Efficient Bayesian Hierarchical Population Estimation Modelling using INLA-SPDE: Integrating Multiple Data Sources, Spatial Autocorrelation, and Investigating Spatial Misalignment

Chibuzor Christopher Nnanatu[1,3], Ortis Yankey[1], Anaclet Désiré Dzossa[2], 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]National Institute of Statistics (NIS)- Cameroon

[3]Nnamdi Azikiwe University, Awka-Nigeria

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

SUPPLEMENTARY MATERIAL

Introduction

This supplementary material contains useful information to support the understanding of the main manuscript. This includes the steps, tables and graphs that are necessary for a better understanding of both the methodology and the results in the main paper. These are summarised under three main sections – methodology, simulation study, and application.

S1: Posterior Simulation and grid cell prediction (Methodology)

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.

Motivation

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 , $\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_8)$, and the estimates of the random effects, predict the population density $\widehat{D}_{g,t}$ $(g = 1, \dots, G)$ for each pixel based on equation (14) and (15) of the main manuscript. The random effects values are simulated from their respective posterior means, that is,
   $$\gamma_{gt} \sim Normal(0, \hat{\sigma}_\gamma) \qquad (SM1)$$
   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
   $$\hat{y}_{g,t} = \widehat{D}_{g,t} \times B_g \qquad (SM2)$$
   (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.
   $$\overline{D}_g = \frac{1}{T}\sum_{t=1}^{T} \widehat{D}_{g,t} \qquad (SM3)$$
   (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.

$$\bar{y}_g = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_{g,t} \qquad (SM4)$$

(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 estimates of uncertainties could be obtained as confidence intervals using for example, the lower bound can be calculated as

$$lower^{(g)} = \bar{w}_g - 2\sigma_g \qquad (SM5)$$

while the upper bound is given by

$$upper^{(g)} = \bar{w}_g + 2\sigma_g \qquad (SM6)$$

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

$$\sigma_g = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{w}_{g,t} - \bar{w}_g\right)^2} \qquad (SM7)$$

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

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

$$uncertainty^{(g)} = \frac{upper^{(g)} - lower^{(g)}}{\bar{w}_g} \qquad (SM8)$$

(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,

$$\hat{y}_t = \sum_{g=1}^{G} \hat{y}_{g,t} \qquad (SM9)$$

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

$$\bar{y}_{admin} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)} \qquad (SM10)$$

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

$$\sigma_{admin} = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(\hat{y}_t - \bar{y}_{admin})^2} \qquad (SM11)$$

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,

$$lower = \bar{y}_{admin} - 2\sigma_{admin} \qquad (SM12)$$

while the upper bound is given by

$$upper = \bar{y}_{admin} + 2\sigma_{admin} \qquad (SM13)$$

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 in equations (25) to (30). In addition, a unique measure of $\boldsymbol{uncertainty}$ may be obtained for the lower administrative units as

$$uncertainty = \frac{upper - lower}{\bar{y}_{admin}} \qquad (SM14)$$

Similar to 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.

S2: Simulation Study

Table S2.1. Posterior Estimates of the total population based on $N = 64$ Areal Units, 14400 prediction grid cells (Total observed count = 19118579)

Data Coverage	**Prediction**	**Posterior Estimates of Counts**			
		Mean	*Lower (2.5% quantile)*	*Median (50% quantile)*	*Upper (97.5% quantile)*
100% (full)	Grid to Grid	19,119,453	19,109,819	19,119,471	19,128,861
	Area to Grid	15,243,147	14,278,053	15,083,408	16,886,263
80%	Grid to Grid	19,104,631	19,083,557	19,104,894	19,122,065
	Area to Grid	15,393,810	14,342,845	15,279,095	16,891,035
60%	Grid to Grid	19,073,375	19,052,517	19,074,209	19,095,162
	Area to Grid	15,708,623	13,805,954	15,448,622	18,488,537
40%	Grid to Grid	19,090,994	19,056,995	19,092,608	19,114,628
	Area to Grid	15,493,404	14,125,171	15,372,288	17,472,298
20%	Grid to Grid	19,110,699	19,060,715	19,110,128	19,166,727
	Area to Grid	16,210,147	14,758,077	16,038,636	18,627,232

Note. Grid to Grid predictions is more accurate than that of the Area to Grid predictions.

