Article

Not peer-reviewed version

Efficient Bayesian Hierarchical Small Area Population Estimation Using INLA-SPDE: Integrating Multiple Data Sources and Spatial-Autocorrelation

Chibuzor Christopher Nnanatu *, Ortis Yankey, Anaclet D. Dzossa, Thomas Abbott, Assane Gadiaga, Attila Lazar, Andrew Tatem

Posted Date: 8 January 2025

doi: 10.20944/preprints202501.0588.v1

Keywords: Population Model; Bayesian Inference; Satellite Imagery; Geospatial Covariates; Census-Independent data; Multiple data sources

Article

Efficient Bayesian Hierarchical Small Area Population Estimation Using INLA-SPDE: Integrating Multiple Data Sources and Spatial-Autocorrelation

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

[1] WorldPop, School of Geography and Environmental Science, University of Southampton, SO17 1BJ, UK
[2] Nnamdi Azikiwe University, Awka-Nigeria
[3] National Institute of Statistics (NIS)- Cameroon
* Correspondence: cc.nnanatu@soton.ac.uk; @ChibuzorNnanatu

Abstract: Statistical modelling approaches which produce fine spatial resolution population estimates have been developed to fill data gaps in resource-poor countries where census data are either outdated or incomplete. These population modelling methods often draw upon recent georeferenced sample population enumeration datasets to predict population density and distribution at both sampled and non-sampled locations, based on their correlation with a set of carefully selected geospatial covariates. These modelled population estimates are increasingly used to support governance, health surveillance, equitable resource allocation, and humanitarian response. However, methodological challenges remain. For example, the georeferenced sample enumeration data are usually disparate and patchy in their distributions, with a high proportion of non-sampled locations that result in highly uncertain estimates. Here, we present a model-based Bayesian geostatistical small area population estimation approach which simultaneously: • Combines multiple sample population enumeration datasets and • Explicitly integrates spatial autocorrelation within a single modelling framework. Findings from a simulation study show varying levels of accuracy in the posterior parameter estimates over different levels of spatial variance and data missingness. The methodology, which was further validated using five nationally representative household listing datasets in Cameroon, provides a valuable methodological development in small area population estimation modelling from sparsely distributed sample enumeration data.

Keywords: population model; bayesian inference; satellite imagery; geospatial covariates; census-independent data; multiple data sources

Specifications Table

Subject area	Environmental Science
More specific subject area	Population density and distribution modelling/estimation
Name of your method	Bayesian Hierarchical Small Area Population modelling, which integrated multiple data sources and spatial autocorrelation within the Integrated Nested Laplace Approximations and Stochastic Partial Differential Equations (INLA-SPDE).

 doi:10.20944/preprints202501.0588.v1

Name and reference of original method	1) Bottom-up Population modelling: Wardrop N.A., Jochem W.C., Bird T.J., Chamberlain H.R., Clarke D., Kerr D., Bengtsson L., Juran S., Seaman V., Tatem A.J. (2018). "Spatially disaggregated population estimates in the absence of national population and housing census data." Proceedings of the National Academy of Sciences 115, 3529–3537. https://www.pnas.org/doi/10.1073/pnas.1715305115 2) INLA: Rue, Havard, Sara Martino, and Nicolas Chopin. (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." Journal of the Royal Statistical Society, Series B 71 (2): 319–92 3) SPDE: Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423–498
Resource availability	All the R codes and datasets used in this study including the simulation study and methods validation/application data are found in this GitHub repository.

Background

Small area population count data support decision-making across all areas of governance. Estimating population numbers affected by disasters, delivering health interventions, planning for elections and allocating resources equitably all require reliable estimates of population distributions at small area scales (UNFPA 2020). Such data are typically collected through a national population and housing census, but these can become quickly outdated in settings with substantial population movements and spatially heterogeneous patterns of fertility and mortality that are hard to predict (Tatem 2022). In addition, in some areas of certain countries, it is sometimes not possible to directly collect such population data due to poor access, conflicts or other security challenges. To fill these data gaps, geospatial methods have recently been developed that leverage advances in satellite imagery, computer vision, geospatial computation and spatial statistics to produce small area population estimates across national extents (e.g., Leasure et al., 2020; Boo et al., 2022; Darin et al, 2022; Nnanatu et al., 2024).

'Bottom-up' population models leverage the statistical relationships between population density measures in incomplete enumerations of an area of interest and a set of geospatial datasets capturing features known to correlate with how humans distribute themselves on the landscape. Predictions of numbers of residents for 100 by 100m grid cells are then typically made, and the use of Bayesian statistical inference methods for the estimation of the population model parameters means that estimates of uncertainties can be provided (Wardrop et al., 2018). However, the input enumeration data which can come from purposely designed 'microcensus' surveys (e.g. Leasure et al, 2020; Boo et al., 2022), incomplete census enumeration (e.g. Darin et al, 2022), or listings from household surveys (e.g. Dooley et al, 2021), are typically sparsely distributed and often exhibit spatial autocorrelation (Chan-Golston et al., 2022). In such situations, the integration of spatial autocorrelation within the analytical framework is highly recommended (Anselin, 1990; Chi & Voss, 2011; Chan-Golston et al., 2022).

Existing bottom-up population models (e.g., Leasure et al., 2020; Boo et al., 2022; Darin et al, 2022), use Bayesian hierarchical regression models to more accurately represent levels of variabilities within a single source of observed enumeration data as random effects, and quantify uncertainties in the parameter estimates in a more straightforward manner. Here, with an aim of improved accuracy

 doi:10.20944/preprints202501.0588.v1

in small area population predictions, we extend the existing approach to allow for the integration of multiple disparate enumeration data sources to increase sample size and obtain larger statistical power. We do this while simultaneously accounting for spatial autocorrelation within the observations to borrow strength (e.g., Chi & Voss, 2011) from nearby locations and more accurately predict population counts in non-sampled locations.

Motivated by the need to rapidly produce small area population data for Cameroon using multiple household listing datasets, we used geostatistical modelling frameworks (Cressie, 1993; Wakefield, 2007; Diggle & Giorgi, 2016; Giorgi et al., 2018), to imply spatial autocorrelation as a distance dependent covariance matrix, such that population distribution between nearby locations is more similar than those further apart (Tobler, 1970). To increase computational efficiency, the integrated nested Laplace approximation (INLA; Rue & Held, 2005; Rue et al., 2009) was used in conjunction with the stochastic partial differential equation (SPDE; Lindgren et al., 2011).

Method Details

Method

Within the context of the bottom-up population modelling (Wardrop et al., 2018), we are often faced with the problem of population prediction at high resolution regular grid cells (pixels) in order to build a set of estimates that can be flexibly summarised and aggregated to other decision making using, for example, administrative units, health zones, wards, or facility catchment areas, including areas where little or no data are observed. In most cases, population enumeration data are only available at some locations, for example, census units (CUs), primary sampling units (PSUs) or enumeration areas (EAs), across a given geographical domain of interest.

Figure 1 shows the schematic representation of the entire population modelling process developed here to address this problem. Specifically, in step 1, the input datasets were first assembled from the disparate sources. These datasets include the enumeration data (containing population counts of people within geographically defined small areas), the gridded geospatial covariates, (e.g., night-time lights intensity, road density, topography, land cover, distance to markets (Nieves et al., 2017)), and the settlement data (e.g. gridded data summarising buildings mapped from satellite imagery (Chamberlain et al., 2024), containing counts of building, building height estimates and other derived metrics). In step 2, these datasets were explored, cleaned, and prepared for the next steps. Part of the exploratory data analysis was testing for the presence of spatial autocorrelation in the observed data using Moran's I statistics (Moran 1950) under the null hypothesis of no spatial clustering. Then, a statistically significant test indicates the presence of spatial autocorrelation.

 doi:10.20944/preprints202501.0588.v1

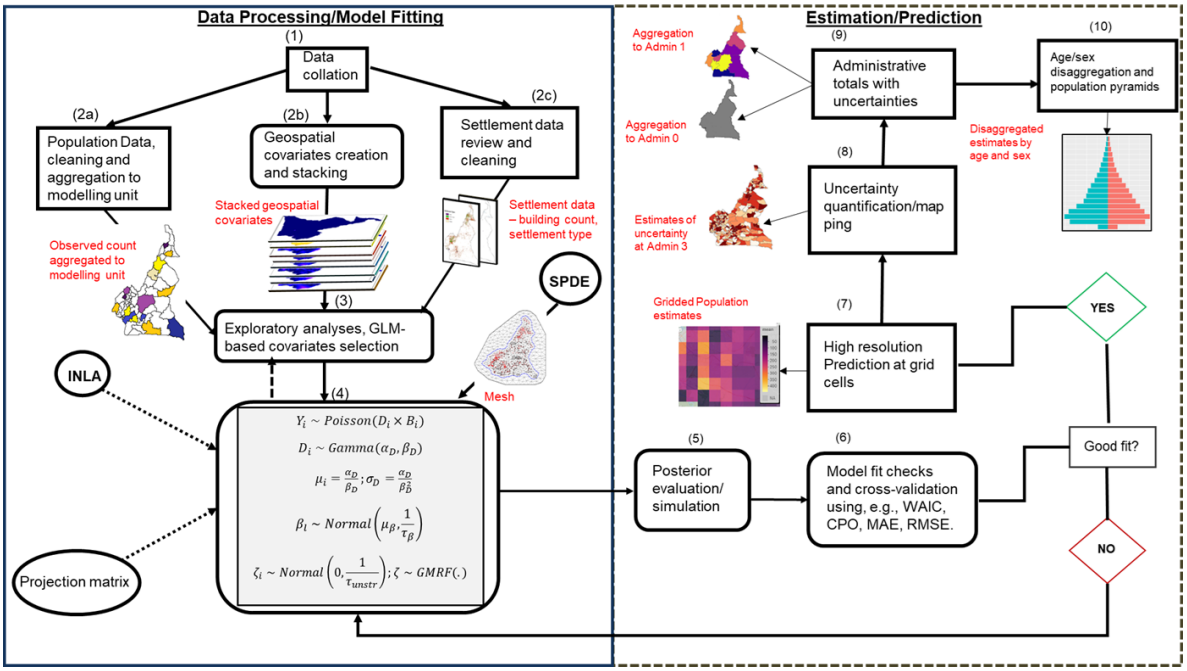

Figure 1. Schematic representation of the Bayesian hierarchical geostatistical bottom-up population modelling steps. INLA – Integrated Nested Laplace Approximation; SPDE – Stochastic Partial Differential Equation; WAIC – Widely Acceptable Information Criterion; CPO – Conditional Predictive Ordinate; MAE – Mean Absolute Error; RMSE – Root Mean Square Error.

