
MODELLED POPULATION ESTIMATES FOR PAPUA NEW GUINEA

A TWO-STEP BAYESIAN HIERARCHICAL GEOSTATISTICAL MODEL

Modelled Population Estimates for Papua New Guinea

A TWO-STEP BAYESIAN HIERARCHICAL GEOSTATISTICAL MODEL

Executive Summary

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 overarching aims of the project are to:

- Develop and test methods and procedures for predicting population numbers at sub-national scales, with age and sex breakdowns and associated uncertainty measures.
- Produce sex- and age-disaggregated population estimates using the methods developed.
- Ensure approaches are coordinated and integrated into government national census efforts and can be readily updated when new data becomes available.
- Undertake training of NSO and other relevant departments to ensure that methods are understood and that the outputs can be effectively used within the government.

All the technical aspects of the project were led by WorldPop at the University of Southampton including the generation of estimates of population for PNG. The technical processes used included bespoke population modelling techniques that combine multiple data sources along with key geospatial covariates, satellite observations and population datasets. WorldPop is a global leader in the production of modelled population numbers with a track record for reliably estimating country-wide population numbers and distributions using geospatial datasets and statistical models.

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.

Initial modelling work produced a wide range of population estimates, with some large values. However, the three final models all estimated the national population size to be between 11 and 12 million. The best fit model indicated that the total population of Papua New Guinea based on the compiled datasets is 11.78M with a 95% credible interval of 11.64-12.03M. 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. Thus, it is important to bear in mind that the estimates of population reported here are only as good as the input datasets.

Although the final model provided the best fit to the data with small margin of error even at the census unit level, the validation of these modelled estimates using a more detailed population and housing census data is highly recommended. In addition, these estimates could be improved by using datasets from upcoming household surveys in conjunction with new satellite-derived settlement data and geospatial covariates.

Definition of Terms

CU – Census Unit

DHS – Demographic and Health Surveys

DRC – Democratic Republic of Congo

GAMM – Generalized Additive Mixed Model

GLM – Generalized Linear Model

GMRF – Gaussian Markov Random Field

INLA – Integrated Nested Laplace Approximation

LLG – Local Level Government

LLIN – Long-Lasting Insecticidal Nets

NSO – National Statistical Office

PNG – Papua New Guinea

RAM – Rotarians Against Malaria

SPDE – Stochastic Partial Differential Equation

UNFPA – United Nations Population Fund

Table of Contents

Executive Summary	1
Definition of Terms	2
List of Tables	4
List of Figures	4
1. INTRODUCTION	5
2. DATA SOURCES	6
2.1 Survey data	6
2.2 Administrative boundaries	7
2.3 Settlement data	8
2.4 Geospatial covariates	9
2.4.1 Sourcing covariate data	9
2.4.2 Creating 100m resolution geospatial covariates	11
3. METHODS	11
3.1 The statistical modelling process	11
3.2 Data preparations	13
3.2.1 Topological checks on administrative boundaries.....	13
3.2.2 Matching survey data cluster ids with CU ids.....	13
3.2.3 Aggregating the Planet settlement data	14
3.2.4 Selection of the potential model covariates	15
3.3 Model building	17
3.3.1 Sampling from the joint posterior distribution	22
3.3.2 Model code	22
3.4 Age/sex disaggregation.....	23
4. RESULTS	24
4.1 Model Fit Assessments and Cross-Validation.....	25
4.1.1 Posterior Simulations	25
4.1.2 Model Cross-validation.....	26
4.2 National-level Estimates	28
4.3 Province-level Estimates.....	29
4.4 District-level Estimates	30
4.3 Age/sex disaggregation results.....	32
5 Discussion	34
5.1 Issues with the previous model results explained	34
5.2 Limitations	35
8. CONCLUSIONS AND RECOMMENDATIONS	36
CONTRIBUTIONS	36
LICENSE	37
SUGGESTED CITATION	37
APPENDIX 1: Initial covariate characteristics and source links	40
APPENDIX 2: Age-Sex population pyramids of provinces	43
APPENDIX 3: Sampling path plots and the associated histograms	47

List of Tables

Table 1: Data source and data description.	6
Table 2: The final model covariates selected via stepwise regression	17
Table 3: Model fit assessment and cross validation	27
Table 4: Model-based and simulation-based national totals across the three models.....	28

List of Figures

Figure 1: A schematic representation of the bottom-up modelling techniques. It combines (usually incomplete sampling-based data) with geospatial covariates along with geospatial observations.	6
Figure 2: Map of Papua New Guinea showing the 24 Provinces	8
Figure 3: An example of settlement and roads (right image) extracted from satellite imagery (left image) by Planet (2021). Extracted buildings (green) and roads (red) overlaid on the satellite image on the right image.	9
Figure 4: Example covariates used in the population modelling.....	10
Figure 5: Modelling process chart. The iterative development arrow shows how the process was repeated multiple times to refine and improve model design before selecting final models.	12
Figure 6: Map of Papua New Guinea showing all the 16,903 CUs with observations (green) and the 15,197 unsampled CUs (grey).	14
Figure 7: Correlation matrix for A) Final covariates for building intensity model, B) Final covariates for population count and density models. The colour green indicates positive and purple negative correlation. The size of the squares indicates the strength of the correlation.	16
Figure 8: The Delaunay Triangulation of PNG (Mesh) used for the INLA-SPDE implementation with 695 nodes.	22
Figure 9: Observed age/sex proportions in the Malaria dataset	23
Figure 10: Sampling paths plots of six randomly selected CUs for the posterior simulation based on the GG model.....	25
Figure 11: Posterior estimates of the mean (left) and standard deviations (right) of the spatial random effects across the three models.....	26
Figure 12: Model fit and Cross validation across the three models. The GG and QP models provided better fit to the data than the NB model which did well at both model-based and simulation-based in-sample prediction but not as good at out-of-sample cross validation.	27
Figure 13: Comparing provincial population estimates across the various models with the RAM data	29
Figure 14: Comparisons of the spatial surfaces of total population estimates across the three models and the RAM data. Estimates show relatively similar spatial trends	29
Figure 15: Comparing district level population estimates across the three models and the RAM data	30
Figure 16: CU level total population (left panel), Esri satellite image base layer with white CU boundary overlay (right panel).....	31
Figure 17: Age and sex pyramid of PNG national estimated population	32
Figure 18: Age and sex pyramid of National Capital District, PNG estimated population	33

1. INTRODUCTION

It is now more than 11 years since the fourth National Population and Housing Census was conducted in Papua New Guinea (PNG) in July 2011. Although preparations for the fifth census are advanced, it has been deferred due to the COVID-19 pandemic and funding gaps, possibly until 2024. Thus, there is a need for intercensal national and subnational demographic data on populations across various administrative units in PNG to enable robust and accurate planning for upcoming development decisions. Recent geospatial data exist (e.g., Pilot Census, Demographic and Health Surveys (DHS) 2016-18, 2013-2021 malaria long-lasting insecticidal net (LLIN) campaign data, Urban Structural Listing 2021, Copernicus data products/layers, Health facility location survey, and WFP mVAM food security and livelihoods survey data) that could be utilised for constructing population estimates.

WorldPop at the University of Southampton has developed capacity over recent years in estimating population numbers and distributions across multiple countries using geospatial datasets and statistical models. Such operational population estimates are valuable in supporting updated planning efforts for the upcoming census, while also providing interim data for national decision making until the fifth housing and population census is completed. Moreover, the models have potential for use in producing small area inter-censal updates.

With the financial support of the Australian Government and PNGAus partnership, UNFPA and the PNG National Statistical Office initiated the population estimation project in 2021. The project aims to create operational small area estimates of population distribution and to ensure that the project deliverables, both the methods and data, can be of sustainable use by the Government of PNG, academics and researchers, development and aid agencies, civil society and other concerned parties long after project completion. The specific objectives of the work are to:

1. Develop and test methods and procedures for predicting population numbers at sub-national scales, with age and sex breakdowns and associated uncertainty measures.
2. Produce sex- and age-disaggregated population estimates using the methods developed.
3. Ensure approaches are coordinated and integrated into government national census efforts and can be readily updated when new data becomes available.
4. Undertake training of NSO and other relevant departments to ensure that methods are understood and that the outputs can be effectively used within the government.

Geospatial modelling approaches have recently been developed and used to provide subnational-scale estimates of population numbers with associated confidence intervals in a range of countries (UNFPA, 2017). Such 'bottom-up' approaches leverage the spatial relationships at local scales between enumerated population densities from samples and a range of geospatial datasets (Figure 1, <https://youtu.be/Z1XrHOt8w2A>). These geospatial datasets are derived from satellite imagery, government mapping and other sources, to build and validate statistical models that are then used to estimate population numbers in areas where only the geospatial datasets are available.

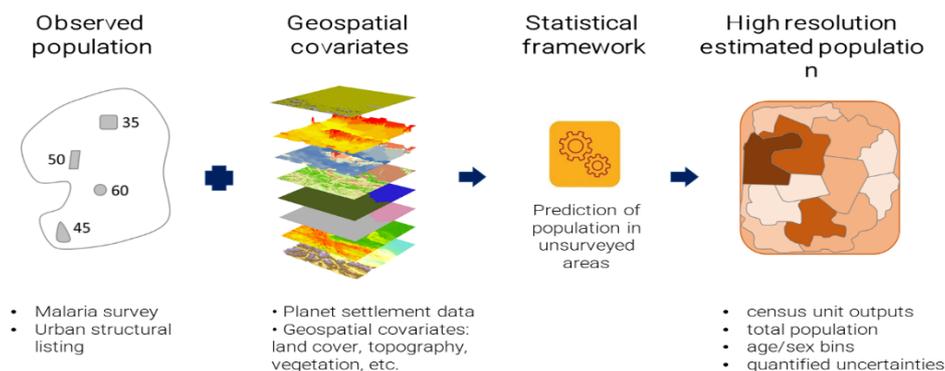


Figure 1: A schematic representation of the bottom-up modelling techniques. It combines (usually incomplete sampling-based data) with geospatial covariates along with geospatial observations.

Application of these approaches have included census-independent population estimation in Nigeria (Leasure et al., 2020) and DR Congo (Boo et al., 2022), complementing census with estimated populations in inaccessible areas in Burkina Faso (WorldPop and INSD, 2019) and national population mapping across both government- and non-government-controlled areas in Afghanistan (unpublished dataset). Active collaborations with national statistical offices in Thailand, Brazil, Colombia, Cameroon, Zambia, South Sudan, Mali, Kenya and Mozambique, among others, are also ongoing.

WorldPop's 'Bottom-up' modelling approach has been applied here for PNG. This method combines geospatial covariates available for the entire region of study with observed population data available from household surveys in a hierarchical Bayesian framework (Wardrop et al, 2018; <https://www.worldpop.org/methods/populations/>). The data used and the methods applied are described below.

2. DATA SOURCES

2.1 Survey data

Table 1: Data source and data description.

Survey	Dates	Coverage	Spatial resolution	Measures
Demographic Housing Surveys (DHS)	2016-18	767 clusters	'Cluster'-level: based on Census Unit areas (whole or segments), ~24 households enumerated per cluster.	Basic indicators of fertility, fertility preferences, family planning practices, childhood mortality, maternal and child health, knowledge and awareness of HIV/AIDS, domestic violence, and other related health issues. Includes household count & age/sex.
Urban Structural Listing	2021	1,959 CUs	Census Unit	Characteristics of structures within urban areas. Includes household counts.

Survey	Dates	Coverage	Spatial resolution	Measures
Malaria Long-Lasting Insecticidal Net (LLIN) survey data	2019-21	15,468 CUs	Census Unit	Household counts and age/sex.
Pilot census	2019-20	12 CUs	Census Unit	Full census data including household counts and age/sex.
Mobile Vulnerability Analysis and Mapping Food Security & Livelihoods Validation Assessment	2016 - 17	3708 (19 households per LLG)	Individuals randomly contacted	Phone-based survey to assess the impact on the 2015-2016 El Nino drought on food security and wellbeing. Includes information on the health impacts, qualitative assessment of hunger, help received by government/NGO's, food currently available, etc.
Malaria Long-Lasting Insecticidal Net (LLIN) survey - Milne Bay	2020	821 villages	Village	Population estimations based on last RAM survey (2016) and 2000 census using annual growth rates for Milne Bay, plus actual population surveyed, and number of nets distributed.

The NSO provided several survey datasets for use in this study (Table 1) including DHS, Urban Structural Listing, Malaria LLIN survey and a pilot census. Of these, the most recent and most reliable data, the Urban Structural Listing (1,959 CUs surveyed in 2021) and Malaria surveys (15,468 CUs, manually digitised for the 2019-21 period) were used in the population estimation model. This is because they were full enumerations of the lowest administrative boundaries (CUs) and had good spatial coverage. The DHS survey was not a full enumeration of CUs and thus has large inherent uncertainties when used for estimating the population size. Similarly, the pilot census had low spatial coverage, having surveyed only 12 CUs.

There was a 524 CUs overlap between the malaria LLIN and urban structural listing datasets with no clear relationship between the values (i.e., some were higher and some were lower than the other). Therefore, in these CUs, it was decided that an average of the two would be used in the population modelling. Only the malaria LLIN survey had age/sex disaggregated observations.

2.2 Administrative boundaries

From smallest to largest, the administrative levels for Papua New Guinea are: CUs (Census Units), Wards, LLGs (Local Level Governments), Districts and Provinces. There are 24 provinces in Papua New Guinea (Figure 2).

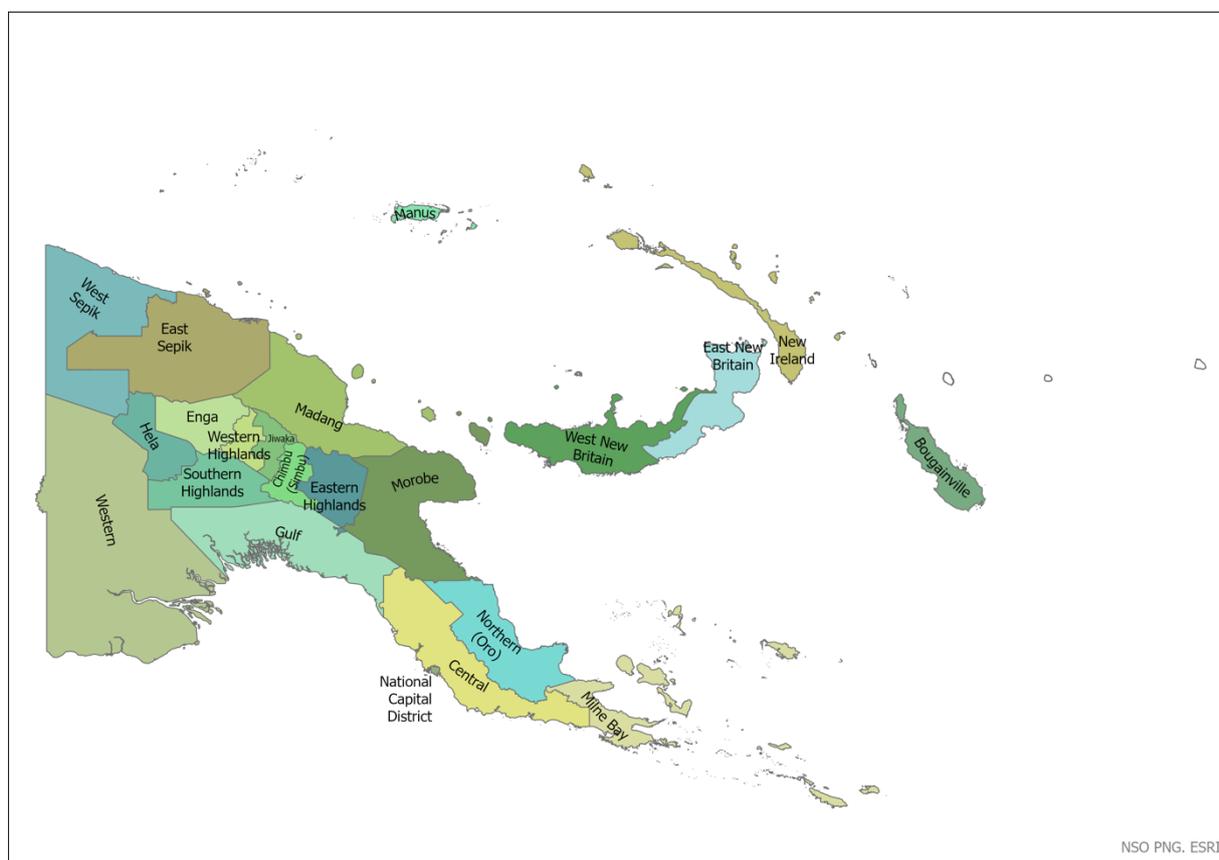


Figure 2: Map of Papua New Guinea showing the 24 Provinces

Two administrative boundary shapefiles were provided for Papua New Guinea by the NSO – one showing CU boundaries and the other showing LLG-level boundaries. A point shapefile was also provided where each point represents a CU. However, following advice from the NSO this was disregarded because of its limited use in the model. Shapefiles showing ward, district, province and national level boundaries were produced using the ‘Dissolve’ data management tool in ArcGIS Pro, based on their respective codes/names provided in the attribute table of the CU level shapefile. Any topological errors were then cleaned (see section 3.2.1 for further details). The CU shapefile consists of a combination of well-defined boundaries (particularly in urban areas) and Voronoi polygons (largely in rural areas) produced from settlement-derived points. It is assumed the settlements from which these points were generated sit entirely within each Voronoi polygon.

