A Two-Step Bayesian Hierarchical Modelling Approach for Estimating Population Density when Settlement Data are Partially Observed

Chibuzor Chris Nnanatu[1,4]*, Amy Bonnie[1], Josiah Joseph[2], Ortis Yankey[1], Duygu Cihan[1], Assane Gadiaga[1], Mercedita Tia[3], Marielle Sander[3], Attila Lazar[1], Andrew Tatem[1]

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

[2]National Statistical Office, Papua New Guinea

[3]United Nations Population Fund, Papua New Guinea

[4]Nnamdi Azikiwe University, Nigeria

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

Background

The growing demand for reliable population estimates at small area units to address several developmental and humanitarian needs necessitates the development of more robust statistical solutions. Model-based population estimation methods (PEMs) such as the bottom-up or top-down geospatial population modelling methods (Stevens et al, 2015; Leasure et al., 2020) provide estimates of population at finer spatial resolutions than census projections, using multiple geospatial data sources and advanced statistical modelling techniques (UNFPA, 2020). These often serve to provide more rapid and up-to-date population estimates using recent datasets (UNFPA, 2020), thereby taking any recent changes in population dynamics into account. Specifically, the bottom-up model leverages advances in satellite-imagery and improved computing efficiency to integrate recent satellite-dependent settlement datasets and other key geospatial covariates with sample population data (e.g., surveys) to predict estimates of population numbers across entire countries at small area scales (usually 100m by 100m grid cells) including in unsampled or partially observed locations (Leasure et al., 2020). However, satellite-based datasets are often only partially observed due to factors such as tree canopy and cloud covers meaning that the settlement data can be systematically biased (Moazami et al., 2022; Yang et al., 2022). When these biases are not identified and corrected for, they can be propagated into parameter estimation, potentially leading to biased population estimates and inaccurate predictions.

This study was motivated by the lack of census enumeration in Papua New Guinea since 2011, and the need to estimate the population numbers, demographics and distributions at small area scales. This was undertaken by integrating nationally representative datasets from Malaria survey and Urban Listing data with satellite-observed settlement data and other geospatial covariates. One key potential modelling challenge here is that many of the rural populations in PNG are under tree canopy cover, and as a result, the available satellite-based settlement data only partially observed the full set of rural structures, thus, there was need to develop an approach that ensures minimum biases in the population estimates. This paper presents a novel two-step approach based on robust Bayesian hierarchical geostatistical modelling framework which adjusts for the potential biases in the partially observed satellite-based settlement data in the first step, and then uses the bias-adjusted data to calculate estimates of population numbers in the second step.

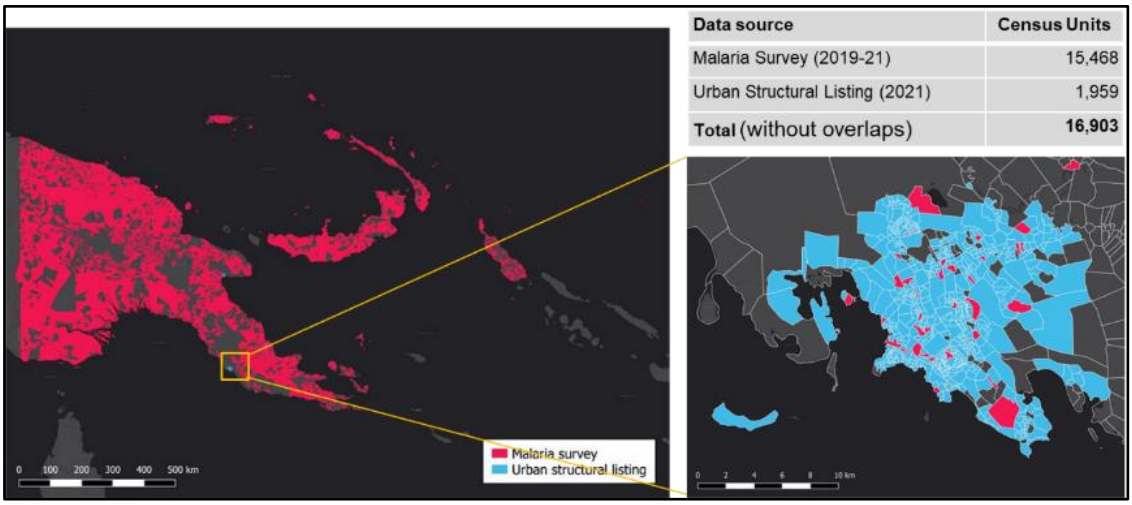

Figure 1. Map of Papua New Guinea showing the Census Units across the country where recent datasets from the Malaria LLIN and UL are available as well as where they overlap.