To ensure that spurious effects of redundant geospatial covariates are eliminated in the model parameter estimates, a rigorous covariate selection process is carried out in step 3, where only the covariates that significantly predicted population density are retained for the final analysis. Fior a given location i, the population density variable D_i was obtained as the number of people (N_i) per building (B_i), that is, $D_i = N_i/B_i$. The continuous geospatial covariates are scaled using z-score so that the parameter estimates based on the datasets emanating from disparate measurement scales can be compared and interpreted in terms of standard deviation. The covariates selection is done using a robust stepwise regression scheme implemented within the Generalized Linear Model (GLM) framework (McCullagh & Nelder, 1989) with the stepAIC function of the 'MASS' package in R. Then the selected covariates were further tested to ensure that the potential effects of multicollinearity are drastically reduced. To do this, we used the 'vif' function of the 'car' package in R to calculate the variance inflation factor (vif) values of each covariate and those with vif < 5 are retained (e.g., James et al.,2013). Finally, the GLM model was refitted and only the statistically significant covariates were retained for the next steps.

In step 4, the geospatial covariates selected in step 3 were used to train Bayesian hierarchical population models using the INLA-SPDE approach. The INLA-SPDE approach provides computational efficiency by using a mesh which is a triangulation of the entire spatial domain of interest allowing the use of sparse covariance matrix on a discrete space instead of a dense covariance on a continuous space (Lindgren et al., 2011).

Steps 5 to 10 follow immediately after model fitting and involved the collation and testing of the model results, posterior predictions at high resolution (approximately 100m by 100m) grid cells, aggregation to various administrative units of interest, and disaggregation of the population totals by age/sex classes.

The model fit assessments and cross-validation of the statistical models were performed by comparing a constellation of model fit metrics. Specifically, for model selection, we relied on the Deviance Information Criterion (DIC), the Widely Applicable Information Criterion (WAIC; Watanabe, 2013) and Conditional Predictive Ordinate (CPO; Pettit 1990). The predictive ability of the

 doi:10.20944/preprints202501.0588.v1

selected models were further evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute Bias (BIAS), and the Pearson correlation coefficients (CORR) of the observed versus predicted population counts. Smaller values of the DIC, WAIC, and CPO indicate a better fit model. Also, smaller values of MAE, RMSE, BIAS, and larger values of CORR indicate model with better predictive ability. Posterior simulations and grid cell predictions were based on the best-fit model. Finally, by dividing the observed data into train (80%) and test (20%) sets, k-fold cross-validation was carried out.

Statistical Modelling

Let Y_i denote the response variable, the count (population) of people in each small area i $(i = 1, 2, \dots, N)$, such that

$$Y_i \sim Poisson(\lambda_i) \quad (1)$$

with equal mean and variance equal to $\lambda_i > 0$ (McCullagh and Nelder 1989). However, it is well known that within the context of population modelling, the data are almost always over-dispersed in that the variance of the response is often larger than the mean and the assumption of equal mean and variance is rarely met (Leasure et al., 2020; Boo et al., 2022).

To circumvent this analytical challenge and improve estimates of population whilst accounting for potential sources of variability, the response variables is redefined in terms of population density (e.g., Leasure et al. 2020) so that

$$population = \frac{people}{settlement} \times settlement \quad (2)$$

where the term $settlement$ is generic and represents any variable that provides an indication of human settlement intensity within a given area, such as, the total built-up area, number of buildings, number of households, and building intensity, all typically obtained from satellite imagery feature extraction. However, the values of any such settlement variable must be available throughout the country for country-wide model prediction purposes. Thus, the term $people/settlement$ represents the population density, D. For ease of exposition, from now on, we will use building count (number of buildings in each area of interest) B as the $settlement$ variable and using a Poisson-Gamma two-stage model (e.g., Wakefield, 2007). Equation (1) becomes

$$Y_i \sim Poisson(\mu_i B_i) \quad (3)$$

where the mean and variance parameter λ_i of Equation (1) is now respecified in terms of the expected density μ_i and building counts B_i of area i, that is, $\lambda_i = \mu_i B_i$, and D_i is the population density which gives the Maximum Likelihood Estimator (MLE) for μ_i. Thus, the model specification in equation (3) allows us to model explicitly the potential overdispersion within the data via the mean density parameter by assuming a Gamma distribution with shape and rate parameters given by μ_i^2/ϕ and μ_i/ϕ, respectively. That is,

$$D_i \sim Gamma(\mu_i^2/\phi\ , \mu_i/\phi) \quad (4)$$

where $E[D_i] = \mu_i$ and $var(D_i) = \phi$. Note that the choice of the Poisson-Gamma two-stage model is because it allows flexibility to explicitly model the inherent overdispersion via the parameter ϕ. Other positively skewed long tail distribution such as the LogNormal distribution (e.g., Leasure et al, 2020) could also be used, so that, $D_i \sim LogNormal(\mu_i, \sigma_D^2)$, where μ_i and σ_D^2 are the log of the expected population density and the random variations in the population density due to overdispersion, respectively.

Despite the simplicity of the specification implied by equation (3), it is important to note that the variance of D_i increases for very small values of B_i which could arise from sparse observations (e.g., Wakefield, 2007). Thus, to avoid inflated population estimates, care must be taken while using this model specification to account for either overdispersion or as aggregation weights to account for potential aggregation error (e.g., Paige et al. 2022). In any case, the expected population density μ_i is

 doi:10.20944/preprints202501.0588.v1

linked to the geospatial covariates (e.g., nighttime lights, distance to healthcare facilities) through the linear predictor η_i given by

$$h(\mu_i) = \eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i \quad (5)$$

where $h(.)$ is an appropriate link function (e.g., log-link), β_0 is the intercept parameter representing the average population density when there is zero effect of the other covariates; $(\beta_1, \dots, \beta_K)$ are the unknown fixed effect coefficients of the K geospatial covariates $(x_1, x_2, \dots, x_K)$ found to significantly predict the population density; ε_i is a Gaussian noise or nugget effect (Cressie, 1993), which accounts for the observation level variability (also known as the fine scale variability, e.g., Paige et al., 2022) not captured by the geospatial covariates, that is, $\varepsilon_i \sim Normal(0, \sigma_\varepsilon^2)$. To ensure that the estimates of the fixed effects parameters $(\beta_1, \dots, \beta_K)$ are interpretable and comparable, it is recommended that the corresponding continuous geospatial covariates which are potentially on different measurement scales be rescaled using for example the z-score such that

$$Z_i = \frac{x_i - \bar{x}_i}{\sigma_i} \quad (6)$$

where Z_i is the scaled version of the geospatial covariate x_i.

Specifically, we extended equation (5) to include the spatial autocorrelation term $\xi(s_i)$ such that the geographical units that share common boundaries are more like each other in terms of population distribution than those further apart (Tobler, 1970). Thus,

$$\eta(s_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \xi(s_i) + \varepsilon_i \quad (7)$$

where $\eta(s_i)$ is the linear predictor, $s_i \in \{s_1, s_2, \dots, s_N\}$ is the i-th spatial unit (e.g., enumeration areas) of the N geolocated spatial units within the study domain. The term $\xi(s_i)$ is the i-th realisation of the Gaussian Random Field (GRF), that is, $\xi(\boldsymbol{s}) \sim GRF(0, \Sigma)$, with the distance dependent Matérn covariance function

$$C(s_i, s_j) = \frac{\sigma_\xi^2}{\Gamma(\nu)2^{\nu-1}} (\kappa d_{ij})^\nu K_\nu (\kappa d_{ij}) \quad (8)$$

where Γ is a gamma function; K_ν is the modified Bessel function of the second kind, order ν; $d_{ij} = ||s_i - s_j||$ is the Euclidean distance between spatial locations s_i and s_j; ν is the smoothness parameter; $\kappa = \frac{\sqrt{8\nu}}{\rho}$ is the scale parameter where ρ is the spatial distance at which the correlation is approximately 0.13; σ is the marginal variance. One computational challenge of geostatistical models of this form especially those implemented via the Markov chain Monte Carlo (MCMC) methods is that the computation of the dense covariance matrix Σ becomes very expensive as the sample size increases (e.g., Bakka et al., 2018). However, the use of the integrated nested Laplace approximation in conjunction with the stochastic partial differential equation (INLA-SPDE; Rue and Held, 2005; Rue et al., 2009; Lindgren et al., 2011) approach provide significant computational advantage. With the INLA-SPDE approach we only need to compute the sparse precision matrix $\boldsymbol{Q} = \Sigma^{-1}$, and the continuously indexed GRF, $\xi(\boldsymbol{s})$ is approximated by a discretely indexed Gaussian Markov Random Field (GMRF) using a piecewise linear basis function representation on a triangulation of the entire study domain also known as 'mesh'. Thus,

$$\xi(\boldsymbol{s}) = \sum_{h=1}^{H} \varphi_h(s)\, \tilde{\zeta}_h \quad (9)$$

 doi:10.20944/preprints202501.0588.v1

where $\varphi_h \in \{0,1\}$ is the value of the piecewise linear function which takes the value of 1 at the d-th node of the mesh and 0 elsewhere for a mesh with a total of H nodes; $\zeta = (\tilde{\zeta}_1, \tilde{\zeta}_2, \ldots, \tilde{\zeta}_H)$ is a GMRF with sparse correlation matrix parameters κ and σ_ζ^2 (see, Lindgren et al., 2011; Blangiardo et al., 2013). Thus, under the INLA-SPDE approach, equation (7) is respecified as:

$$\eta(s_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{h=1}^{H} \widetilde{\boldsymbol{A}}_{ih}\zeta + \varepsilon_i \tag{10}$$

where $\widetilde{\boldsymbol{A}}_{ih}$ is the i,d-th element of the $N \times H$ sparse projection matrix $\widetilde{\boldsymbol{A}}$ which maps the N observations to the H nodes of the mesh. Note that it is straightforward to extend equation (10) to include other random effects terms to capture other unobserved sources of variability like those due to settlement type (rural-urban), regions, data source, interacting random effect terms, etc. Thus, the hierarchical regression-modelling framework is specified below:

$$Y_i \sim Poisson(\mu_i B_i)$$

$$D_i \sim Gamma\left(\frac{\mu_i^2}{\phi}, \frac{\mu_i}{\phi}\right)$$

$$h\big(\mu(s_i)\big) = \eta(s_i)$$

$$\eta(s_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{i,k} + \sum_{m=1}^{M} f_m(z_{m,i}) + \sum_{h=1}^{H} \widetilde{\boldsymbol{A}}_{ih}\zeta + f_p(type) + f_r(reg) + f_{r,p}(reg \times type) + \varepsilon_i \tag{11}$$

where $\{f_m\}_{m=1}^{M}$ are the random effect functions of the M different data sources; while $f_p(type)$, $f_r(reg)$, and $f_{r,p}(reg \times type)$ capture the variabilities due to differences in population distributions across different settlement types, regions, and their interactions, respectively. As stated above, the framework allows the incorporation of as many random effects as possible, however, care must be taken to avoid overfitting the data.