2.3 Settlement data

The Planet (www.planet.org) experimental non-public settlement data product was downloaded from the Planet API, where it was made available to WorldPop through a sharing agreement as a tiled product. The settlement data product provides information on the locations of buildings/settlements in gridded (raster) format (Figure 3). The settlement raster was created through the classification of available cloud-free satellite imagery from a 7-month period (July 2021 – January 2022), with classified outputs compiled to create a single gridded output. Values in the settlement raster range from 0 to 254 with higher values indicating a higher likelihood of settlement. The gridded dataset in geoTiff format has a spatial resolution of approximately 4.77 metres and is projected in World Mercator Auxillary Sphere (EPSG: 3857).

A known issue with satellite-derived building/settlement mapping datasets is the permanent canopy or cloud cover that prevent known settlements from being seen on images. Comparison with Facebook HRSL data (<https://ciesin.columbia.edu/data/hrsl/>) for PNG revealed that both suffer from the same issue and the Planet data was generally a higher quality mapping of buildings/settlements, with substantially finer resolution available too.

The project also had access to Esri-generated building footprints (n.d.). However, as this source lacked building data for large areas, it was not considered suitable for population modelling.



Figure 3: An example of settlement and roads (right image) extracted from satellite imagery (left image) by Planet (2021). Extracted buildings (green) and roads (red) overlaid on the satellite image on the right image.

2.4 Geospatial covariates

2.4.1 Sourcing covariate data

Geospatial covariates are data related to population density or the spatial distribution of the population. They are needed for population estimation modelling, as relationships between the covariates in locations with enumerated population data can be used to estimate population in locations without enumeration population counts. This requires that geospatial covariate datasets must cover the entire study area for which population is being estimated (Wardrop et al., 2018).

A wide range of covariates were considered in the development of modelled population estimates for Papua New Guinea, ranging from topography through climate and land use/land cover to settlement information (two examples are shown in Figure 4). In total, 52 geospatial covariates were considered at the start of the covariate selection process (see Appendix 1), with the initial list of covariates informed by previous WorldPop projects. All geospatial covariates were created as gridded (raster) datasets with a harmonised spatial resolution and grid cell alignment, from which spatially aggregated summary statistics were calculated.

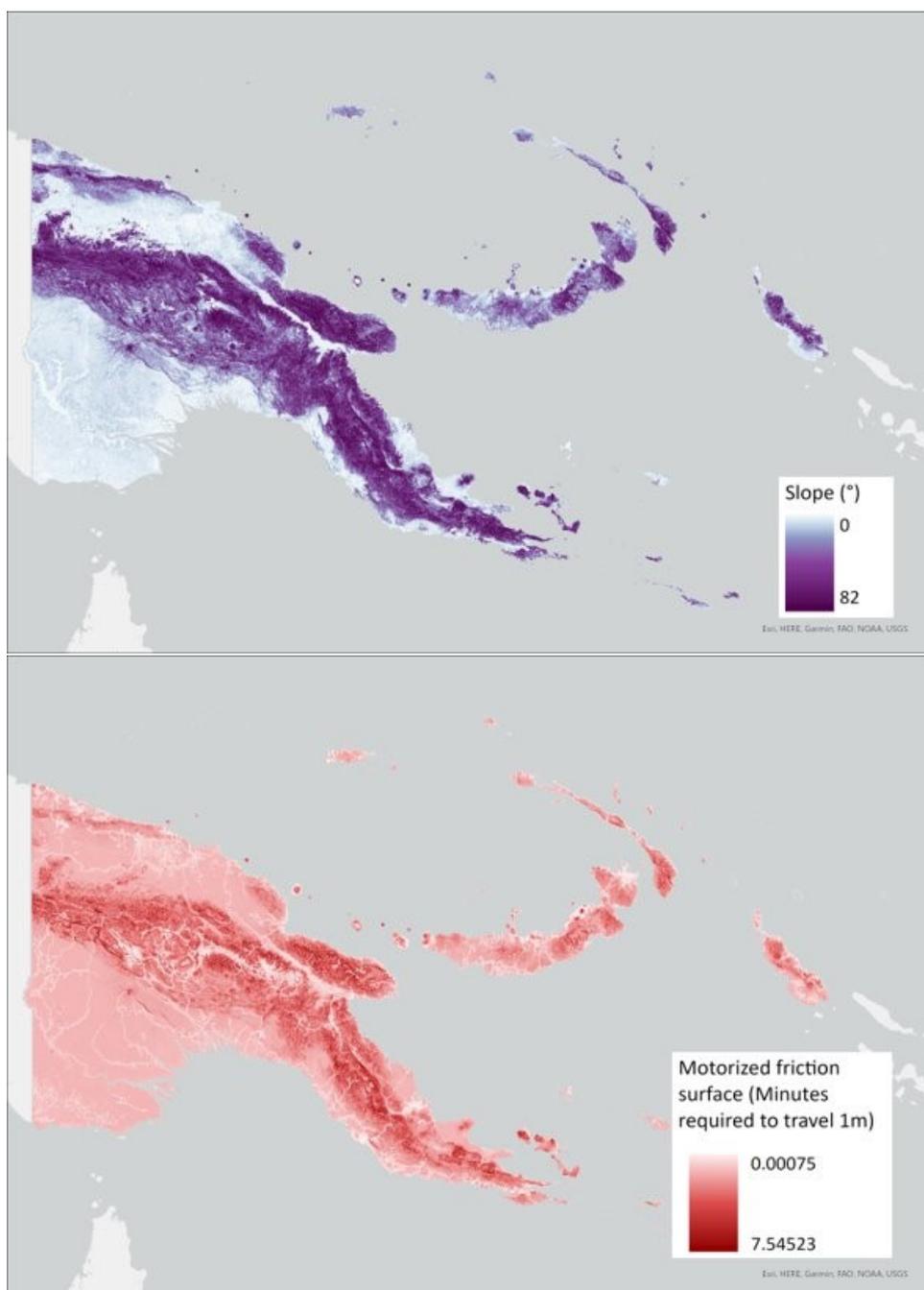


Figure 4: Example covariates used in the population modelling

The geospatial covariates came from a number of different data sources, drawing on publicly available datasets including products and data on features such as roads, rivers and points of interest (POIs). The diverse nature of the geospatial covariates together provide data on a wide range of factors considered to be related to the density and spatial distribution of population. For example, the distribution and density of population is likely to be related to the distance to the nearest major road, with higher population density and counts potentially closer to major roads. Using roads data from OpenStreetMap (OSM), the distance to roads were calculated using a Euclidean distance calculation. For factors such as slope and topography, global datasets providing information on elevation were included. Climatic

factors such as precipitation totals were calculated from satellite-based observations, as were landcover classifications, from which distance-to metrics were calculated.

Satellite-based datasets are collected from sensors measuring in a range of wavelengths, including the visible and microwave wavelengths of the electromagnetic spectrum. For the subset of covariates that are reliant on satellite imagery collected in the visible portion of the spectrum, consideration is given to features of interest that may be obscured on the imagery in some locations. For example, a road in a forested area, may be partially obscured in locations with particularly dense canopy cover.

The geospatial covariates are generated from various sources and they each contribute some information related to the spatial distribution of population. All covariates are explored and interrogated in the covariate selection process. When combined within the statistical model, they enable reliable estimation of population.

2.4.2 Creating 100m resolution geospatial covariates

The main purpose of the covariates is to describe the landscape and environment at fine resolution. For this reason, we have generated them at 100m resolution to capture the fine spatial variations, but in the statistical modelling, these were aggregated up to CUs (i.e., to the level of the analyses).

Initial covariate processing was carried out in Esri's ArcGIS Pro (v.2.7). All covariates were standardized to a mastergrid, produced from a boundary of the national boundary obtained from the CU level administrative boundary shapefile provided by the NSO. This was buffered by ~100m to account for any discrepancies in the boundaries.

As we produced a buffer, we needed to interpolate the values of missing raster cells from the input covariates where coverage did not match. In several cases, this was required to cover the national boundary, as many input covariate datasets did not match the boundary provided, an issue largely driven by Papua New Guinea's complex coastline and numerous islands. The ArcGIS tool 'nibble' was utilised to interpolate missing coverage areas.

Raster layers were then reprojected (where required) to match the projection and cell size of the mastergrid (100m resolution).

As the final outputs for this projection were sub-national population estimates, the 'Zonal Statistics' tool in ArcGIS was used to calculate CU-level statistics – including mean, median, minimum, maximum and standard deviation.

3. METHODS

3.1 The statistical modelling process

The modelling process consists of several steps and tasks (Figure 5). The first step is to collate the available and accessible survey and/or administrative boundary datasets. The tasks include cataloguing, data exploration, data cleaning, and variable recoding.

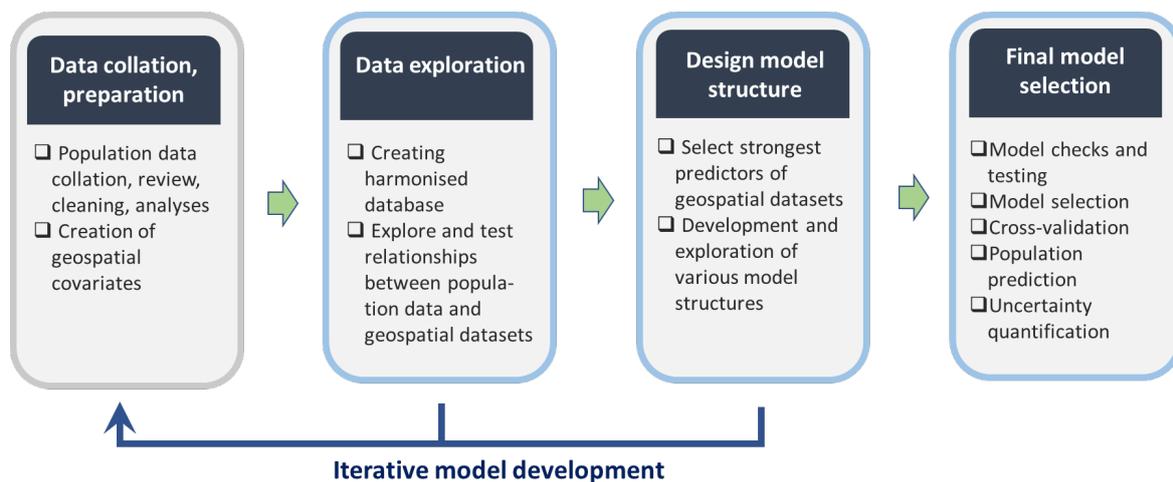


Figure 5: Modelling process chart. The iterative development arrow shows how the process was repeated multiple times to refine and improve model design before selecting final models.

The second step is the preparation and stacking of predictor inputs such as the settlement data and the geospatial covariates. Settlement data (e.g. settled locations, settlement density or intensity) is one of the most important predictor types as they are known to directly influence the population distribution (Leasure et al., 2020). Other geospatial covariates describe the fine spatial variation of the landscape, which can be associated with or directly influence population densities. For these covariates to be included in the modelling, certain criteria must be met – they must have a significant influence on the population spatial distribution, they must have values for the entire area, and they must include geographical information on location. Land use/land cover, topographical and climatic variables for example are often used as geospatial covariates. When these are prepared, the covariate selection process starts to find the best predictors and thus to create a parsimonious model (i.e., the simplest model with the best explanatory predictive power).

The third step starts building the population estimation model. It begins with data integration and testing of alternative model structures (usually from simple to complex), using various sets of geospatial covariates. Population modelling is an iterative process, and thus, steps 1-2-3 are repeated (i.e., finding new observations to fill the gaps, refining the settlement data, creating better covariates) until the model design becomes satisfactory.

As soon as a few well performing model structures are created, the fourth step starts. Here, the models are scrutinised with statistical performance tests, and the best model is identified. Further tests are done on this final model, including cross-validation to test that the model structure is robust and is not under- or over-fitted to the observations. Predictions into the unsurveyed areas are completed using the trained final model, and the model uncertainties are quantified. Thus, in summary, specific modelling steps utilised in this project are:

1. Select the best geospatial covariates for each of the response variables - building intensity, population count and population density using stepwise regression
2. The selected covariates are then used to fit the final models.
3. Predicted building intensity is used to calculate the population density thereby accounting for the variability and gaps within the dataset. Models for negative Binomial and Quasi-Poisson do not need building intensity as input as the response variable is the population count and not the population density.

3.2 Data preparations

The 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).

3.2.1 Topological checks on administrative boundaries

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.

Identified topological errors included gaps, overlaps, duplications and missing polygons. In the original census-unit shapefile, there existed a single ward (Bonkembil/Maskabil) that extended beyond the national boundary given in the LLG shapefile. Communications with the NSO confirmed this was incorrect and it was removed accordingly. There were also cases where the LLG and census-unit shapefiles did not agree. All CUs should fit within the LLG boundaries however there were instances where the LLG boundaries cut through some CU boundaries. Communications with the NSO clarified the correct boundaries.

3.2.2 Matching survey data cluster ids with CU ids

Each of the 32,100 CUs of PNG is identified by a geocode – an 11-digit code consisting of the province, district, LLG, ward and CU code:

14 06 24 11 001
 Province District LLG Ward Census unit

Given the frequent re-naming and re-coding of administrative boundaries in PNG it was found that geocodes and names across administrative boundary shapefiles and population survey datasets did not always match. This presented an issue joining the malaria survey data (which was shared in CSpPro format) to the shapefile.

Unfortunately, not all administrative level identification code issues could be resolved. For example, there remain some CUs in the census-unit shapefile that share the same geocode as another (i.e., 10040282007, 14010108401, 11060226001, 11020102002, 08040202002, 15040481007, 15020104003). To mitigate this issue unique IDs (from 1 – 32,100) were assigned to all CUs in both shapefiles and survey datasets to ensure there would be no double joins and give each CU a unique identifier.

When joining the malaria dataset to the administrative shapefile, names were also used to join. Where the name was of highly similar spelling and the province code was the same, a join was confirmed. However, where names were similar and province codes were different (e.g., 01 & 10, 12 & 13) joins were not included. This emphasises the need to review and harmonise the administrative boundary identification system.

Data for 17,788 CUs were received from the malaria survey, of which 15,468 could be matched to the administrative boundary shapefile. Therefore, together with the 1,959 Urban Structural Listing CUs, 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 6).

