Efficient approach for integrating spatial-autocorrelation and multiple data sources in small-area population estimation

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

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

[2]Department of Statistics, Nnamdi Azikiwe University, PMB 5025, Awka-Nigeria

[3]National Institute of Statistics, Siège social : 20, Rue 3025, Quartier du Lac, Yaoundé, Cameroon

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

SUPPLEMENTARY MATERIAL

S1: Application to Cameroon Household listing Datasets

Below, we outline the various steps undertaken with respect to the data cleaning activities.

The outcome of interest in the household listing dataset was household size. For each household listing dataset, first we selected only the households with observed household size for recruiting the training dataset. This means that all other households (or rows) within a given EA which have no household size information or simply NA, where NA means 'Not Available' were dropped at this stage for the purpose of model training. The assumption here was that the households were simply not visited and that the missingness is completely at random.

Other forms of missingness were addressed using mean imputation (e.g., Little & Rubin, 2002). For example, 99 or 98 were commonly used as missing data codes within the data. According to the Cameroon NIS, these codes were used for inaccessible households which were not visited by the enumerators. Thus, the missing household sizes were imputed using mean imputation which assumes that missingness is either missing completely at random (MCAR) or simply missing at random (MAR). This simply entails calculating the average household size within each EA and then imputing the missing household size with the average household size of the EA. In the end, a total of 348 791 people were imputed across the entire dataset, i.e., 15% of the total surveyed population. These were implemented in R statistical programming software.

Given that the spatial scale or unit of our model is at the EA (or cluster) level, we summed the total number of people in each household for each cluster to form the cluster total as the response variable of our model.

To create uniform EA id that would be matching with those within the shapefiles across the various surveys, we concatenated the arrondisement (subdivision) ID and cluster ID of each cluster as unique ids. This makes it a lot straightforward to join the shapefiles of the datasets.

The shapefiles for the household listing were cleaned separately. First, the shapefiles were mapped and assessed for any potential boundary or information anomaly. Although there were no major issues in the shapefiles with respect to their geometry and they were also not corrupted. A few points with invalid geometry mainly due to misalignment was deleted from the shapefiles.

The joining of the shapefiles was done just by using the ids of the clusters and arrondisements within the shapefiles and concatenating them to create a unique cluster level id that would match those of the household listing dataset.

The summarised household listing datasets were joined to their respective EA shapefiles. All the datasets were combined into one. For duplicated clusters, i.e. exactly same cluster surveyed across the 5 datasets, only the household size value for most recent data was used. For example, the most recent data was CMIS, followed by EESI 3, and ECAM5, hence where we have observations for CMIS and EESI 3, we maintained only CMIS since CMIS is more recent. Also, where we have observations for EESI 3 and ECAM5, we have maintained EESI 3 and dropped ECAM5.

Table S1: Input data. Description of the final household listing datasets used for the population estimation

Survey	Reference Year	Brief Description
ECAM5 Phase1	2021	The sampling frame used for ECAM5 is the list of all EAs resulting from the cartographic work carried out in 2017 for the fourth General Census of Population and Housing (RGPH4) by the Central Bureau of Censuses and Populations Studies in Cameroon (BUCREP). The sampling frame consisted of 21826 EAs, out of which 402 EAs were surveyed in ECAM5 Phase 1
ECAM5 Phase 2	2021	This data collection was carried out in 395 EAs out of 21826 EAs in the country. ECAM5 Phase 2 is a subset of ECAM5 and the survey was conducted after ECAM5 Phase 1
ECAM5 Phase 3	2021	This data is a subset of ECAM5 and 411 EAs were surveyed out of 21826 total EAs in the country. ECAM5 Phase 3 was conducted after ECAM5 Phase 2.