Bayesian Inference for Hierarchical Population Models

In Bayesian inference context, interest is on the joint posterior distribution of the latent field $\boldsymbol{w} = (\eta, \beta_0, \boldsymbol{\beta}, f_m, f_p, f_{rp}, \zeta, \varepsilon)$ and the hyperparameters $\boldsymbol{\theta} = (\tau_\beta, \tau_m, \tau_p, \tau_{rp}, \tau_\beta, \tau_\varepsilon)$ given by

$$\pi(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{i \in I} \pi(y_i|w_i\boldsymbol{\theta}) \tag{12}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution, $\pi(\boldsymbol{w}|\boldsymbol{\theta})$ is a latent Gaussian model (LGM), and $\pi(\boldsymbol{y}|w, \theta)$ is the likelihood function of observing the data given the latent field and the hyperparameters which are assumed to be conditionally independent. The posterior distribution is then approximated and evaluated using INLA-SPDE as already stated above with prior distributions given by:

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

$$\beta_k \sim Normal\left(\mu_\beta, \frac{1}{\tau_\beta}\right)$$

$$\zeta \sim GMRF\left(0, \psi(\kappa, \sigma_\zeta^2)\right)$$

$$f_k \sim Normal\left(0, \frac{1}{\tau_k}\right)$$

 doi:10.20944/preprints202501.0588.v1

$$\varepsilon_i \sim Normal(0, 1/\tau_\varepsilon)$$

$$\tau_w \sim Gamma(\alpha_w, \beta_w) \quad (13)$$

where $\alpha_w > 0$ and $\beta_w > 0$ are hyperparameters and $k, w \in \{\beta, m, p, r, rp, \varepsilon\}$. Then the predicted density $\widehat{D}_{(s_i)}$ is obtained as the back transformed values of the predicted linear predictor $\hat{\eta}_i$, that is, $\widehat{D}_i = \exp(\hat{\mu}_i)$.

Finally, the predicted population count is given as a weighted product of the population density and the building count, that is, $\hat{y}_i = \widehat{D}_i \times B_i$.

Model Fit Checks and Cross-Validation Metrics

Conditional Predictive Ordinates (CPO)

The CPO is a cross-validatory criterion which calculates the probability of observing a held-out observation not used in the model training set such that given the i-th observation y_i. Thus, the CPO is the posterior probability of observing y_i when the model is fit using all data but y_i, that is,

$$CPO_i = \pi(y_i | y_{-i}) \quad (14)$$

Large values of CPO indicate a better fit of the model to the observation, while small values indicate a bad fitting of the model to that observation, which may be an outlier.

Then, the negative sum of the log of the CPO given in equation (14) provides a measure of predictive ability of the model with the smaller the better, that is,

$$-\sum_{i=1}^{n} \log(CPO_i) \quad (15)$$

Mean Absolute Error (MAE)

The mean absolute error (MAE) provides a measure of the average magnitude of errors within a set of predictions irrespective of the direction. It is calculated using

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (16)$$

where, y_i and $\hat{y}_i$ are the observed and predicted values, respectively. The model with the smaller MAE value provides a better fit.

Root Mean Square Error (RMSE)

The root mean square error (RMSE) is similar to the MAE in that they both provide an idea on the average magnitude of prediction error. However, the RMSE is found to be more useful when large errors are not desirable. RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (17)$$

Similar to the MAE, models with lower RMSE values provide better fit.

Pearson Correlation

The Pearson correlation coefficient $r(-1 \le r \le 1)$ is the coefficient of correlation between the observed counts and predicted counts.

$$r = \frac{\sum_i (y_i - \bar{y})\left(y_i^{(pred)} - \bar{y}^{(pred)}\right)}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i \left(y_i^{(pred)} - \bar{y}^{(pred)}\right)^2}} \quad (18)$$

 doi:10.20944/preprints202501.0588.v1

where $y_i, \bar{y}, y_i^{(pred)}$ and $\bar{y}^{(pred)}$ are the observed values, mean of the observed values, the predicted values, and the mean of the predicted values, respectively. Note that equation can be simply written as

$$r = \frac{Cov\left(Y, Y^{(pred)}\right)}{\sigma_Y \sigma_{Y^{(pred)}}} \tag{19}$$

where, $Cov\left(Y, Y^{(pred)}\right)$ is the covariance between the observed y and the predicted value $y^{(pred)}$, and σ_z, $z \in \left\{y, y^{(pred)}\right\}$, are the corresponding standard deviations.

Absolute Bias (BIAS)

This measures the average deviation of the predicted value from the observed value:

$$\text{BIAS} = \left| \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i) \right| \tag{20}$$

Smaller values of BIAS indicate better fit model. The closer the value to zero the better the model.

Coefficient of Variation

For each posterior sample, we computed the coefficient of variation as a measure of uncertainty in the posterior parameter estimation. This was done by dividing the standard deviation with the.mean.

Motivating Dataset

This study was motivated by the lack of a reliable up-to-date small area population data to support healthcare campaigns and other intervention programmes in Cameroon, and to build an alternative sample frame given that the most recent census at the time of writing was conducted in 2005. Completely anonymized versions of seven (7) nationally representative but disparate household listings conducted between 2018 and 2022, were obtained from the Cameroon National Institute of Statistics (NIS), also known in French as *Institut National de la Statistique* (INS, https://ins-cameroun.cm/en/). Following rigorous data cleaning and exploratory activities, five of these datasets conducted between 2021 and 2022 were selected for population modelling (Figure 2). There were 2,587,569 people counted across the 509,628 households with an average of ~5 people per household, across 2,290 Enumeration Areas (EAs). Eventually, the datasets were combined and aggregated up to the EA level which served as the population modelling unit. Further details on how the datasets were explored, cleaned and combined are provided within Section S1 of the Supplementary document.