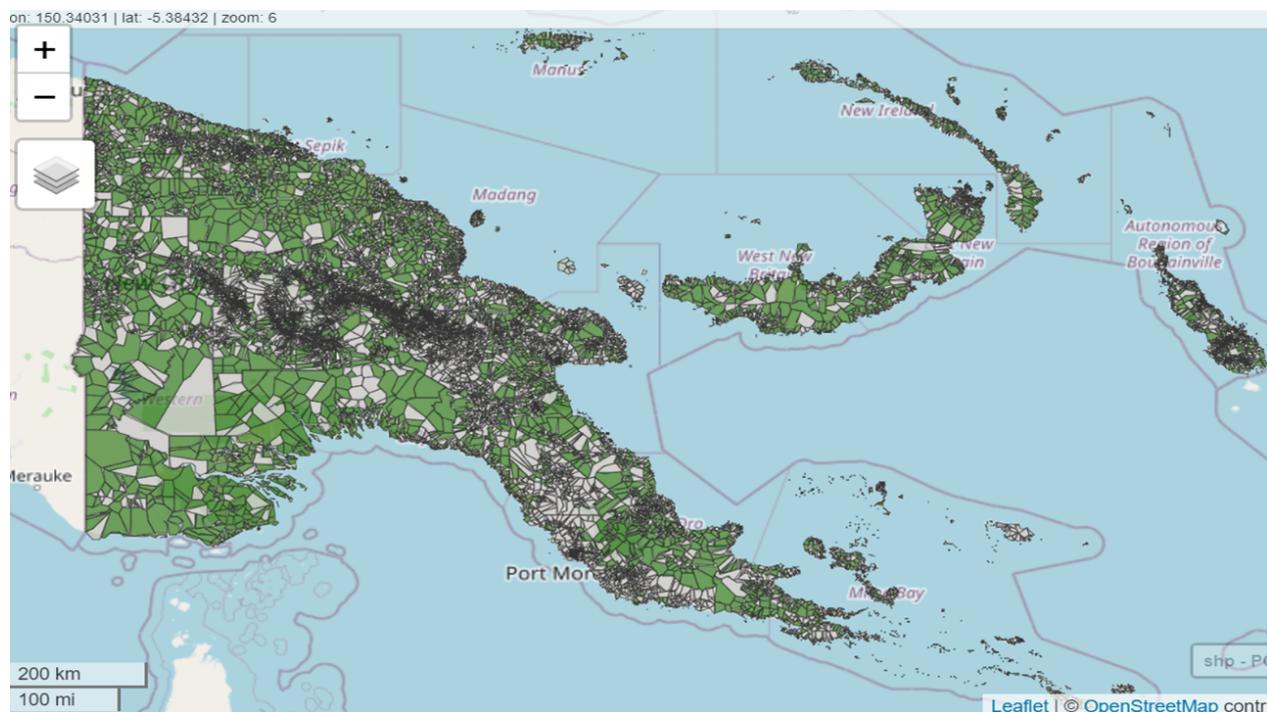


Figure 6: Map of Papua New Guinea showing all the 16,903 CUs with observations (green) and the 15,197 unsampled CUs (grey).

3.2.3 Aggregating the Planet settlement data

The tiled settlement data were first mosaiced into a single raster dataset covering the entirety of the PNG map. In the raw data, raster values run from 0 to 254, where 0 is NoData, 1 is all non-settled grid cells and values increase with likelihood/frequency of a grid cell being classified as settled. To ensure that non-settled grid cells had a value of 0 and avoid confusion during modelling, a value of 1 was subtracted from all raster values, so that they ranged from 0 to 253. Settlement grid cell values (0 to 253) were summed to calculate an "intensity" of settlement measure, with the spatial resolution and grid cell alignment matching the mastergrid.

Given the difference in coordinate systems and spatial resolution between the PNG mastergrid and the Planet settlement data, it was necessary to both re-project and resample the settlement data as part of this process. In resampling the Planet settlement data from 4.77 m to ~100 m, grid cell values in the settlement data product were summed within each grid cell of the mastergrid. This was done by first creating a zonal raster where each grid cell in the mastergrid had a unique ID value. The coordinate system for the Planet settlement raster was also changed to WGS84 (EPSG: 4326) to be in the same as the mastergrid coordinate system. Zonal statistics (sum) were then calculated using the mastergrid zonal raster and the re-projected Planet settlement raster (both in WGS84). These summed values provide a settlement "intensity" measure at ~100m spatial resolution.

3.2.4 Selection of the potential model covariates

As stated earlier, to achieve model parsimony we took steps to ensure that only the best set of covariates that significantly predicted population/population density were included in the final model. This was done in four steps. First, we carried out an extensive covariate selection process using a stepwise regression approach with both forward and backward movements (James et al., 2014; Bruce and Bruce, 2017). Here, covariate selection was based on fitting the appropriate generalized linear model (GLM; McCullagh and Nelder, 1989) to the data, first combined with all the 52 geospatial covariates prepared for the population modelling.

For the model utilising settlement information through building intensity, we assumed a Gaussian distribution for the natural logarithm of the building intensity B_i that is, $\log(B_i) \sim Normal(\bar{B}_i, \sigma_B^2)$. Then, within the context of the GLM, the mean parameter \bar{B}_i was linked to the K geospatial covariates with the identity link function through the linear predictor

$$\bar{B}_i = \eta_i^{(B)} = \beta_0^{(B)} + \sum_{k=1}^K \beta_k x_{i,k} \quad (1)$$

where $\beta_0^{(B)}$ is the intercept or the baseline building intensity when the effect of the geospatial covariates is zero; and $(\beta_1, \dots, \beta_K)$ are the fixed effects coefficients of the K geospatial covariates (x_1, \dots, x_K) assumed to significantly influence the response (e.g., building intensity).

Similarly, the population density $D_i (= N_i/\hat{B}_i)$ is assumed to be Gamma distributed, that is, $D_i \sim Gamma(\alpha, \beta)$, where N_i is the number (count) of people within the i th unit; \hat{B}_i is the predicted building intensity for the i th unit; and $\alpha > 0$ and $\beta > 0$, are the shape and rate parameters, respectively. Then, the mean parameter $\vartheta = \alpha/\beta$ is linked to the J geospatial covariates with the log link function through the linear predictor

$$\ln(\vartheta) = \eta_i^{(D)} = \beta_0^{(D)} + \sum_{j=1}^J \beta_j x_{i,j} \quad (2)$$

where $\beta_0^{(D)}$ is the intercept or the baseline population density when the effect of the J geospatial covariates is zero; and $(\beta_1, \dots, \beta_J)$ are the fixed effects coefficients of the geospatial covariates (x_1, \dots, x_J) .

Note that the model description above is a two-step process used for modelling the building intensity as Gaussian and the population density as Gamma, hence the name Gamma-Gaussian. The model borrows strength from the building intensity, which provides a proxy measure of the settlement intensity.

An alternative modelling approach is to consider using modelling structures which do not employ settlement intensity information. In this case, the model is defined in terms of the population count (response variable). Considering the inherent overdispersion within the data, we assumed that the population count C_i is Negative Binomial distributed, that is, $C_i \sim NegativeBinomial(\mu_i, \phi)$, where μ_i and ϕ are the mean and the overdispersion parameters, respectively. Then, the mean parameter μ_i is linked to the M geospatial covariates with the log link function through the linear predictor

$$\ln(\mu_i) = \eta_i^{(C)} = \beta_0^{(C)} + \sum_{m=1}^M \beta_m x_{i,m} \quad (3)$$

where $\beta_0^{(C)}$ is the intercept or the baseline population count when the effect of the M geospatial covariates is zero; and $(\beta_1, \dots, \beta_M)$ are the fixed effects coefficients of the geospatial covariates (x_1, \dots, x_M) . The models defined above are fitted in R using `glm()` function (equations 1 & 2) and `glm.nb()` function for equation 3. Then, variable selection based on each model was carried out via the `stepAIC()` function of the 'MASS' package using the direction option 'both'. By using both the forward and backward directions, the likelihood of missing any key covariate becomes significantly small.

In the second step, further checks were carried out on the covariates identified as best in step 1, focussing on the issue of multicollinearity which could inflate variance and ensuring that this does not arise. To do this, we used the 'vif' function of the 'car' package. In line with the rule of thumb, vif values less than 5 should be deemed acceptable, while values above 5 are indicative of potential multicollinearity (James et al., 2013). High 'vif' value covariates were then discarded before the further assessments were made.

Furthermore, we refitted the generalized linear model (GLM) in equation (1) using the selected covariates in the third step and examined the coefficients of the parameter estimates for redundancy and any covariates that did not significantly influence response variable were discarded.

Finally in step four, we visually inspected the correlation matrix of the covariates as a final check for multicollinearity. Note that throughout the analyses, covariates selection was carried out before the actual model fitting.

Ultimately, 13 covariates were chosen as the best predictors of building intensity, while 15 covariates were identified as providing the best fit for both the population density and population count models (Table 2). In Figure 7, we show the correlation matrix of the 15 covariates adopted in the final density and population count models, which indicate moderate correlations among the covariates. Further descriptions of the final covariates are provided in Table 2.

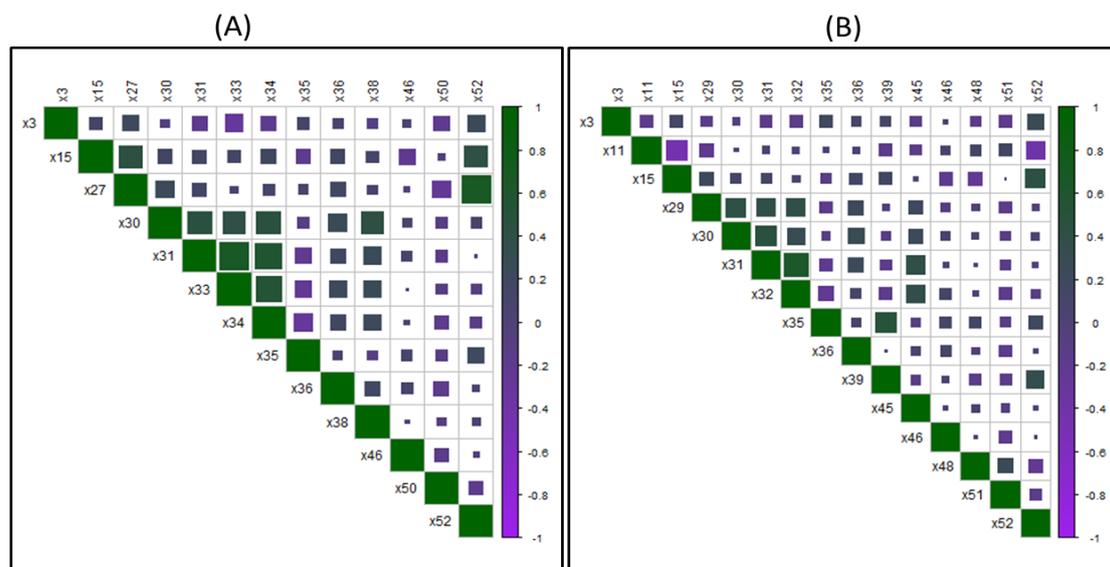


Figure 7. Correlation matrix for A) Final covariates for building intensity model, B) Final covariates for population count and density models. The colour green indicates positive and purple negative correlation. The size of the squares indicates the strength of the correlation.

Table 2: The final model covariates selected via stepwise regression

Variable	Description	Population count model	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	✓	✓	✓

Note: 15 covariates were selected for both the population count and density models, while 13 covariates were selected for the building intensity model.

3.3 Model building

Bottom-up population modelling methods use geo-located survey data of population from a sample of locations with the goal of estimating and predicting the population into other unsampled areas. A statistical model is fitted to these data to estimate population size in unsampled areas, based on the association with spatial covariate information (Leasure et al., 2020 and Wardrop et al., 2018).

In its simplest form, the population of people in a given area of interest is defined as

$$population = \frac{number\ of\ people}{settled\ area} \times settled\ area \quad (4)$$

where the first term, number of people per settled area, is the population density. Note that the 'settled area' in equation (4) can be any covariate that sufficiently defines the population density, e.g., the total area occupied, number of buildings, number of households, building intensity.

Given that population data are count data, a natural assumption is that the data follow a Poisson distribution with equal mean and variance. Then the geostatistical hierarchical model is given by

$$\begin{aligned}
Y_i &\sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \beta_0 + \sum_{l=1}^L \beta_l x_{il} + f_{struct}(s_i) + f_{unstr}(s_i) \\
\pi(\beta_0) &\propto 1 \\
\beta_l &\sim \text{Normal}(\mu_\beta, 1/\tau_{\beta_l}) \\
f_{unstr} &\sim \text{Normal}\left(0, \frac{1}{\tau_{unstr}}\right) \\
\tau_j &\sim \text{Gamma}(\alpha_\tau, \beta_\tau), \text{ where } j \in \{\beta, unstr\} \quad (5)
\end{aligned}$$

where Y_i and λ_i are the total and average number of people in area i ; β_0 and $(\beta_1, \dots, \beta_L)$ are the intercept and the fixed effects coefficients of some L geospatial covariates (x_1, \dots, x_L) ; $f_{struct}(s_i)$ and $f_{unstr}(s_i)$ are spatially correlated and spatially independent random effects with respect to a given location. Note that for point-level observations, s_i is the longitude-latitude of the observation point, while for areal level observations s_i is the centroid. Including the decomposed spatial random effects is in line with the first law of Geography, which states that locations that are close to each other are more similar in characteristics than those that are further apart (Tobler, 1970). Thus, this ensures that random effects due to areas that share common boundaries and are therefore more likely to be similarly affected by factors such as migration/displacement due to unrest, unfriendly climate, etc. are explicitly accounted for.

Within the context of the integrated nested Laplace approximation (INLA, Rue et al., 2009), the structured or correlated spatial random effect $f_{struct}(\cdot)$ is a Gaussian random field (GRF) given by

$$f_{struct} \sim \text{Normal}(0, \Sigma(\psi)) \quad (6)$$

where $\Sigma(\psi)$ is a dense isotropic distance dependent covariance matrix of the Matérn family

$$C_\nu(s) = \frac{\sigma^2}{\Gamma(\lambda)2^{\nu-1}} (\kappa \|s_i - s_j\|)^\nu K_\nu(\kappa \|s_i - s_j\|) \quad (7)$$

where Γ is a gamma function; K_ν is the modified Bessel function of the second kind; ν and κ are the parameters of the covariance; $\|s_i - s_j\|$ is the Euclidean distance between spatial locations s_i and s_j . The INLA-SPDE approach circumvents the well-known computation big 'n' problem associated with $\Sigma(\psi)$ and reduces the computational cost from $O(n^3)$ by approximating the GRF using a more sparse Gaussian Markov random field (GMRF) by discretizing the continuous space using a mesh (Rue and Held, 2005). However, one key condition under which equation (5) is valid is that the mean and the variance of the data are equal (McCullagh and Nelder 1989). This condition is rarely met within the context of population modelling due to overdispersion of data, that is, $E(Y) \neq \text{var}(Y)$. In other words, population data are inherently overdispersed and the use of statistical models, which explicitly account for this, is often required. In which case, we have

$$Y_i \sim \text{Poisson}(D_i \times B_i) \quad (8)$$

where D_i and B_i are the population density and the building intensity, respectively. Usually, the density variable is assumed to follow a lognormal distribution (Leasure et al., 2020),

however, a Gamma distribution has also been shown to work well (Nnanatu et al., 2023). It is always recommended that the various models be tested and the best model is selected based not only on the structure of the dataset and the processes that generated it, but also on the statistical performance. For the lognormal distribution, $D_i \sim \text{LogNormal}(\bar{D}_i, \sigma_D^2)$ where \bar{D}_i and σ_D^2 are the location and scale parameters respectively. $D_i \sim \text{Gamma}(\alpha, \beta)$, where α, β are the shape and rate parameters $E[D_i] = \alpha/\beta$, $\text{var}(D_i) = \alpha/\beta^2$, then, \bar{D}_i is linked to the geospatial covariates through the linear predictor

$$\log(\bar{D}_i) = \eta_i^{(D)} = \beta_0 + \sum_{k=1}^{15} \beta_k x_{ik} + f_{struc}(s_i) + f_{unstr}(s_i) \quad (9)$$

where $(\beta_1, \dots, \beta_{15})$ are the 15 geospatial covariates listed in Table 2 as the best predictors for population density.

