Disagregating Census Data for Population Mapping Using a Bayesian Additive Regression Tree Model

Ortis Yankey[1], Chigozie E. Utazi[1], Christopher C. Nnanatu[1], Assane N. Gadiaga[1], Thomas Abbot[1], Attila N. Lazar[1], Andrew J. Tatem[1]
[1]University of Southampton, Worldpop Research Group, Highfield, Southampton, SO17 1BJ

INTRODUCTION

Population data is crucial for policy decisions, but fine-scale population numbers are often lacking due to the challenge of sharing sensitive data. To address population data needs at small geographic scales, Top-down population modelling approach has been one of the methods used to disaggregate census data to small area scales. Top-down population mapping is a census-dependent process that uses census data to estimate and redistribute population numbers from larger administrative areas, such as provinces or districts, to smaller area units, typically at 100 m resolution (McKeen et al., 2023; Tatem, 2022).

Dasymetric population mapping (Stevens et al., 2015) involving the use of the Random Forest (RF) approach has been widely used in top-down population disaggregation. This involves combining ancillary geospatial covariates with observed population data to produce a weight layer that is used to disaggregate the population totals into grid cells using the RF model.

The RF model's major limitation is its inability to quantify the uncertainties associated with the predicted populations. Estimates of predicted population uncertainty are useful in addressing inherent biases and variability in population estimates and offer a degree of confidence in the estimated population, which can be useful for policy decisions.

In this study, we used a new approach using a Bayesian Additive Regression Tree (BART) to disaggregate population data and quantify the predicted population's uncertainties. We also compared the BART model with the RF model to determine the relative performance of the two models. The study adopted a simulation study involving both methods and also used both methods to disaggregate the 2021 census data for Ghana to compare the two approaches.

METHOD

Data Source

We obtained population census data from Ghana's recent national population census in 2021 from the Ghana Statistical Service. For administrative purposes, Ghana is divided into 261 districts, and the population data was aggregated at these districts. The modelling also included 24 geospatial covariates associated with population distribution

Procedure

- The process of disaggregating the population into small areas involved the following steps:
- The objective is to predict population density, which is used as a weighting layer. To obtain population density, we divided the observed population count for a given administrative unit by its total area, which serves as a response variable.
- We then log-transformed the response variable and combined it with district-level geospatial covariates to fit two separate models using the RF approach and the BART approach.
- Both models were subsequently used to make a prediction on a set of geospatial covariates at 100 m to obtain a predicted population density, which was used as a weighting layer to disaggregate the 2021 total population at 100 m gridcells.

Simulation Study

- We also conducted a simulation study to investigate the predictive performance of the RF model and the BART model.
- We used a regression tree model to simulate pixel-level population count. We aggregated the simulated pixel population to the district level to obtain the simulated total population at district level
- We then used both BART and the RF models to disaggregate the simulated district population back to the pixel level.
- We measured the predictive performance of both models by comparing the simulated pixel population with the disaggregated population.
- We calculated a variety of model metrics, including correlation, root mean square error (RMSE), bias, and mean square error (MSE), to assess the performance of both models.