Table S1.2. Posterior estimates of the total population based on $N = 100$ Areal Units, 14400 prediction grid cells (Total observed count = 10395302

Data Coverage	**Prediction**	**Posterior Estimates of Counts**

		Mean	*Lower (2.5% quantile)*	*Median (50% quantile)*	*Upper (97.5% quantile)*
100% (full)	Grid to Grid	10,402,481	10,395,066	10,402,831	10,409,833
	Area to Grid	8,590,232	8,163,939	8,556,368	9,388,491
80%	Grid to Grid	10,385,503	10,372,516	10,385,487	10,398,788
	Area to Grid	8,675,551	8,340,256	8,622,415	9,266,889
60%	Grid to Grid	10,423,318	10,410,727	10,422,779	10,438,665
	Area to Grid	8,472,879	8,046,329	8,431,964	9,016,358
40%	Grid to Grid	10,443,508	10,422,765	10,444,718	10,462,687
	Area to Grid	8,672,709	8,107,152	8,607,997	9,669,746
20%	Grid to Grid	10,481,319	10,456,043	10,480,259	10,509,090
	Area to Grid	8,990,035	8,680,986	8,944,019	9,476,545

Note. Area to Grid predictions improved slightly and better than when it was 64 area units. While Grid to Grid predictions is more accurate as expected.

Table S2.2. Posterior estimates of the total population based on N=400 Areal Units, 14400 prediction grid cells (Total count = 19957503)

Data Coverage	**Prediction**	**Posterior Estimates of Counts**			
		Mean	*Lower (2.5% quantile)*	*Median (50% quantile)*	*Upper (97.5% quantile)*
100% (full)	Grid to Grid	19,966,091	19,955,140	19,966,378	19,979,000
	Area to Grid	19,523,902	19,296,458	19,509,124	19,864,786
80%	Grid to Grid	19,887,778	19,869,220	19,887,256	19,906,814
	Area to Grid	19,492,079	19,128,334	19,453,845	20,013,142
60%	Grid to Grid	20,001,695	19,973,916	20,002,072	20,022,995
	Area to Grid	19,700,622	19,427,604	19,649,180	20,130,003
40%	Grid to Grid	19,923,207	19,890,887	19,925,108	19,947,416
	Area to Grid	19,049,675	18,862,114	19,017,409	19,318,493
20%	Grid to Grid	20,076,543	20,026,726	20,079,470	20,122,859
	Area to Grid	19,692,229	19,184,861	19,684,399	20,315,293

Note. Grid to Grid performs slightly better than the Area to Grid predictions. However, Area to Grid predictions show a lot more improvements.

Table S2.3. Posterior estimates of the total population based on N=900 Areal Units, 14400 prediction grid cells (Total count = 24515238)

Data Coverage	**Prediction**	**Posterior Estimates of Counts**			
		Mean	*Lower (2.5% quantile)*	*Median (50% quantile)*	*Upper (97.5% quantile)*
100% (full)	Grid to Grid	25,011,751	24,911,323	25,010,218	25,168,242
	Area to Grid	24,293,037	23,952,433	24,240,198	24,800,990
80%	Grid to Grid	25,016,752	24,961,623	25,016,888	25,075,222
	Area to Grid	24,215,112	23,922,103	24,144,284	24,605,412
60%	Grid to Grid	24,798,470	24,736,246	24,799,485	24,857,764
	Area to Grid	24,419,701	23,990,224	24,387,149	25,085,133
40%	Grid to Grid	24,790,742	24,731,747	24,792,190	24,853,478
	Area to Grid	24,331,337	23,891,122	24,300,630	24,839,227
20%	Grid to Grid	24,499,765	24,315,894	24,499,331	24,678,119
	Area to Grid	24,317,293	23,975,376	24,308,843	24,764,599

Note. Both the Grid to Grid and Area to Grid predictions performed almost equally well in predicting the total population.