Third Employment and Informal Sector Survey(EESI 3)	2021	The *Third Employment and Informal Sector Survey* was implemented in 2021. he main objective of the survey was to collect data for measuring labour market indicators in the country. A total of 759 EAs from the above-mentioned sampling frame were surveyed for this data collection.
Cameroon Malaria Indicator Survey (CMIS)	2022	The Malaria Indicator Survey was implemented in 2022. The survey was conducted in 438 EAs in the country from the above-mentioned sampling frame. The main objective of the survey was to provide information on malaria prevention, treatment, and prevalence in Cameroon.

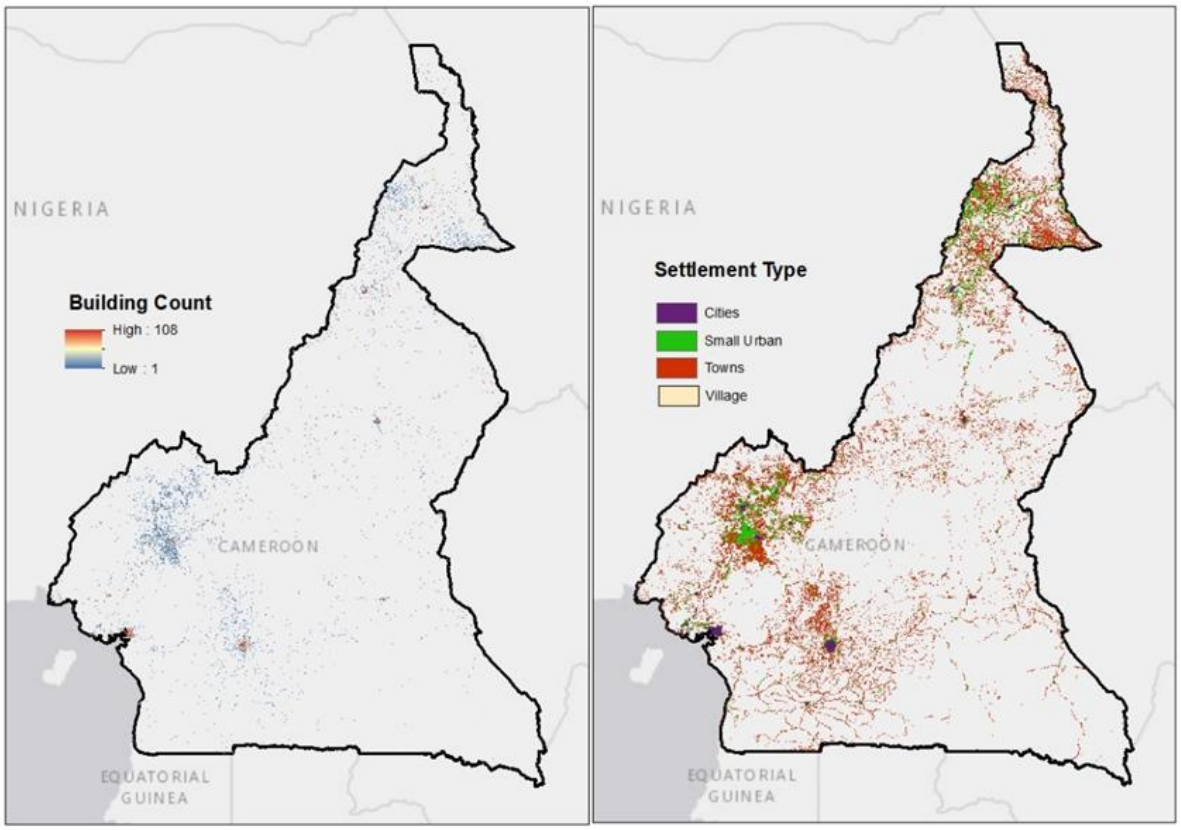

Fig S1. Distribution of Building count (left) and Settlement types (right) across Cameroon. The building counts were obtained from the building footprint layer provided by the Digitize Africa project of Ecopia AI and Maxar Technologies ('year 2', 2020/2021; Ecopia.AI and Maxar Technologies, 2020). The 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.