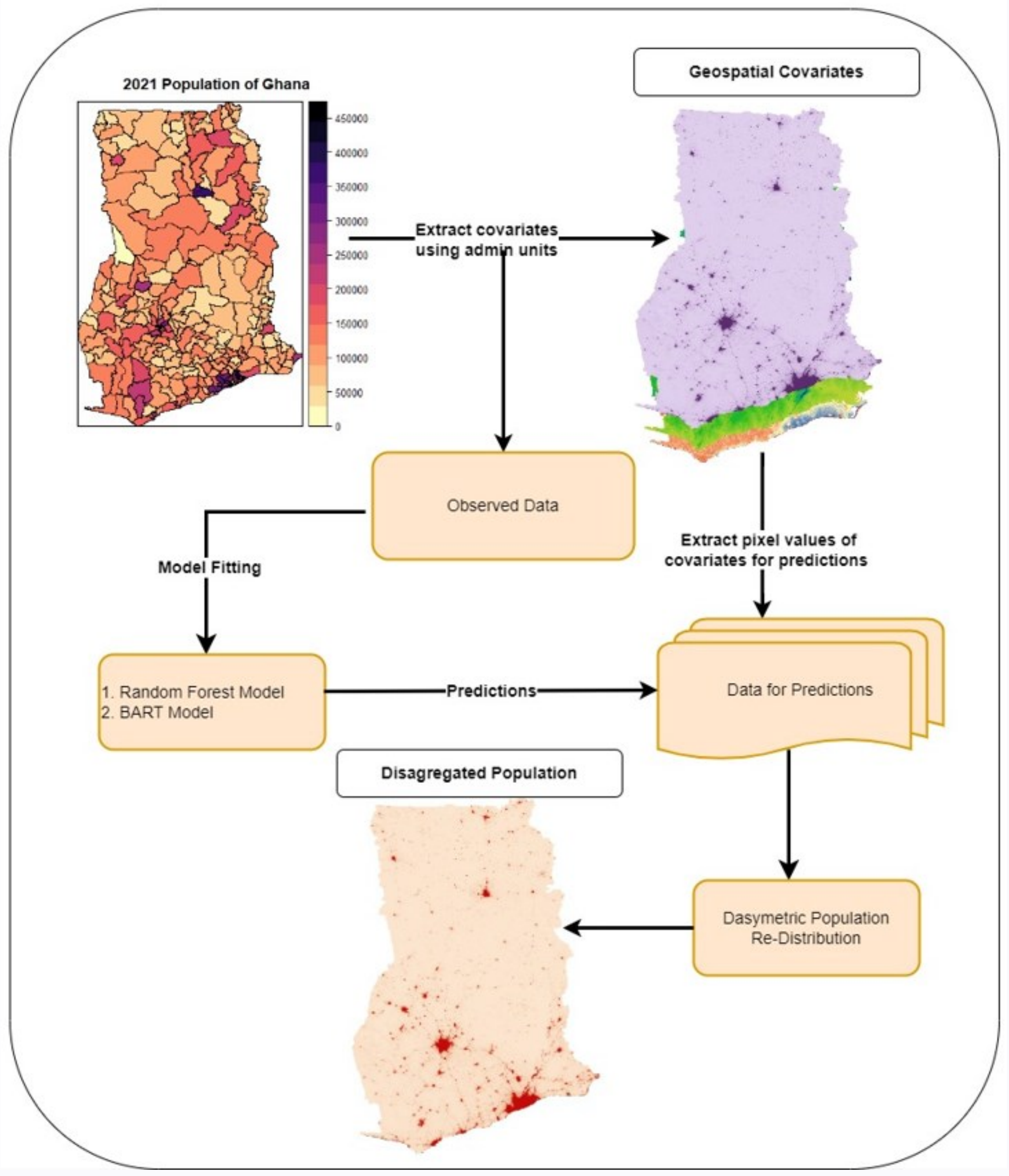


Figure 1: Illustrating the process of geospatial covariate processing and model fitting

RESULTS

Simulation Study

Table 1. Goodness of fit metrics of simulated data

Models	Predictions	Bias	Imprecision	MSE	RMSE	Pearson r	R^2	% Coverage
Random-Forest	In-sample (district)	-0.04	0.17	0.03	0.17	0.93	0.96	
	Out-of-sample (district)	-0.06	0.28	0.08	0.28	0.86		
	Pixel-Predictions	0.00	28.7	826	28.7	0.66		
BART	In-sample (district)	-0.003	0.05	0.003	0.05	0.99	0.99	99.45
	Out-of sample (district)	0.002	0.02	0	0.02	0.99		93.59
	Pixel Predictions	0.00	22.44	503.42	22.44	0.81		

- From the simulation study, the BART model performed better than the RF model across all model metrics.
- The $\mathbf{R^2}$ for the BART model was almost 100%, whereas the RF model was 96%.
- The correlation between the disaggregated pixel level population count and the simulated pixel level population count was 0.66 for the RF model compared to 0.81 in the BART model.
- Various other evaluation metrics, such as imprecision, MSE, and RMSE, exhibited substantially lower values in the BART model in comparison to the RF model.

