliminary results, combining collected data and estimates resulting from the previously presented model, were communicated in November 2020 (Institut national de la statistique et de la démographie, 2019). These population estimates represent 10.2% of the national population and include the majority of municipalities whose population has decreased compared to the 2006 census (in purple on

Figure 5. Municipal populations, estimated and enumerated in the census, and their population growth since 2006



Note: Population growth since the 2006 census is represented by colour (purple for negative growth and green for growth above 30%). The coverage of the 2019 census is represented by areas with either no outline (fully enumerated municipalities) or a thin (partial coverage) or thick (no coverage) outline.
Source: Authors' mapping based on census data collected by INSD in 2019 and 2006.

the map) due to security problems. A significant increase (> 30%) in the population of certain estimated municipalities (in green) can also be seen. These mainly consist of partially enumerated municipalities (thin outline), which probably received members of the population of neighbouring areas that could not be enumerated at all (thick outline) due to generalized insecurity.

2. Top-down modelling: an opportunity for small-scale validation

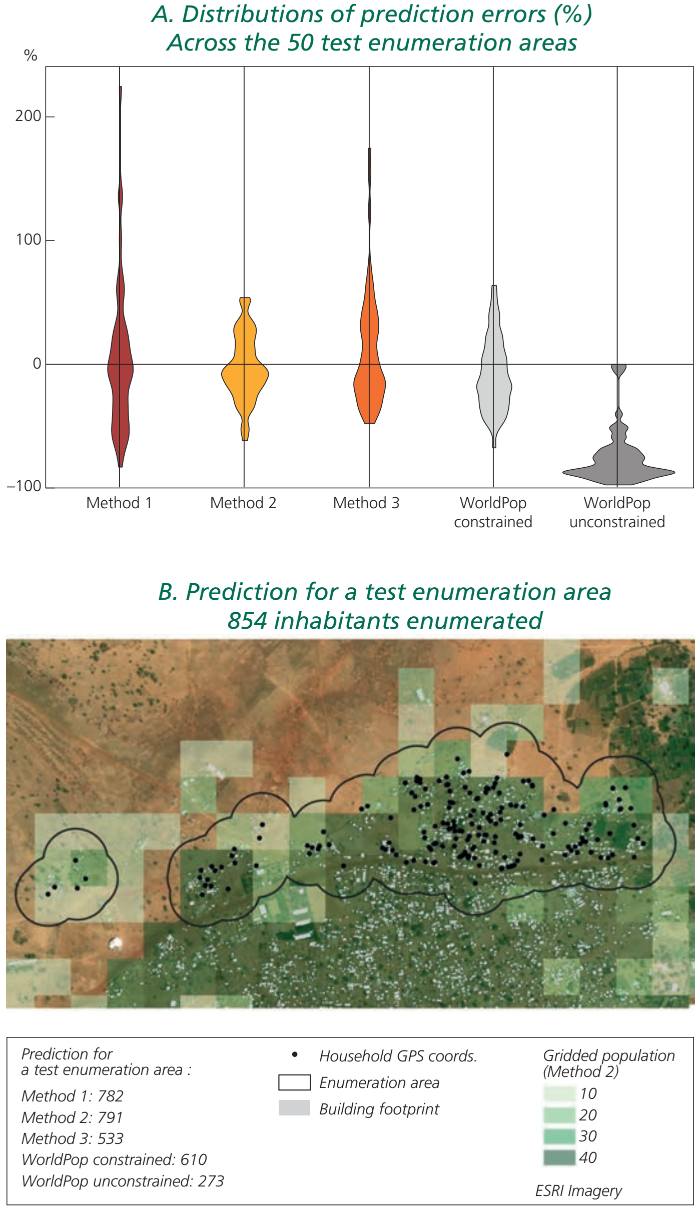
Disaggregating municipal population counts involves estimating the relationship between population density and all chosen geographic variables at the municipal level, and applying this relationship at the grid-cell level. An aim of the present study was also to measure the impact on disaggregation of the denominator used to calculate population density. The first denominator chosen is the total area of the administrative units, which is identical to the definition used for WorldPop's gridded populations (WorldPop Research Group et al., 2018). But the recent production of building footprints (Ecopia.AI and Maxar Technologies, 2019) has made it possible to spatially constrain human settlement to built areas, and thus to work with the number of people per unit of built area. Two methods exist for calculating the built area: either by summing the built cells, as for the constrained WorldPop estimates (Bondarenko et al., 2020), or by summing the area of all buildings. Using open-access scripts from Bondarenko et al. (2018), we assessed the impact on disaggregation of using the different denominators to calculate municipal population densities: (a) the total area of the municipality (Method 1); (b) the area of built cells (Method 2); and (c) the area of buildings (Method 3).