Note that equation 9 can be extended to include random effects for settlement type, provinces as well as their nesting structures, so that the full hierarchical model structure with population density is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(D_i \times B_i) \\ D_i &\sim \text{Gamma}(\alpha_D, \beta_D) \text{ and } \bar{D}_i = \alpha/\beta \\ \log(\bar{D}_i) = \eta_i^{(D)} &= \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{struc}(s_i) + f_{unstr}(s_i) + f_t(t) + f_{t,p}(t, p) \\ \pi(\beta_0) &\propto 1 \\ \beta_l &\sim \text{Normal}(\mu_\beta, 1/\tau_\beta) \\ f_{unstr} &\sim \text{Normal}\left(0, \frac{1}{\tau_{unstr}}\right) \\ f_t &\sim \text{Normal}(0, 1/\tau_t) \\ f_{t,p} &\sim \text{Normal}(0, 1, \tau_{t,p}) \\ \tau_j &\sim \text{Gamma}(\alpha_\tau, \beta_\tau), \text{ where } j \in \{\beta, unstr, t, tp\} \end{aligned} \quad (10)$$

where $f_t(\cdot)$ and $f_{t,p}(\cdot)$ are the settlement type and province-settlement type nested random effects respectively. Similarly, the full hierarchical model structure with building intensity is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(D_i \times B_i) \\ \log(B_i) &\sim \text{Normal}(\bar{B}_i, 1/\tau_B) \\ \bar{B}_i = \eta_i^{(B)} &= \beta_0 + \sum_{k=1}^{13} \beta_k x_{ik} + f_{struc}(s_i) + f_{unstr}(s_i) + f_t(t) + f_{t,p}(t, p) \\ \pi(\beta_0) &\propto 1 \end{aligned}$$

$$\begin{aligned}
\beta_l &\sim \text{Normal}(\mu_\beta, 1/\tau_\beta) \\
f_{unstr} &\sim \text{Normal}\left(0, \frac{1}{\tau_{unstr}}\right) \\
f_t &\sim \text{Normal}(0, 1/\tau_t) \\
f_{t,p} &\sim \text{Normal}(0, 1, \tau_{t,p}) \\
\tau_j &\sim \text{Gamma}(\alpha_\tau, \beta_\tau), \text{ where } j \in \{B, \beta, unstr, t, tp\} \quad (11)
\end{aligned}$$

Then, the predicted population count is given by

$$\hat{y}_i = \hat{D}_i \times \hat{B}_i \quad (12)$$

where

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

and

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

Furthermore, for the models with population count as the response variable, two further assumptions were made about the variance of the observations.

First, we assumed that the mean of the observed counts has a quadratic relationship with the variance in which case the hierarchical structure of the model is described based on a Negative Binomial model as

$$\begin{aligned}
Y_i &\sim \text{NegativeBinomial}(\mu_i, \phi) \\
\log(\mu_i) = \eta_i^{(C)} &= \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{struc}(s_i) + f_{unstr}(s_i) + f_t(t) + f_{t,p}(t, p)
\end{aligned}$$

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

$$\begin{aligned}
\beta_l &\sim \text{Normal}(\mu_\beta, 1/\tau_\beta) \\
f_{unstr} &\sim \text{Normal}\left(0, \frac{1}{\tau_{unstr}}\right) \\
f_t &\sim \text{Normal}(0, 1/\tau_t) \\
f_{t,p} &\sim \text{Normal}(0, 1, \tau_{t,p}) \\
\tau_j &\sim \text{Gamma}(\alpha_\tau, \beta_\tau), \text{ where } j \in \{\beta, unstr, t, tp\} \quad (13)
\end{aligned}$$

where μ_i and ϕ are the mean and the dispersion parameters, respectively.

The second assumption is that the mean population count has a linear relationship with the variance of the observation. We specified this using a Poisson distribution with an

observation-level random effect, also known as, quasi-Poisson with the full hierarchical modelling structure given by

$$\begin{aligned}
Y_i &\sim \text{Quasi-Poisson}(\lambda_i, \phi_i) \\
\log(\lambda_i) = \eta_i^{(C)} &= \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + f_{unstr}(s_i) + f_t(t) + f_{t,p}(t,p) + \xi_i \\
\pi(\beta_0) &\propto 1 \\
\beta_l &\sim \text{Normal}(\mu_\beta, 1/\tau_\beta) \\
\xi_i &\sim \text{Normal}(0, \phi) \\
f_t &\sim \text{Normal}(0, 1/\tau_t) \\
f_{t,p} &\sim \text{Normal}(0, 1, \tau_{t,p}) \\
\tau_j &\sim \text{Gamma}(\alpha_\tau, \beta_\tau), \text{ where } j \in \{\beta, unstr, t, tp\} \quad (14)
\end{aligned}$$

Note that for both the Negative Binomial and Quasi-Poisson models, the predicted population is obtained by back-transforming the linear predictor as exponential. For each of the models defined in equations 9 through to 14, the response variable (e.g., building intensity, population count, population density) is assumed to be conditionally independent of the generic latent field $\mathbf{w} = (\eta, \beta_0, \boldsymbol{\beta}, f_t, f_{tp}, f_{struc}, f_{unstr})$ and the hyperparameters $\boldsymbol{\theta} = (\tau_t, \tau_{t,p}, \tau_\beta)$ so that joint posterior distribution of the latent fields and the hyperparameters given the data is given by

$$\pi(\mathbf{w}, \boldsymbol{\theta} | y) \propto \pi(\boldsymbol{\theta}) \pi(\mathbf{w} | \boldsymbol{\theta}) \prod_{i \in I} \pi(y_i | w_i \boldsymbol{\theta}) \quad (15)$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution, $\pi(\mathbf{w} | \boldsymbol{\theta})$ is a latent Gaussian model (LGM), and $\pi(y | w, \boldsymbol{\theta})$ is the likelihood function of observing the data given the latent field and the hyperparameters. The posterior distribution is then approximated and evaluated using INLA-SPDE as already stated above.

To implement the INLA-SPDE approach, we first built the triangulation of the entire spatial domain also known as a mesh. The 695 nodes mesh used for our models is given in Figure 8.

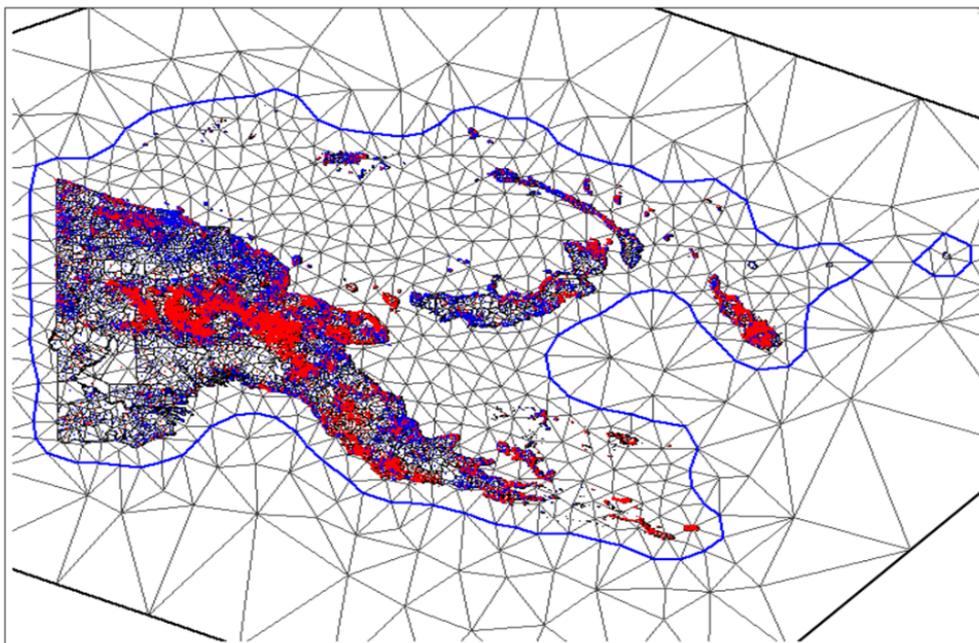


Figure 8: The Delaunay Triangulation of PNG (Mesh) used for the INLA-SPDE implementation with 695 nodes.

The observations are then projected to the mesh nodes using the projection matrix. See Rue and Held (2005) for more information on the use of GMRF via mesh. For full Bayesian inference, we assigned the following prior distributions to the parameters and hyperparameters

$$\beta \sim \text{Normal}(0, 0.01)$$

$$\tau_k \sim \text{Gamma}(1, 0.00005)$$

where $k \in \{t, tp, \beta, \text{unstr}, \text{struc}\}$.

3.3.1 Sampling from the joint posterior distribution

To improve estimates of the posterior marginal distribution $\pi(\theta_i | \theta_{-i}, y)$, we drew =2000 samples from the joint posterior density $\pi(\boldsymbol{\theta} | y)$. In line with the Rao-Blackwellization theory (Robert & Roberts, 2021; Blackwell, 1947; Rao, 1945), drawing samples from the conditional distribution of the parameter θ_1 , say, and then averaging over all the iterations will normally improve the estimation of the marginal distribution $\pi(\theta_1 | \theta_2, y)$.

In addition, within the context of population modelling, interest is usually focussed on having estimates of population totals at various administrative levels, ideally with the corresponding measures of uncertainties that can be valuable for policy design and implementation. While INLA automatically calculates the 95% credible intervals for the mean values, it is not straightforward to obtain the same for the sum of the means (total) because the sum of quantiles is not the same as the quantile of totals. To obtain the quantile of the totals, we need to generate a distribution of the totals and then calculate the quantities of interest.

3.3.2 Model code

All model calculations were done in 'R'. The scripts of the final model can be downloaded from GitHub (https://github.com/wpgp/PNG_Bottom_Up_Modelling).

3.4 Age/sex disaggregation

Population pyramids for administrative units (i.e., LLG, district and province levels) were initially produced using age/sex observations from the malaria survey data. The survey data contained missing age and sex information, which was denoted as both “99” and “NA” in the dataset. Following NSO confirmation that “99” in the dataset represented missing values, all these values were changed into “NA”. NA values were then removed before the age-sex proportions were calculated.

Besides the missing age or sex values of the observations, the initial survey dataset was also spatially incomplete. As age/sex pyramids were required for every LLG, a method for estimating the LLG level age/sex distributions was developed.

To do this, the survey data and the administrative units from the shape files were combined by administrative unit identifiers to create a complete list of LLGs. The joined file had the complete set of the LLGs, however only 16,903 CUs had any age and sex data.

To take account of this missing data, the population pyramid of these administrative units were replaced by the next level up population pyramid, 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. The malaria data based observed age/sex pyramid at national level is shown in Figure 9.

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.

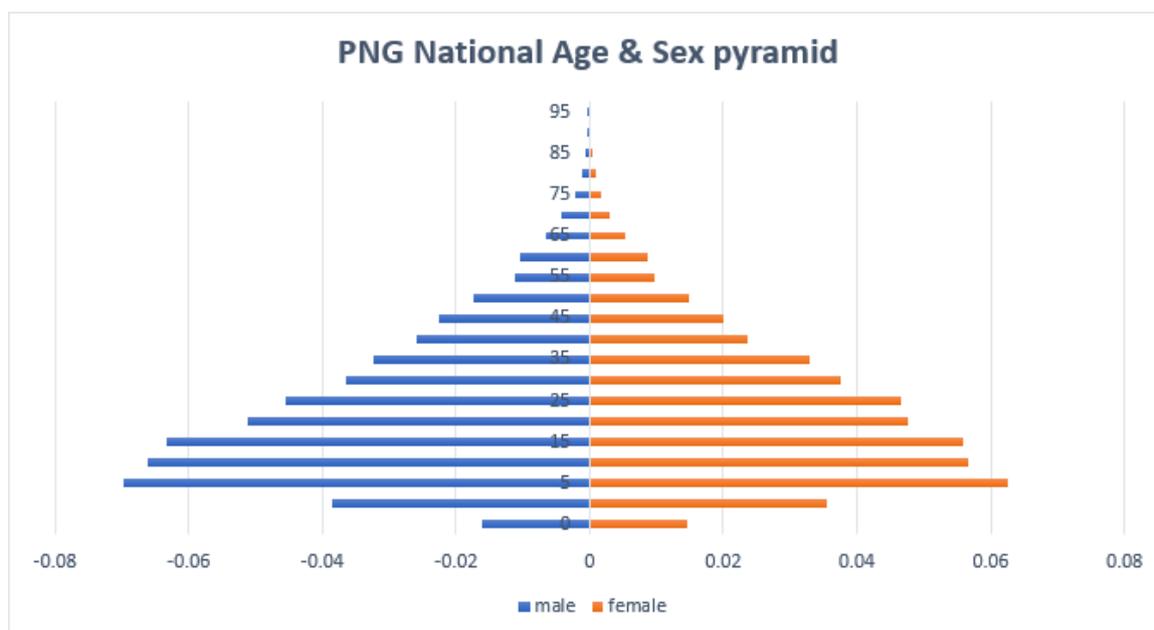


Figure 9: Observed age/sex proportions in the Malaria dataset

4. RESULTS

The following subsections will summarise the performance and outputs of the three INLA model applications:

- Gamma-Gaussian (abbreviated as GG) – which uses a two-step process by first modelling and predict building intensities in inhabited areas that were covered by tree canopies and could not be captured by the camera. Then using the predicted building intensity, more reliable estimates of population density and hence population counts, are obtained.
- Quasi-Poisson (abbreviated as QP) – which accounted for overdispersion within the data by assuming that the variance of the observations is linearly related to the geospatial covariates. Observation-level random effect was included to account for the variabilities in the observation due to differences in census units.
- Negative Binomial (abbreviated as NB) – which assumes that the observed count has a negative binomial distribution and that there is a quadratic relationship between the observation variance and the model covariates. The NB model has inbuilt parameter for estimating data overdispersion.

The estimates will also be compared with the National Rotarians Against Malaria (RAM) data as an alternative estimate and with satellite imagery.

4.1 Model Fit Assessments and Cross-Validation

4.1.1 Posterior Simulations

The model chains are sufficiently exploring the posterior parameter space (Figure 10) which indicates good mixing. Sampling path plots and the associated histograms based on QP and NB are provided in the Appendix.

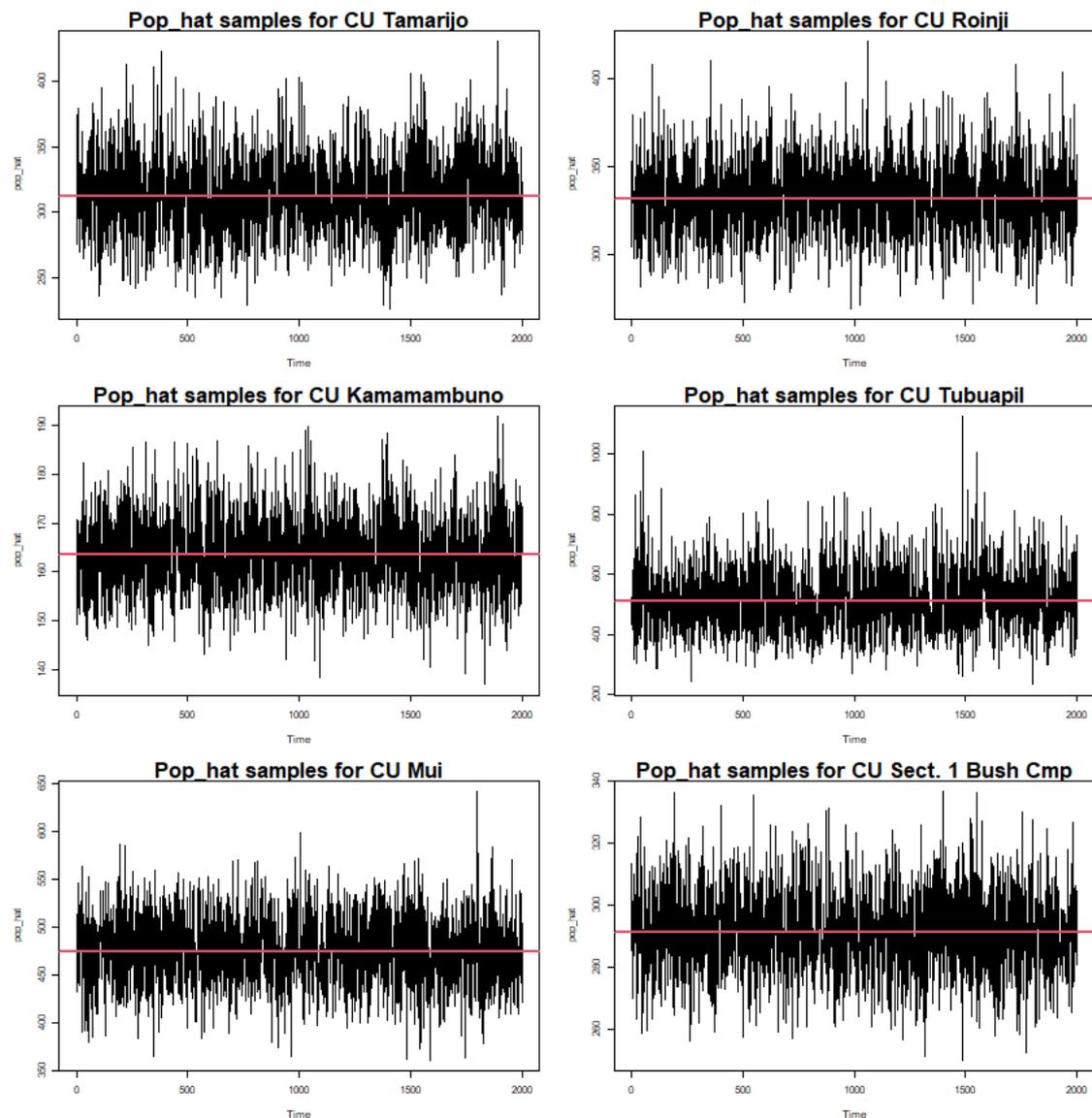


Figure 10: Sampling paths plots of six randomly selected CUs for the posterior simulation based on the GG model.

