Small area population estimation from health intervention campaign surveys and partially observed settlement data

Chibuzor Christopher Nnanatu[1,2]*, Amy Bonnie[1], Josiah Joseph[3], Ortis Yankey[1], Duygu Cihan[1], Assane Gadiaga[1], Hal Voepel[1], Thomas Abbott[1], Heather Chamberlain[1], Mercedita Tia[4], Marielle Sander[4], Justin Davis[5], Attila Lazar[1] and Andrew J. Tatem[1]

1.WorldPop, School of Geography and Environmental Science, University of Southampton, SO17 1BJ, UK; 2. Nnamdi Azikiwe University, Nigeria;
4. National Statistical Office, Papua New Guinea; 5.United Nations Population Fund, Papua New Guinea; 5.Planet Labs, San Francisco, USA

Background

Recent and reliable small area population numbers are required for effective governance, but financial and logistical challenges mean that national censuses are typically only undertaken every ten years or more. Geospatial modelling approaches have been developed that utilise bespoke microcensus surveys linked with satellite-derived settlement maps and other spatial datasets to fill population data gaps across countries with outdated or incomplete census data. However, microcensus surveys can be complex logistically and expensive, while satellite-based settlement maps can often be incomplete in tropical rural areas where tree canopies and cloud cover can obscure them. These factors limit the wider application of geospatial modelling approaches. Here, we present a novel two-step Bayesian hierarchical modelling approach that can integrate routinely collected health intervention campaign data and partially observed settlement data to produce reliable small area population estimates. Reductions in relative error rates were 32-73% in a simulation study, and ~32% when applied to malaria survey data in Papua New Guinea. The results highlight the value of demographic data that is collected routinely through health intervention campaigns or household surveys for improving small area population estimates, and how biases introduced through satellite data limitations can be overcome.

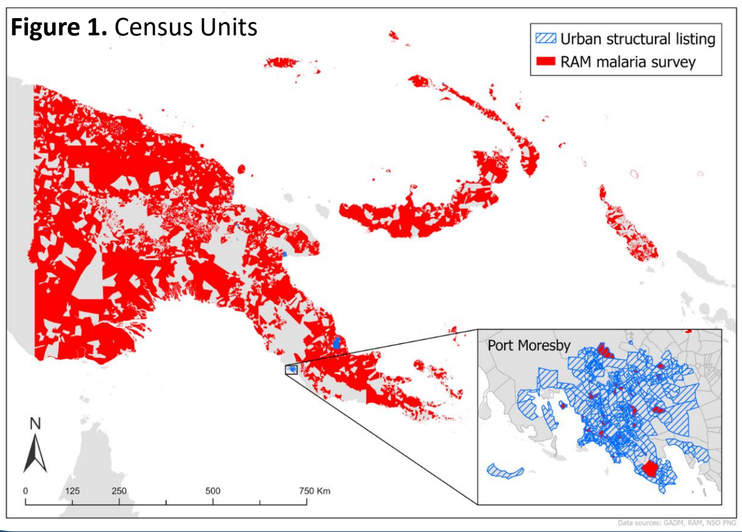

Figure 1. Census Units

Data Description

Altogether, data were available for 16,903 census units (CU) out of the 32,100 CUs in PNG. The input datasets which provided the household counts used for the population models included nationally representative Malaria Long Lasting Insecticidal Net (LLIN) Survey (2019 to 2021), and Urban Listing (UL) datasets at census units level in Papua New Guinea (Figure 1). Satellite-based settlement data and the geospatial covariates raster files were sourced from Planet (www.planet.org) and WorldPop (www.worldpop.org), respectively. However, there are indications that buildings under tree canopy covers were missed by the satellite (Figure 2).

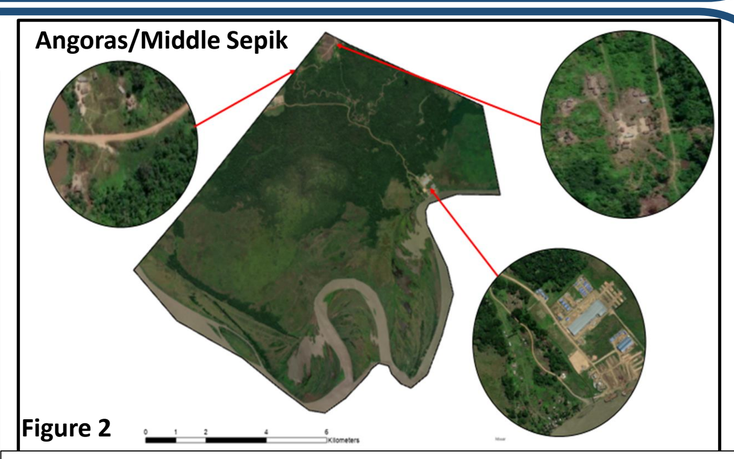

Figure 2

Malaria survey recorded a disproportionately large 891 people compared to the satellite-observed settlements.