Table S2. Descriptions of the geospatial covariates considered in the statistical models.

Covariate name	Source	Format	Resolution	Link
Distance to ACLED battles data 2021	ACLED	Raster	100m	https://acleddata.com/
Distance to ACLED conflict data 2021	ACLED	Raster	100m	https://acleddata.com/
Distance to ACLED explosions 2021	ACLED	Raster	100m	https://acleddata.com/
Distance to ACLED protests 2021	ACLED	Raster	100m	https://acleddata.com/
Distance to ACLED riots 2021	ACLED	Raster	100m	https://acleddata.com/
Distance to ACLED strategic developments 2021	ACLED	Raster	100m	https://acleddata.com/
Motorized friction surface 2020	MAP	Raster	100m	https://malariaatlas.org/research-project/accessibility-to-healthcare/
Walking friction surface 2020	MAP	Raster	100m	https://malariaatlas.org/research-project/accessibility-to-healthcare/
Distance to cities 2015	MAP	Raster	100m	https://malariaatlas.org/research-project/accessibility-to-cities/
Distance to places of education 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to Health providers 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to marketplaces 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to places of worship 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to local roads 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to main roads 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html

Distance to Water bodies 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to railway stations 2022	OSM	Raster	100m	https://www.geofabrik.de/data/download.html
Distance to primary road intersections 2016	WorldPop	Raster	100m	https://www.worldpop.org/geodata/listing?id=33
Distance to cultivated areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to woody areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to shrub area edges 2015 (130)	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to herbaceous areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to sparse vegetation areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to aquatic vegetation areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to Urban area 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Distance to bare areas 2015	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Current average total annual precipitation 2020	Copernicus	Raster	100m	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=form
Current average annual temperature 2020	Copernicus	Raster	100m	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=form
Slope 2000	Worldpop	Raster	100m	https://www.worldpop.org/project/categories?id=14
Elevation 2000	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=15
Distance to coastline 2000-2020	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=16

Nighttime lights 2020 VIIRS	WorldPop	Raster	100m	https://www.worldpop.org/project/categories?id=17
Distance to CAT1 protected areas 2017	Worldpop	Raster	100m	https://www.worldpop.org/project/categories?id=18
Buildings area 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings length 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings mean area 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings mean length 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings total area 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings total length 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Buildings density 2020	WorldPop/ Ecopia	Raster	100m	https://wopr.worldpop.org/?CMR/Buildings/v1.1
Distance to built settlement areas worldpop 2020	WorldPop	Raster	100m	https://hub.worldpop.org/geodata/summary?id=17090