Results from a simulation study carried out under various survey coverage versus satellite observation coverage scenarios show that our methodology caused between ~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. Our methodology provides an important model-based bias correction framework within the Bayesian hierarchical population modelling contexts which could be easily adapted in other settings.

Data and Methods

This study is motivated by Malaria Long Lasting Insecticidal Net (LLIN) Survey of 2019 to 2021, and 2021 Urban Listing (UL) datasets which are both nationally representative at census units level in Papua New Guinea (Figure 1). The datasets contain information on the household counts which would serve as input population data for our models. Altogether, data were available for 16,903 census units (CU) out of the 32,100 CUs that are spread across the country.

A schematic representation of the 2-step method modelling workflow is provided in Figure 2. In step 2, potential biases in satellite-observed building intensity (a proxy for settlement intensity) was corrected for using a robust Bayesian hierarchical regression modelling framework. In the second step, the bias-corrected building intensity was integrated with key geospatial covariates and the observed population datasets to calculate estimates of population numbers at census unit levels. The satellite data and the geospatial covariates raster files were sourced from Planet (www.planet.org) and WorldPop (www.worldpop.org), respectively. The settlement data from Planet which came in gridded format at ~4.77m spatial resolution is an AI/machine-learning derived classification of satellite imagery where buildings/settlement structures have been detected. As a result, human settlement structures that are under tree canopy or cloud covers will be missed, thereby, posing a data challenge.

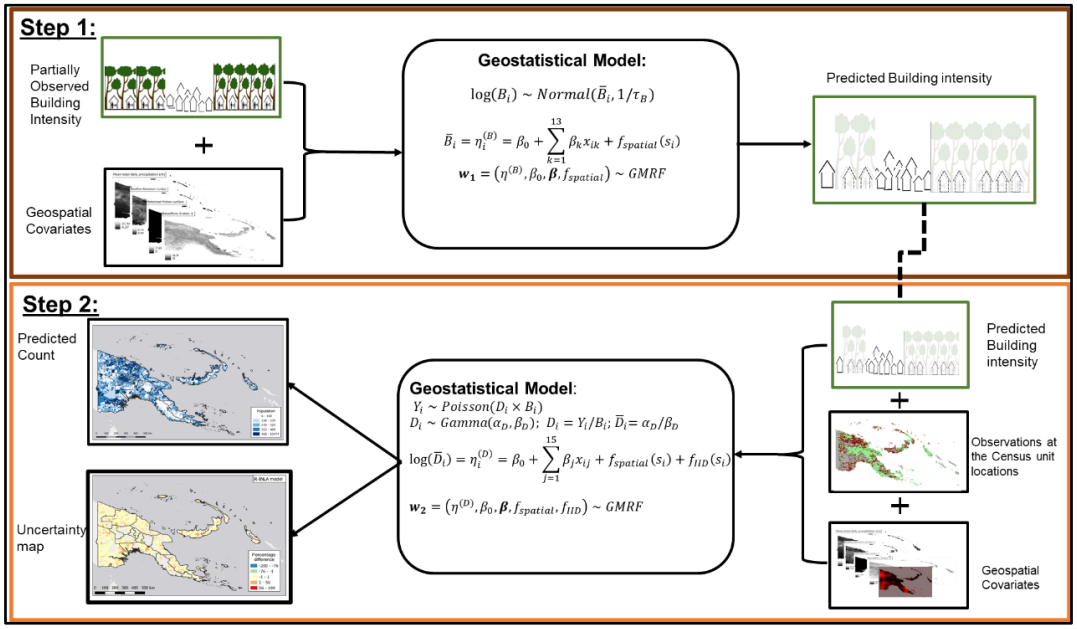

Figure 2. Two-Step Hierarchical Regression Modelling workflow

As shown in Figure 2, the partially observed settlement data (building intensity) is the outcome variable in step 1 allowing us to predict estimates of building intensity at locations covered by tree canopies and cloud. By assuming that observations within neighbouring locations share common characteristics and are more similar than those further apart at both steps one and two, our methodology accounted for spatial autocorrelation within the data which allowed us to borrow strengths from contiguous locations with more observations to predict observations in areas with little or no observations (Leasure et al., 2020). We used R statistical programming language as the implementation software while model parameter inference was based on Bayesian statistical inference via the integrated nested Laplace approximations in conjunction with the stochastic partial differential equations (INLA-SPDE; Rue et al., 2009). The INLA-SPDE approach is faster and more accurate when compared with sampling-based methods such as the Markov Chain Monte Carlo (MCMC) techniques which are known to be computationally expensive for multi-dimensional datasets and model results are not usually