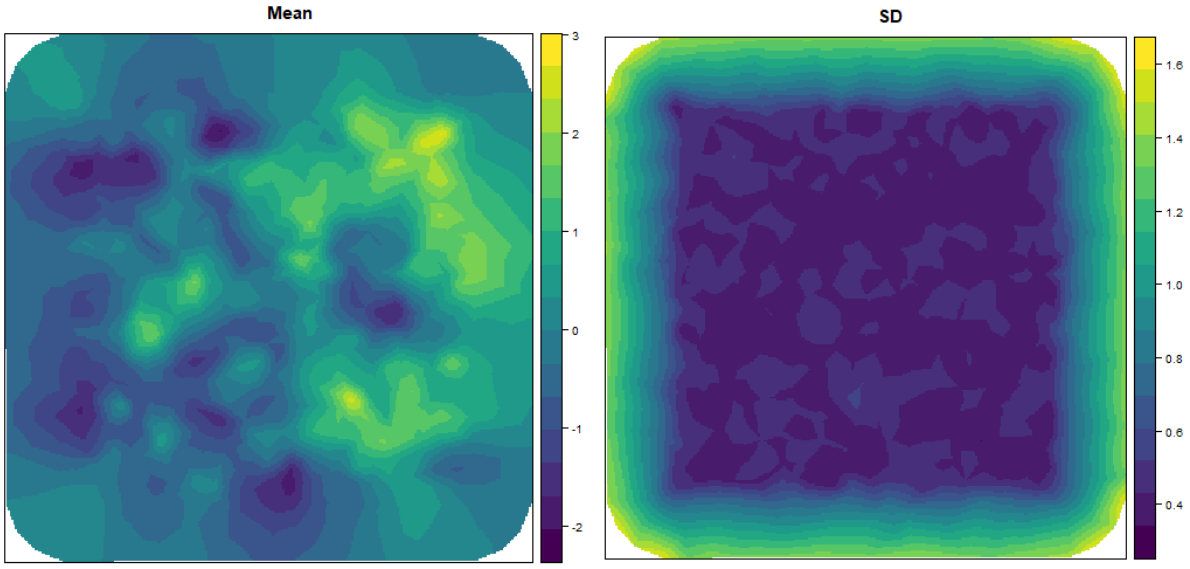

Figure S2.1 Posterior estimates of mean (left) and standard deviation (SD; right) of the spatial random effects for 400 area units. The mean posterior estimates of random effects map (left) clearly show patterns of spatial

clustering. The SD values are moderately low especially in areas with large observations and there were high values of SD within the extended mesh.