Methodology

We developed a novel two-step approach based on a Bayesian hierarchical geostatistical modelling framework which first adjusts for the potential biases in the partially observed satellite-based settlement data, and then uses the bias-adjusted data to calculate estimates of population numbers in the second step (Figure 3). At each step, models were implemented using the Integrated Nested Laplace Approximation (INLA) modelling framework [4] using R statistical programming software.

To test the efficacy of our methodology, we first carried out an extensive simulation study where we examined how accurately our method was able to estimate population numbers under different proportions of satellite-based settlement data observation coverages (65%, 70%, 75%, 80%, 85%, 90%, 95%, 100%) versus population data observation coverage (20%, 40%, 60%, 80%, 100%) combinations. Afterwards, we applied the methodology to estimate the subnational population estimates of PNG.

Model fit indices and model cross validation relied on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute Bias, and the correlation between the observed and the predicted population count.

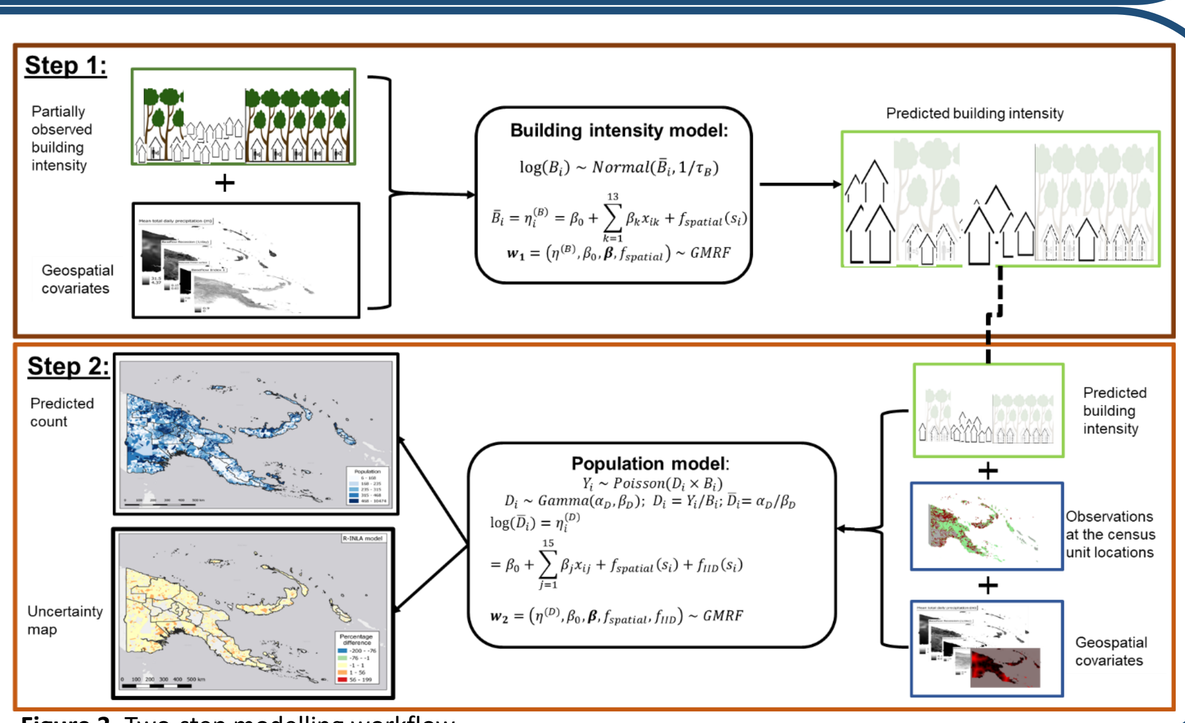

Figure 3. Two-step modelling workflow

Results and Discussion

Results from the simulation study indicated that the two-step Bayesian hierarchical model (TSBHM) outperformed the contemporary BHM approach across all the model fit metrics (Figure 4) and brought about ~63% and ~88% reduction in relative bias. When applied to estimate population numbers at subnational scales in PNG, the two-step model solution reduced relative bias by ~33% leading to more precise estimates and accurate predictions. The population data for which the total national population estimates are given in Table 1, are now published on the website of the national statistical office of Papua New Guinea - https://www.nso.gov.pg/statistics/population/. Figure 5 shows the spatial distribution of the estimates of uncertainty in the predicted counts of people at the census unit level.