The spatial surfaces of means and standard deviations of the QP and NB models are almost identical, unlike that of the GG (Figure 11). By including settlement information, the GG model had the advantage of fully exploring the heterogeneities across the various settled areas.

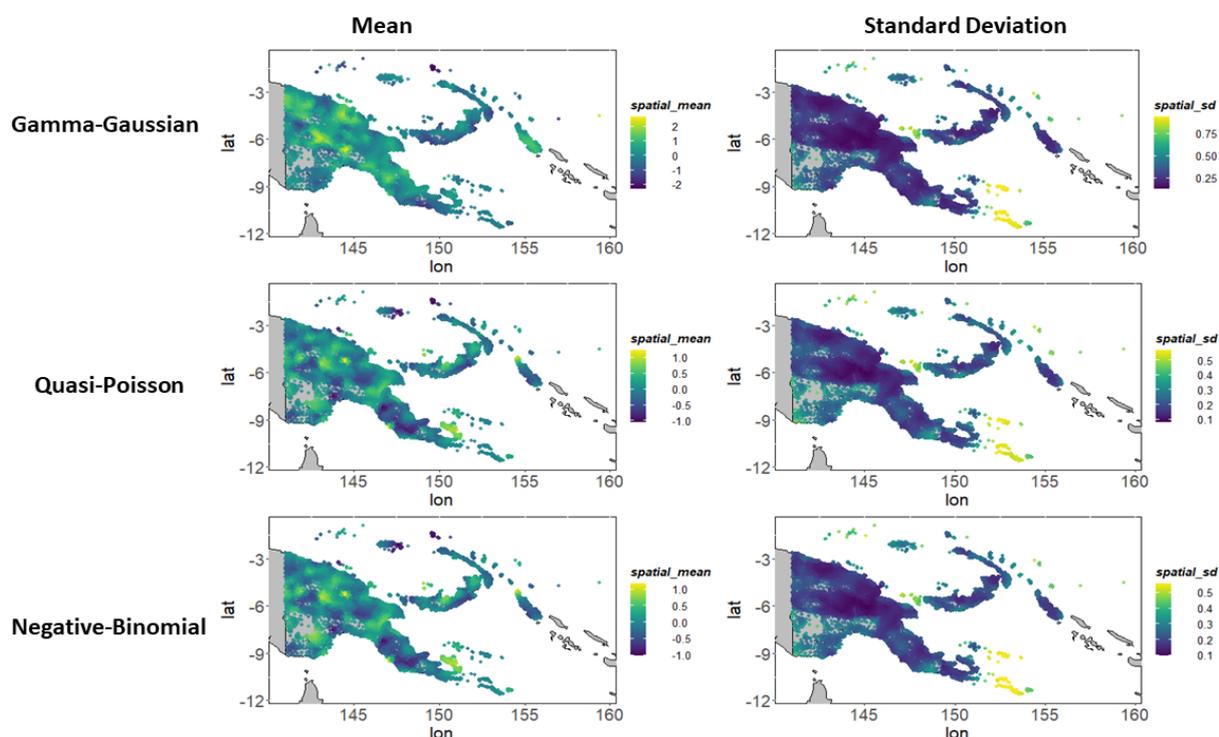


Figure 11: Posterior estimates of the mean (left) and standard deviations (right) of the spatial random effects across the three models

4.1.2 Model Cross-validation

Cross-validation is when the model is trained multiple times using a randomly selected proportion of the observations, and the predictions compared against the remaining observations. Since the observations have inherent uncertainties, cross validation tests the robustness of the model to ensure that it is not over- or under-fitted. Overfitting occurs when the model does not just reproduce the information from the observations but also includes noise. Testing the robustness of the model is important as the model may be required to accurately predict population for additional study sites.

Cross-validation metrics for training ten times using a randomly selected 80% of the observations are shown below in Table 3 and in Figure 12. In this study model validation was undertaken at three-levels – model, simulation and train-test level. First, how well the population estimates from the model predicted actual observed values was assessed. Second, the fit of the posterior estimates of population (based on the 2×10^3 simulations) across the three models to the observed data was evaluated. Finally, out-of-sample cross validation was achieved by randomly dividing the data into training (80% of the samples) and testing (20% of the subsamples). Models trained using the training subsamples were then used to predict the population values of the testing data.

Best model selection was based on the various model fit metrics used in this study – Accuracy, Absolute Bias, Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE). For all but accuracy, the smaller the metric value, the better the model fits the data.

All three models (GG, QP, NB) performed well in both in-sample and posterior simulations, however, the performance of the NB model in the cross validation dropped significantly (Table 3, Figure 8). The GG model slightly outperformed the QP model, therefore, the GG model was chosen as the best model.

Table 3: Model fit assessment and cross validation

Metric	Model								
	Model-based (in-sample)			Posterior simulation			Cross-validation		
	GG	QP	NB	GG	QP	NB	GG	QP	NB
% Accuracy	99.93	99.88	82.76	99.93	99.11	85.05	99.91	99.91	95.49
Δ Bias	0.41	1.2E-6	43.97	2.11	0.03	46.49	0.11	0.92	39.39
RMSE	1.81	1.36	272.60	4.21	1.96	270.69	1.68	2.30	292.42
MAE	0.76	1.05	148.18	2.45	1.33	152.49	0.96	1.67	184.75
Imprecision	1.76	1.38	269.04	3.65	1.96	266.68	1.68	2.11	289.79

Note. GG – Gamma-Gaussian; QP – Quasi-Poisson; NB – Negative Binomial. Posterior estimates were based on 2×10^3 posterior draws. Both the model-based and posterior simulation-based totals are similar and also within the 95% credible interval across the three models.

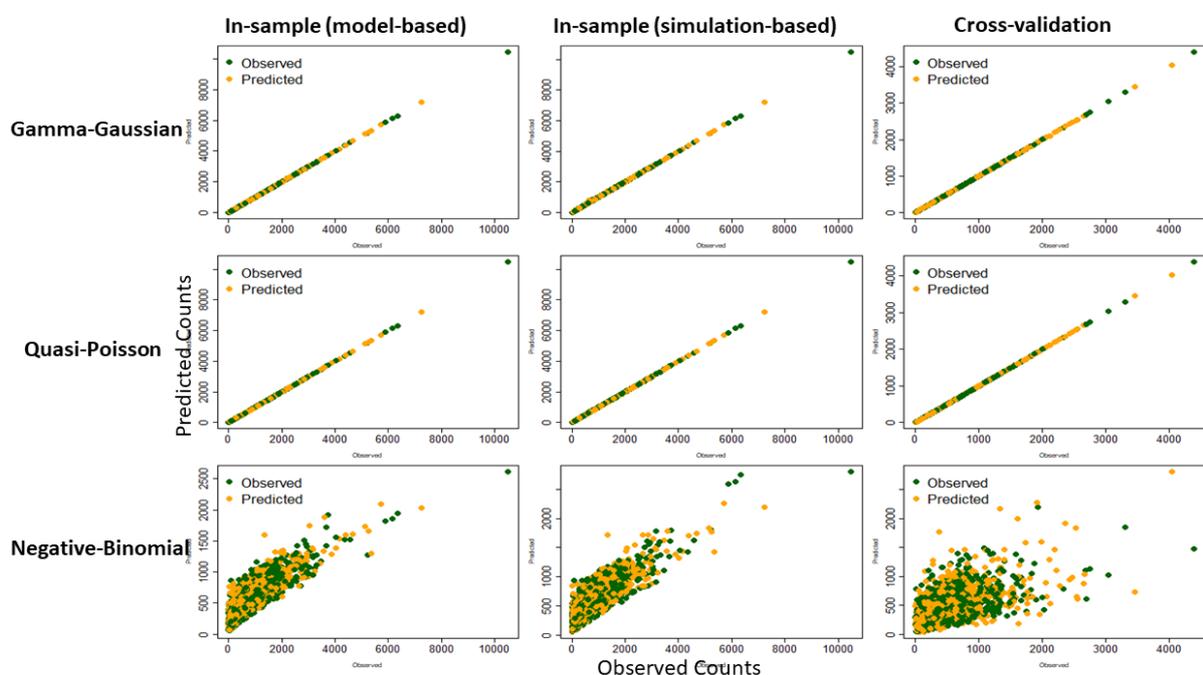


Figure 12: Model fit and Cross validation across the three models. The GG and QP models provided better fit to the data than the NB model which did well at both model-based and simulation-based in-sample prediction but not as good at out-of-sample cross validation.

4.2 National-level Estimates

The current projections for PNG for 2022 are 9.6 million (US Census Bureau, 2022) and 10.1 million (UNDESA 2022).

Table 4 shows the estimated national population for the three models. All estimates are very similar, between 11 and 12 million with quite narrow credible intervals. The best model (GG) estimates the national population as ~11.78M with 95% credible interval of (lower = ~11.64M, upper = ~12.03M).

The population model provides LLG level population estimates with national coverage for PNG, along with the estimate of the number of people belonging to different age-sex groups. These model-based population estimates most likely represent the time period around 2020-21, corresponding to the malaria survey and Urban Structural Listing survey observations (2019-2021; median year: 2020) and the period when the satellite imagery used to generate settlement footprints was captured (2021).

The national estimate of RAM is 9,776,518 (2021) for PNG. RAM visits most areas of PNG every three years, although they only focus on rural areas. Some areas are inaccessible due to conflicts and some very rural areas have geographical challenges to access. In these cases, RAM uses the last census results as a baseline and applies a spatially varying annual growth rate of 2.1-3.9 percent. Although this dataset is not officially approved, it is based on the most frequent observations collected during regular malaria bednet campaigns, and thus is shown as an alternative estimate.

Table 4: Model-based and simulation-based national totals across the three models

Model	Population Estimates			
	Model-based	Posterior simulation-based		
	Total	Total	Lower	Upper
GG	11,705,452	11,781,559	11,644,772	12,028,038
QP	11,430,876	11,429,300	11,347,392	11,532,996
NB	11,225,746	11,185,397	11,060,366	11,327,769

Note: GG - Gamma-Gaussian; QP - Quasi-Poisson; NB - Negative. Posterior estimates were based on 2×10^3 posterior draws. Both the model-based and posterior simulation-based totals are similar and also within the 95% credible interval across the three models.

4.3 Province-level Estimates

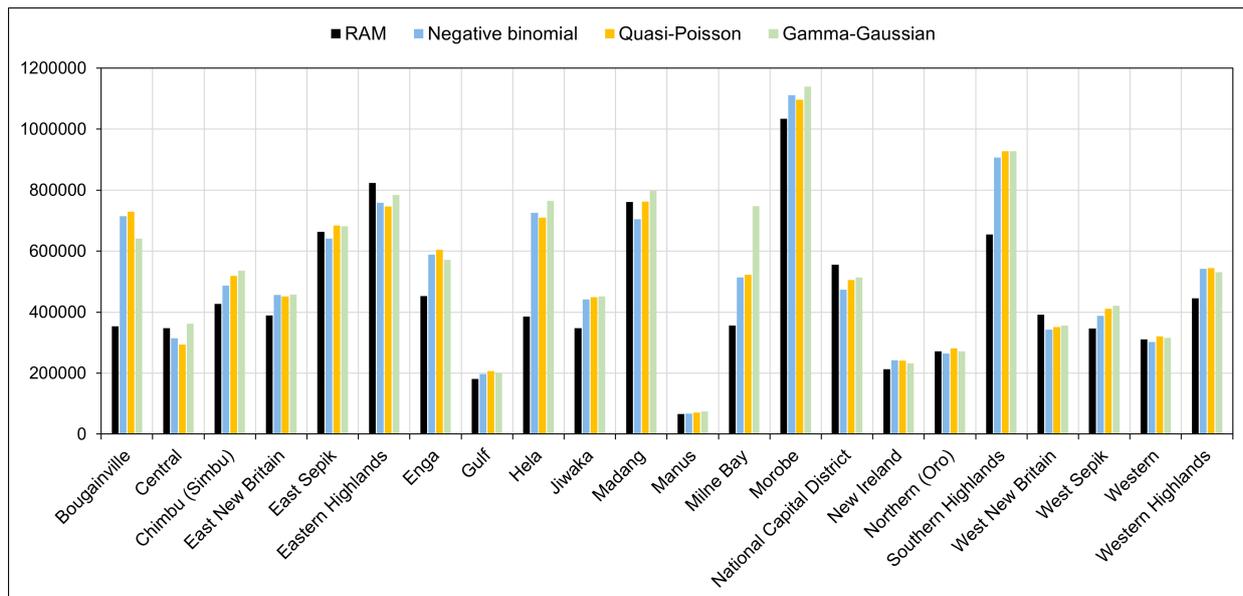


Figure 13: Comparing provincial population estimates across the various models with the RAM data

Population estimates are similar for 20 out of 24 provinces across the four models (Figures 13 and 14). This somewhat increases the confidence in the model results. However, estimates of population totals across the various provinces differed widely between the RAM data and the three models for Hela, Southern Highlands, Milne Bay and Enga. These provinces would benefit from additional field surveys to confirm the true population size.

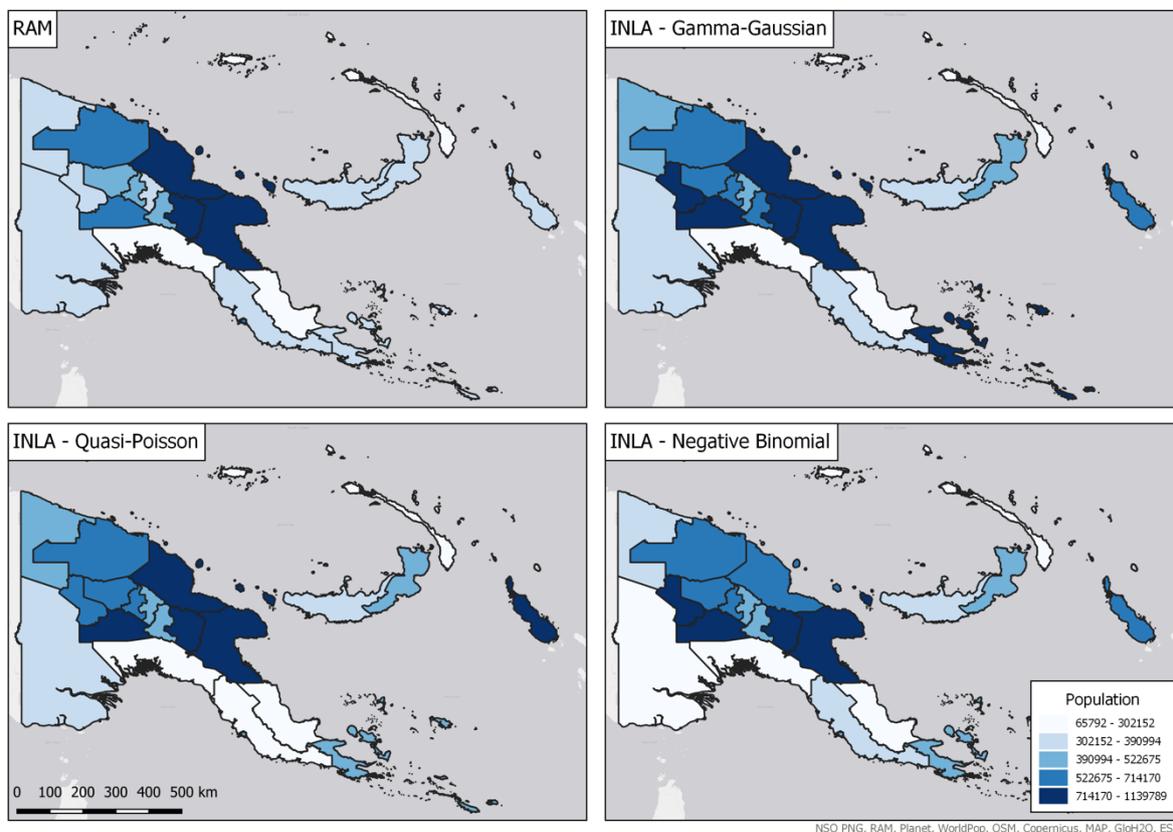


Figure 14: Comparisons of the spatial surfaces of total population estimates across the three models and the RAM data. Estimates show relatively similar spatial trends

4.4 District-level Estimates

The district level estimates for the three models are mostly similar to that of the RAM data, but with some exceptions (Figure 15).

