Release Statement

Modelled Population Estimates for Papua New Guinea, version 1.0

27 July 2023

Original Release: 27 July 2023

ABSTRACT

This project was initiated in 2021 to generate modelled population estimates for Papua New Guinea (PNG) to support their census preparations. It was powered by the Australian Government through the PNGAus partnership, the United Nations Population Fund (UNFPA) and the PNG National Statistical Office.

The project team combined recent 2019-2021 malaria bednet campaign data, urban structural listing 2021 data, and geospatial covariates to model and estimate population numbers at census unit level, and aggregate at other relevant administrative units (e.g., national, province, and districts) using a Bayesian statistical hierarchical modelling framework. The approach facilitated simultaneous accounting for the multiple levels of variability within the data hierarchy. It also allowed the quantification of uncertainties in parameter estimates.

These model-based population estimates can be considered as most accurately representing the years 2020-21. This time period corresponds to the malaria survey and urban structural listing survey observations (2019-2021; median year: 2020) and the period of the satellite imagery used to generate settlement footprints (2021). Although the methods were robust enough to explicitly account for key random biases within the datasets, it is noted that systematic biases, which may arise from sources other than random errors within the observed data collection process, are most likely to remain.

These data were produced by the WorldPop Research Group at the University of Southampton in collaboration with the National Statistical Office of PNG and UNFPA under the project called “Population-modelled estimation for Papua New Guinea in collaboration with the National Statistical Office, 2021-22” (PNG40-0000004504). The final statistical modelling was designed, developed, and implemented by Chris Nnanatu. Data processing was done by Amy Bonnie with additional support from Tom Abbott, Tom McKeen, Heather Chamberlain, Ortis Yankey, Duygu Cihan and Assane Gadiaga. Project oversight was done by Attila Lazar and Andy Tatem. Household survey listing data were provided by the National Statistical Office, and the settlement footprint was generated by Planet.

Please, note that the same modelled population data (with minor rounding difference of 41 in the national total) can also be downloaded from the NSO’s website: https://www.nso.gov.pg/statistics/population/

The authors followed rigorous procedures designed to ensure that the used data, the applied method and thus the results are appropriate and of reasonable quality. If users encounter apparent errors or misstatements, they should contact WorldPop at release@worldpop.org.

WorldPop, University of Southampton, and their sponsors offer these data on a "where is, as is" basis; do not offer an express or implied warranty of any kind; do not guarantee the quality, applicability, accuracy, reliability or completeness of any data provided; and shall not be liable for incidental, consequential, or special damages arising out of the use of any data that they offer.

RELEASE CONTENT

1. PNG_population_v1_0_adminTotals.zip
2. PNG_population_v1_0_agesex.zip
3. PNG_population_v1_0_methods.zip

LICENSE

SUGGESTED CITATIONS

WorldPop and National Statistical Office of Papua New Guinea. 2022. Census-independent population estimates for Papua New Guinea (2020-21), version 1.0. WorldPop, University of Southampton. DOI: 10.5258/SOTON/WP00763

FILE DESCRIPTIONS

The projection for all GIS files is the geographic coordinate system WGS84 (World Geodetic System 1984).

PNG_population_v1_0_adminTotals.zip

This zip file contains the following seven files:

LLG_estimates_main_gamma_gaussian.csv
District_estimates_main_gamma_gaussian.csv
Province_estimates_main_gamma_gaussian.csv

These csv files contain estimates of total population size at LLG, District and Province levels across PNG. Fields: names (admin unit names), total (estimated most likely total population within the admin unit), lower (2.5% Credible Interval of the estimates total population), upper (97.5% CI of the estimates total population).