A first observation is that when density is calculated based on the area of buildings, calculated residential density is higher in rural areas than in urban areas. This effect is explained by the greater presence of non-residential buildings in urban areas, which decreases the ratio of population to built area. Using this predicted residential density at the grid-cell level as a weight to disaggregate population counts (WorldPop Research Group et al., 2018) would yield lower counts in urban areas than in rural areas. To remedy this, the predicted density is multiplied by the built area within the grid cell, representing a prediction of population size, which is then used as a weight in disaggregating municipal population counts.

Disaggregation is traditionally evaluated using two administrative levels of population data: the higher-level geographic unit is disaggregated and its predictions compared to data from the lower administrative level (Stevens et al., 2015). However, finer-grained population data are available to us: namely, the counts from individual enumeration areas. Considering the problems with the quality of the GPS data collected (listed in Section III.1), 50 enumeration areas distributed across the country and the various residential contexts were selected.

Figure 6 offers a clear demonstration of the importance of the building footprints for disaggregation. Unconstrained WorldPop estimates lead, on average,

Figure 6. Comparison of disaggregation models across 50 test enumeration areas



A. Distributions of prediction errors (%) Across the 50 test enumeration areas

B. Prediction for a test enumeration area 854 inhabitants enumerated

Note: The graph on the left shows the distribution of relative prediction errors ((prediction – observation)/observation × 100) on a test sample with the three methods used to calculate the spatial denominator as well as the constrained and unconstrained WorldPop estimates. The image on the right is a visualization, at the enumeration area level, of the circling of GPS points to define the limits of the area, the building footprint, and the final gridded population (Method 2). The predicted population sizes with different methods are also presented.
Source: Authors' calculations based on census data collected by INSD in 2019.

to a prediction error of 90%, while Method 1 produces prediction errors of up to +200%. The method of calculating the built area has an impact on the disaggregation, with a greater dispersion of errors when the area of buildings is used (Method 3). Finally, applying this methodology with the municipal population counts from the enumeration, coupled with those estimated by the top-down approach for non-enumerated municipalities (Section III.1) as well as a specific set of geographical covariates (Method 2), produces the lowest prediction errors on average (–3%) and the lowest error dispersion (±27%).[5]

IV. Advantages and limitations of gridded estimates

The gridded population estimates for Burkina Faso presented in this article are the result of two successive modelling steps. The first (probabilistic model) was aimed at obtaining population sizes for areas not covered by the census. The second (deterministic model) aimed to break down hybrid municipal population estimates at the grid-cell level. This study highlights the importance of the gridding of demographic data. In addition to offering a precise image of the distribution of the population (Figure 4), it provides a coherent statistical basis for linking enumerated and non-enumerated areas, as well as demographic characteristics and geographical information.

1. An appeal at the crossroads between geography and demography

INSD’s official publication of gridded demographic estimates is a step in the direction of putting richer spatial analysis within the reach of all, in a context where it is rare for demographic or even geographical data to be freely available at a submunicipal level. Gridding compensates for the absence of a finer-grained level of administrative units. Grid cells can be used flexibly with any division of the territory (Thomson et al., 2020). However, to take full advantage of the gridded format requires the mastery of techniques specific to geographic information systems that are not among the traditional methods of demographic analysis. The mobilization of this type of resource therefore must be accompanied by dissemination and outreach activities. The development of a multilingual application that facilitates the visualization and aggregation of estimates is an initial step[6] (Leasure et al., 2020b).

2. Responding to the challenges of the 2020 round of censuses

The series of censuses taking place between 2015 and 2024 has seen the completion of classical censuses in certain regions of the world, particularly sub-Saharan Africa, jeopardized by rising insecurity (Jhamba et al., 2020).

