Integrating multiple household-based data with satellite-based observations to produce high-resolution population estimates in LMICs using INLA-SPDE

Chibuzor Christopher Nnanatu[1,3], Ortis Yankey[1], Anaclet Désiré Dzossa[2], Thomas Abbott[1], Assane Gadiaga[1], Attila Lazar[1], Andrew J Tatem[1]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, SO17 1BJ, UK

[2]National Institute of Statistics (NIS)- Cameroon

[3]Nnamdi Azikiwe University, Awka-Nigeria

Corresponding Author's email: cc.nnanatu@soton.ac.uk

Background

Census projections which rely on datasets from national population and housing censuses provide population data required for policymaking and implementation across various countries including low- and middle-income countries (LMICs) (UNFPA 2020). However, censuses are often constrained by paucity of resources which means that census could be delayed longer than the usual ten (10) years intervals in some countries and this has been further exacerbated by the recent COVID-19 pandemic. In some contexts where there are significant changes in population size and distribution due to heterogeneity in migration, fertility and mortality patterns, census projections could easily become outdated and misleading thus requiring alternative sources of population numbers within the intercensal period (Tatem, 2022).

Modelled population estimates which integrate population data (e.g., Microcensus, household survey) with satellite-based settlement data (e.g., building footprints) and geospatial covariates (e.g., nighttime lights) using advanced statistical models, provide high-resolution population data. The datasets which are often available in raster formats at a very fine spatial scale usually 100m by 100m offer the flexibility to obtain population counts at any small area unit of interest. Thus, modelled population estimates are alternative sources of population count where census projections are outdated or lacked the level of granularity required for supporting decision-making, equitable resource allocation, disasters response, disease surveillance, healthcare interventions, and planning of censuses and elections (UNFPA 2020).

Materials and Methods

Within the context of 'bottom-up' population modelling (e.g., Leasure et al., 2020; Boo et al., 2022; Darin et al, 2022), input population data are combined with a stack of geospatial covariates and satellite-observed settlement data such as building count or built-up area to produce estimate of population density. Population predictions are then produced at grid cell level while aggregation to the desired areal unit is easily.

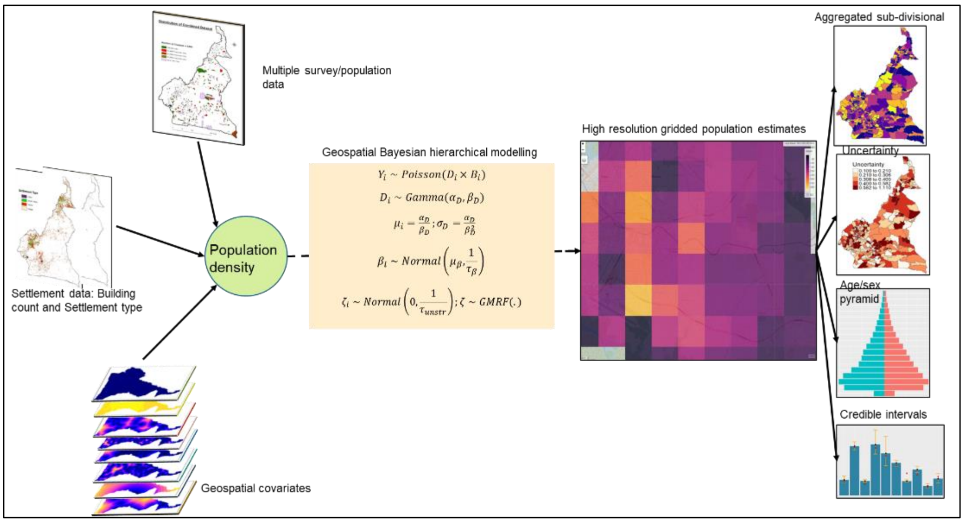
Multiple survey/population data
Settlement data: Building count and Settlement type
Geospatial covariates
Population density
Geospatial Bayesian hierarchical modelling
$Y_i \sim Poisson(D_i \times B_i)$
$D_i \sim Gamma(\alpha_D, \beta_D)$
$\mu_i = \frac{\alpha_D}{\beta_D}; \sigma_D = \frac{\alpha_D}{\beta_D^2}$
$\beta_i \sim Normal\left(\mu_\beta, \frac{1}{\tau_\beta}\right)$
$\zeta_i \sim Normal\left(0, \frac{1}{\tau_{unstr}}\right); \zeta \sim GMRF(.)$
High resolution gridded population estimates
Aggregated sub-divisional
Uncertainty
Age/sex pyramid
Credible intervals

Figure 1 Schematic representation of the key components of bottom-up population modelling and estimation.