National_estimates_main_gamma_gaussian.csv

This csv file contains estimates of total population size of the PNG. Fields: total (estimated most likely total population within the LLG, lower.2.5% (2.5% Credible Interval of the estimates total population), median.50% (50% CI of the estimates total population), upper.97.5% (97.5% CI of the estimates total population).

gamma_gaussian_LLG.shp
gamma_gaussian_district.shp
gamma_gaussian_province.shp

These polygon shapefile contain estimates of total population size at LLG, District and Province levels across PNG. Fields: names (admin unit names), total (estimated most likely total population within the admin unit), lower (2.5% Credible Interval of the estimates total population), upper (97.5% CI of the estimates total population), Uncertaint (estimated uncertainty – see below for further details).

The uncertainty values are the difference between the upper and lower 95% credible intervals of the posterior prediction divided by the mean of the posterior prediction: (upper – lower)/mean. As a consequence, cells with a mean prediction of 0 result in NA uncertainty values. These numbers provide a comparable measure of uncertainty in population estimates across the country. Uncertainty estimates cannot be summed across admin units to produce an uncertainty measure for a multi-cell area. Uncertainty for multiple admin units can only be calculated using the admin units' posterior predictions.

PNG_population_v1_0_agesex.zip
This zip file contains 4 csv files:

agesex_gamma_gaussian_llg.csv
agesex_gamma_gaussian_province.csv
Each file provides population estimates for an age-sex group for each LLG and Province across PNG. Fields:

- xxx_Name: LLG or Province names
- age: the age classes (< 1; 1 to 4; 5 to 9; 10 to 14; 15 to 19; 20 to 24; 25 to 29; 30 to 34; 35 to 39; 40 to 44; 45 to 49; 50 to 54; 55 to 59; 60 to 64; 65 to 69; 70 to 74; 75 to 79; 80 +)
- sex: These are labelled with either an “m” (male) or an “f” (female)
- total: total population within the admin unit
- lower: 2.5% Credible Interval of the estimates total population
- upper: 97.5% CI of the estimates total population

agesex_gamma_gaussian_llg18.csv
agesex_gamma_gaussian_prov18.csv
Each file provides population estimates for an age-sex group for each LLG and Province across PNG. Fields:

- xxx_Name: LLG or Province names
- age: the age classes (<18; 18+)
- sex: These are labelled with either an “m” (male) or an “f” (female)
- total: total population within the admin unit
- lower: 2.5% Credible Interval of the estimates total population
- upper: 97.5% CI of the estimates total population

While this data represents population counts, values contain decimals, i.e. fractions of people. This is because both the input population data and age-sex proportions contain decimals. For this reason, it is advised to aggregate the rasters at a coarser scale. For example, if four grid cells next to each other have values of 0.25 this indicates that there is 1 person of that age group somewhere in those four grid cells.

PNG_population_v1_0_methods.zip
PNG_population_v1_0_methods.pdf
This pdf file contains a report describing the statistical methods developed to produce these population estimates.

gg_model_covariates_selection.R
gg_model_final_main.R
gg_model_cross_validation.R
These ‘R’ scripts contains the model code. These codes are also published on GitHub (https://github.com/wpgp/PNG_Bottom_Up_Modelling).

RELEASE HISTORY

Version 1.0 (27 July 2023) [doi: 10.5258/SOTON/WP00763]

- Original release of the data set.

ASSUMPTIONS AND LIMITATIONS

These population estimates most likely represent the 2020-21 time period, but because of the different years of the input data used to build the model, a precise time point cannot be allocated. Most of the population observations came from 2020, but the most recent data were from 2021. The settlement data also reflected 2021. This settlement data primarily determined the spatial distribution of the gridded population estimates, whereas the observations defined the magnitude of population. This model assumes that population densities and age/sex distributions observed during the earlier time period are still representative of the more recent period.

Since the survey data were not geolocated (i.e., there were no GPS points or cluster boundaries), the NSO's CU boundaries were adopted as the most accurate representation of survey locations. There was an overlap of 524 CUs within the two survey datasets. As they did not exactly match and none of them were consistently higher or lower than the other, thus an average in the overlapping CUs total count was calculated and used in the population model. This represents an area of uncertainty that requires further investigation.