[5] The final gridded population estimate from Method 2, an overview of which can be seen in Figure 6, is available for download at: https://data.worldpop.org/repo/wopr/BFA/population/v1.0

[6] https://apps.worldpop.org/woprVision

Cameroon, which has been experiencing attacks from the Islamist sect Boko Haram since 2015, as well as secessionist pressures in its English-speaking areas, has not been able to complete its census mapping (Ebolé Bola, 2019). In this context, the UNFPA has been promoting the use of statistical methods for estimating the population (UNFPA, 2020). While our Bayesian estimation framework allows different sources of population data to be combined while quantifying uncertainty (Leasure et al., 2020a), it only allows for the reconstruction of one of the variables yielded by the census: population size. It is more difficult to use remote sensing data to establish many other characteristics of the population, such as composition by sex and age, socio-economic level, housing conditions, and migration. Current techniques for high-resolution mapping of social indicators, such as access to drinking water (Local Burden of Disease WaSH Collaborators, 2020) and school attendance (Local Burden of Disease Educational Attainment Collaborators, 2020), use geostatistical modelling techniques that require sampling which covers the entire territory. The approach taken here, using a hierarchical model with nested geographic levels to model the spatial distribution of the population, allows for extrapolation to non-covered areas. It does rely, however, on the hypothesis of similar building occupation in accessible and inaccessible areas, after controlling for the type of buildings and the administrative structure—a similarity locally nuanced by geographical covariates.

3. An analytical weakness: population displacements

The insecurity that prevents complete census coverage also leads to internal population movements on a large scale, which are difficult to quantify and map (Carr-Hill, 2014). Of the 414,000 migrants identified during Burkina Faso's fifth PHC (i.e. who reported that they had resided in another municipality the previous year), 30% were from municipalities with security challenges. To take this into account in estimating the numbers of individuals not reached by the census, we use an accounting method at the municipal level (see Section III.1). This does not, however, take into account displacements within the same municipality – for example, to the municipal seat. And it only includes displacements that were recorded or took place in 2019, whereas insecurity began to emerge in 2015 (but specifically accelerated in 2019). The predicted population can be modified, however, by estimating a surface called the weighting layer, which redistributes the gridded population numbers according to internal migration. Such a model was recently developed to estimate the population of South Sudan, based on the Armed Conflict Locations and Events Database and assessments by the International Organization for Migration (Dooley et al., 2021).

4. A technical dependency: the building footprints

A primary assumption concerning the quality of these predictions is the accuracy of the building footprints, i.e. how successful is the AI's extraction of data on building outlines from satellite images. Ecopia.AI and Maxar Technologies (2019) guarantee less than 5% false positives and negatives in a randomly selected sample of sites. The dating of the images also plays a role. For Burkina Faso, 20% of the images date from before 2015. However, these images are of rural areas, where major changes in urbanization are less likely to have occurred. For inaccessible municipalities, this proportion is only 15%, whereas 50% of images were taken between 2018 and 2020. If a locality was recently developed, then our model will not be able to accurately predict its current number of inhabitants. Conversely, if a village has been emptied of its inhabitants while its built structures remain in place, its population will be overestimated. An additional model of the correspondence between observed buildings and detected building footprints can be developed by including the date of the satellite image. But this requires a dataset where buildings are identified, typically drawn from census mapping. Moreover, in the present study, any building detected through satellite imagery is considered potentially residential. This leads to an overestimation of the number of residents in institutional or industrial areas. One solution for detecting non-residential buildings is to define a threshold for their size, but this creates a considerable risk of false negatives leading to the removal of residents from the map. The other possibility is to refine the building typology, going beyond the urban/rural binary to enable the model to detect differences in population density (Jochem et al., 2021).

Conclusion

The hybrid census, which combines field enumeration with high-resolution estimates, represents an undeniable technological advance. It means that in contexts where security challenges entail incomplete coverage of the population, geographical data from satellite imagery and other sources can be used to supplement population data. However, the objectives of a population census extend well beyond simply counting the population. The use of spatial modelling to estimate other demographic trends, such as gender and age composition, socio-economic level, and migration, remains an open challenge.