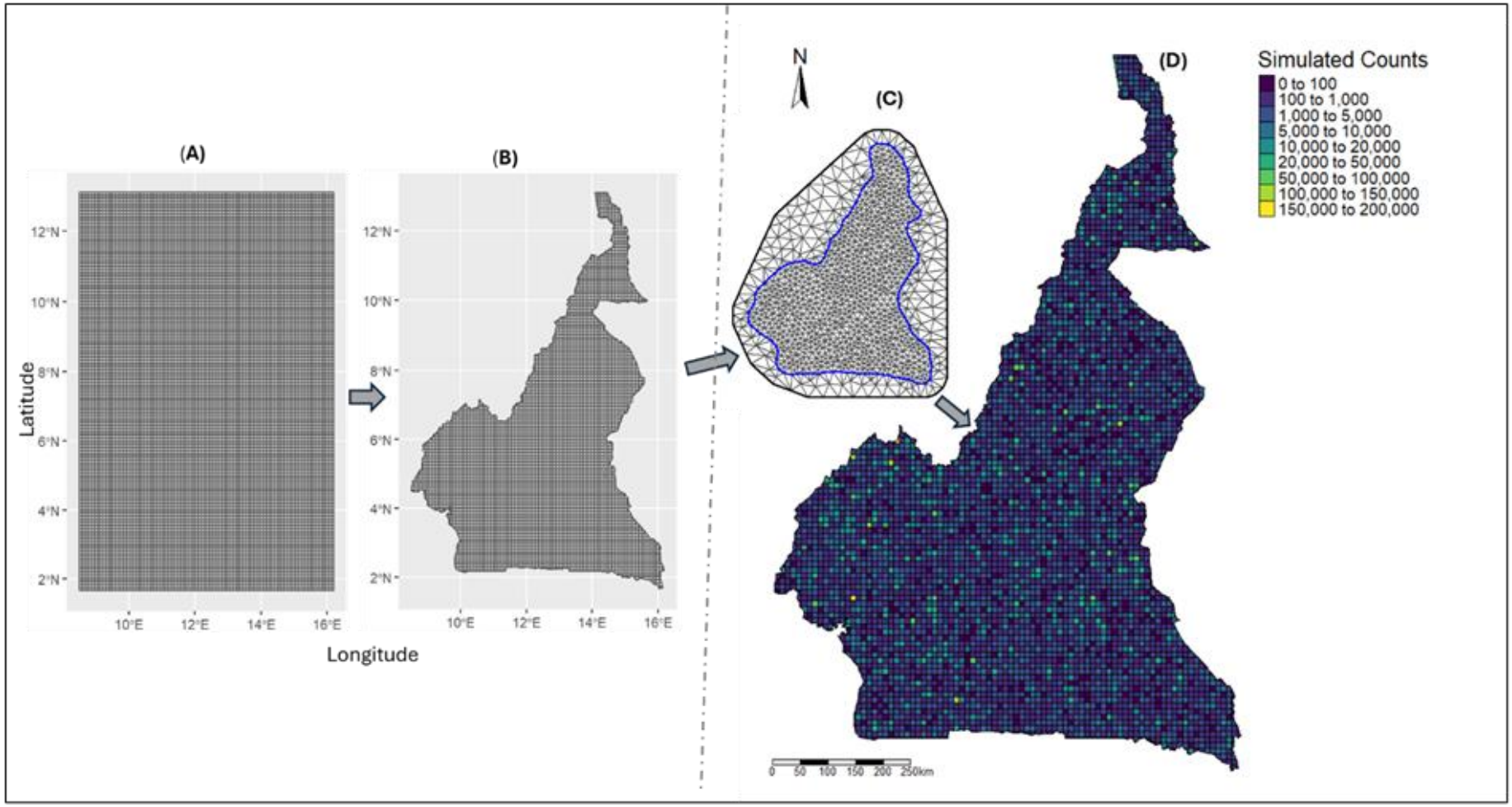

Fig S2. Simulation study scheme showing the A) 11,008 10km-by-10km grid cells, B) grid cells cropped to the national boundary of Cameroon C) Mesh – discretization of the spatial domain with 534 triangular vertices (D) Simulated counts across the entire grid cells with 100% observations and low spatial variance ($\sigma_\xi = 1.25$)

Table S3: 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	0.2
Marginal variances, σ_ξ	1.25, 1.75, 2.75, 3.75
Data source variance	Equal: 0.001 Unequal: 0.0015,0.002, 0.0017, 0.001, 0.0032
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

S2. Posterior Simulation and grid cell prediction

In this Section, we give the motivation for the posterior resampling and later provide the step-by-step approach utilised in the posterior simulation and uncertainty quantifications for both mean values and aggregated estimates. **Background**

Let $\pi(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y})$ denote the (approximate) joint posterior distribution of the latent field and hyperparameters given the data, in Bayesian statistical inference, the main goal is usually to evaluate the desire summary statistics such as the mean or the standard deviation and quantify uncertainty in these estimates as 95% credible interval, say, of the posterior samples from $\pi(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y})$. However, while it could be quite straightforward to obtain the estimates of uncertainty using relevant INLA functions, it is more complicated to obtain estimates of aggregated total populations in area/administrative units of interest because quantiles cannot be summed since the total of quantiles is not the same as the quintile of the totals.