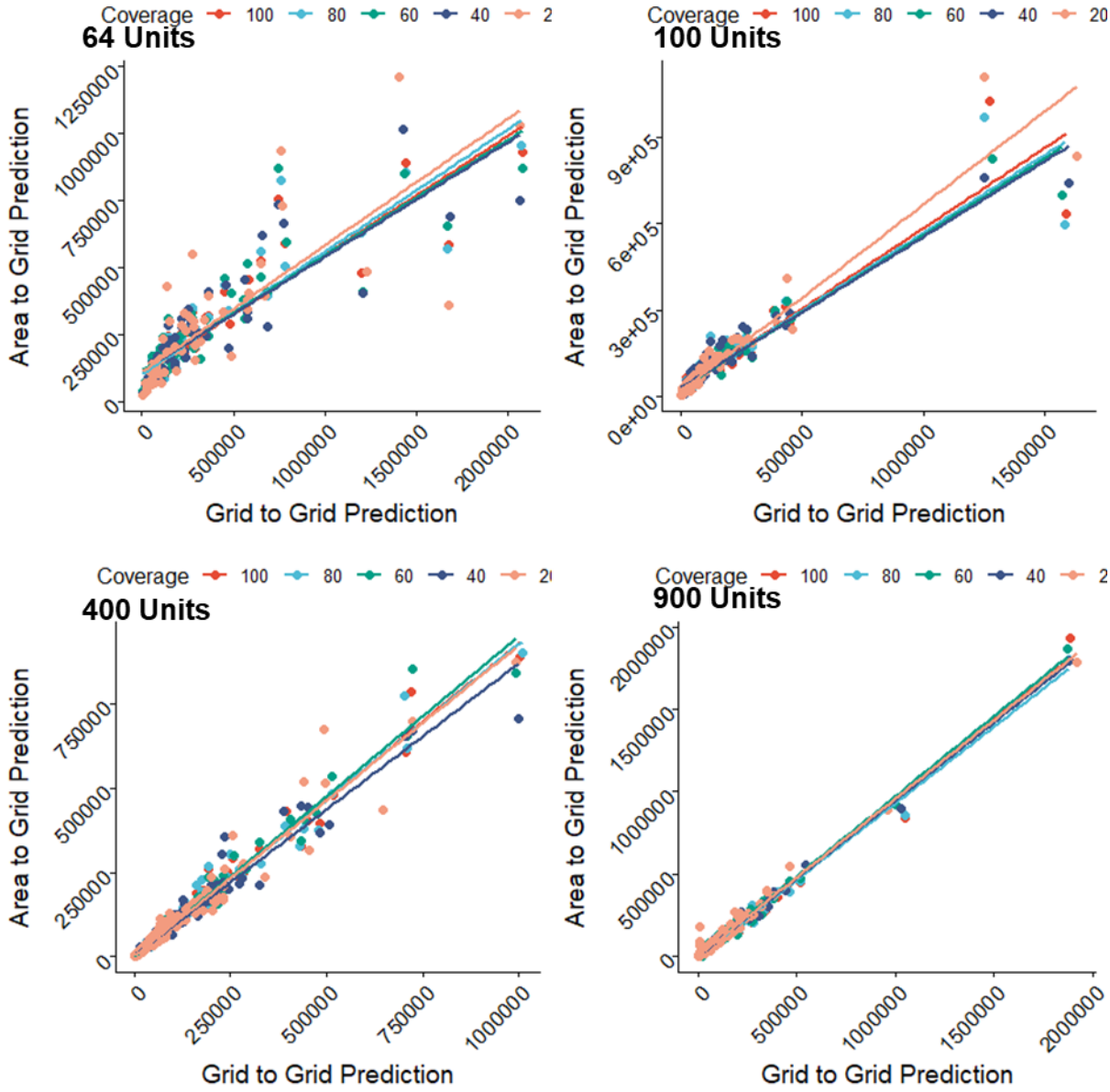

Figure S2.2. Scatter plots of the area units aggregation of the population predictions based on Grid to Grid versus Area to Grid models. Estimates became a lot similar as number of area units increased