Moreover, the theoretical consequences of the inclusion of estimates in the population census extend beyond the scientific production of statistical data. It reopens the argument of the State's powers and responsibilities in carrying out an exhaustive count of its population. Regarding the 1990 census, the United States Supreme Court rejected the use of statistical methods to adjust for the undercounting of marginalized populations using a sample survey (Anderson and Fienberg, 1996). This highlights the difficult articulation of legal questions (responsibility for the production of population counts) with metrological questions (what is the best statistical method of estimating the population?) (Desrosières, 2000). In situations where data collection is complicated, as in Burkina Faso, coupled with both the production of increasingly precise auxiliary data and increasingly sophisticated statistical techniques, scientific advances enable the production of complete and up-to-date population data. This opportunity should not be overlooked, particularly given the increasingly degraded security context seen in the countries of the Sahel at the opening of the 2020 round of censuses.

Appendix

Appendix A: Bayesian hierarchical population model

The objective is to model the population in enumeration areas N_i, a variable that is discrete by definition and that can thus be described by a Poisson probability distribution. Poisson's distribution has a single parameter that governs both the mean and the variance of the variable. To represent the over-dispersion of the observations, the population is broken down using population density, a continuous variable, multiplied by the observed variable A_i, the built area. Because population density is continuous, it can be defined using a log-linear regression (log because it is a positive variable) integrating the predictive spatial variables $x_{i,k}$ and a y-intercept α_l estimated hierarchically by region l, i.e. with a hyperparameter $a_{national}$ that constrains each regional estimate to be consistent with the national estimate. Finally, the variance σ of the density (which corresponds to the variance of the error term in a frequentist presentation, i.e. $\log(D_i) = \alpha + \beta x_i + \varepsilon$ with $\varepsilon \sim Normal(0,\sigma)$, and which describes the uncertainty around the estimated density) is also estimated hierarchically by region l.

More precisely, for each enumeration area i belonging to region l:

Base model:

$$N_i \sim Poisson(D_i\, A_i) \tag{1}$$

$$\text{Log}\ (D_i) \sim Normal\ (\overline{D}_i, \sigma_l) \tag{2}$$

$$\overline{D}_i = \alpha_l + \textstyle\sum_{k=1}^{K} \beta_k\, x_{i,k} \tag{3}$$

Prior distributions :

$$\beta_k \sim Normal(0,1) \tag{4}$$

$$\alpha_l \sim Normal(a_{national}, s_{national}) \tag{5}$$

$$a_{national} \sim TruncNormal(11,3)$$

$$s_{national} \sim TruncNormal(0,1)$$

$$\sigma_l \sim Normal(b_{national}, v_{national}) \tag{6}$$

$$b_{national} \sim TruncNormal(0,1)$$

$$v_{national} \sim TruncNormal(0,1)$$

where:

(1) models the population size N_i of the enumeration area as a Poisson distribution because a population size is a positive discrete event, with the population density D_i as a parameter, multiplied by the built area within the enumeration area A_i

(2) models the log population density according to a normal distribution because density is a positive event, with mean $\overline{D}_i$ and variance σ_l, estimated hierarchically

(3) estimates $\overline{D}_i$ as a linear regression on K spatial variables x_k with coefficients β_k and intercept α_l, estimated hierarchically

(4) indicates the prior distributions of the β_k, which are assumed to be independent and centred at 0 so as not to assume an impact of the spatial variables

(1) and (6) hierarchically structure the prior distributions of α_l and σ_l by making them each depend on two parameters estimated at the national level: for the first, $a_{national}$ and $v_{national}$, and for the second, $b_{national}$ and $v_{national}$. To restrict these two hyperparameters to positive values, the prior distributions that govern them are truncated normal distributions. While overall the prior distributions are not very informative, the mean of $a_{national}$ is 11, as suggested by the sample of population densities, to speed up the estimation procedure, while its variance is 3 to relax the associated constraint.

In using hierarchical estimation, we assume that the variance and population density at the origin differ by geography. It can be assumed, for example, that variance will be comparatively lower in rural areas than in urban areas, which contain highly contrasting residential environments. The model presented here has a single geographical level l, but it has been extended in practice to multiple nested geographical levels.