 doi:10.20944/preprints202501.0588.v1

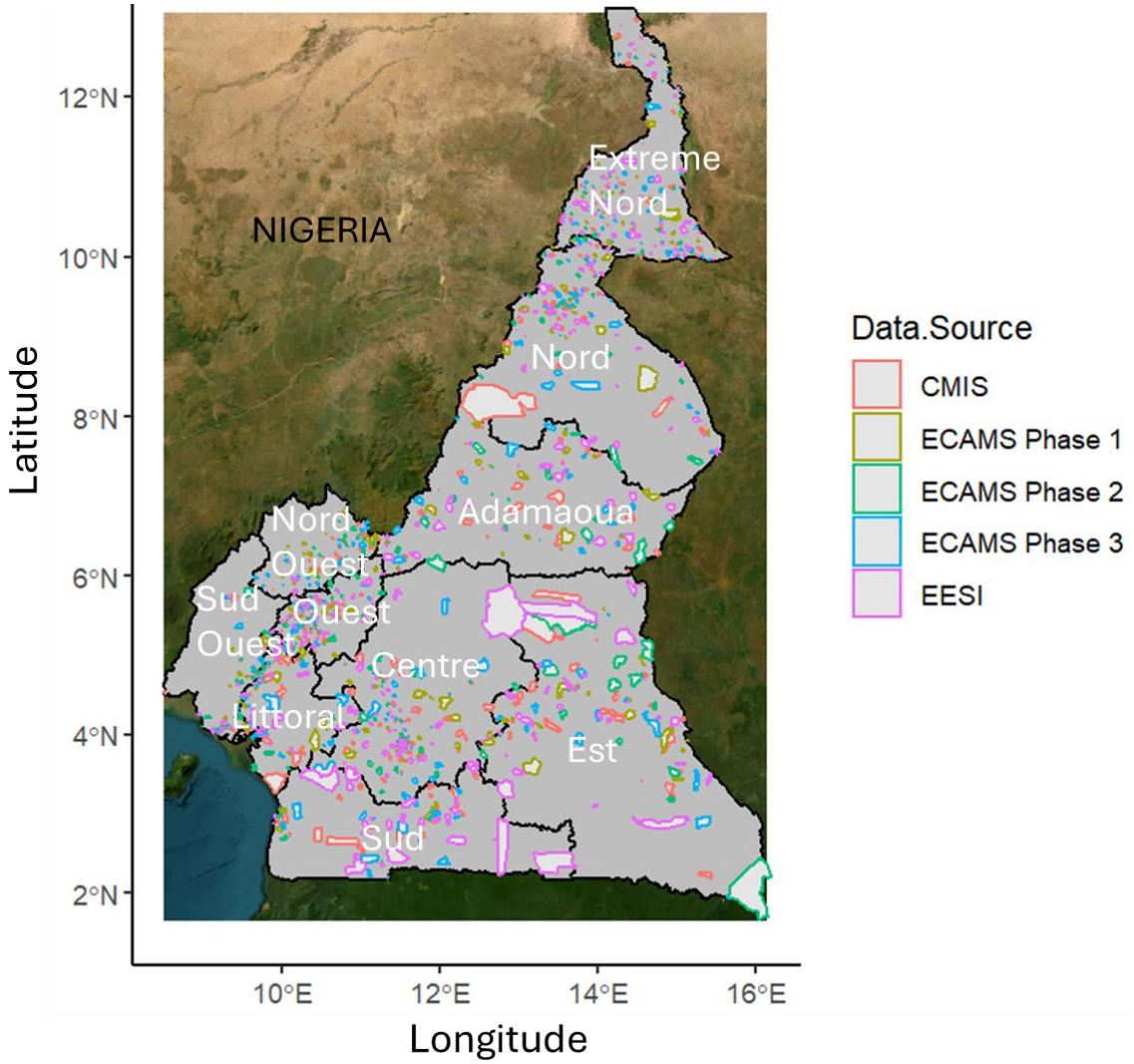

Figure 2. Map of Cameroon showing the distribution of the 2290 Enumeration Areas (EAs) observed across the 10 regions of the country for the 5 household listing datasets. *CMIS - Cameroon Malaria Indicator Survey (CMIS 2022); ECAM5_Phase1-The 2021/2022 fifth Cameroon Households Survey phase 1; ECAM5_Phase2-The 2021/2022 fifth Cameroon Households Survey phase 2; ECAM5_Phase3-The 2021/2022 fifth Cameroon Households Survey phase 3; EESI3 – 2021 third Employment and Informal Sector Survey.*

Testing for Spatial Clustering

Before proceeding to analyse the data, we first carried out Moran's I test for the existence of spatial autocorrelation using the 'moran.test' function of the 'spdep' package in R, after defining the neighbourhood structure using the 'queen' option. However, the Moran's I test statistic returned a statistically significant test with p-value < 0.01 which indicated the presence of spatial autocorrelation within the data.

Statistical Model Implementation

Following from equations (1) – (11) above, the observed data Y_i (the total number of people observed per EA) is Poisson distributed random variable with mean/variance parameters $\lambda_i = \mu_i B_i$, where μ_i is the average population density per EA, and B_i is the total number of buildings (total building counts) per EA. The building counts were obtained from the building footprint layer provided by the Digitize Africa project of Ecopia AI and Maxar Technologies ('year 2', 2020/2021; Ecopia.AI and Maxar Technologies, 2020). The building footprint layer contains polygons

 doi:10.20944/preprints202501.0588.v1

representing visible individual buildings on satellite images. This was rasterised using WorldPop's 3-arc-sercond resolution mastergrid to calculate number of buildings, building area, building perimeter, coefficient of variation, and other related metrics for each grid cell.

Figure S1.1 of the Supplementary document shows the rasterised building footprints and the settlement type classifications of the structures in Cameroon. The settlement type classification was obtained from the Global Human Settlement (GHS) degree of urbanization layer (Schiavina, 2022), which was re-classified into four settlement types namely: cities, small urban, towns and villages.

Model Fitting

Of the 43 geospatial covariates initially identified (Table S1 of the Supplementary document), after covariates selection using the GLM-based stepwise selection methods (McCullagh & Nelder, 1989; James et al., 2013), only 8 were eventually retained as providing the best fit for the population density model (Table 1).

Table 1. List of the last 8 geospatial covariates included in the final model.

Covariate	Description	Source	Year	Original format	Resolution
Cov1	Distance to ACLED conflict data	https://acleddata.com/	2021	Raster	100m
Cov2	Distance to ACLED explosions	https://acleddata.com/	2021	Raster	100m
Cov3	Distance to waterbodies	https://www.geofabrik.de/data/download.html	2022	Raster	100m
Cov4	Distance to herbaceous areas	https://www.worldpop.org/project/categories?id=14	2022	Raster	100m
Cov5	Distance to local roads	https://www.geofabrik.de/data/download.html	2022	Raster	100m
Cov6	Distance to marketplaces	https://www.geofabrik.de/data/download.html	2022	Raster	100m
Cov7	Slope	https://www.worldpop.org/project/categories?id=14	2000	Raster	100m
Cov8	Night-time light brightness	https://www.worldpop.org/project/categories?id=17	2020	Raster	100m

Note. ACLED- Armed Conflict Location & Event Data (www.acleddata.com).

The 8 final geospatial covariates were then used to test various nested Bayesian hierarchical models, and the four top competing nested models are specified below:

 doi:10.20944/preprints202501.0588.v1

$$\textbf{Model1}: \eta(s_i) = \beta_0 + \sum_{k=1}^{8} \beta_k Cov_{i,k} + \sum_{d=1}^{534} \widetilde{\boldsymbol{A}}_{id}\zeta + f_p(type) + \varepsilon_i$$

$$\textbf{Model 2}: \eta(s_i) = \beta_0 + \sum_{k=1}^{8} \beta_k Cov_{i,k} + \sum_{d=1}^{534} \widetilde{\boldsymbol{A}}_{id}\zeta + f_p(type) + f_r(reg) + \varepsilon_i$$

$$\textbf{Model 3}: \eta(s_i) = \beta_0 + \sum_{k=1}^{8} \beta_k Cov_{i,k} + \sum_{d=1}^{534} \widetilde{\boldsymbol{A}}_{id}\zeta + f_{r,p}(reg \times type) + \varepsilon_i$$

$$\textbf{Model 4}: \eta(s_i) = \beta_0 + \sum_{k=1}^{8} \beta_k Cov_{i,k} + \sum_{d=1}^{534} \widetilde{\boldsymbol{A}}_{id}\zeta + f_p(type) + f_{r,p}(reg \times type) + \varepsilon_i \qquad (21)$$

where, the terms β_0, $\{\beta_k\}_{k=1}^{8}$, $\boldsymbol{A}$, ζ, f_p, $f_{r,p}$, and ε are intercept, fixed effect coefficients of the 8 geospatial covariates (Table 1) $Cov_{i,k}\,(i = 1, \dots, 2290; k = 1, \dots, 8)$, 2290×534 projection matrix, spatial autocorrelation term, settlement type (4 classes – cities, small urbans, towns and villages), settlement type – region interaction term, and the zero mean Gaussian nugget effect, respectively. The 10 regions in Cameroon along with the 4 settlement classes constituted a total of $4 \times 10\ (= 40)$ settlement type versus region interaction effects. Figure S3.1C of the supplementary material shows the mesh with 534 vertices which was employed for the model implementation. More details about the design of mesh can be found in Gomez-Rubio (2020) and some of the references therein. The projection matrix $\boldsymbol{A}$ maps the observations unto the mesh nodes to facilitate computational efficiency at the mesh nodes.

Prior Distribution

Initial sensitivity analyses which involved the testing of various priors and hyperprior values indicated that the following INLA default priors and hyperpriors provided no worse fit:

$$\beta_0 \sim \text{Uniform}(0,1)$$
$$\beta_k \sim \text{Normal}(0, 0.01)$$
$$\vartheta \sim \text{Normal}(0, 1000), \text{ where } \vartheta \in \{f_m, f_p, f_{rp}\}$$
$$\tau_\varepsilon \sim \text{Gamma}(0.01, 0.01)$$
$$\tau_w \sim \text{Gamma}(1, 0.00005) \text{ where } w \in \{\beta, m, p, rp\} \qquad (22)$$

However, an alternative prior specification using a joint penalized complexity (PC) prior (Simpson et al. 2017) could still be used.

Finally, the predicted population density $\widehat{D}_i$ is obtained by as the exponent of the linear predictor, that is,

$$\widehat{D}_i = \exp\left(\hat{\beta}_0 + \sum_{k=1}^{8} \hat{\beta}_k x_{i,k} + \sum_{d=1}^{534} \widetilde{A}_{id}\zeta + \hat{f}_p(type) + \hat{f}_{r,p}(reg \times type) + \hat{\varepsilon}_i\right) \qquad (23)$$

So that the predicted population count is obtained deterministically as $\hat{y}_i = \widehat{D}_i \times B_i$.

Model Fit Checks

When implemented to the Cameroon combined household listing data, model 4 (which included spatial autocorrelation, settlement type and settlement type - region interactions random effects) has the lowest DIC and lowest CPO values according to Table 2, thus, provided the best fit for the data. This underscores the importance of jointly accounting for spatial autocorrelation and the random effects of data hierarchy across settlement types within various regions/administrative units within a bottom-up population modelling framework.

 doi:10.20944/preprints202501.0588.v1

Table 2. Model fit indices for the top competing models.

Model	DIC	WAIC	CPO
Model 1	1953.635	1453.671	6143.235
Model 2	1944.475	1486.583	6108.889
Model 3	1922.432	1636.743	6326.046
Model 4	1921.678	1501.331	5990.725

Posterior sampling and GRID Cell Predictions

The INLA-SPDE approach utilized here allowed us to generate posterior marginal distributions of the best fit model (Model 4) which allowed us to draw more samples from the stationary distribution and carry out Bayesian statistical inference. This was implemented by using the 'inla.posterior.sample' function of the 'INLA' package. However, to use the 'inla.posterior.sample' function, it needs to be activated during the model fitting by setting **config = TRUE** within the **control.compute** argument of the inla() function. The posterior samples were then used to predict population densities/numbers at high resolution prediction pixels. Further details of the posterior simulation and grid cell prediction approach are provided in Section S2 of the Supplementary document.