It is known that some settlements are under permanent canopy cover and were not captured in the Planet settlement data. This is a limitation common to all population modelling efforts of this type that are based on imagery, though the statistical modelling approaches put forward here recognise this and aim to limit the impacts. To remedy this, CU was adopted as the lowest spatial scale in the modelling. Settlement locations were used instead of Planet data as a direct input, and alternative model estimations were implemented with and without settlement data to check the validity of the model results.

Among the limitations, it is important to note that due to lack of data on such factors, the estimates provided do not explicitly account for population migration.

SOURCE DATA

- **Urban Structural Listing (2021):** 1,959 Census Units containing the characteristics of structures within urban areas. Includes household counts.
- **Malaria Long-Lasting Insecticidal Net (LLIN) survey data (2019-21):** 15,468 Census Units containing household counts and age/sex.
- **Administrative boundary shapefiles** were provided for Papua New Guinea by the NSO (CU boundaries and LLG-level boundaries).
- **Planet settlement raster** (www.planet.org): experimental non-public settlement data product providing information on the locations of buildings/settlements in gridded (raster) format (spatial resolution of approximately 4.77 metres) from cloud-free satellite imagery from a 7-month period (July 2021 – January 2022).

Other geospatial covariates: please see Appendix 1 of the methods report (PNG_population_v1_0_methods.pdf).

METHODS OVERVIEW

The methods report contains the full documentation of the PNG model application. Here, we will focus on the final two-stage modelling approach only. Further information on other alternative models tested are available in the technical report (WorldPop and NSO-PNG. 2022).
Overall, six key steps are involved in the entire modelling processes that produced the population estimates (Figure 1).

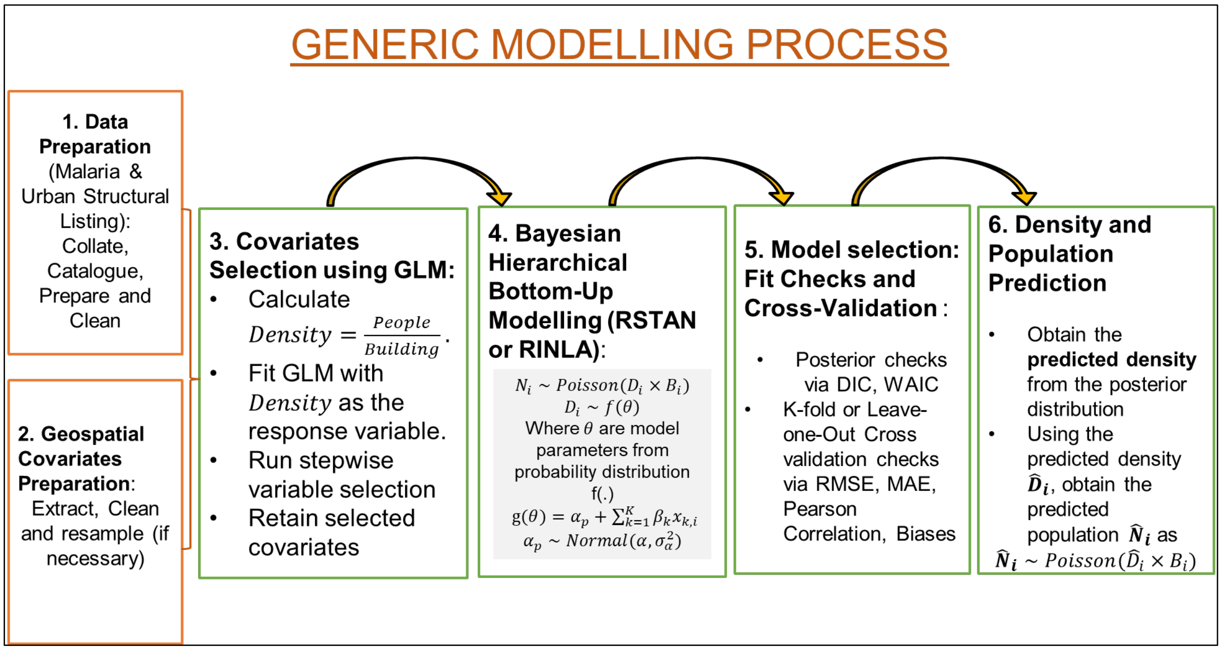
GENERIC MODELLING PROCESS
1. Data Preparation (Malaria & Urban Structural Listing): Collate, Catalogue, Prepare and Clean
2. Geospatial Covariates Preparation: Extract, Clean and resample (if necessary)
3. Covariates Selection using GLM:
• Calculate $Density = \frac{People}{Building}$.
• Fit GLM with $Density$ as the response variable.
• Run stepwise variable selection
• Retain selected covariates
4. Bayesian Hierarchical Bottom-Up Modelling (RSTAN or RINLA):
$N_i \sim Poisson(D_i \times B_i)$
$D_i \sim f(\theta)$
Where θ are model parameters from probability distribution f(.)
$g(\theta) = \alpha_p + \sum_{k=1}^{K} \beta_k x_{k,i}$
$\alpha_p \sim Normal(\alpha, \sigma_\alpha^2)$
5. Model selection: Fit Checks and Cross-Validation :
• Posterior checks via DIC, WAIC
• K-fold or Leave-one-Out Cross validation checks via RMSE, MAE, Pearson Correlation, Biases
6. Density and Population Prediction
• Obtain the predicted density from the posterior distribution
• Using the predicted density $\boldsymbol{\hat{D}_i}$, obtain the predicted population $\boldsymbol{\hat{N}_i}$ as
$\boldsymbol{\hat{N}_i} \sim Poisson(\hat{D}_i \times B_i)$