References

Alderman H., Babita M., Demombynes G., Makhatha N., Özler B., 2002, How low can you go? Combining census and survey data for mapping poverty in South Africa, *Journal of African Economies*, 11(2), 169–200. https://doi.org/10.1093/jae/11.2.169

Anderson M., Fienberg S. E., 1996, An adjusted census in 1990: The Supreme Court decides, *Chance*, 9(3), 4–9. https://doi.org/10.1080/09332480.1996.10542491

Backer L., Holt Bloch V. V., 2011, *GEOSTAT 1A – Representing census data in a European population grid*, Kongsvinger, The European Forum for GeoStatistics.

Bondarenko M., Kerr D., Sorichetta A., Tatem A., 2020, *Census/projection-disaggregated gridded population datasets for 51 countries across sub-Saharan Africa in 2020 using building footprints*, University of Southampton.
https://doi.org/10.5258/SOTON/WP00682

Bondarenko M., Nieves J., Sorichetta A., Stevens F. R., Gaughan A. E. et al., 2018, *wpgpRFPMS: WorldPop random forests population modelling R scripts* (Version 0.1.0), University of Southampton.

Breiman L., 2001, Random forests, *Machine Learning*, 45(1), 5–32,
https://doi.org/10.1023/A:1010933404324

Buchhorn M., Lesiv M., Tsendbazar N.-E., Herold M., Bertels L. et al., 2020, Copernicus global land cover layers – collection 2, *Remote Sensing*, 12(6), 1044.
https://doi.org/10.3390/rs12061044

Buettner T., Garland P., 2008, *Preparing population estimates for all countries of the world: Experiences and challenges*, Rome meeting of Committee for the Coordination of Statistical Activities.

Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B. et al., 2017, Stan: A probabilistic programming language, *Journal of Statistical Software*, 76(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Carr-Hill R., 2014, Measuring development progress in Africa: The denominator problem, *Canadian Journal of Development Studies / Revue canadienne d'études du développement*, 35(1), 136–154. https://doi.org/10.1080/02255189.2014.884969

Checchi F., Stewart B. T., Palmer J. J., Grundy C., 2013, Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations, *International Journal of Health Geographics*, 12(art. 4).
https://doi.org/10.1186/1476-072X-12-4

Chilès J.-P., Delfiner P., 2009, *Geostatistics: Modeling spatial uncertainty*, New York, John Wiley & Sons.

Claeson C.-F., 1963, Co-ordinate system map of population distribution in Sweden 1960, *Geografiska annaler*, 45(4), 282–287.
https://doi.org/10.1080/20014422.1963.11881036

Deichmann U., Eklundh L., 1991, *Global digital datasets for land degradation studies: A GIS Approach* (GRID Case Study Series No. 4), Nairobi, GEMS.
https://wedocs.unep.org/20.500.11822/29349

Desrosières A., 2000, L'histoire de la statistique comme genre: style d'écriture et usages sociaux, *Genèses*, 39(2), 121–137. https://doi.org/10.3917/gen.039.0121

Dooley C., Jochem W., Sorichetta A., Lazar A., Tatem A. et al., 2021, *Description of methods for South Sudan 2020 gridded population estimates from census projections adjusted for displacement* (Version 2.0), University of Southampton. https://doi.org/10.5258/SOTON/WP00710

Ebolé Bola F. C., 2019, *Cameroun: le 4e* recensement de la population dans l'impasse, Agence de Presse Africaine.

Ecopia.AI, Maxar Technologies, 2019, *Digitize Africa data*.

Elith J., Leathwick J. R., 2009, Species distribution models: Ecological explanation and prediction across space and time, *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elvidge C. D., Baugh K., Zhizhin M., Hsu F. C., Ghosh T., 2017, VIIRS night-time lights, *International Journal of Remote Sensing*, 38(21), 5860–5879. https://doi.org/ 10.1080/01431161.2017.1342050

Eyinga Dimi E. C., 2019, *Du recensement classique au recensement numérique: l'expérience du Cameroun dans le cadre du 4e Recensement Général de la Population et de l'Habitat*, Yaoundé, Bureau central des recensements et des études de population.