Posterior predictions of the mean population count per 100m square grid cell (or pixel) across the entire spatial domain along with the corresponding inset maps are provided in Figure 3A. The inset maps focused on the two major cities in Cameroon, namely, Yaoundé which is the administrative capital located in the Central region of Cameroon, and Douala which is the commercial capital and located within the Littoral region of Cameroon (see Figure 2). The zoomed in (inset) maps show overall higher population density and distribution per grid cell in Douala than in Yaoundé.

 doi:10.20944/preprints202501.0588.v1

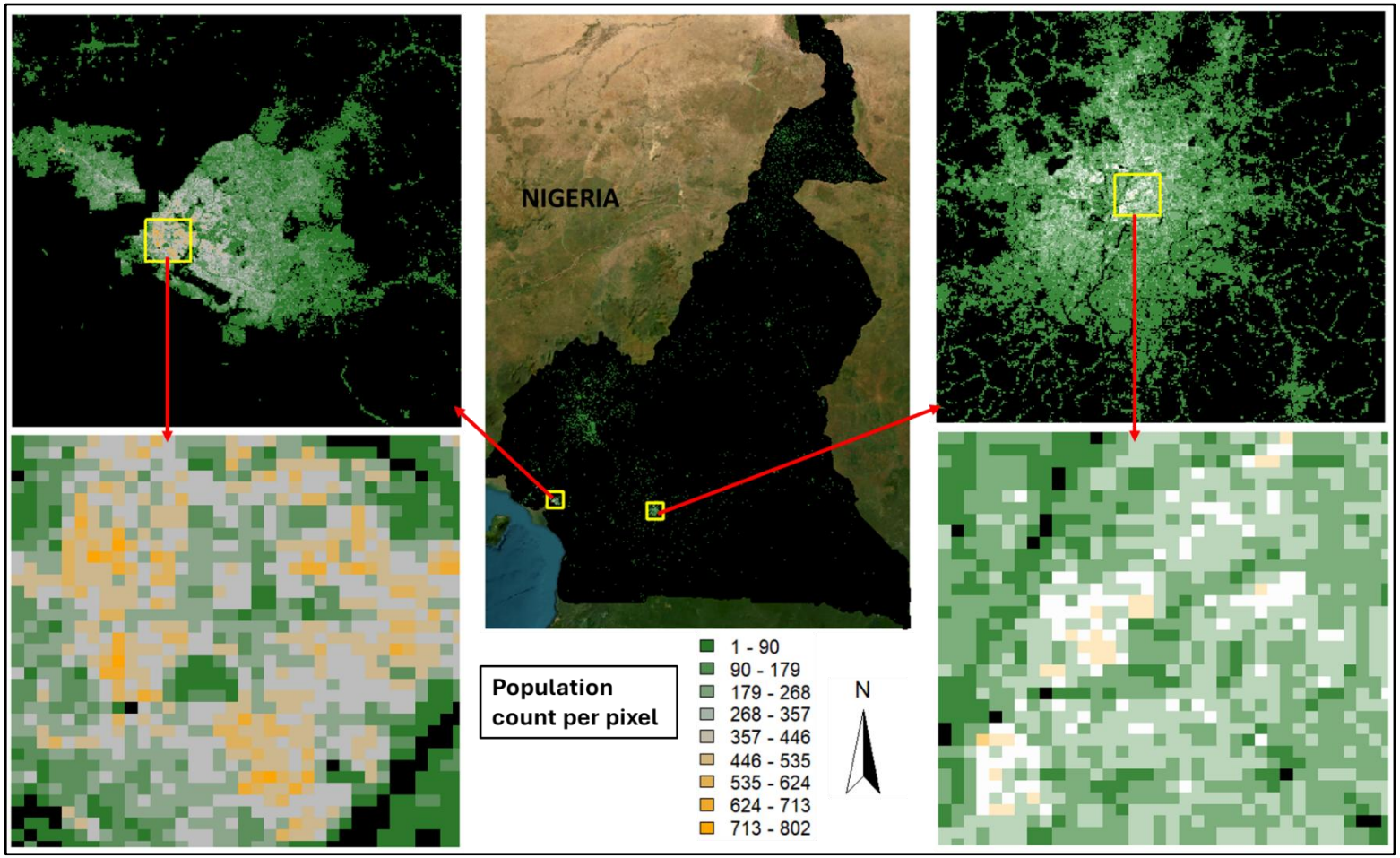

Figure 3A. Predicted population counts across Cameroon at 100m-by-100m square, with corresponding inset maps created for the two major cities in Cameroon – Douala and Yaoundé. A minimum of ~1 and a maximum of ~799 people per grid cell was predicted across the country.This suggests the existence of more clustered but heterogeneous settlement patterns or higher concentration of various forms of residential high-rise buildings per grid cell in Douala than in Yaoundé. It could also mean that more high-rise buildings in Douala are used for residential purposes.

 doi:10.20944/preprints202501.0588.v1

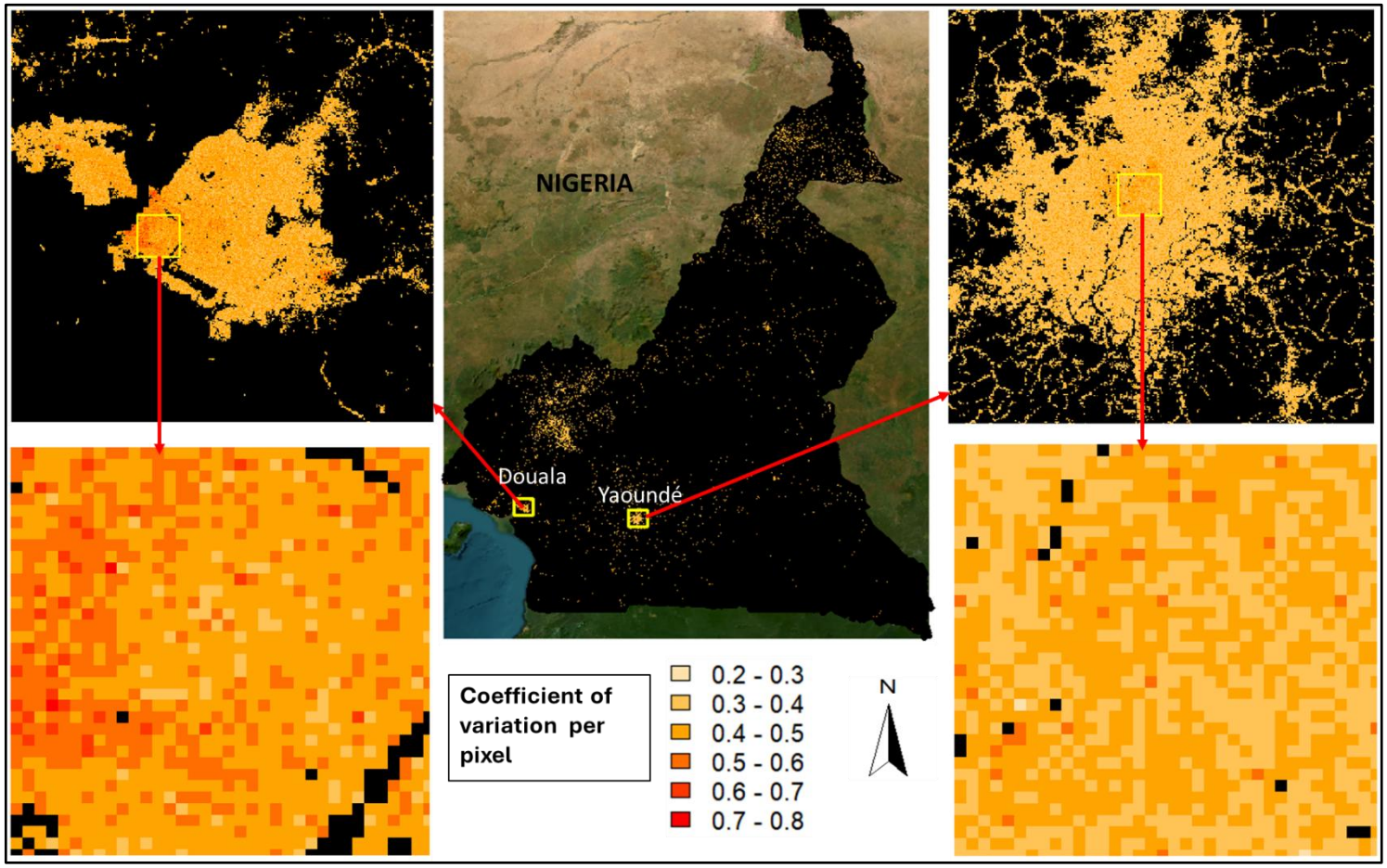

Figure 3B. Coefficient of variation (CV) of the predicted mean of population counts across Cameroon at 100m-by-100m square, with corresponding inset maps created for and within the two major cities in Cameroon – Douala and Yaoundé.

 doi:10.20944/preprints202501.0588.v1

Estimates of uncertainties in the predictions of population counts across the grid cells were quantified through the coefficient of variations (CV) which was calculated as the ratio of the standard deviation of the predicted population count σ_g and the predicted mean count $\bar{x}_g$ per grid cell. The CV provides the relative measures of variability across the grid cells and was provided across the entire spatial domain of Cameroon (Figure 3B). The values of the CV ranged from 0.2 to 0.8 with the highest values obtained in the highest population density areas of Doula but in mostly lowest population density areas of Yaoundé (Figure 3B, inset maps). The high variabilities in the high population density areas of Douala further reinforce the existence of more heterogeneous settlement patterns in Douala than in Yaoundé.

Method Validation

We used a two-pronged approach to validate our methods:

1) A simulation study
2) K-fold cross-validation using the combined real household listing datasets.

Simulation Study

First, the entire spatial domain was taken as a regular rectangle and divided into 11,008 grid cells at 10km-by-10km resolution (Figure S3.1 of the Supplementary document). Note that the 10km square grid cell resolution was chosen for computational convenience while maintaining spatial detail, but any other resolution could be used. We used the Cameroon boundary file, which was provided by the Cameroon National Institute of Statistics (NIS) to crop the grid cells to align perfectly with Cameroon boundary (Figure S2.1C). To check the impacts of spatial autocorrelation, first, we assumed that the entire population was completely observed, i.e. 100% survey coverage provided through five (5) different data sources. We simulated 3 datasets using different spatial variance parameter values set at $\sigma_\xi \in \{0.01, 0.1, 1\}$, where $\sigma_\xi = 0.01$ - low spatial variance; $\sigma_\xi = 0.1$ – moderate spatial variance; and $\sigma_\xi = 1$ for high spatial variance. Other input parameter values used in the simulation are presented in Table 3.