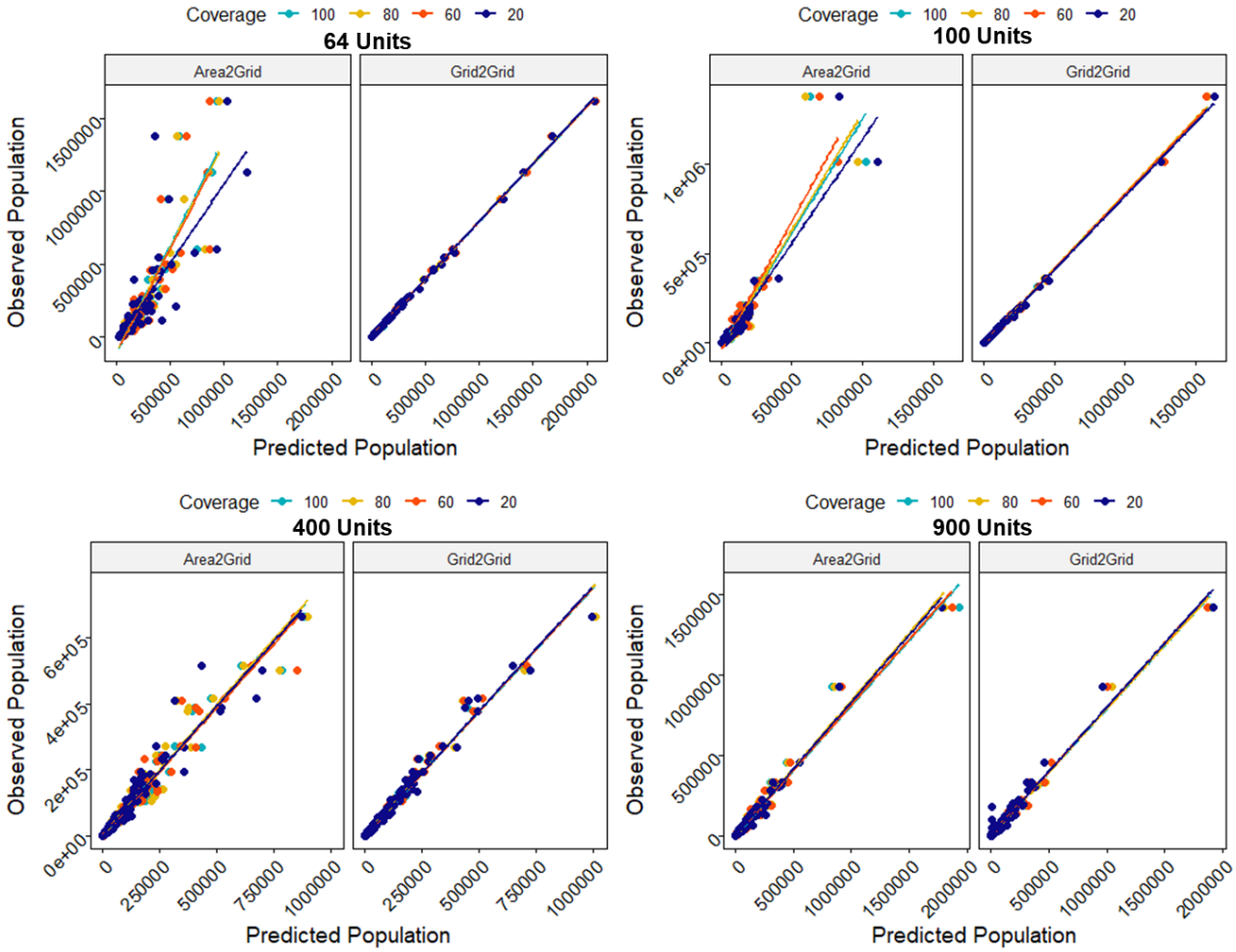

Figure S2.3. Scatter plots of the area units aggregation of the population predictions based on Grid to Grid versus Area to Grid models. Estimates became a lot similar as number of area units increased.

S3: APPLICATION TO CAMEROON HOUSEHOLD LISTING DATASETS

Model fit assessments and cross-validation

Goodness-of-fit

Model fit indices based primarily on the WAIC and CPO show that among the four top models specified in equation (26) of the main manuscript, **Model IV**) which has the lowest WAIC (=1350.10), and the lowest negative sum of log CPO (=5766.08) values provided the best fit to the data (Table 5). This suggests that the variability in Cameroon's population density is well captured by the differences in settlement types as well as their differences across the 10 regions of the country.

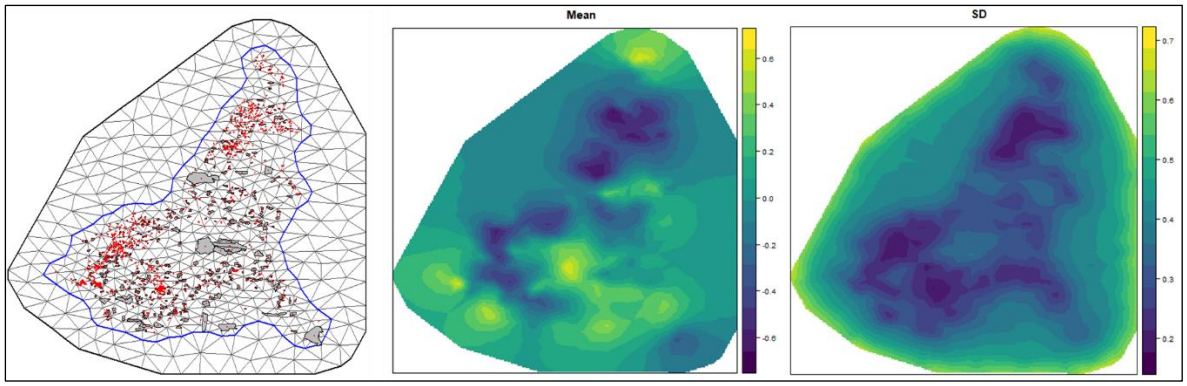

Figure S3.1. Triangulation mesh built for the entire country (left) with the centroids of the EAs with observations in red. The two maps on the right are the mean and standard deviations of the posterior random effects based on the best fir Model (IV). Areas with low or no observations had high variance while variance was low for areas with large observations.

k-fold Cross-validation

Results from the k-fold cross-validation metrics of Model IV indicated that the model predictive ability remained relatively stable and consistent across the $k\ (=5)$-folds used as the test set (Table S6). In particular, the correlation between the observations and the predicted counts remained high at $\geq 98\%$ across the folds. Also, estimates of percentage coverage $\%\mathrm{INCI}$ indicated that at least 99.96% of the observations are within the 95% credible interval of the predicted counts.