Ferreira L. Z., Blumenberg C., Utazi C. E., Nilsen K., Hartwig F. P. et al., 2020, Geospatial estimation of reproductive, maternal, newborn and child health indicators: A systematic review of methodological aspects of studies based on household surveys, *International Journal of Health Geographics*, 19(1), 41. https://doi.org/10.1186/s12942-020-00239-9

Friedl M. A., McIver D. K., Hodges J. C., Zhang X. Y., Muchoney D. et al., 2002, Global land cover mapping from MODIS: Algorithms and early results, *Remote Sensing of Environment*, 83(1–2), 287–302. https://doi.org/10.1016/S0034-4257(02)00078-0

Friedman J. H., 2001, Greedy function approximation: A gradient boosting machine, *Annals of statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Guengant J.-P., Lankoande M., Tapsoba E., 2009, *Projections démographiques 2007-2050*, INSD, Ouagadougou.

Harris I., Jones P .D., Osborn T. J., Lister D. H., 2014, Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 dataset, *International Journal of Climatology*, 34(3), 623–642. https://doi.org/10.1002/joc.3711

Harris I., Osborn T. J., Jones P., Lister D., 2020, Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, *Scientific Data*, 7(1), 1–18. https://doi.org/10.1038/s41597-020-0453-3

Hillson R., Alejandre J. D., Jacobsen K. H., Ansumana R., Bockarie A. S. et al., 2014, Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: A case study of Bo City, Sierra Leone, Noor A. M.(ed.), *PLOS One*, 9(11). https://doi.org/ 10.1371/journal.pone.0112241

INSD, 2009, *Projections démographiques de 2007 à 2020 par région et province*, Ouagadougou, INSD.

INSD, 2017, *Projections démographiques des communes du Burkina Faso de 2007-2020*, Ouagadougou, INSD.

INSPIRE, 2014, *D2.8.I.2 data specification on geographical grid systems–Technical guidelines*, INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems.

Institut Géographique du Burkina Faso, 2015, *Base nationale de données topographiques*, Ouagadougou, IGB.

Institut National de la Statistique et de la Démographie, 2019, *Recensement général de la population et de l'habitation de 2019 du Burkina Faso – Résultats provisoires*, Ouagadougou, INSD.

Jhamba T., Juran S., Jones M., Snow R., 2020, UNFPA Strategy for the 2020 round of population and housing censuses (2015–2024), *Statistical Journal of the IAOS*, 36(1), 43–50. https://doi.org/10.3233/SJI-190600

Jochem W. C., Leasure D. R., Pannell O., Chamberlain H. R., Jones P. et al., 2021, Classifying settlement types from multi-scale spatial patterns of building footprints, *Environment and Planning B: Urban Analytics and City Science*, 48(5). https://doi.org/10.1177/2399808320921208

Kuhn M., 2008, Building predictive models in R using the caret package, *Journal of Statistical Software*, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Langford M., 1991, The areal interpolation problem: Estimating population using remote sensing in a GIS framework, in Masser I., Blakemoore M., *Handling geographical information: Methodology and potential applications*, New York, Longman, 55–77.

Lawson A. B., 2013, *Statistical methods in spatial epidemiology*, New York, John Wiley & Sons.

Leasure D. R., Jochem W. C., Weber E. M., Seaman V., Tatem A. J., 2020a, National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty, *Proceedings of the National Academy of Sciences*, 117(39), 24173–24179, https://doi.org/10.1073/pnas.1913050117

Leasure D. R., Tatem A. J., Bondarenko M., Darin E., 2020b, *wopr: An R package to query the WorldPop Open Population Repository, version 0.4.0*, WorldPop Research Group, University of Southampton.