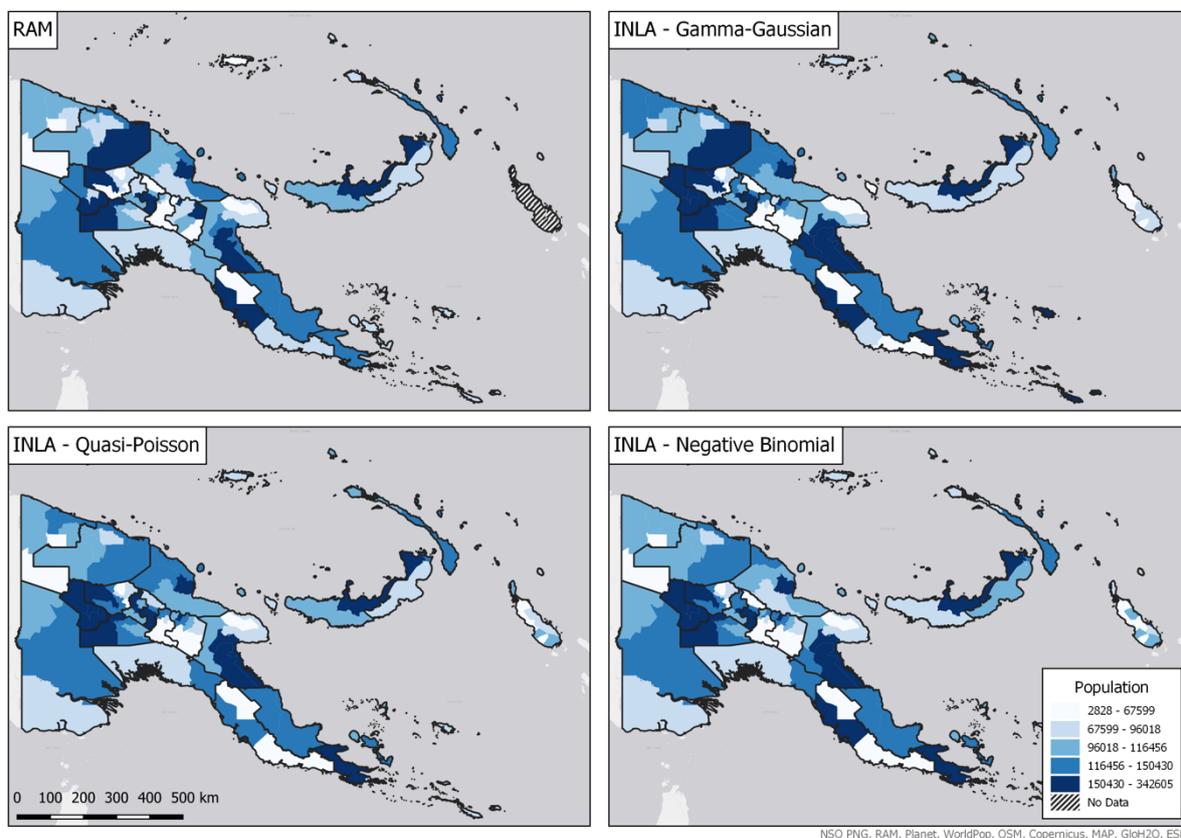


Figure 15: Comparing district level population estimates across the three models and the RAM data

4.2.5 Comparison with satellite images

The CU level model results were compared to satellite images at a few locations in Port Moresby to check whether the residential / non-residential variations were captured well in the model. Figure 16 shows examples, and throughout the country these were typically well captured.

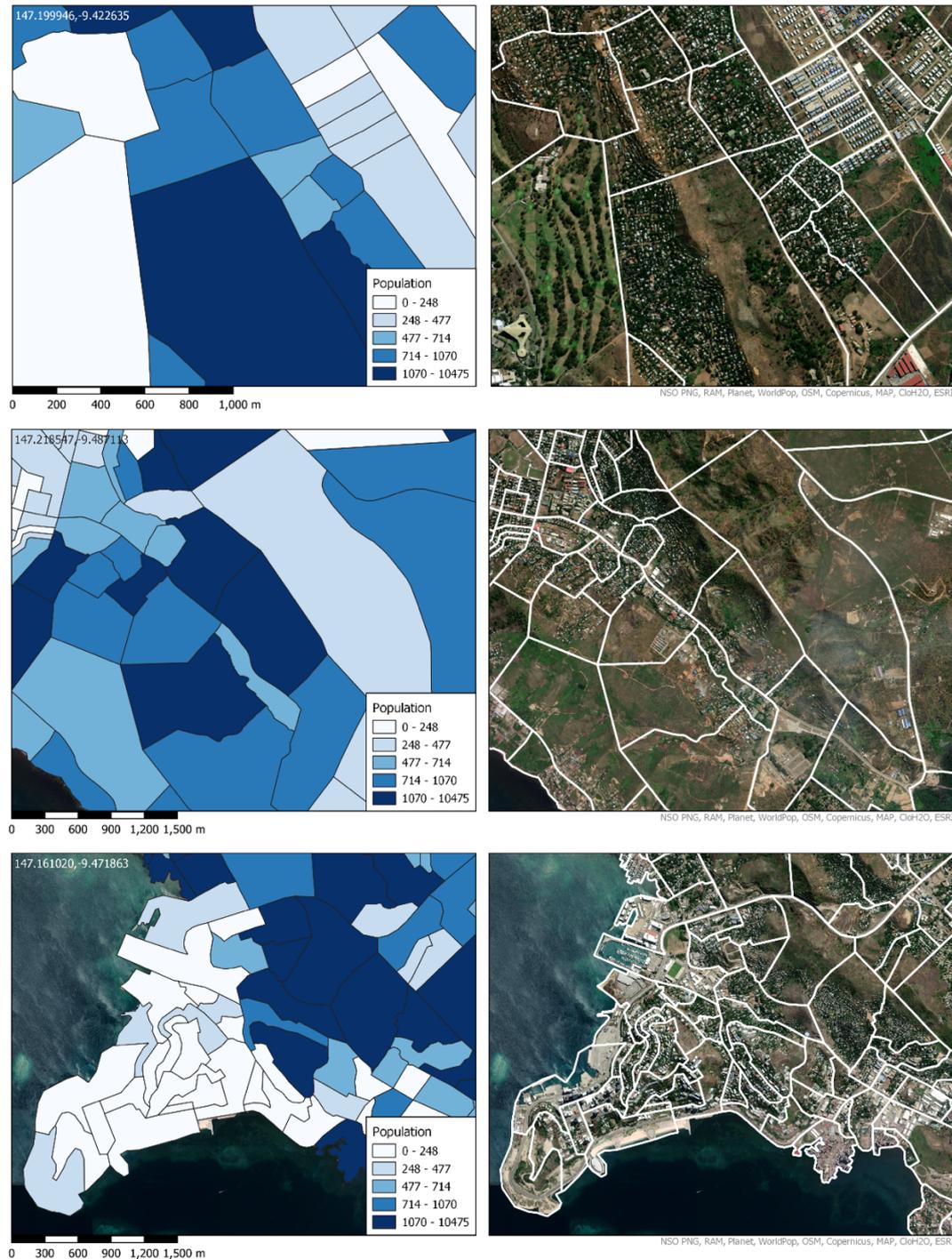


Figure 16: CU level total population (left panel), Esri satellite image base layer with white CU boundary overlay (right panel)

4.3 Age/sex disaggregation results

The population modelling work produced pyramids that characterise PNG's populations as being predominantly young and growing ('Expansive'). These types of pyramids are larger at the bottom and shrink with the increasing age groups. The estimated results indicate a slightly smaller under-five population than the next age group (Figure 17), which may suggest decreasing birth and death rates.

The percentage of children and adolescents (under 15 years old), working age population (15 to 65 years old) and elderly population (above 65 years old) to the total population are estimated to be 35.5%, 61.8% and 2.6% respectively. This is in line with the World Bank 2021 estimations for the country (WorldBank WDI 2022).

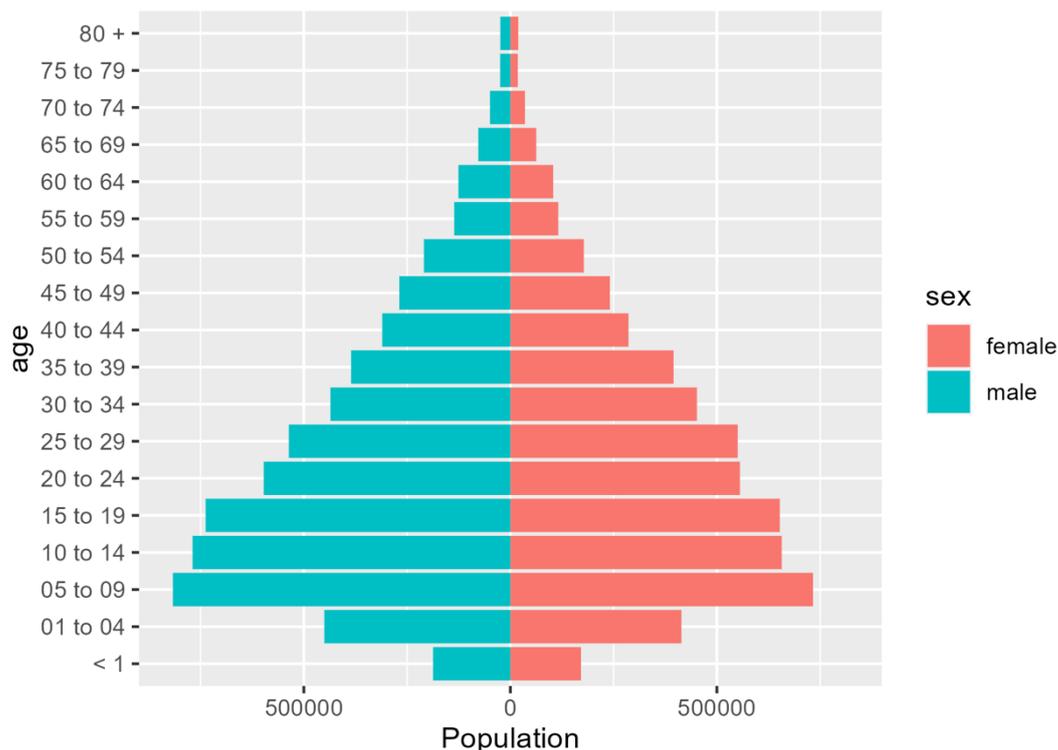


Figure 17: Age and sex pyramid of PNG national estimated population

Estimated provincial population pyramids show similar results to the national trend, where younger age groups are more prevalent than the elderly population (see Appendix 4 for provincial population pyramids). Some provinces have a slightly deformed shape of an expansive population pyramid, particularly National Capital District (NCD), which displays more aging population trends than other provinces (Figure 20). NCD is assumed to be more developed and urbanised than the other provinces, which may have resulted in lower birth, death and fertility rates than others.

NCD's population pyramid estimate also shows a higher proportion of 25 to 29 year olds than the national pyramid or pyramids for other provinces. It is unclear whether this deformed shape is a result of the small sample size, or if the district is in different phase of demographic transition, or if the district attracts young adult workers.

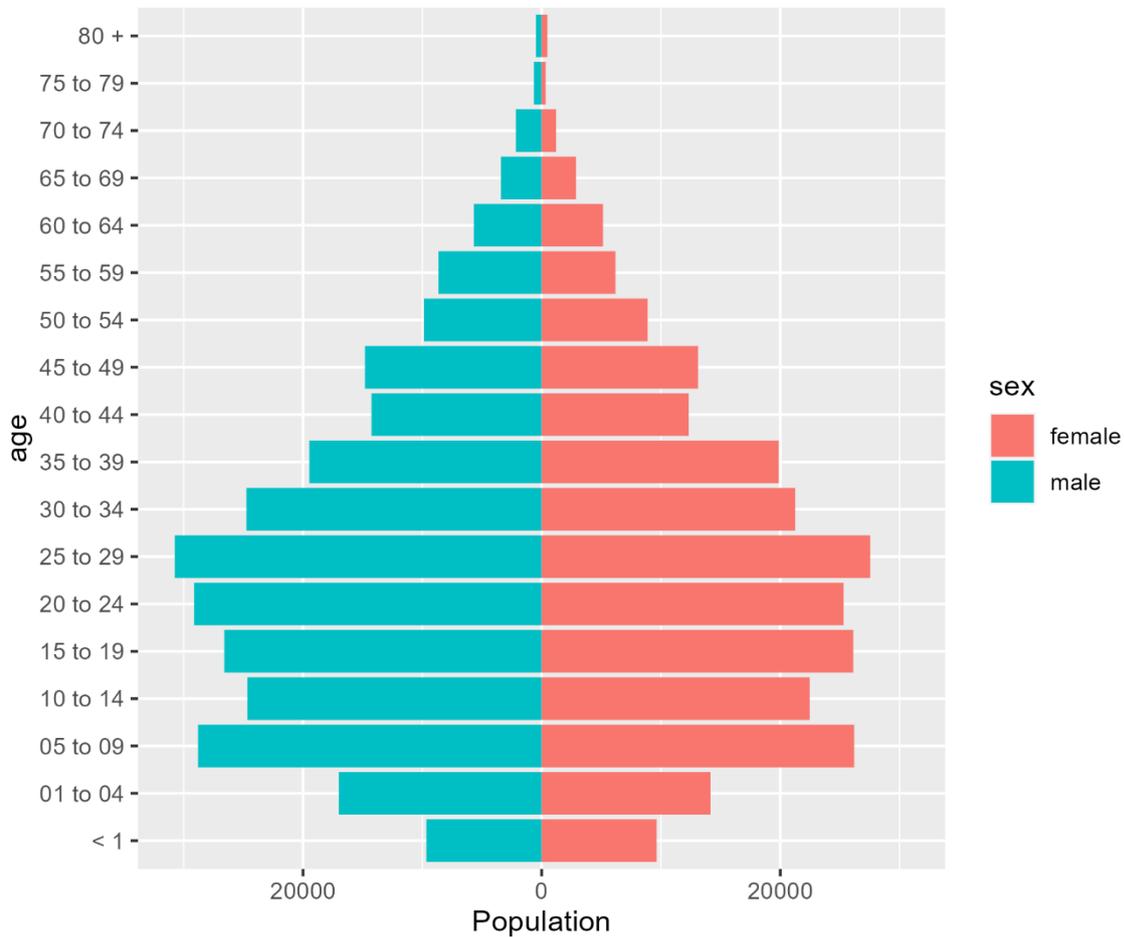


Figure 18: Age and sex pyramid of National Capital District, PNG estimated population

5 Discussion

A two-step Bayesian hierarchical geostatistical model was applied to the PNG integrating 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. The approach facilitated simultaneous accounting for the multiple levels of variability within the data hierarchy. The best fit model indicated that the total population of PNG based on the compiled datasets is 11.78M with a 95% credible interval of 11.64-12.03M, representing the years 2020-21.

The modelling approach used in estimating the population of Papua New Guinea (PNG) initially gave rise to inflated population numbers (e.g. a national total of 17 million). It is believed that these inflated figures were derived from overdispersion of observations and the permanent canopy cover of some areas which obscured satellite imagery. See section 5.1 below for more details. These initial findings suggested the need for the adoption of more advanced statistical techniques, and to continue with the iterative modelling process (Figure 5).

Three alternative statistical population models were explored:

- A negative Binomial (NB) based model which includes an overdispersion parameter and can be independent of settlement information. This model thus did not use the Planet imagery-based settlement data.
- A quasi-Poisson (QP) based model which is a modified version of the regular Poisson density but also includes an overdispersion parameter explicitly specified within the modelling framework.
- A two-step model solution also called Gamma-Gaussian (GG) model which uses a model-based validated settlement information to calculate population density.

These models were fitted within a Bayesian hierarchical modelling framework using the INLA-SPDE approach for efficiency. Following rigorous model fit assessments, posterior simulations and cross-validations, the GG model, which accounted for unobserved effect of settlement type, was selected to be the best fit model, largely due to its highest predictive ability.