For each of the 3 initially simulated datasets, different levels of missingness or proportions of survey coverages were allowed – 100%, 80%, 60%, 40%, 20%. Thus, altogether, 15 datasets were simulated and tested. And for ease of exposition, variabilities across the different data sources were assumed to be similar with a variance parameter of 0.01. The R scripts used for the implementation of the simulation study are available on the GitHub repository here: https://github.com/wpgp/Efficient-Population-Modelling-using-INLA-SPDE.

Table 3. Simulation study parameters.

Parameter	Value
Grid cell size	10,000
Percentage spatial coverage, $P\%$	100; 80; 60; 40; 20
Smoothness parameter, ν	1
Range of spatial dependence, ρ	0.3
Marginal variances, σ_ξ	0.01, 0.1, 1
Intercept and Coefficients of 5 geospatial covariates, $\boldsymbol{\beta}$ **for** building count simulation	$Intercept, \beta_0$=2.21, β_1=0.06 β_2=0.15, β_3=-0.21, β_4=-0.18, β_5=0.27
Intercept and Coefficients of 5 geospatial covariates, $\boldsymbol{\beta}$ **for** population count simulation	$Intercept, \beta_0$ =3.5, β_1 =0.41 β_2 =0.08, β_3=-0.04, β_4=-0.15, β_5=0.22

Model Fit Metrics (Simulation Study)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2025 doi:10.20944/preprints202501.0588.v1

Figure 4 shows the model fit metrics calculated from the various combinations of the simulation parameter values. Apart from the absolute BIAS, model predictive ability increased with lower spatial variance across all fit metrics regardless of the proportion of missingness within the observed enumeration data. Interestingly, model predictive ability based on BIAS appears to be more sensitive to missingness proportions with model predictions becoming less accurate over lower spatial variance as the proportion of missing data increased. This finding underscores the importance of accounting for spatial variance/spatial autocorrelation while dealing with spatially clustered demographic datasets. But it also highlights the advantage of using multiple model fit metrics to validate and evaluate model performance.

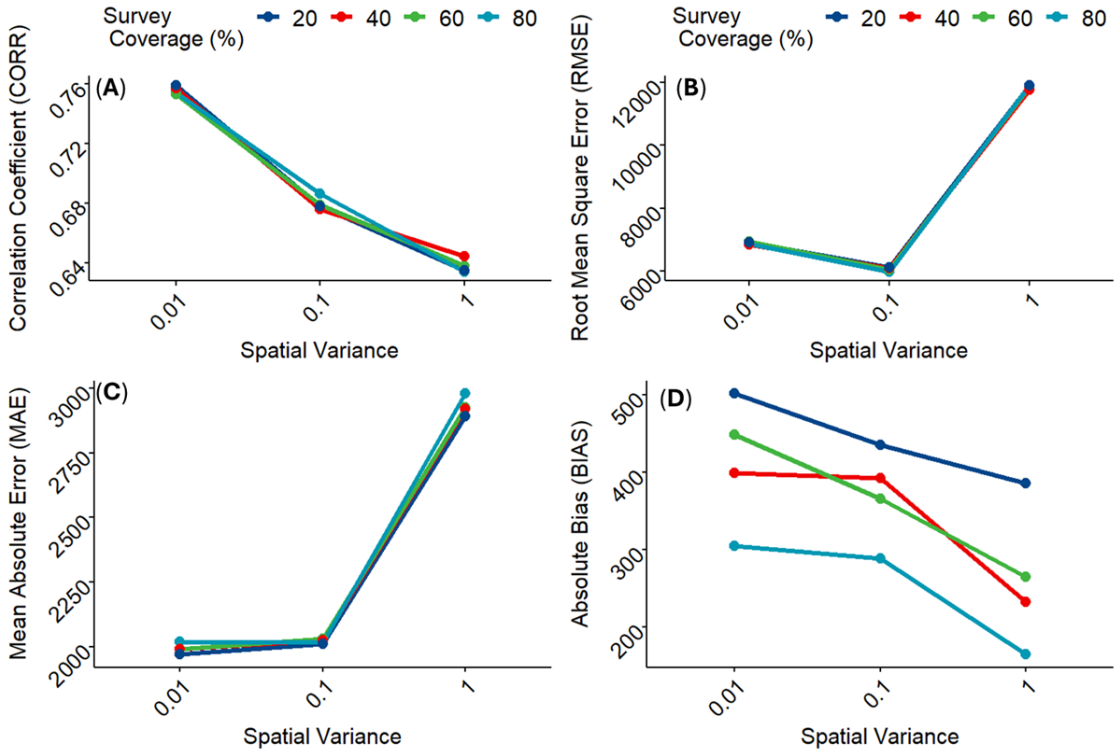

Figure 4. Model fit metrics calculated across the different data scenarios within the simulation study – low, medium and high spatial variance and different levels of survey coverage. These are A) the Pearson correlation coefficient (CORR), B) Root Mean Square Error (RMSE), C) Mean Absolute Error (MAE), and D) absolute bias.

Further evaluation of the simulation study outputs was done by examining the correlations between the simulated population counts and the predicted population counts. Figure 5 shows the scatter plots of the simulated counts versus predicted counts produced across the different levels of data missingness over different spatial autocorrelation structures (i.e., different levels of spatial variances). Overall, the predicted population counts across the various levels of survey coverage (or missingness) within each level of spatial variance, correlated nicely with the corresponding simulated population counts. However, there are indications of higher prediction accuracy of population count for higher spatial dependence (i.e., lower spatial variance). For example, Figure 5B (moderate spatial variance) and Figure 5C (high spatial variance) show evidence of overestimation of the population counts when compared to the low spatial variance scenario in Figure 5A. This suggests a widening variability in the estimates of the population counts as the magnitude of spatial variance increased, thereby underscoring the importance of taking the potential effects of spatial autocorrelation into account in population modelling.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2025 doi:10.20944/preprints202501.0588.v1

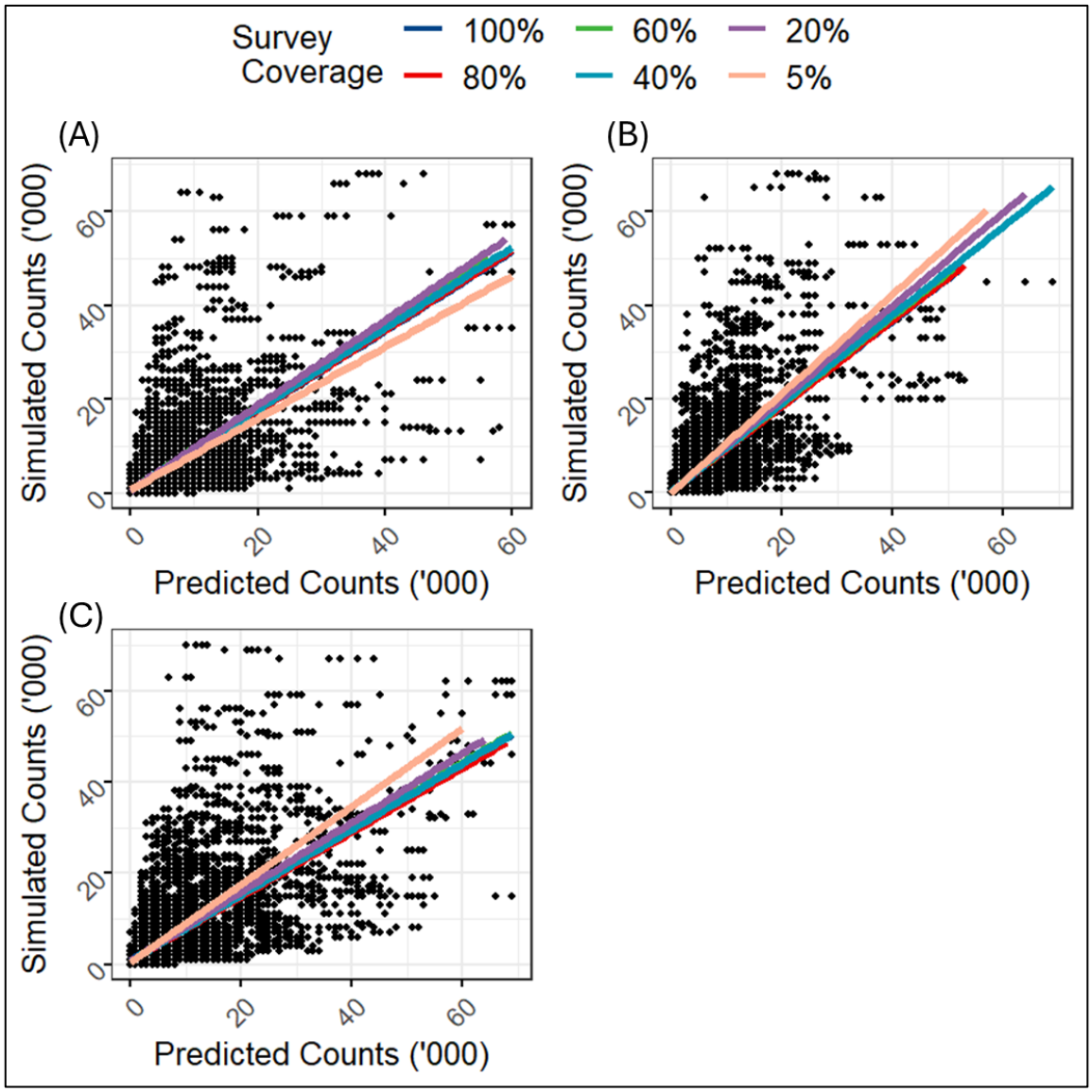

Figure 5. Scatter plots of the simulated population count versus predicted population counts at A) low spatial variance ($\sigma_\xi^2 = 0.01$); B) medium spatial variance ($\sigma_\xi^2 = 0.1$), and C) high spatial variance ($\sigma_\xi^2 = 1$).Low spatial variance (high spatial dependence) produced the highest prediction accuracy. The log-transformed version of Figure 5 is shown in Figure S3.2 of the supplementary document.

k-Fold Cross-Validation (Real Data Application)

k-fold cross validation was used to validate the proposed methodology and test the stability and the predictive ability of the model. Here, the observed motivating Cameroon dataset was used to investigate two forms of cross-validation approaches – in-sample and out-of-sample cross-validation. For the in-sample cross validation, all the datasets were used to train the model as described above. Then 20% of the data was randomly selected and used as a test set. In this case, the test set were random subsets of the training set. This was repeated for 5 times with a different set of test samples selected each time. For each fold, the values of the test samples were set to NA and then predicted using the model parameters. Model fit metrics were then computed and stored.