Leyk S., Gaughan A. E., Adamo S. B., de Sherbinin A., Balk D. et al., 2019, The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use, *Earth System Science Data*, 11(3), 1385–1409. https://doi.org/10.5194/essd-11-1385-2019

Leyk S., Nagle N. N., Buttenfield B. P., 2013, Maximum entropy dasymetric modeling for demographic small area estimation, *Geographical Analysis*, 45(3), 285–306, https://doi.org/10.1111/gean.12011

Local Burden of Disease Educational Attainment Collaborators, 2020, Mapping disparities in education across low- and middle-income countries, *Nature*, 577(7789), 235–238, https://doi.org/10.1038/s41586-019-1872-1

Local Burden of Disease WaSH Collaborators, 2020, Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17, *The Lancet. Global Health*, 8(9), E1162–E1185. https://doi.org/10.1016/S2214-109X(20)30278-3

Longva S., Thomsen I., Severeide P. I., 1998, Reducing costs of censuses in Norway through use of administrative registers, *International Statistical Review / Revue Internationale de Statistique*, 66(2), 223–234. https://doi.org/10.2307/1403491

Martin D., 2011, Directions in population GIS, *Geography Compass*, 5(9), 655–665. https://doi.org/10.1111/j.1749-8198.2011.00440.x

Petrov A., 2012, One hundred years of dasymetric mapping: Back to the origin, *The Cartographic Journal*, 49(3), 256–264. https://doi.org/10.1179/1743277412Y.0000000001

R Core Team, 2020, *R: A language and environment for statistical computing*, Vienna, R Foundation for Statistical Computing.

Scrope G. P., 1833, *Principles of political economy*, London, Longman, Rees, Orme, Brown, Green, & Longman.

Sorichetta A., Hornby G. M., Stevens F .R., Gaughan A. E., Linard C. et al., 2015, High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020, *Scientific Data*, 2(1), 1–12. https://doi.org/10.1038/sdata.2015.45

Stevens F. R., Gaughan A. E., Linard C., Tatem A. J., 2015, Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data, *PLOS One*, 10(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

Stevens F. R., Gaughan A. E., Nieves J. J., King A., Sorichetta A. et al., 2020, Comparisons of two global built area land cover datasets in methods to disaggregate human population in eleven countries from the global South, *International Journal of Digital Earth*, 13(1), 78–100. https://doi.org/10.1080/17538947.2019.1633424

Tatem A. J., 2017, WorldPop, open data for spatial demography, *Scientific Data*, 4(1), 170004. https://doi.org/10.1038/sdata.2017.4

Thomson D. R., Rhoda D. A., Tatem A. J., Castro M. C., 2020, Gridded population survey sampling: A systematic scoping review of the field and strategic research agenda, *International Journal of Health Geographics*, 19(1), 34. https://doi.org/10.1186/s12942-020-00230-4

Tusting L. S., Bisanzio D., Alabaster G., Cameron E., Cibulskis R. et al., 2019, Mapping changes in housing in sub-Saharan Africa from 2000 to 2015, *Nature*, 568(7752), 391–394. https://doi.org/10.1038/s41586-019-1050-5

United Nations, 2017, *Principles and recommendations for population and housing censuses, Revision 3*, New York, United Nations.

United Nations Population Fund, 2017, *New Methodology: A hybrid census to generate spatially disaggregated population estimates*, New York, UNFPA.

United Nations Population Fund, 2019, *Technical guidance: Post enumeration surveys in population and housing censuses*, New York, UNFPA.

United Nations Population Fund, 2020, *The value of modelled population estimates for census planning and preparation*, New York, UNFPA.

Utazi C. E., Thorley J., Alegana V. A., Ferrari M. J., Takahashi S. et al., 2019, Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies, *Nature Communications*, 10(1), 1–10. https://doi.org/10.1038/s41467-019-09611-1

Wardrop N. A., Jochem W. C., Bird T. J., Chamberlain H. R., Clarke D. et al., 2018, Spatially disaggregated population estimates in the absence of national population and housing census data, *Proceedings of the National Academy of Sciences*, 104(14), 3529–3537. https://doi.org/10.1073/pnas.1715305115

Weber E. M., Seaman V. Y., Stewart R. N., Bird T. J., Tatem A. J. et al., 2018, Census-independent population mapping in northern Nigeria, *Remote Sensing of Environment*, 204, 786–798. https://doi.org/10.1016/j.rse.2017.09.024

Weiss D. J., Nelson A., Gibson H., Temperley W., Peedell S. et al., 2018, A global map of travel time to cities to assess inequalities in accessibility in 2015, *Nature*, 553(7688), 333–336. https://doi.org/10.1038/nature25181

WorldPop Research Group, Department of Geography and Geosciences, University of Louisville, Département de Géographie, Université de Namur, Center for International Earth Science Information Network (CIESIN), Columbia University, 2018, *Global High Resolution Population Denominators Project – Funded by The Bill and Melinda Gates Foundation (OPP1134076)*, University of Southampton.