Figure 1. Bottom-up population modelling workflow.

Data preparations

The first and second steps of the process which may happen simultaneously involve the preparation of both input population data and geospatial covariates. Data preparation steps included (i) topological checks on administrative boundaries, (ii) matching survey data cluster ids with CU ids, (iii) aggregating the Planet settlement data to 100m resolution grids, (iv) selecting the best performing geospatial covariates, and (v) integrating all data sources to a joint database. All data preparations were done in ArcGIS pro (Esri, 2018), QGIS (QGIS Development Team, 2022) using R (R Core Team, 2021).

Topological checks were carried out in both Esri's ArcGIS Pro (v.2.7) and QGIS (v.3.20) on the administrative boundary shapefiles – both those provided for us and those we generated by aggregating them. The 'Topology Checker' plugin and 'Fix geometries' tool in QGIS, and the 'Check geometry' tool in ArcGIS Pro were utilised. A topology was also created within the ArcGIS Pro project geodatabase to identify errors.

Of the total of 32100 CUs, malaria survey for 17,788 CUs, of which 15,468 could be matched to the administrative boundary shapefile. Therefore, together with the 1,959 CUs which have Urban Structural Listing data, there were 16,903 CUs available for training the population model.

This model was then used to predict (estimate) populations in the 15,197 unsampled CUs (Figure 2).