The out-of-sample cross-validation approach is slightly different and more rigorous. Here, the full data was randomly split into a 20% test set and 80% training set. The model was trained with the 80% training set, while the withheld values of the 20% test set were predicted using the trained model parameters. In this case, the test set was not part of the training set. As in the in-sample strategy, the k-folds test sets were non-overlapping, and the values of model fit metrics were calculated after

 doi:10.20944/preprints202501.0588.v1

model predictions with each test set. Thus, for each of the in-sample and out-of-sample cross-validations we used k {=5} folds. The model fit metrics calculated across all the folds for each method along with their average values are presented in Table 3. Model fit metrics based on the two strategies were stable and similar, and there is an adequate average correlation coefficient of at least 98% for both strategies. The R scripts used for the implementation of the cross-validation strategies are available in the GitHub repository: https://github.com/wpgp/Efficient-Population-Modelling-using-INLA-SPDE.

Additionally, we carried out a visual inspection of the cross-validation outputs by displaying the violin plots with embedded notched boxplots of the predicted population counts for each test set fold (Figure 6). The findings further reinforced the outputs in Table 4, which highlighted the high level of accuracy and efficiency of our methodology. There was good agreement between the predicted values across each corresponding folds for both the in-sample and out-of-sample strategies.

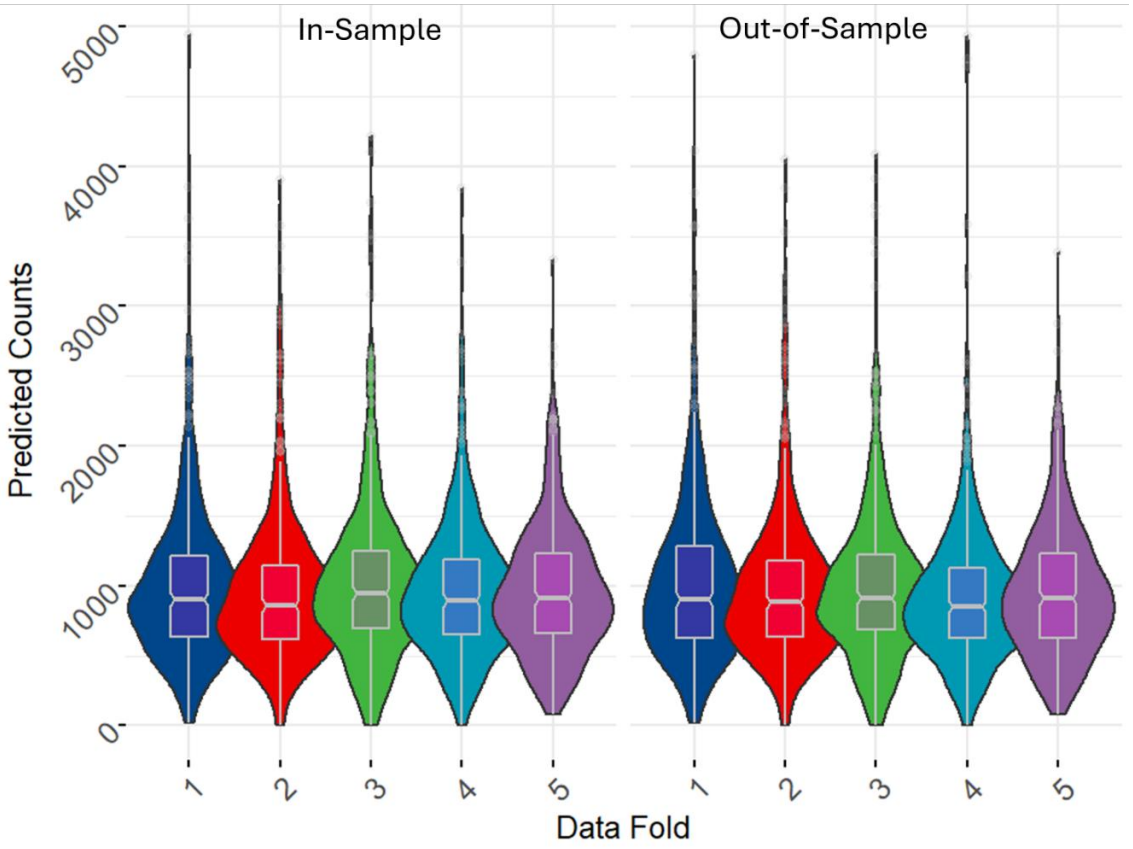

Figure 6. Violin plots with notched box plots of the predicted population counts for each of the 5 folds used for the real data cross-validation for both in-sample and out-of-sample cross-validations. The in-sample fit figures are looking similar to the out-of-sample ones indicating a good performance of our methodology.

 doi:10.20944/preprints202501.0588.v1

Table 4. Model validation metrics obtained from the k{=5}-fold cross validation (in-sample and out-of-sample).

		METRICS			
DATA	FOLD	MAE	RMSE	BIAS	CORR
In-Sample	Fold 1	148.2161	314.2054	88.0447	0.9909
	Fold 2	133.7209	187.8360	76.3253	0.9895
	Fold 3	137.3771	216.0733	68.3691	0.9813
	Fold 4	137.5897	203.0070	73.1012	0.9796
	Fold 5	136.8034	214.6483	68.7344	0.9875
	Mean	**138.7414**	**227.1540**	**74.9150**	**0.9858**
Out-of-Sample	Fold 1	160.5572	350.7214	67.3685	0.9838
	Fold 2	135.8985	189.0198	38.2595	0.9870
	Fold 3	171.1974	292.6777	107.5242	0.9726
	Fold 4	157.2633	247.1928	121.7557	0.9774
	Fold 5	135.7502	199.6707	70.2284	0.9868
	Mean	**152.1333**	**255.8565**	**81.02725**	**0.9815**

Comparing the Modelled Estimates with Projected Estimates

The last population and housing census in Cameroon was in 19 years ago in 2005 and since then, official population data for Cameroon have been provided through projections that use the 2005 census as the baseline. The projections were made using the cohort component approach (Preston et al., 2001). In Figure 7, we compare the 2022 official projections received from the Cameroon National Institute of Statistics (NIS) with the modelled estimates produced from our methodology for administrative unit 1 (regions). Given the different input data sources and modelling approaches used, we do not expect the output population estimates to agree, but the comparisons can be informative.

 doi:10.20944/preprints202501.0588.v1

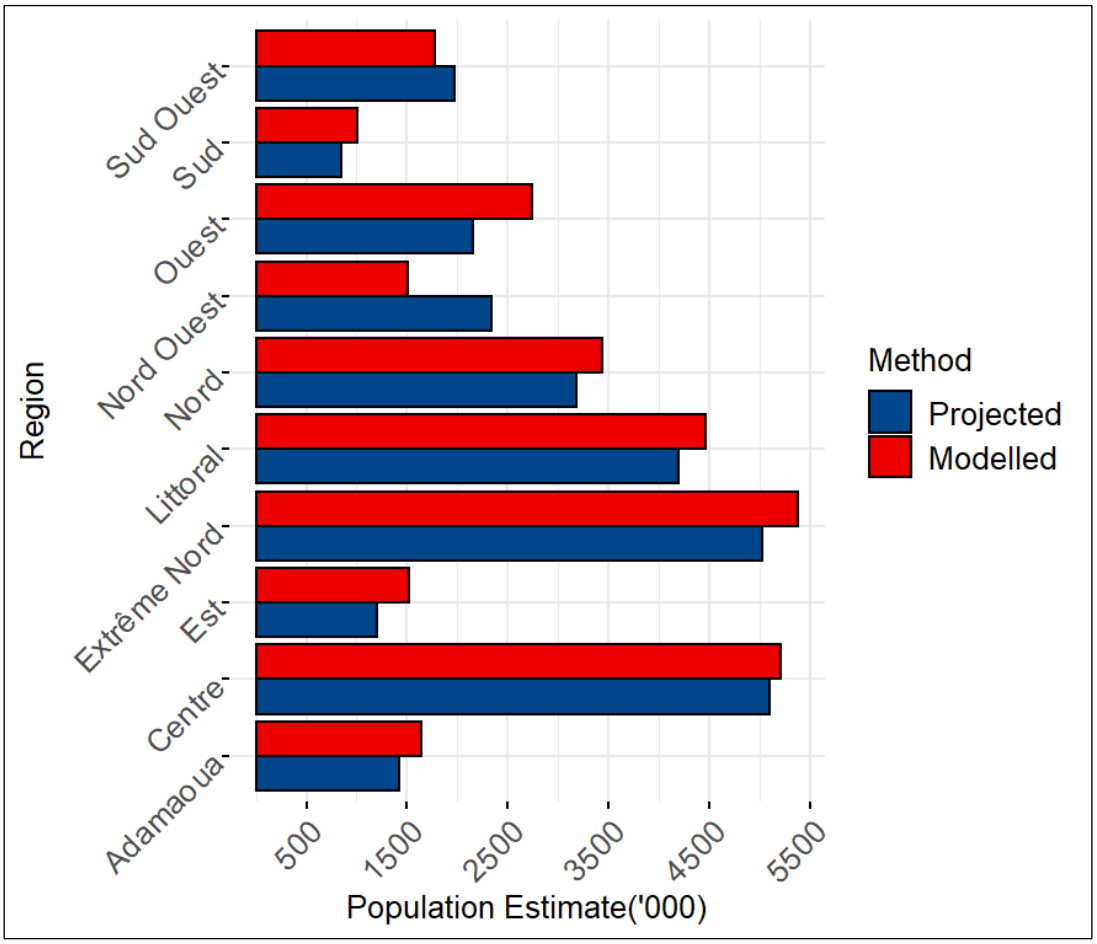

Figure 7. A bar plot for the comparison of the aggregated modelled population estimates for Cameroon with the 2022 official population projections (National Institute of Statistics, 2016) provided by the Cameroon National Institute of Statistics.