Moreover, bottom-up population modelling uses Bayesian hierarchical regression modelling framework to account for the variabilities in the observed input population data due to complex survey design and heterogeneity in population density and distribution. The Bayesian inference approach also means that the quantification of uncertainties in parameter estimates is straightforward and implemented via the 95% credible interval. However, the potential effects of spatial autocorrelation within the observed data on the model estimates is yet to be explored. Thus, here we seek to fill this gap by developing a robust statistical population modelling technique based on the integrated nested Laplace approximation and stochastic partial differential equations frameworks (INLA-SPDE, Rue et al 2009; Lindgren et al, 2011). The INLA-SPDE approach offers a fast and efficient approach for including spatial autocorrelation defined through a triangulation of the study domain or mesh.

Simulation study

We carried out a simulation study designed to assess the accuracy of the population estimates over different number of small areal units versus survey coverage combinations. Specifically, how well can we estimate the population numbers as the sample size and coverage proportions decrease? To explore this, first, for a given grid cell g, we simulated the grid cell-level data (population $p_{ig} \sim Poisson(\lambda_g)$ and building $B_{ig} \sim Poisson(\bar{B}_g)$ counts) for the entire study area on a regular unit square and then aggregate to area units i such that $p_i = \sum_g p_{ig}$ and $B_i = \sum_g B_{ig}$ are the areal-level population and building counts respectively. Then the population density $D_k = p_k/B_k$ is assumed to follow Gamma distribution with parameters α_k and β_k ($D_k \sim Gamma(\alpha_k, \beta_k)$) with mean $\mu_k = \alpha_k/\beta_k$ and variance $\phi_k = \alpha_k/\beta_k^2$, where k is a generic index representing either a grid cell or an aerial unit. Note that here, area units are subnational units of model training which could be a local government area, enumeration area, etc. The mean population density μ_k is then assumed to relate to a set of geospatial covariates and spatially varying random effects through the link function defined in equation (1)

$$log\ (\mu_k) = \beta_0 + \sum_{J=1}^{J} \beta_j x_{ij} + \xi(s_k) + \zeta_k \qquad (1)$$

where β_0 is the intercept parameter which represents the baseline (average) population density when the effect of the other predictors is zero; $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$ is a vector of unknown fixed effects coefficients of the K (linear) geospatial covariates that could predict the distribution of the population density such as nighttime lights, distance from healthcare facilities, schools, ζ_k is the IID random effects and $\xi(s_k)$ is the spatially varying random effects for capturing spatial autocorrelation. The model parameters are separately trained at grid cell and aggregated areal levels. They are then used to predict population numbers at regular grid cells. Here, for clarity, grid cell predictions based on grid unit trained models are called Grid to Grid (or Grid2Grid), while grid cell predictions based on models trained at the areal units are called Area to Grid (or Area2Grid). We compared the accuracy of the predicted population numbers based on Grid2Grid and Area2Grid models using a constellation of the model fit indices in Figure 2. The results indicate that our methodology was robust over different number of areal unit-observation coverage combinations with more accurate predictions over larger survey coverages.

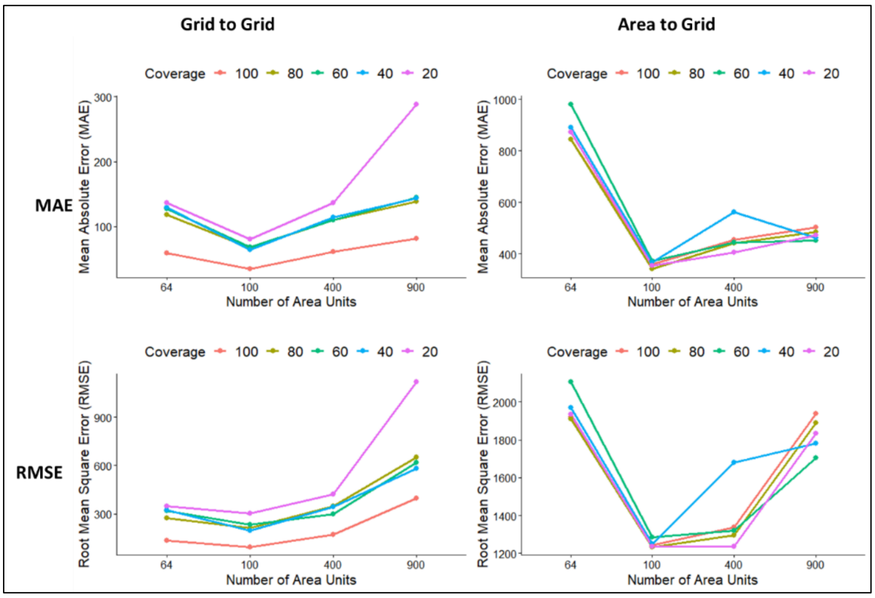

Figure 2. Estimates of the MAE and RMSE values of the models across various levels of missing values for Grid2Grid and Area2Grid models

Application to Household-based data in Cameroon

Population data provided by the household listing datasets from five nationally representative household-based surveys were used to produce small area population estimates in Cameroon, as a proof of concept (Figure 3). These model input datasets along with the administrative boundaries/ shapefiles were obtained from the Cameroon National Institute for Statistics (NIS). The data were collected between 2021 and 2022 based on a 2-stage stratified sampling design across 2290 Enumeration Areas (EAs) with a total of 2587569 individuals in 509628 households after data cleaning.