To address this challenge, first, we obtain a sampling distribution of the mean totals generated from the joint posterior distribution $\pi(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y})$. Once a large enough mean totals sample has been constructed, it becomes straightforward to obtain the desired statistics such as mean, standard deviation, and 95% credible intervals. The 95% credible interval is obtained as quantiles at 2.5% for lower bound and at 97.5% for the upper bound. The idea here is that drawing samples from the conditional distribution of a given parameter θ_1, say, and then averaging over all the iterations will normally improve the estimation of the posterior marginal distribution $\pi(\theta_1|\theta_2, y)$, and may be viewed as synonymous to the Rao-Blackwellization theory since group averaged estimators cannot increase the variance (Robert & Roberts, 2021; Blackwell, 1947; Rao, 1945).

The five key steps for drawing samples (simulation) from the joint posterior density are listed below:

Posterior Simulation steps

1) Specify and fit the INLA model using the **inla()** function. Setting **config = TRUE** within the **control.compute** argument allows INLA to draw samples from the (approximate) joint posterior distribution $\pi(\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{y}})$ obtained, whenever required.

2) Using the (approximate) joint posterior distribution obtained in step 1, draw T samples of the latent field $\{\widetilde{\boldsymbol{w}}_t\}_{t=1}^{T}$ and hyperparameters $\{\widetilde{\boldsymbol{\theta}}_t\}_{t=1}^{T}$ via the **inla.posterior.sample()** function of the R-INLA package. This entails refitting the INLA model N times whilst allowing some variabilities within the random parameters. Note that the number of iterations T is variable, however, the larger the better. Note that computational cost would also increase with T.

3) For the G $100m \times 100m$ square grid cells, define a $G \times D$ projection matrix $\widetilde{\boldsymbol{A}} = (\widetilde{\boldsymbol{A}}_{g,d})$ to project the grid cell level values at the D mesh nodes.

4) For t in 1 to T, do the following:

   (a) Select the t-th sample from the joint posterior sample $\{\widetilde{\boldsymbol{w}}_t, \widetilde{\boldsymbol{\theta}}_t\}$

   (b) Using the scaled grid cell values of the covariates which were used in the model fitting stage at step 1, along with the intercept and the associated estimates of

fixed effects coefficients , $\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_8)$, and the estimates of the random effects, predict the population density $\widehat{D}_{g,t}$ $(g = 1, \ldots, G)$ for each pixel as outlined in the main manuscript. The random effects values are simulated from their respective posterior means, that is,

$$\gamma_{gt} \sim Normal(0, \hat{\sigma}_\gamma) \qquad (S1)$$

where $\gamma \in \{f_m, f_p, f_{rp}, \varepsilon\}$ and $\hat{\sigma}_\gamma > 0$ is the estimated variance for each random effect obtained in step 1.

(c) Using the predicted density and the observed building counts B_g within a given grid cell, obtain the predicted population counts as

$$\hat{y}_{g,t} = \widehat{D}_{g,t} \times B_g \qquad (S2)$$

(d) Repeat sub-steps (a) to (c) until a sample of the desired size is obtained.

(e) Store the $G\ by\ T$ matrices of the sampling distributions of the predicted population density and population count across the grid cells.

5) From the stored simulated posterior samples, generate the various summary statistics such as:

(i) **Mean grid cell level population density** is obtained as the average predicted population densities across the entire sample of size T for the g-th grid cell.

$$\overline{D}_g = \frac{1}{T}\sum_{t=1}^{T} \widehat{D}_{g,t} \qquad (S3)$$

(ii) **Mean grid cell population count** is obtained as the average predicted population count across the entire sample of size T for the g-th grid cell.

$$\bar{y}_g = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_{g,t} \qquad (S4)$$