repeatable. In addition, the use of Bayesian inference approach meant that prior knowledge on the parameter values can be incorporated and that uncertainties in parameters estimations are readily quantifiable. These estimates of uncertainties play a vital role in the design and implementation of national policies targeted at small area units (UNFPA, 2020).

Simulation Study

We carried out an extensive simulation study to explore the performance of our proposed methodology under different scenarios. Specifically, we considered different percentage survey coverage ($p\%$) versus percentage satellite observation coverage ($b\%$) permutations with p and b taking values from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\{0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00\}$, respectively. For example, one of the combinations is when population data for the 60% of the population is available ($p = 0.6$) versus satellite data coverage of only 85% of the population ($b = 0.85$). Thus, in all, 40 different datasets of these permutations were simulated and tested. Note that $p = 1.0$ implies 100% observation of every individual within the population, that is, census. Similarly, when $b = 1.00$, it means that the satellite-based settlement data was fully (100%) observed and assumed bias free, thus, may not require any bias correction. For each dataset, we first fitted normal Bayesian hierarchical model (BHM) without first correcting for the potential biases in the settlement data. Next, we first corrected the settlement data biases via the two-step Bayesian hierarchical modelling (TSBHM) approach using the same datasets and evaluated model performances using a suit of statistical modelling fit indices. In particular, we tested and compared the model performances and predictive abilities using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute Bias (AB), Pearson's Correlation (PC), and Relative Error Rate (RER). Apart from the PC in which higher values indicate better fit, smaller values based on the other fit metrics indicate a better fit model.

Findings

Simulation study results indicate that the model which first adjusted for the potential biases in the settlement data provided a better fit than the model which ignored these biases. Specifically, Figure 3 shows that the TSBHM consistently provided lower MAE, RMSE, AB, and Normalized RER than the BHM models. In addition, the TSBHM more accurately predicted estimates of population than the BHM as can be shown with higher PC values for TSBHM than for BHM.