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**

The size of the population, the denominator of many statistical indicators, is crucial for public policy. National statistical offices organize the collection of this information, most often through a census. But what happens when parts of a country are not accessible to census enumerators? Today, spatial data extracted from satellite imagery offer high-resolution geographical information with complete coverage. When combined with a partial population count, they offer an unprecedented opportunity to estimate the size of the population in inaccessible areas. The spatial precision of these data also makes possible the production of a high-resolution gridded population estimate, an innovative data format at the intersection of geography and demography. Based on the case of Burkina Faso, this article analyses how, by dividing a country into 100 m by 100 m cells, a Bayesian hierarchical model can be used to estimate the population of areas with security challenges which could not be enumerated during the 2019 census. This gridding allows the resulting counts to be disaggregated using a statistical learning model, yielding unparalleled spatial precision in population estimates.

Edith Darin, Mathias Kuépié, Hervé Bassinga, Gianluca Boo, Andrew J. Tatem • **La population vue du ciel : quand l'imagerie satellite vient au secours du recensement**

Le dénombrement de la population, dénominateur de nombreux indicateurs statistiques, est crucial pour les politiques publiques d'un pays. Il est du ressort des instituts nationaux de statistique d'en organiser la collecte, le plus souvent par le biais d'un recensement. Que se passe-t-il lorsqu'une partie du territoire n'est pas accessible aux agents recenseurs ? Actuellement, les données spatiales, telles qu'extraites de l'imagerie satellite, offrent une information géographique complète et de haute résolution, qui représente, lorsque combinée à un dénombrement partiel de la population, une opportunité sans précédent pour estimer les effectifs des territoires manquants. Leur précision spatiale rend également possible une estimation carroyée de la population en haute résolution, un format de données innovant à la croisée de la géographie et de la démographie. À partir du cas du Burkina Faso, cet article analyse comment le découpage du pays en carreaux de 100m sur 100m permet dans un premier temps de développer un modèle pour estimer, par le biais d'une approche hiérarchique bayésienne, la population des zones caractérisées par des problèmes sécuritaires n'ayant pas pu être dénombrées lors du dernier recensement de 2019. Ce découpage permet dans un second temps de désagréger les effectifs obtenus, par le biais d'un modèle d'apprentissage statistique pour obtenir une précision spatiale d'estimation de la population inégalée.

Edith Darin, Mathias Kuépié, Hervé Bassinga, Gianluca Boo, Andrew J. Tatem • **La población vista desde el cielo: la imaginería satélite al rescate del censo**

El recuento de la población, denominador de numerosos indicadores estadísticos, es crucial para las políticas públicas de un país. Corresponde a los institutos nacionales de estadística el organizar la recogida de datos, la mayoría de las veces a través de un censo. ¿Qué ocurre cuando una parte del territorio no es accesible para los agentes encargados de elaborar el censo? Hoy día, los datos espaciales, tal y como se extraen de las imágenes satélite, proporcionan una información geográfica completa y de alta resolución que, al ser combinada con un recuento parcial de la población, representa una oportunidad sin precedentes para estimar los efectivos de los territorios inaccesibles. La precisión espacial hace igualmente posible una estimación por cuadrantes de la población en alta resolución, un formato de datos innovador a caballo entre la geografía y la demografía. A partir del estudio de Burkina Faso, este artículo analiza cómo la división del país en cuadrantes de 100 m por 100 m permite desarrollar en una primera fase un modelo para estimar, mediante una aproximación jerárquica bayesiana, la población de las zonas con problemas de seguridad de las que no pudo hacerse el recuento en el último censo de 2019. En una segunda fase, el recuento permite desagregar los efectivos obtenidos, mediante un modelo de aprendizaje estadístico, para obtener una precisión espacial de estimación de la población inigualable.

Keywords: gridded population, geospatial data, census, hierarchical model, Bayesian statistics, building footprint, remote sensing, Burkina Faso

Translated by Paul Reeve