Table S3.1. Output metrics from k-fold cross-validation

Metrics	**Fold_1**	**Fold_2**	**Fold_3**	**Fold_4**	**Fold_5**	**Average**
MAE	131.90	140.41	122.93	163.73	153.13	142.42
RMSE	287.32	220.81	180.39	231.25	234.75	230.90
%INCI	100.00	99.78	100.00	100.00	100.00	99.96
Correlation (r)	0.99	0.98	0.98	0.98	0.98	0.98

Figure S3.2 compares the distribution of the posterior estimates across the 5-folds cross-validation sets. The violin plots with embedded boxplots in Figure S7show very similar patterns across the datasets indicating model estimation stability. This is supported by the scatterplot in Figure 11 of the main manuscript which shows similar correlation patterns between the observed and predicted values across the folds.

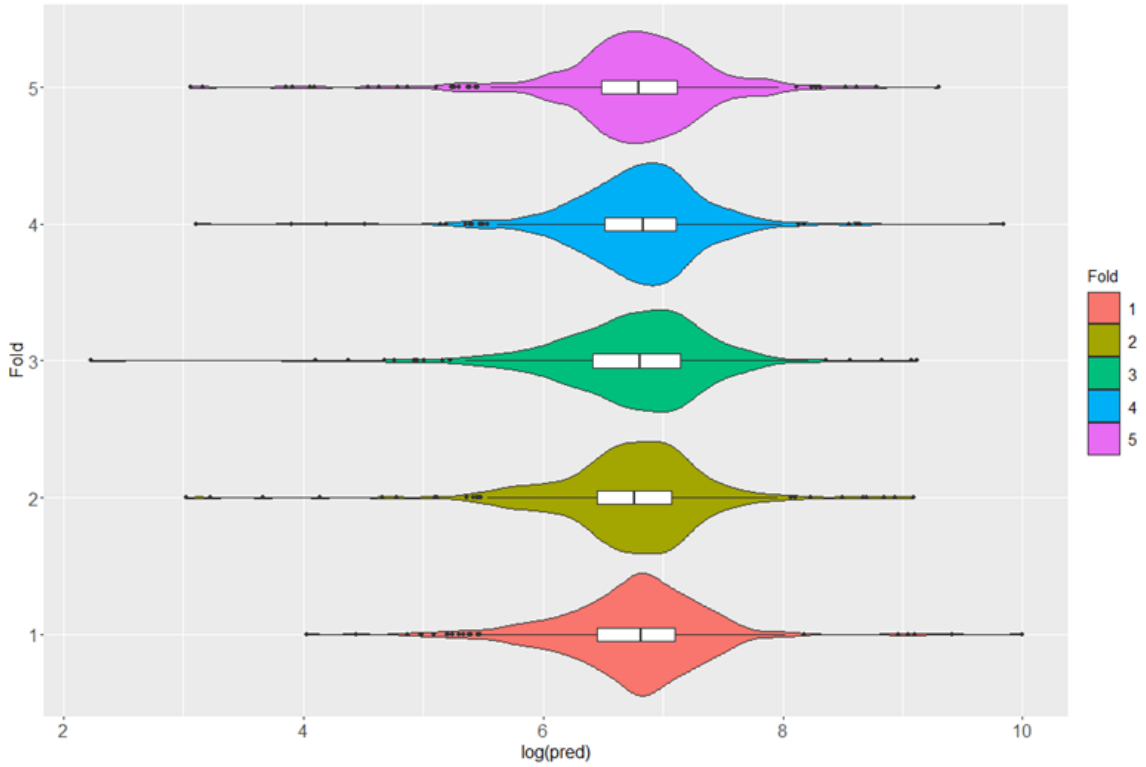

Figure S3.2. Violin plots of the posterior estimates of population across the 5-folds cross validation. Estimates were quite similar throughout indicating that the model is robust across the entire study areas.