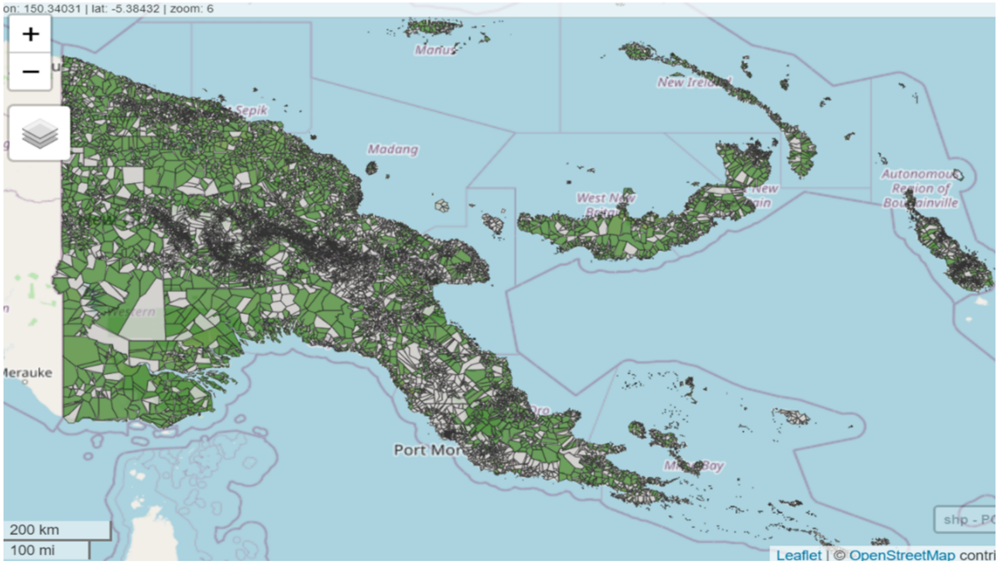
on: 150.34031 | lat: -5.38432 | zoom: 6
+
−
Manus
New Ireland
Sepik
Madang
West New Britain
New Britain
Autonomous Region of Bougainville
Merauke
Port Mor
Oro
Milne Bay
200 km
100 mi
shp - P
Leaflet | © OpenStreetMap contri

Figure 2: Map of Papua New Guinea showing all the 16,903 CUs with observations (green) and the 15,197 unsampled CUs (grey).

Geospatial Covariates Selection

The selection of the most important geospatial covariates (predictors) that most influenced population density and/or spatial distribution was based on (forward-backward) stepwise regression (James et al., 2013; Bruce and Bruce, 2017) using Generalized Linear Model (GLM; McCullagh and Nelder, 1989). Note that these covariates must have values for the entire area, and they must include geographical information on location. Further checks were carried out on the selected covariates to ensure that the issue of multicollinearity that could give rise to variance inflation does not arise. This was done by using the vif() function of the 'car' package such that covariates with vif values are retained as a rule of thumb (James et al., 2013). These processes were carried out for both the two-stage modelling processes when correcting biases in the building count and when using the corrected bias to estimate population density. Eventually, 13 and 15 covariates were selected for the building count and population density models, respectively (Table 1).

Table 1: The final model covariates selected via stepwise regression

Variable	Description	Population density model	Building intensity model
X3	Mean total daily precipitation	✓	✓
X11	Baseflow Index 1	✓	
X15	Baseflow Recession	✓	✓
X27	Motorized friction surface		✓
X29	Distance to health providers	✓	
X30	Distance to local roads	✓	✓
X31	Distance to main roads	✓	✓
X32	Distance to marketplace	✓	
X33	Distance to places of education		✓
X34	Distance to places of worship		✓
X35	Distance to aquatic vegetation areas	✓	✓
X36	Distance to artificial surface edges	✓	✓
X38	Distance to cultivated areas		✓
X39	Distance to ESA-CCI-LC inland water	✓	
X45	Distance to OSM major waterways	✓	
X46	Distance to shrub area edges	✓	✓
X48	Distance to woody areas	✓	
X50	Resampled DMSP-OLS night-time lights		✓
X51	Resampled VIIRS night-time lights	✓	
X52	Slope	✓	✓

Two-stage Bayesian Hierarchical Bottom-Up Modelling

Here, we briefly describe the two-step bottom-up modelling (Leasure et al.,2020 and Wardrop et al., 2018) approach employed here. As noted above, in the first stage, we corrected biases in the building intensity derived from the settlement data via the following model specifications:

STAGE ONE: Let B_i denote the building intensity for the ith area unit so that the log-transformed value is normally distributed with mean and precision $\bar{B}_i$ and τ_B, respectively. Then the full hierarchical model is given by,

$$\log(B_i) \sim Normal(\bar{B}_i, 1/\tau_B)$$

$$\bar{B}_i = \eta_i^{(B)} = \beta_0 + \sum_{k=1}^{13} \beta_k x_{ik} + f_{struc}(s_i) + f_t(t) + f_{t.p}(t, p)$$

$$\pi(\beta_0) \propto 1$$

$$\beta_k \sim Normal(\mu_\beta, 1/\tau_\beta)$$

$$f_{struc} \sim GMRF$$

$$f_t \sim Normal(0, 1/\tau_t)$$

$$f_{t,p} \sim Normal\big(0, 1, \tau_{t,p}\big)$$

$$\tau_j \sim Gamma(\alpha_\tau, \beta_\tau), \text{ where } j \in \{B, \beta, unstr, t, tp\} \qquad (1)$$

Where η_i is the linear predictor; β_0 is the intercept which is the baseline average building intensity when the effect of the geospatial covariates $x_{i,1}, \ldots, x_{i,13}$ on the building intensity value is zero; $\beta_1, \ldots, \beta_{13}$ are the corresponding fixed effects coefficients to be estimated; f_{struc} is the spatially correlated random effect which captures shared characteristics among neighbouring spatial units and allows us to more accurately estimate response in areas with little or no observations; $f_t(.)$ and $f_{t,p}(.)$ are the settlement type and province-settlement type nested random effects respectively. For Bayesian inference, uniform prior is assigned to the intercept term β_0; while normal priors are assigned to the fixed effects term β_k, and the random effects for settlement type and the settlement type – province nested effect; τ_j are the corresponding hyperparameters which are assigned Gamma priors.

Furthermore, The structured or correlated random effect f_{struc} is modelled as a Gaussian Markov Random Fields with sparse distance-dependent covariance matrix for computational efficiency via the integrated nested Laplace approximation – Stochastic partial differential equation (INLA-SPDE; Rue et al., 2009; Rue and Held, 2005; Lindgren et al., 2011) framework. Then the predicted building count is given by

$$\hat{B}_i = h(\bar{B}_i) \qquad (2)$$

where $h(.)$ is an appropriate inverse of the link function $g(.)$ which is identity in this case. This predicted building intensity which has now taken into account the satellite partial observations, and shared spatial homogeneity and heterogeneity is now used as a corrected model input for estimating the population density and population count in stage two.

STAGETWO: Let Y_i and D_i denote the population count and the population density of the ith spatial unit, respectively. Then, the full hierarchical model is given by,

$$Y_i \sim Poisson(D_i \times \hat{B}_i)$$

$$D_i \sim Gamma(\alpha_D, \beta_D) \text{ and } \bar{D}_i = \alpha/\beta$$

$$\log(\bar{D}_i) = \eta_i^{(D)} = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{struc}(s_i) + f_{unstr}(s_i) \;\; + f_t(t) + f_{t.p}(t, p)$$

$$\pi(\beta_0) \propto 1$$

$$\beta_l \sim Normal(\mu_\beta, 1/\tau_\beta)$$

$$f_{unstr} \sim Normal\left(0, \frac{1}{\tau_{unstr}}\right)$$

$$f_t \sim Normal(0, 1/\tau_t)$$

$$f_{t,p} \sim Normal\big(0, 1, \tau_{t,p}\big)$$

$$\tau_j \sim Gamma(\alpha_\tau, \beta_\tau), \text{ where } j \in \{\beta, unstr, t, tp\} \qquad (3)$$

where $\eta_i^{(D)}$ is the linear predictor; β_0 is the intercept which is the baseline average population density when the effect of the geospatial covariates $x_{i,1}, \dots, x_{i,15}$ is zero; $\beta_1, \dots, \beta_{15}$ are the corresponding fixed effects coefficients to be estimated; f_{struc} and f_{unstr} are the spatially correlated and spatially independent (uncorrelated) random effects, respectively. While the spatially dependent random effect captures shared characteristics among neighbouring spatial units and allows us to more accurately estimate response in areas with little or no observations, the spatially uncorrelated random effect captures variability due to spatial heterogeneity; $f_t(.)$ and $f_{t,p}(.)$ as well as the prior distributions are as defined in stage one.