National estimates based on the Gamma-Gaussian (GG) model are, total = 11.78M (95%CI, lower = 11.64M, upper = 12.03M), while the national estimates based on the other models are 11.43M (11.35M, 11.53M) for Quasi-Poisson (QP) and 11.18M (11.06M, 11.33M) for Negative Binomial (NB) models. Although both QP and NB models provided narrower 95% credible bounds at the national level estimates than the GG model, GG provided the highest overall predictive ability. This is mostly because the GG model has the advantage of borrowing strength from settlement data, unlike the QP and NB models. Thus, the GG model was able to estimate population numbers in areas that both QP and NB models missed due to their model structures. This also explains why the GG-based estimates are slightly higher than that of QP and NB models.

5.1 Issues with the previous model results explained

Some previous iterations of the population model produced inflated national population numbers that received significant attention in the region (e.g. an estimate of 17 million for the country). The size of the variance of population data is usually larger than the size of the average population count. This is known as 'overdispersion', and population estimates based

on any statistical modelling technique that did not account for this inherent overdispersion is most likely to produce misleading results. To circumvent this, we model population counts in terms of population density (using settlement data) where population density is the number of people in a given area divided by the settled area (see previous sections for details on this). This simply ensures that estimates of population are calculated for settled areas only, thereby minimizing potential biases in the estimates.

However, estimates of population density are only as good as the input settlement data used. A too high (or too low) value for settlement data results in a too small (or too large) estimate of population density in the areas of interest. While too low population density estimates will tend to underestimate the population, too high population density estimates will normally overestimate the population numbers. This latter issue arose in early model outputs, where results suggested an inflated estimate of 17 million. This was due to too low (or even zero) population intensity values in the Planet settlement data, mainly for locations that were obscured from the satellite observations due to forest canopy cover.

Following standard scientific processes, a model testing and result triangulation approach was adopted (see figure 5). To provide independent (baseline) estimates of populations, advanced statistical methods were adopted that intrinsically accounted for overdispersion and avoided the use of the Planet settlement data. This was implemented using two different but related model solutions, as outlined above.

Further, a robust two-step modelling technique was used that first corrected the potential biases in the Planet data by estimating under-canopy settlement areas, and then used this updated settlement information to calculate estimates of population density. All three approaches resulted in very similar national and lower admin levels (Provinces, Districts & LLGs) total population, between 11 and 12 million at the national level.

Following rigorous model fit assessments and model validation exercises (i.e. following the iterative processes outlined in figure 5) the best performing model that the project team have the highest level of confidence in was selected as providing the final result, with an estimated 11.78 million population at national level (95% credible interval: 11.64-12.03M).

5.2 Limitations

These population estimates most likely represent the 2020-21 time period, but because of the different ages 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 an average in the overlapping CUs 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.

8. CONCLUSIONS AND RECOMMENDATIONS

Census-independent high resolution population estimates were produced for PNG using household survey datasets and other geospatial data using a novel satellite-image-based statistical methodology. The 2020-21 total estimated population is 11.8M with a 95% credible interval of [11.6-12M]. This relatively narrow interval implies high confidence in the result, but field validation is still recommended due to uncertainties in the input datasets.

Population models have inherent biases and unquantified errors even when a set of models point to the same result. Furthermore, population models will never be able to compete with the accuracy and data richness of a high quality population and housing census, and therefore, this exercise should not be seen as a replacement for the upcoming population and housing census.

Estimates of population totals broadly agreed with the RAM estimates across 20 of the 24 provinces; however, provincial totals differed significantly between the RAM data and the three models for Hela, Southern Highlands, Milne Bay and Enga. Therefore, we recommend extra data collection exercises in these affected provinces.

The analyses represent the first of their kind in PNG and could (i) support the preparations for the upcoming population and housing census, (ii) provide a method for producing estimates in any areas not able to be enumerated in the upcoming census, (iii) and provide a mechanism to make use of routine surveys in inter-censal periods to update population estimates at small area scales. A recent co-authored UNFPA-WorldPop report outlines the opportunities that spatial modelled population estimate methods offer to support the census process: <https://www.unfpa.org/resources/value-modelled-population-estimates-census-planning-and-preparation>.

CONTRIBUTIONS

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). Initial statistical modelling was done by Hal Voepel while 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.

LICENSE

These data may be redistributed following the terms of a [Creative Commons No Derivatives Attribution 4.0 International \(CC BY-ND 4.0\)](#) license.

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.

SUGGESTED CITATION

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.

REFERENCES

- Blackwell D. (1947). "Conditional expectation and unbiased sequential estimation". *The Annals of Mathematical Statistics* 18: 105–110
- Boo, G., E. Darin, D. R. Leasure, C. A. Dooley, H. R. Chamberlain, Lázár, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W. and Tatem, A. J. (2022). "High-resolution population estimation using household survey data and building footprints." *Nature Communications* 13(1): 1330. DOI: 10.1038/s41467-022-29094-x. <https://doi.org/10.1038/s41467-022-29094-x>
- Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O'Reilly Media.
- Esri (2018). ArcGIS Pro 2.1 Redlands, CA: Environmental Systems Research Institute
- Gareth, J., Witten, D., Hastie, T. and Tibshirani, R.(2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.
- Havard, R., Martino, S. and Chopin, N. (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
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

- 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.
- Qader S H, Harfoot A, Kuepie M, Darin E, Juran S, Lazar AN, Tatem AJ. 2022. National automatic preEnumeration Areas (preEAs) in Burkina Faso (2019), version 1.0. WorldPop, University of Southampton. Doi:10.5258/SOTON/WP00731
- Qader, S., Lefebvre, V., Tatem, A. et al. Semi-automatic mapping of pre-census enumeration areas and population sampling frames. *Humanit Soc Sci Commun* 8, 3 (2021).
<https://doi.org/10.1057/s41599-020-00670-0>
- 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>.
- Rao, C. R. (1945). "Information and accuracy attainable in estimation of statistical parameters". *Bulletin of the Calcutta Mathematical Society* 37:81–91
- Robert, C.P., Roberts, G. (2021). –"Blackwellisation in the Markov Chain Monte Carlo Era". *International Statistical Review*, 89(2), pp. 237-249
- Stan Development Team. 2020. RStan: The R interface to Stan. R package version 2.21.2.
<https://mc-stan.org>.
- Stan Development Team. 2021. Stan Modeling Language Users Guide and Reference Manual, 2.28. <https://mc-stan.org>.
- Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* (Supplement: Proceedings, International Geographical Union. Commission on Quantitative Methods), 46: 234–240. DOI:10.2307/143141.
- United Nations, Department of Economic and Social Affairs, Population Division (2022). *World Population Prospects 2022: Data Sources*. (UN DESA/POP/2022/DC/NO. 9).
- UNFPA Technical Brief (2017). *New Methodology: a hybrid census to generate spatially disaggregated population estimates*. December 2017.
https://www.unfpa.org/sites/default/files/resource-pdf/Hybrid_Census_Brief_v9.pdf
- US Census Bureau (2022) Subnational Population Data by Geographic Area. Papua New Guinea - Local Level Governments (ADM3). <https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/subnationalpopulation.html>
- The World Bank, World Development Indicators (2022). [Population ages 15-64, total](https://data.worldbank.org/indicator/SP.POP.1564.TO) [Data file]. Retrieved from <https://data.worldbank.org/indicator/SP.POP.1564.TO>
- 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 (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and 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). <https://dx.doi.org/10.5258/SOTON/WP00644>

WorldPop and Institut National de la Statistique et de la Démographie du Burkina Faso. 2021. Census-based gridded population estimates for Burkina Faso (2019), version 1.0. WorldPop, University of Southampton. doi: 10.5258/SOTON/WP00687. <https://wopr.worldpop.org/?BFA/Population/v1.1>

APPENDIX 1: Initial covariate characteristics and source links

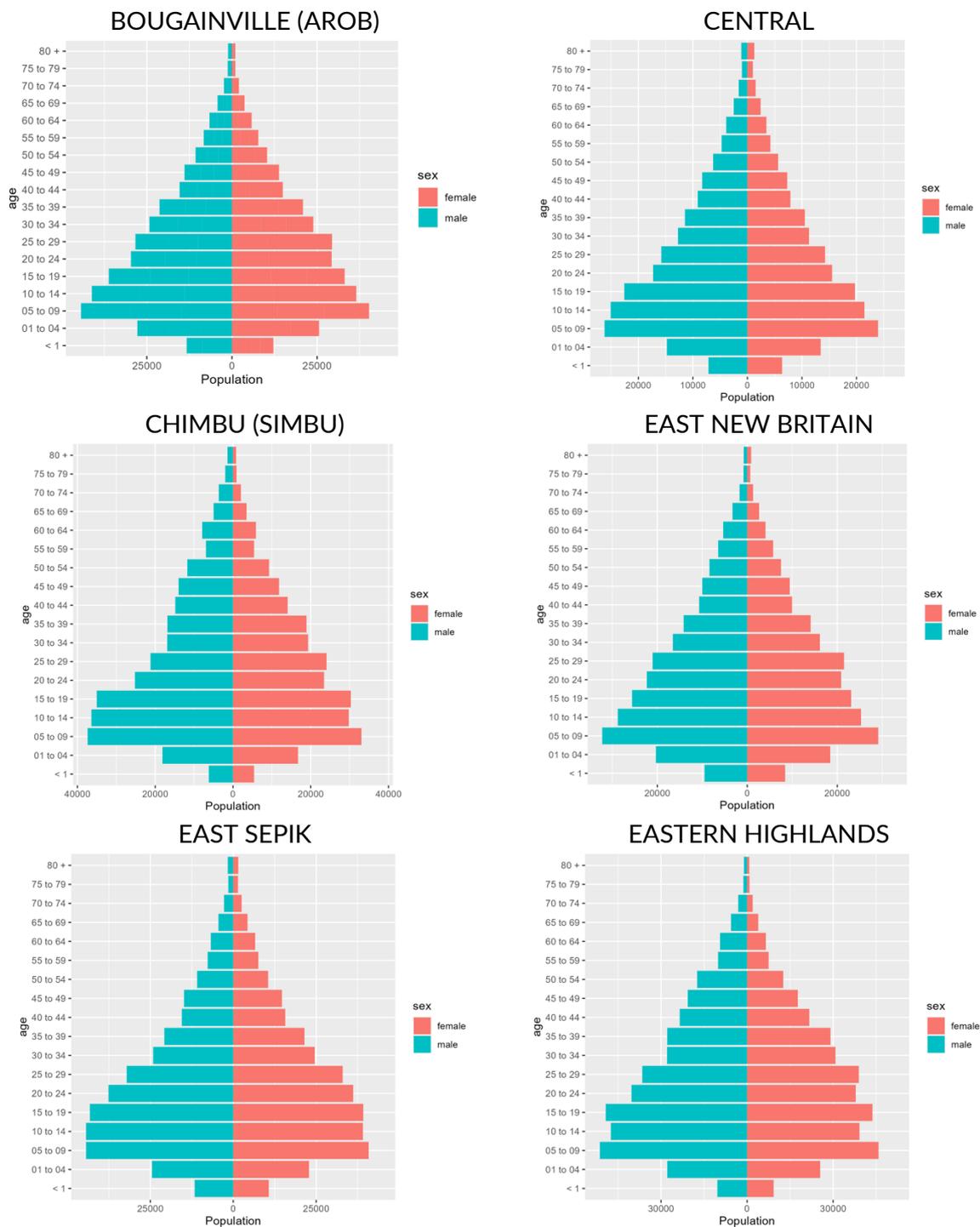
Covariate	Date	Unit	Source	Link
Slope	2000	Degrees	WorldPop	https://www.worldpop.org/geodata/summary?id=23186
Elevation	2000	Metres	WorldPop	https://www.worldpop.org/geodata/summary?id=23435
Resampled VIIRS night-time lights	2016	nanoWatts/cm ² /sr	WorldPop	https://www.worldpop.org/geodata/summary?id=18704
Distance to IUCN strict nature reserve and wilderness area edges	2017	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=18215
Resampled DMSP-OLS night-time lights	2011	Unit of radiance ranging from 0-6300	WorldPop	https://www.worldpop.org/geodata/summary?id=18953
Distance to open-water coastline	2020	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=23933
Distance to ESA-CCI-LC inland water	2012	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=24182
Distance to cultivated areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to woody areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to shrub area edges	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to herbaceous areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to sparse vegetation areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to aquatic vegetation areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to artificial surface edges	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to bare areas	2015	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=22937
Distance to OSM major road intersections	2016	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=17717
Distance to OSM major waterways	2016	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=17966

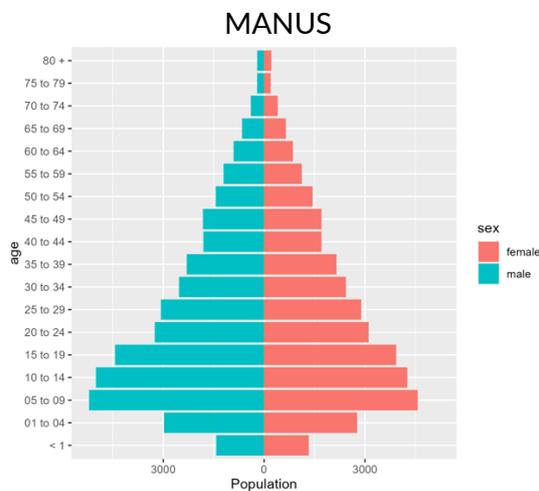
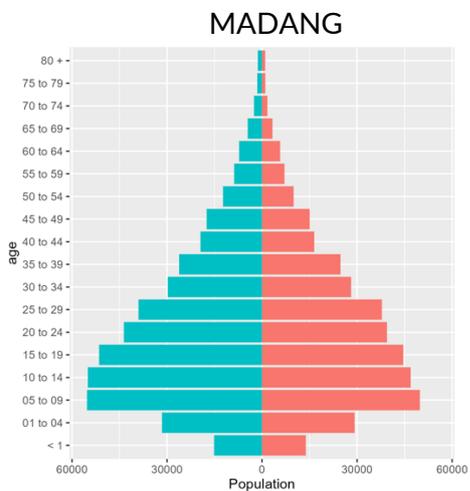
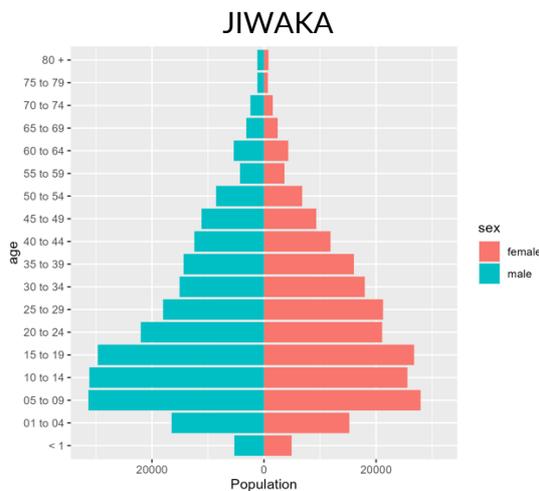
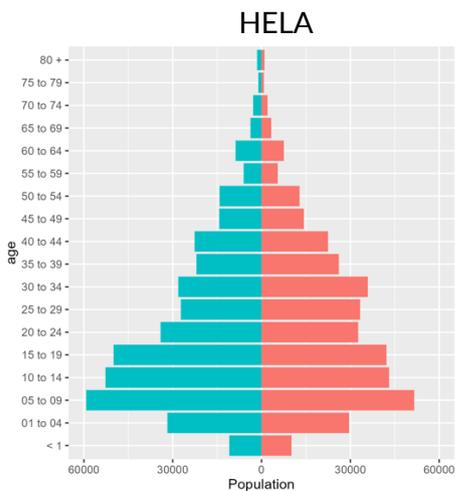
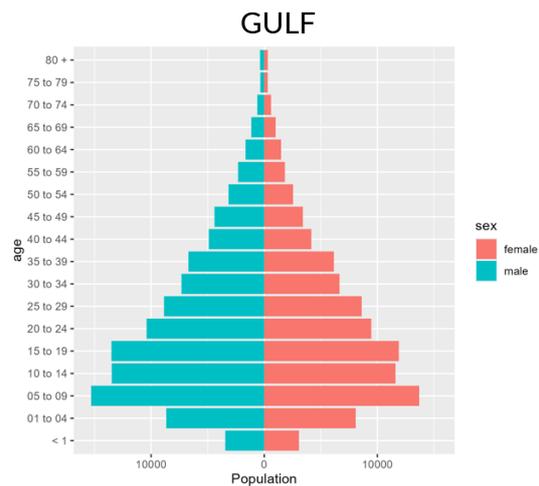
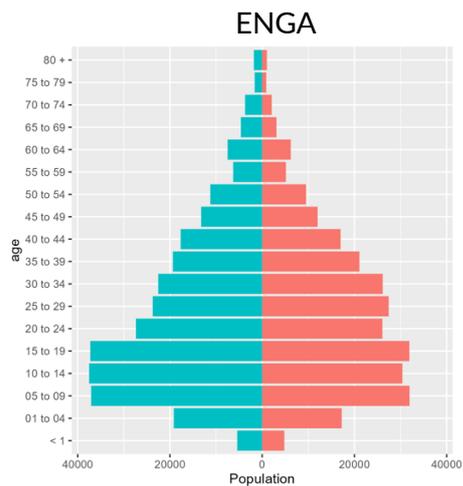
Distance to OSM major roads	2016	Kilometres	WorldPop	https://www.worldpop.org/geodata/summary?id=17468
Distance to main roads	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Distance to local roads	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Distance to places of worship	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Distance to places of education	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Distance to health providers	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Distance to marketplace	2016-2021	Decimal degrees	OSM	https://download.geofabrik.de/australia-oceania/papua-new-guinea.html
Motorized friction surface	2019	Minutes required to travel 1 metre	MAP	https://malariaatlas.org/explorer/#/
Walking friction surface	2019	Minutes required to travel 1 metre	MAP	https://malariaatlas.org/explorer/#/
Mean 2m dewpoint temperature	2011-2021	Celsius	Copernicus	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=form
Mean 2m temperature	2011-2021	Celsius	Copernicus	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=form
Mean total daily precipitation	2011-2021	Metres	Copernicus	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=form
Baseflow index (1), defined as the ratio of long-term base flow to total Q (Smakhtin 2001), computed from daily Q data using the recursive digital filter of Van Dijk (2010) with the “window size” set to 5 days		-	GloH20	http://www.gloh2o.org/gscd/