(iii) **Grid cell level uncertainties** in the estimates of the grid cell population density and population count are based on the 95% credible interval obtained by taking the 2.5% and 97.5% quantiles of the g-th grid cell samples $\hat{y}_{g,1}, \hat{y}_{g,2}, \ldots, \hat{y}_{g,T}$ and $\widehat{D}_{g,1}, \widehat{D}_{g,2}, \ldots, \widehat{D}_{g,T}$ for population count and population density, respectively. Alternatively, the pixel level 95% upper and lower bounds estimate of uncertainties could be obtained as confidence intervals using for example, the lower bound can be calculated as

$$lower^{(g)} = \overline{w}_g - 2\sigma_g \qquad (S5)$$

while the upper bound is given by

$$upper^{(g)} = \overline{w}_g + 2\sigma_g \qquad (S6)$$

The grid cell level standard deviation, σ_g, is given by

$$\sigma_g = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\widehat{w}_{g,t} - \overline{w}_g\right)^2} \qquad (S7)$$

$\widehat{w}_{g,t} \in \{\widehat{D}_{g,t}, \hat{y}_{g,t}\}$ and $\overline{w}_g \in \{\overline{D}_g, \bar{y}_g\}$.

The lower and upper bounds of the estimates provide an indication of the variability around the posterior estimates. A unique measure of $uncertainty^{(g)}$ may then be derived as the average deviation given by

$$uncertainty^{(g)} = \frac{upper^{(g)} - lower^{(g)}}{\overline{w}_g} \qquad (S8)$$

(iv) **Obtain a distribution of the population totals** at the various administrative levels: For each iteration, obtain **the total population** $\hat{y}_t$ by summing the predicted count over all the grid cells, that is,

$$\hat{y}_t = \sum_{g=1}^{G} \hat{y}_{g,t} \qquad (S9)$$

Thus, **the administrative level summary statistics** of interest will then be obtained from the distribution of the total counts for all the T iterations $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T$, so that the **mean total count** within a given administrative unit $\bar{y}_{admin}$ is given by

$$\bar{y}_{admin} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)} \qquad (S10)$$

The **standard deviation** σ_{admin} is given by

$$\sigma_{admin} = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(\hat{y}_t - \bar{y}_{admin})^2} \qquad (S11)$$

Similarly, the measures of uncertainty are then obtained as the 95% credible intervals using the **quantile()** function of base R set at 2.7% for the lower bound and at 97.5% for the upper bound. The **uncertainty** can also be given as confidence intervals using,

$$lower = \bar{y}_{admin} - 2\sigma_{admin} \qquad (S12)$$

while the upper bound is given by

$$upper = \bar{y}_{admin} + 2\sigma_{admin} \qquad (S13)$$

Note that for the lower administrative levels, subsets of samples belonging to the admin level of interest are first obtained from the overall samples below applying the various steps described above). In addition, a unique measure of $uncertainty$ may be obtained for the lower administrative units as

$$uncertainty = \frac{upper - lower}{\bar{y}_{admin}} \qquad (S14)$$

Alternatively, coefficient of variation which is the ratio between the standard deviation and the mean could be used as a measure of prediction uncertainty. Like the usual posterior sampling using methods such as MCMC, one may wish to view the mixing and distribution of the posterior samples across the parameter space using trace plots and histograms. However, unlike in the MCMC where these graphs are used to check convergence of the Markov chains, within INLA the graphs are for descriptive statistics because the samples are drawn from the 'true' joint posterior density.