Then, with the estimated building intensity for the entire locations $\hat{B}_i$, the predicted population count $\hat{y}_i$ is given by

$$\hat{y}_i = \widehat{D}_i \times \hat{B}_i \qquad (4)$$

where

$$\widehat{D}_i = \exp\left(\beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{struc}(s_i) + f_{unstr}(s_i) \;\; + f_t(t) + f_{t.p}(t,p)\right)$$

and

$$\hat{B}_i = \beta_0 + \sum_{k=1}^{13} \beta_k x_{ik} + f_{struc}(s_i) + f_t(t) + f_{t.p}(t,p) \qquad (5)$$

For each of the selected model, posterior sampling was carried out to ensure improved estimates as well as to enable uncertainty quantification for the population totals. This was followed by model cross-validation was carried out using k-fold (with k = 5) in which at each non-overlapping kth-fold, the data are randomly divided into 80% training set and 20% testing set, that is, 80% of the data are used to train the model while 20% are held out and used to test the model by predicting the values of the held data using the model parameters. A summary flow chart for the two-stage (two-step) modelling as described above is given in Figure 3.

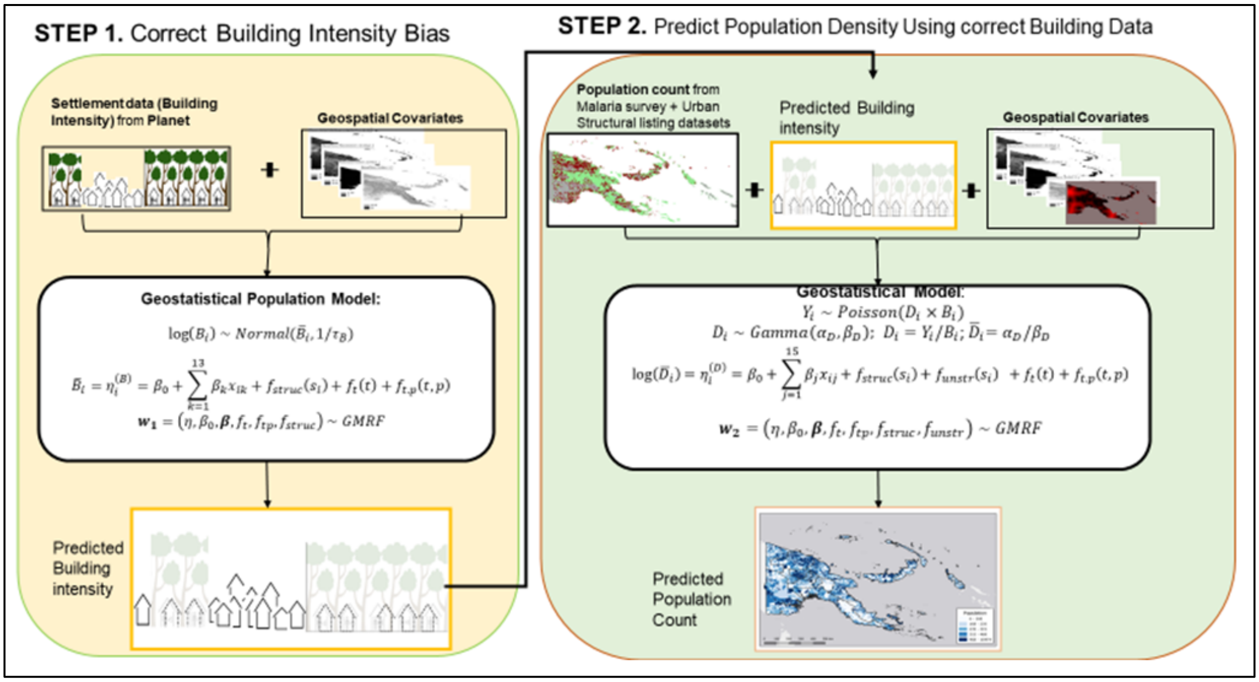
STEP 1. Correct Building Intensity Bias
Settlement data (Building Intensity) from Planet
Geospatial Covariates
Geostatistical Population Model:
$\log(B_i) \sim Normal(\bar{B}_i, 1/\tau_B)$
$\bar{B}_i = \eta_i^{(B)} = \beta_0 + \sum_{k=1}^{13} \beta_k x_{ik} + f_{struc}(s_i) + f_t(t) + f_{t,p}(t,p)$
$\boldsymbol{w_1} = (\eta, \beta_0, \boldsymbol{\beta}, f_t, f_{tp}, f_{struc}) \sim GMRF$
Predicted Building intensity
STEP 2. Predict Population Density Using correct Building Data
Population count from Malaria survey + Urban Structural listing datasets
Predicted Building intensity
Geospatial Covariates
Geostatistical Model:
$Y_i \sim Poisson(D_i \times B_i)$
$D_i \sim Gamma(\alpha_D, \beta_D)$; $D_i = Y_i/B_i$; $\bar{D}_i = \alpha_D/\beta_D$
$\log(\bar{D}_i) = \eta_i^{(D)} = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{struc}(s_i) + f_{unstr}(s_i) + f_t(t) + f_{t,p}(t,p)$
$\boldsymbol{w_2} = (\eta, \beta_0, \boldsymbol{\beta}, f_t, f_{tp}, f_{struc}, f_{unstr}) \sim GMRF$
Predicted Population Count