Baseflow index (2) computed from daily Q data following the local-min method described in Pettyjohn and Henning (1979) and Sloto and Crouse (1996) with the “duration of surface runoff” N set to 5 days.		-	GloH20	http://www.gloh2o.org/gscd/
Baseflow index (3) computed from daily Q data using a 7-day moving min to derive base flow.		-	GloH20	http://www.gloh2o.org/gscd/
Baseflow index (4) computed from daily Q data following the procedure described in Gustard et al. (1992), which takes the minima at 5-day nonoverlapping intervals and subsequently connects the valleys in this series of minima to generate base flow.		-	GloH20	http://www.gloh2o.org/gscd/
Baseflow recession constant, defined as the rate of baseflow decay (Vogel and Kroll 1996), computed from daily Q data as described in Beck et al. (2013b), with the “window size” set to 5 days and days with zero flow ignored.		day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
Daily flow percentiles (exceedance probability) computed from daily Q data. The number refers to the percentage of time that the flow is exceeded.	1	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	5	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	10	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	20	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	50	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	80	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/
	90	mm day ⁻¹	GloH20	http://www.gloh2o.org/gscd/

APPENDIX 2: Age-Sex population pyramids of provinces

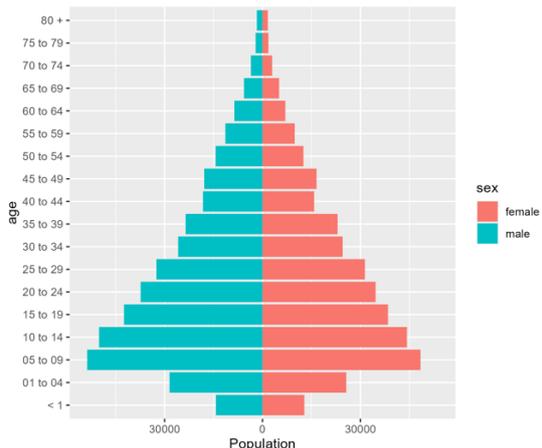
Figure A4.1: Age and sex estimated population pyramids by provinces



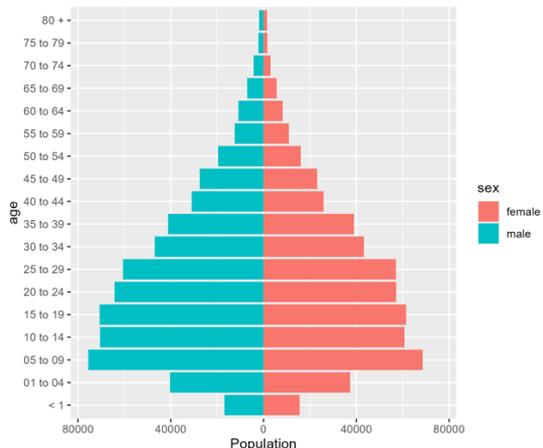


MILNE BAY

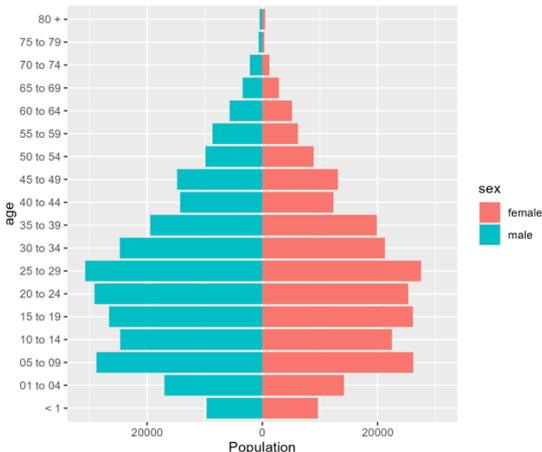
MOROBE



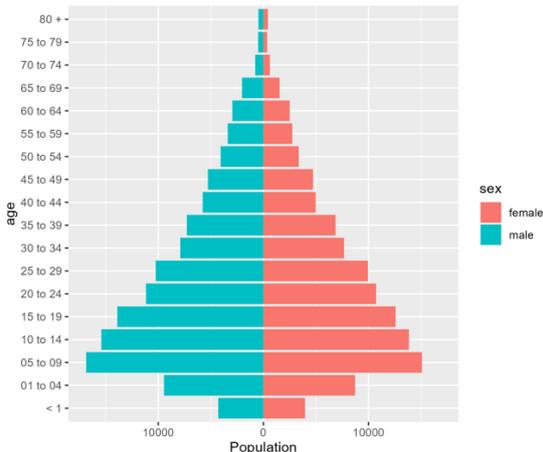
NATIONAL CAPITAL DISTRICT



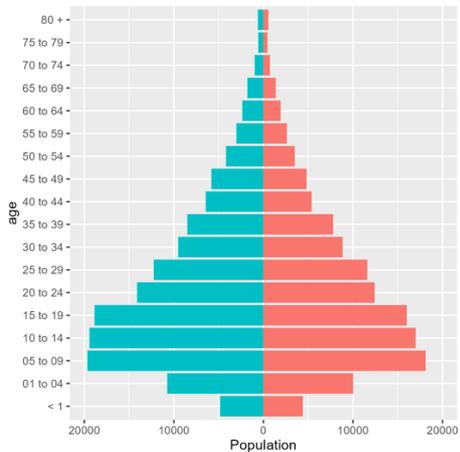
NEW IRELAND



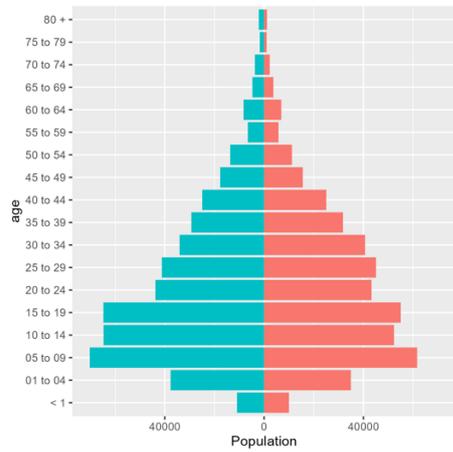
NORTHERN (ORO)



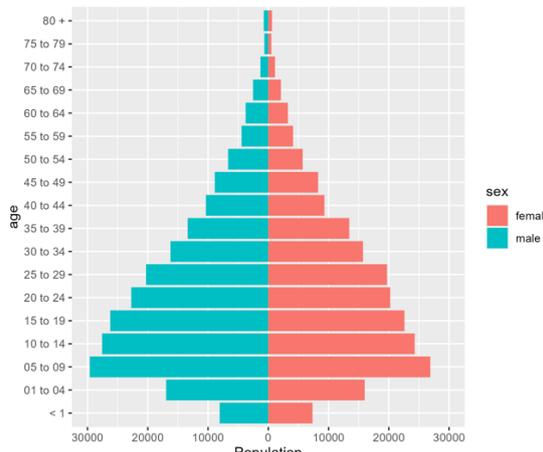
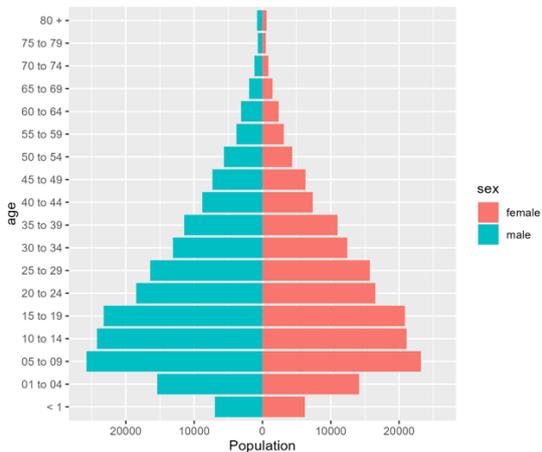
SOUTHERN HIGHLANDS



WEST NEW BRITAIN

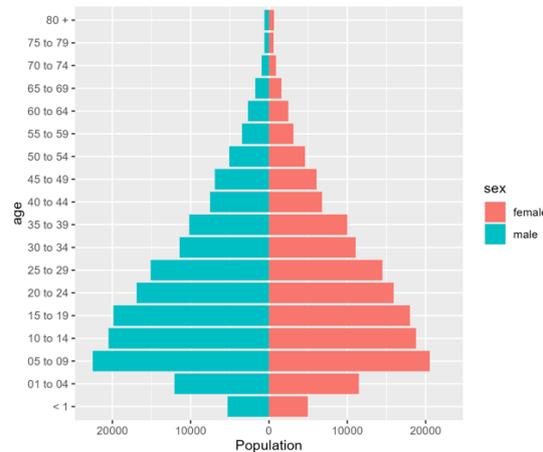
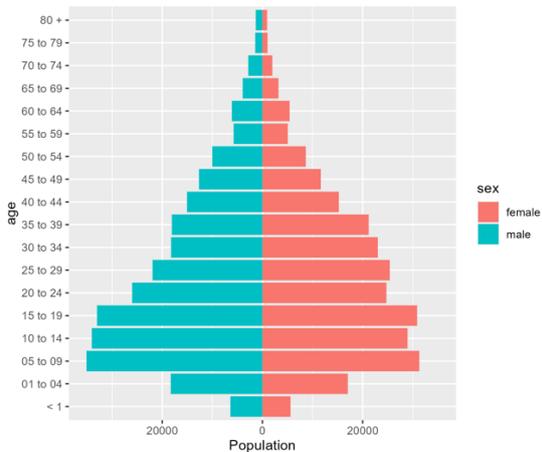


WEST SEPIK



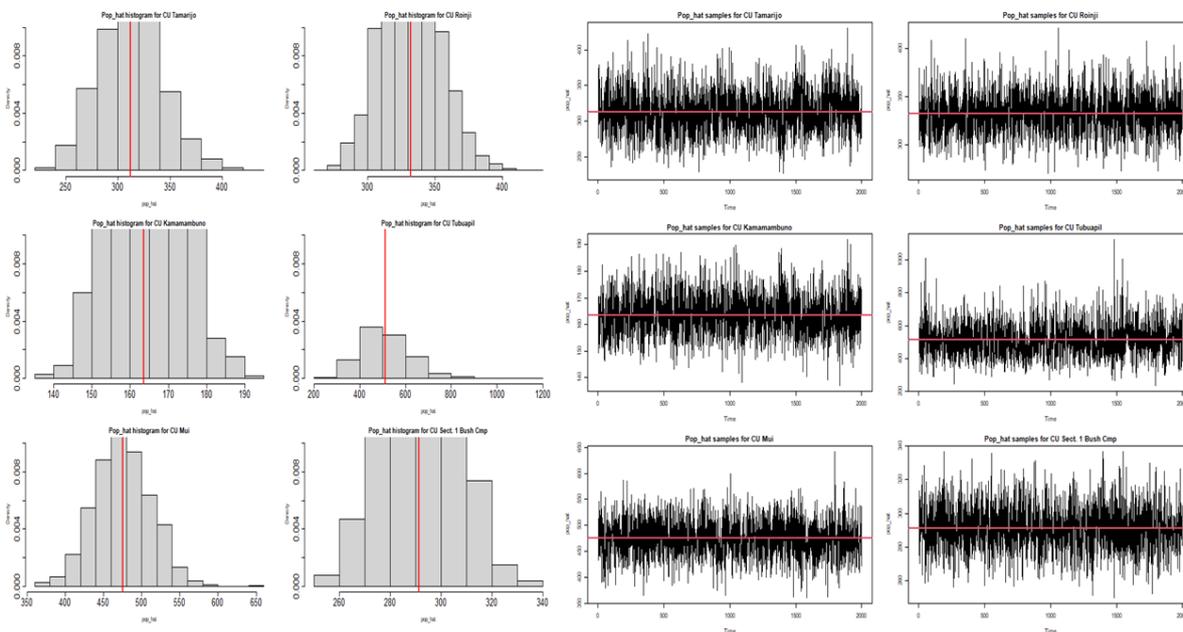
WESTERN HIGHLANDS

WESTERN

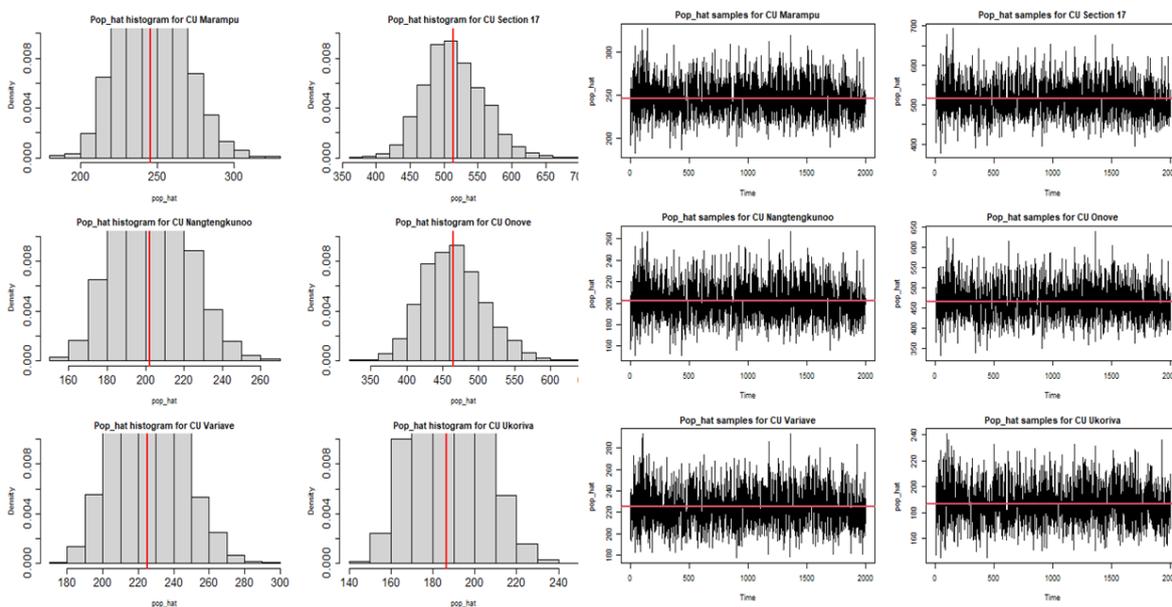


APPENDIX 3: Sampling path plots and the associated histograms

Gamma-Gaussian



Quasi-Poisson



Negative Binomial