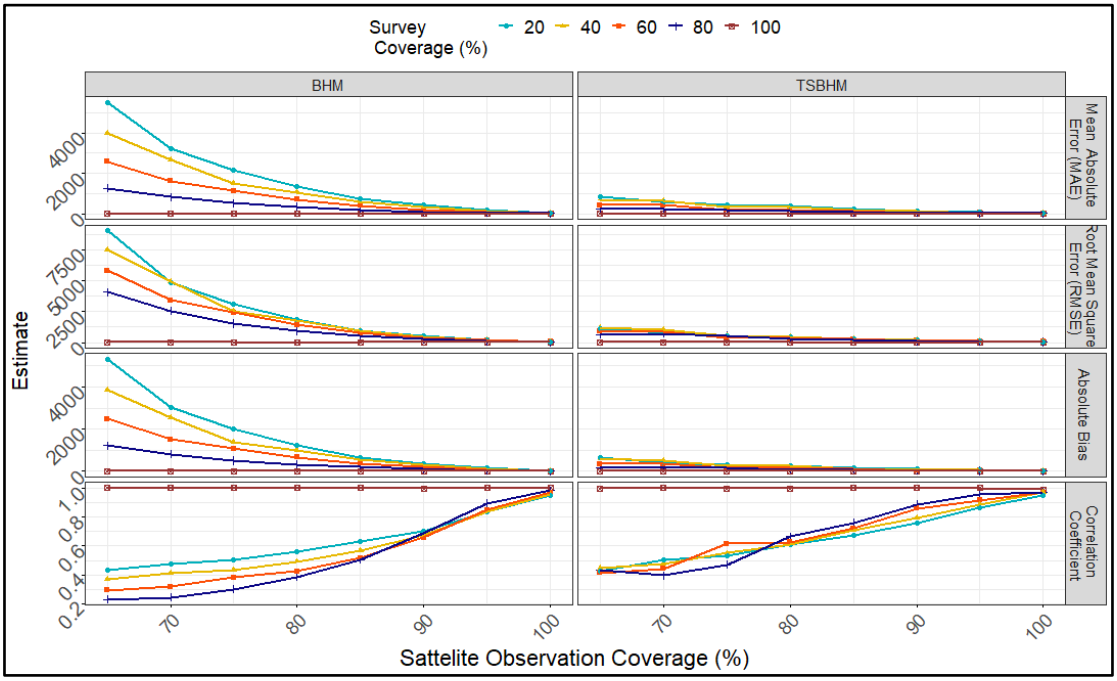

Figure 3. Comparisons in model fit indices of TSBHM and BHM.

Furthermore, Figure 4 shows the variations in percentage reductions in relative biases when the TSBHM is implemented. Percentage reduction in relative biases increased as the proportion of partially observed data increased but varied between 63% and 88% thus indicating that our methodology is very robust and works well as expected.

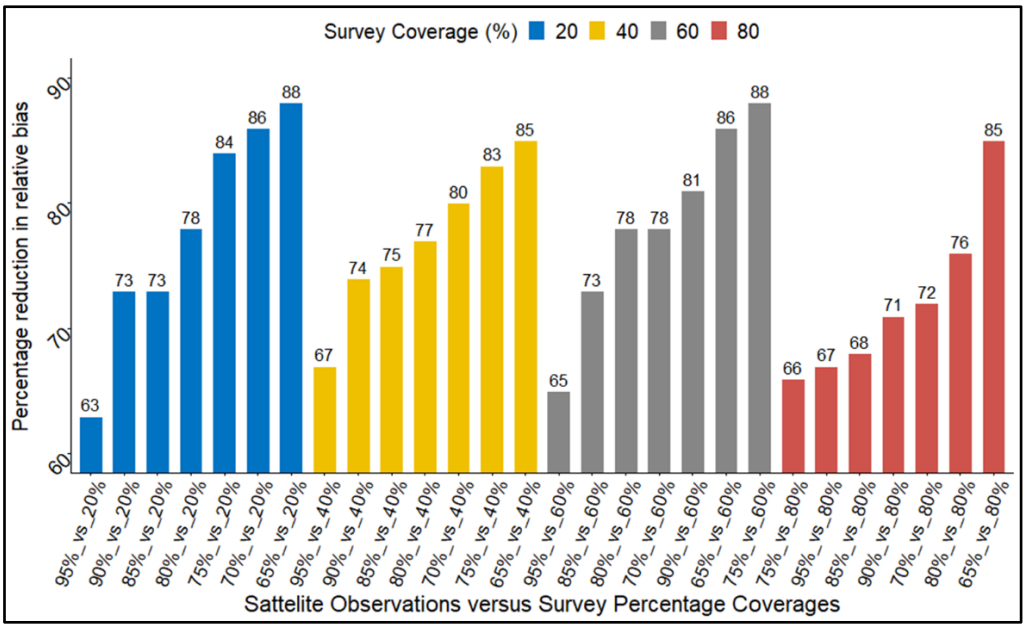

Figure 4. Percentage Reduction in Relative Bias when TSBHM is used. Note that 95%_vs_20% in the x-axis ticks label represents dataset obtained from 95% satellite coverage when only 20% of the observations was observed. Percentage reduction in relative bias increased as the proportion of partially observed satellite-based settlement data increased.

When applied to the PNG datasets, the two-step modelling approach (TSBHM) consistently outperformed the conventional BHM method (Table 1) and reduced biases by approximately 33%.

Table 1 Model fit metrics comparison between BHM and TSBHM.

Model	MAE	RMSE	AB
BHM	3.19	5.73	3.14
TSBHM	2.25	4.21	2.10

These findings reiterate the importance of accounting for potential biases while using satellite-based observations to ensure improved parameter estimates and more accurate predictions. Future research would explore the applications of our methodology in other contexts.

References

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.

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

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

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

UNFPA, The Value of Modelled Population Estimates for Census Planning and Preparation. 2020b, UNFPA: New York, USA. https://www.unfpa.org/resources/value-modelled-population-estimates-census-planning-and-preparation

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)

Moazami, S., Na, W., Najafi, M. R., de Souza, C. (2022). Spatiotemporal bias adjustment of IMERG satellite precipitation data across Canada. Advances in Water Resources, 168 (104300).