Notably, apart from the Nord Ouest and Sud Ouest regions, all of the regions showed slightly higher modelled estimates than the official projections. The lower modelled estimates for the Nord Ouest and Sud Ouest regions could be a result of prolonged high rates of conflicts and insecurity within the regions, thereby leading to high rates of displacement and out-migration that were not accounted for in the cohort component projections.

One key strength of the modelling approach outlined here is that it takes advantage of more recent small area population data and geospatial covariates that can capture recent changes in population distributions and associated drivers, unlike the cohort component population projection approach used by the NIS. This feature of our modelling approach has led the Cameroon NIS to adopt the modelled estimates in supporting census preparation and as a sample frame in the design and implementation of health campaigns.

The full datasets have been published online and can be download freely from WorldPop data repository.

Discussions

In this paper, we presented statistical population modelling method that allowed for the integration of multiple data sources as well as spatial autocorrelation within a bottom-up population modelling framework (e.g., Leasure et al., 2020; Boo et al., 2022; Darin et al, 2022; Nnanatu et al., 2024). For an improved computational efficiency and higher accuracy, we adapted the integrated nested Laplace approximation (INLA; Rue et al. 2009) statistical modelling technique, in conjunction with the stochastic partial differential equation (SPDE; Lindgren et al. 2011) strategies. Bayesian statistical

 doi:10.20944/preprints202501.0588.v1

framework enabled us to integrate prior information whilst simultaneously estimating the hierarchical regression model parameters along with their uncertainties.

The methodology was successfully validated using both an extensive simulation study and real data application. The aim of the simulation study was to evaluate the robustness of the methodology over different combinations of missing data proportions and spatial autocorrelation (spatial variance). Model performance in terms of prediction accuracy increased with lower proportion of missing data values and higher spatial autocorrelation. As a proof of concept, small area estimates of population were produced for Cameroon using 5 nationally representative household listing datasets. Modelled estimates were validated using k-fold cross-validation while model selection was based on the deviance information criterion (DIC; smaller values indicated better fit models).

However, it is important to highlight the key limitations of our methodology: First, within the simulation study, we only the scenario where the different data sources do not differ significantly in terms of their data collection designs. Although, this was the case in the motivating dataset where all five data sources used the same sampling strategy. The data source random effect was found not to be statistically significant in preliminary studies and so it was not included in the later models. However, in contexts where data are collated from different sources with significantly different data collection strategies, it makes sense to explore the potential effects of different levels of data source variabilities using an extensive simulation study. Also, the use of household listing datasets and geospatial covariates from different years without explicitly accounting for the year difference effects, and the various sampling frames is likely to be a source of variability that needs to be investigated in future studies. Additionally, temporarily displaced populations due to insecurity may affect population estimates when building footprints are used as a covariate. These aspects will be explored further in future studies.

Nevertheless, the methodology presented here is an important development within the context of population modelling and will serve to provide more accurate small area population data required to address several population data gaps across many countries. Our modelling approach draws upon more recent small area population data and geospatial covariates to estimate population numbers at unsampled locations thereby capturing recent drivers of population changes/density and distributions. As noted, this is a key strength of our approach over the commonly used population projection methods like the cohort component population projection method (e.g., Smith et al., 2013). These datasets which were produced in close collaboration with the Cameroon National Institute of Statistics are publicly available (https://data.worldpop.org/repo/wopr/CMR/population/v1.0/) and now being used by the Cameroon NIS in supporting census preparation and as a sample frame in the design and implementation of health campaigns in Cameroon.

Finally, we have made the R programming codes used to implement the methodology publicly available (https://github.com/wpgp/Efficient-Population-Modelling-using-INLA-SPDE) and easily accessible to facilitate its reproducibility and easier adaptation in different contexts by students, researchers and policymakers.

Supplementary Materials: Supplementary material which contains a description of the posterior simulation, and some relevant figures is available here: https://github.com/wpgp/Efficient-Population-Modelling-using-INLA-SPDE/blob/main/Bayesian_Geostat_Pop_Mod_MethodsX_supplementary.pdf.

Author Contributions: Conceptualization: CCN; Data Curation: ANL, ADD, OY, CCN; Formal Analysis: CCN, OY, TA, AG; Methodology: CCN; Project Administration: ANL, AJT; Software: CCN; Supervision: ANL, AJT; Writing - original draft: CCN; Writing – review & editing: AJT, CCN, OY, TA, AG, ANL, ADD.

Acknowledgments: We are grateful to the Cameroon National Institute of Statistics (NIS) for providing the household listing datasets, and the projected population estimates used in this research. We appreciate Edith Darin for reviewing the initial draft of manuscript, we thank the Spatial Statistical Modelling (SSPM) team for their very constructive comments on the initial drafts. This work is part of the GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) program funded by the Bill and Melinda Gates Foundation and the United Kingdom Foreign, Commonwealth & Development Office (OPP1182425). Project

 doi:10.20944/preprints202501.0588.v1

partners include WorldPop at the University of Southampton, the United Nations Population Fund (UNFPA), Center for Integrated Earth Science Information (CIESIN) in the Columbia Climate School at Columbia University, and the Flowminder Foundation.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics statements: The datasets utilised in this study were completely anonymised and aggregated to avoid confidentiality issues in accordance with relevant data protection regulations. Additionally, Ethical approval was obtained from the University of Southampton; ethics approval number: ERGO II 72177 (GRID3 - Cameroon pre-EA tool application)

References

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30, 185–207

Bakka, H., Rue, H., Fuglstad, G-A., et al.(2018). Spatial modeling with R-INLA: A review. WIREs *Comput Stat*.;10:e1443. https://doi.org/10.1002/wics.1443

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

Chamberlain, H. R., Darin, E., Adewole, W.A., Jochem, W.C., Lazar, A.N., Tatem, A.J. (2024). Building footprint data for countries in Africa: to what extent are existing data products comparable? Comput. Environ. Urban Syst. 110 102104. https://doi.org/10.1016/j.compenvurbsys.2024.102104

Chan-Golston, A., Banerjee, S., Belin, T.R. *et al.* (2022). Bayesian finite-population inference with spatially correlated measurements. *Jpn J Stat Data Sci* **5**, 407–430. https://doi.org/10.1007/s42081-022-00178-8

Chi, G., Voss, P.R. (2011). Small-area population forecasting: borrowing strength across space and time. *Population, Space, and Place.* **17**, 505 - 520. https://doi.org/10.1002/psp.617

Cressie N. (1993). *Statistics for spatial data*, revised ed.. New York: Wiley.

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

Diggle, P.J., Giorgi, E. (2016). Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings. *Journal of The American Statistical Association*, 515 (111).

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

Ecopia and Digital Globe (2017). Technical specification: Ecopia building footprints powered by DigitalGlobe. Available at: https://dg-cms-uploads-production.s3.amazonaws.com/uploads/legal_document/file/109/DigitalGlobe_Ecopia_Building_Footprints_Technical_Specification.pdf (accessed 11 April 2022)

Giorgi, E., Diggle, P.J., Snow, R.W., Noor, A.M. (2018). Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys. *Int Stat Rev*. Dec;86(3):571-597. doi: 10.1111/insr.12268.

Gomez-Rubio, V. (2020). Bayesian inference with INLA. CRC Press. Boca Raton, FL. https://becarioprecario.bitbucket.io/inla-gitbook

James, G., et al. (2013). *An Introduction to Statistical Learning*. 1st ed., PDF, Springer.

Leasure, D. R., Jochem, W.C., Weber, E. M., Seaman, V., 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

 doi:10.20944/preprints202501.0588.v1

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

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

Moran, P.A.P. (1950), 'Notes on Continuous Stochastic Phenomena', Biometrika, 37, 17–23.

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

Nieves, J. J., Stevens, F. R., Gaughan, A. E., Linard, C., Sorichetta, A., Hornby, G., Patel, N. N., Tatem, A. J. (2017). Examining the correlates and drivers of human population distributions across low- and middle-income countries. J. R. Soc. Interface.1420170401 http://doi.org/10.1098/rsif.2017.0401

Nnanatu, C.C., Bonnie, A., Joseph, J., Yankey. O., Cihan, D., Gadiaga, A., Voepel, H., Abbott, T., Chamberlain, H., Tia, M., Sander, M., Davis, J., Lazar, A.,Tatem, A.J. (2024). Small area population estimation from health intervention campaign surveys and partially observed settlement data. *Research Square* (Preprint). https://doi.org/10.21203/rs.3.rs-5059066/v1

Paige, J., Fuglstad, G-A., Riebler, A., Wakefield, J. (2022). Spatial aggregation with respect to a population distribution: Impact on inference, *Spatial Statistics*, 52, https://doi.org/10.1016/j.spasta.2022.100714.

Pettit, L. I. (1990). "The Conditional Predictive Ordinate for the Normal Distribution." Journal of the Royal Statistical Society. Series B (Methodological) 52 (1): pp. 175–84.

Preston, S. H., Heuveline, P., Guillot, P. (2001) *Demography: Measuring and Modeling Population Processes,* Wiley.

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

Rue, H., Martino, S., 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

Schiavina, M., Melchiorri, M., Pesaresi, M., Politis, P., Freire, S., Maffenini, L., Florio, P., Ehrlich, D., Goch, K., Tommasi, P., Kemper, T. (2022). GHSL Data Package 2022, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-53071-8, doi:10.2760/19817, JRC 129516

Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). Statistical Science, 32(1), 1–28.

Smith, S.K., Tayman, J., Swanson, D.A. (2013). Overview of the Cohort-Component Method. In: A Practitioner's Guide to State and Local Population Projections. The Springer Series on Demographic Methods and Population Analysis, vol 37. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7551-0_3

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(sup1), 234–240. https://doi.org/10.2307/143141

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

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* 8, 158–183. doi:10.1093/biostatistics/kxl008

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

Watanabe, S. (2013). "A Widely Applicable Bayesian Information Criterion." Journal of Machine Learning Research 14: 867–97.

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

 doi:10.20944/preprints202501.0588.v1