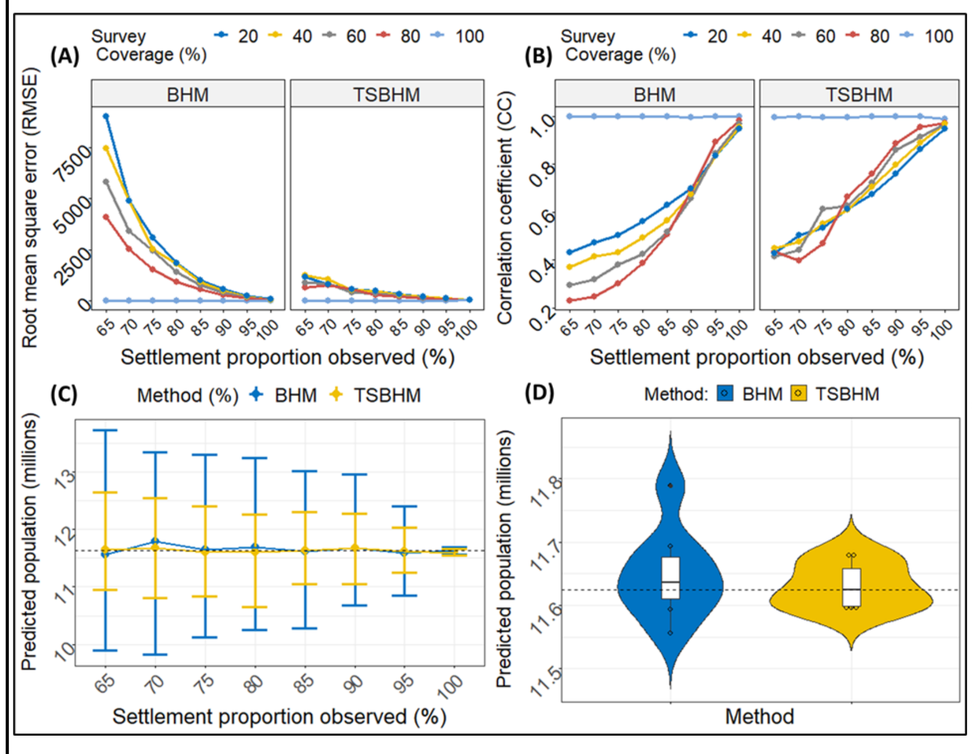

Figure 4. Model fit metrics from the simulation study

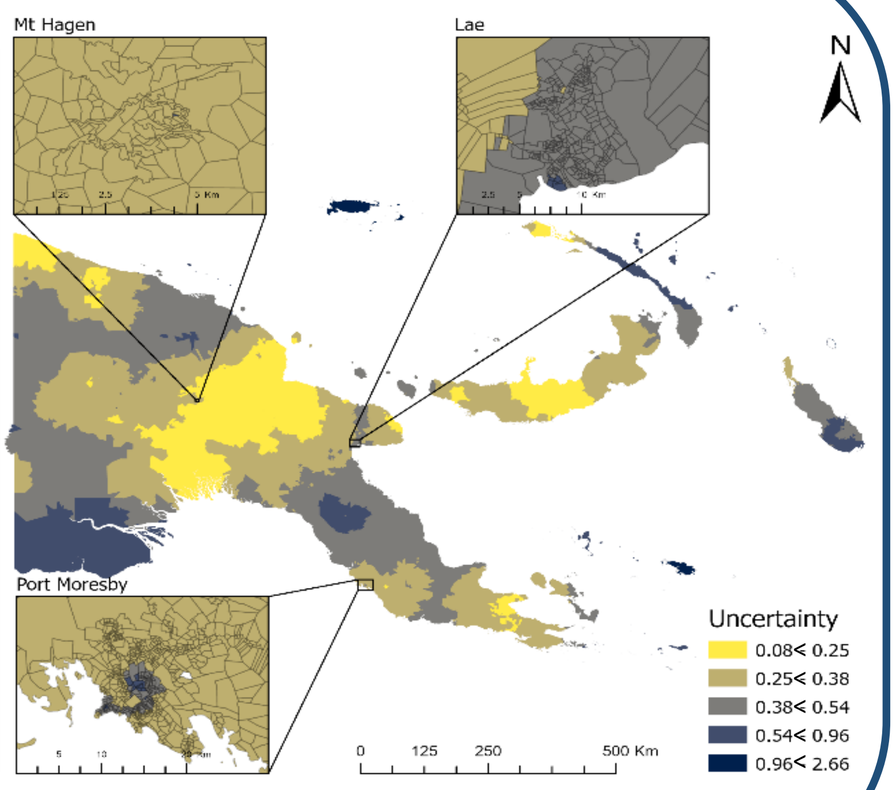

Figure 5. Estimates of uncertainty of the predicted census unit level population counts

Bibliography

[1] UNFPA, The Value of Modelled Population Estimates for Census Planning and Preparation. 2020b, UNFPA: New York, USA.

[2] 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

[3] Yang, X., Yang, S., Tan, M.L., Pan, H., Zhang, H., Wang, G., Wang, Z., 2022. Correcting the bias of daily satellite precipitation estimates in tropical regions using deep neural network. J. Hydrol. 608, 127656. (Bi-LSTM - bidirectional long short-term memory recurrent network)

[4] 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. https://www.unfpa.org/resources/value-modelled-population-estimates-census-planning-and-preparation

[5] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.