The final estimates produced from the posterior simulations can be obtained in various formats as data tables, maps or raster files

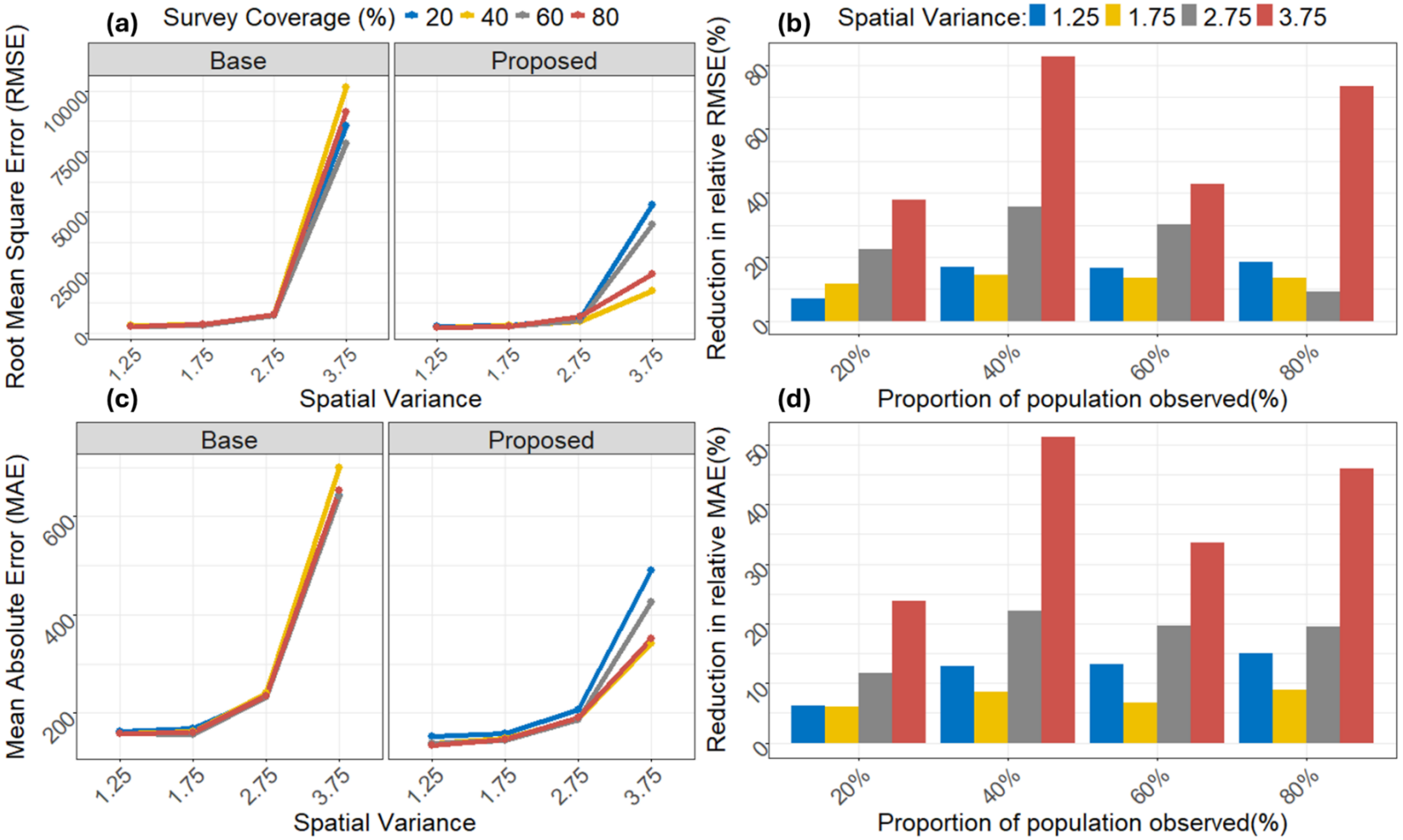

Fig S3. Simulation study results comparing model fit metrics obtained from the 'base' method versus the 'proposed' method over different data missingness-spatial variance combinations, for **equal data source variances**. a) Root Mean Square Error (RMSE); b) Percentage reduction in relative root mean square error (*RRMSE*) produced by the proposed method over the base model; c) Mean Absolute Error (MAE); d) Percentage reduction in relative mean absolute error (*RMAE*) produced by the proposed method over the base model.

Table S4. 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