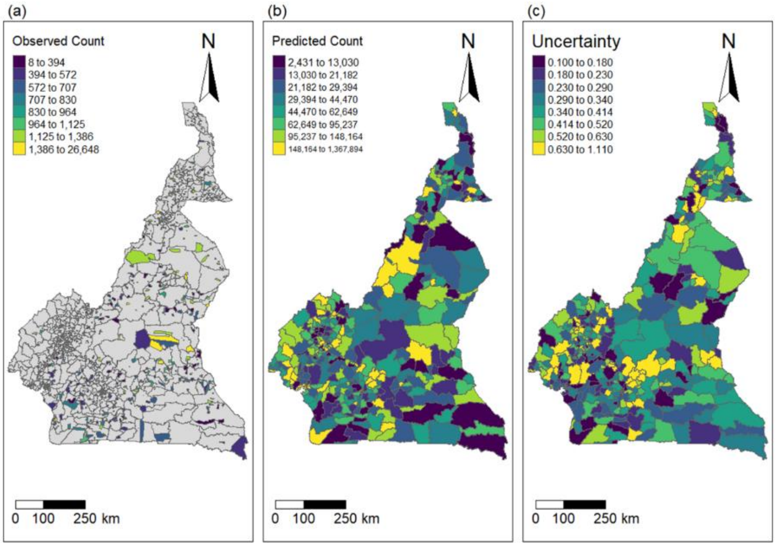

Figure 3. Spatial distribution of the observed counts predicted counts and the uncertainty across Admin level 3 in Cameroon. Uncertainty = (Upper - Lower)/Mean

References

UNFPA. (2020). “The value of modelled population estimates for census planning and preparation.” Technical Guidance Note, August 2020 (updated version 2). https://www.unfpa.org/resources/value-modelled-population-estimates-census-planning-and-preparation.

Tatem, A.J. (2022). Small area population denominators for improved disease surveillance and response. *Epidemics*, 41. https://doi.org/10.1016/j.epidem.2022.100641

Leasure, D. R., W. C. Jochem, E. M. Weber, V. Seaman and A. J. Tatem (2020). "National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty." *Proceedings of the National Academy of Sciences”*: 201913050. DOI: 10.1073/pnas.1913050117. https://www.pnas.org/doi/pdf/10.1073/pnas.1913050117

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lazar, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., Tatem, A. J. (2022). “High-resolution population estimation using household survey data and building footprints.” *Nature Communications*, 13, 1330. https://doi.org/10.1038/s41467-022-29094-x

Darin, E., M. Kuépié, H. Bassinga, G. Boo and A. J. Tatem (2022). "La population vue du ciel : quand l’imagerie satellite vient au secours du recensement." *Population* (french edition) **77**(3): 467-494

WorldPop and Institut National de la Statistique et de la Démographie du Burkina Faso. (2022). Census- based gridded population estimates for Burkina Faso (2019), version 1.1. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00736.

Dooley, C. A., Leasure, D.R., Boo, G. and Tatem, A.J. 2021. Gridded maps of building patterns throughout sub-Saharan Africa, version 2.0. University of Southampton: Southampton, UK. Source of building footprints “Ecopia Vector Maps Powered by Maxar Satellite Imagery”© 2020/2021. doi:10.5258/SOTON/WP00712.

Wardrop N.A., Jochem W.C., Bird T.J., Chamberlain H.R., Clarke D., Kerr D., Bengtsson L., Juran S., Seaman V., Tatem A.J. (2018). “Spatially disaggregated population estimates in the absence of national population and housing census data.” *Proceedings of the National Academy of Sciences 115*, 3529–3537. https://www.pnas.org/doi/10.1073/pnas.1715305115

Rue, Havard, Sara Martino, and Nicolas Chopin. (2009). “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society*, Series B 71 (2): 319–92

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 73(4), 423–498

National Institute of Statistics (2016). Projections Démographiques et Estimations des Cibles Prioritaires des Différents Programmes et Interventions de Santé. Ministère de la Santé Publique, Cameroon, June 2016. 144 pages. https://ins-cameroun.cm/en/statistique/projections-demographiques-et-estimations-des-cibles-prioritaires-des-differents-programmes-et-interventions-de-sante/