Figure 3. Flow chart of the two-step modelling approach. Biases in the building count was first corrected using a model-based solution in the first step. The corrected building data is then used to estimate population density/population count in the second step.

Population pyramids for administrative units (i.e., LLG, district and province levels) were produced using age/sex observations from the malaria survey data. Besides the missing age or sex values of the observations, the initial survey dataset was also spatially incomplete. Incomplete population pyramids were replaced by the next level up population pyramid (e.g. district level), if it was complete, assuming therefore that the demographic characteristics are the most similar. Therefore, LLG pyramids were replaced by district pyramids, and district pyramids were replaced by province pyramids, if needed. Because all population pyramids at province level were complete, these were used as a last resort. Finally, the age-sex proportions (i.e., population pyramids at LLG level) were applied to the population estimates (i.e., total LLG population) to allocate the modelled total populations to the different age-sex classes.

All calculations were undertaken in R. The computer code can be found in the methods.zip file and also downloadable from GitHub (https://github.com/wpgp/PNG_Bottom_Up_Modelling).

WORKS CITED

Esri(2018). ArcGIS Pro 2.1 Redlands, CA: Environmental Systems Research Institute

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

Bruce, Peter, and Andrew Bruce. 2017. Practical Statistics for Data Scientists. O'Reilly Media.

Leasure, D. R., Jochem, W. C. , Weber, E. M. , Seaman, V. and Tatem, A. J. (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

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd Edition. Chapman; Hall/CRC.

QGIS Development Team (2022). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

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

Wardrop, N. A., Jochem, W. C. , Bird, T. J., Chamberlain, H. R. , Clarke, D,. Kerr, D., Bengtsson, L., Juran, S., Seaman, V. and A. J. 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(14): 3529-3537. DOI: 10.1073/pnas.1715305115.

WorldPop and National Statistical Office of Papua New Guinea. 2022. Census-independent population estimates for Papua New Guinea (2020-21), version 1.0. WorldPop, University of Southampton

Rue H, Held L. (2005). Gaussian Markov random fields. Theory and applications. Chapman & Hall.

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