Ghana 2021 Population Disaggregation

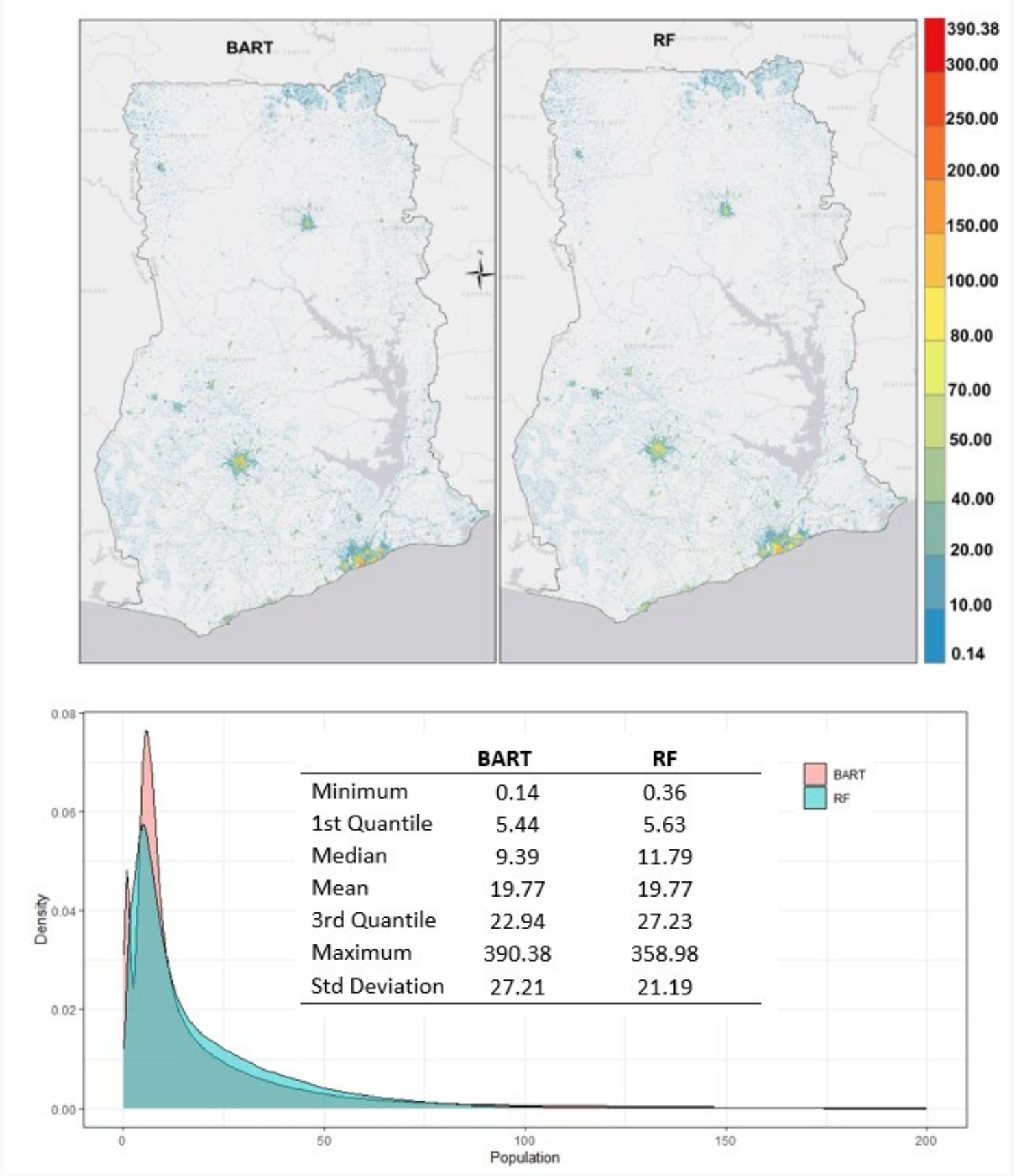
Table 2. Goodness of Fit Metrics of 2021 Population Census Disaggregation

Models	Predictions	Bias	Imprecision	MSE	RMSE	Pearson r	R^2	% Coverage
RF	In-sample (district)	-0.04	0.23	0.05	0.23	0.85	0.96	
	Out-of-sample (district)	-0.03	0.15	0.02	0.15	0.92		
BART	In-sample (district)	-0.01	0.07	0.04	0.07	0.99	0.998	98.91
	Out-of-sample (district)	-0.01	0.05	0.002	0.05	0.96		92.31

Table 2 provides model metrics for disaggregating Ghana's 2021 census data using the BART model and the RF model.

- We observed that the BART model outperforms the RF model in disaggregating the census data. The RMSE, MSE and Bias were lower in the BART model compared to the RF model.

Spatial Distribution of Disaggregated Population and Uncertainty Quantification



	BART	RF
Minimum	0.14	0.36
1st Quantile	5.44	5.63
Median	9.39	11.79
Mean	19.77	19.77
3rd Quantile	22.94	27.23
Maximum	390.38	358.98
Std Deviation	27.21	21.19

Figure 2: Spatial distribution of disaggregated population

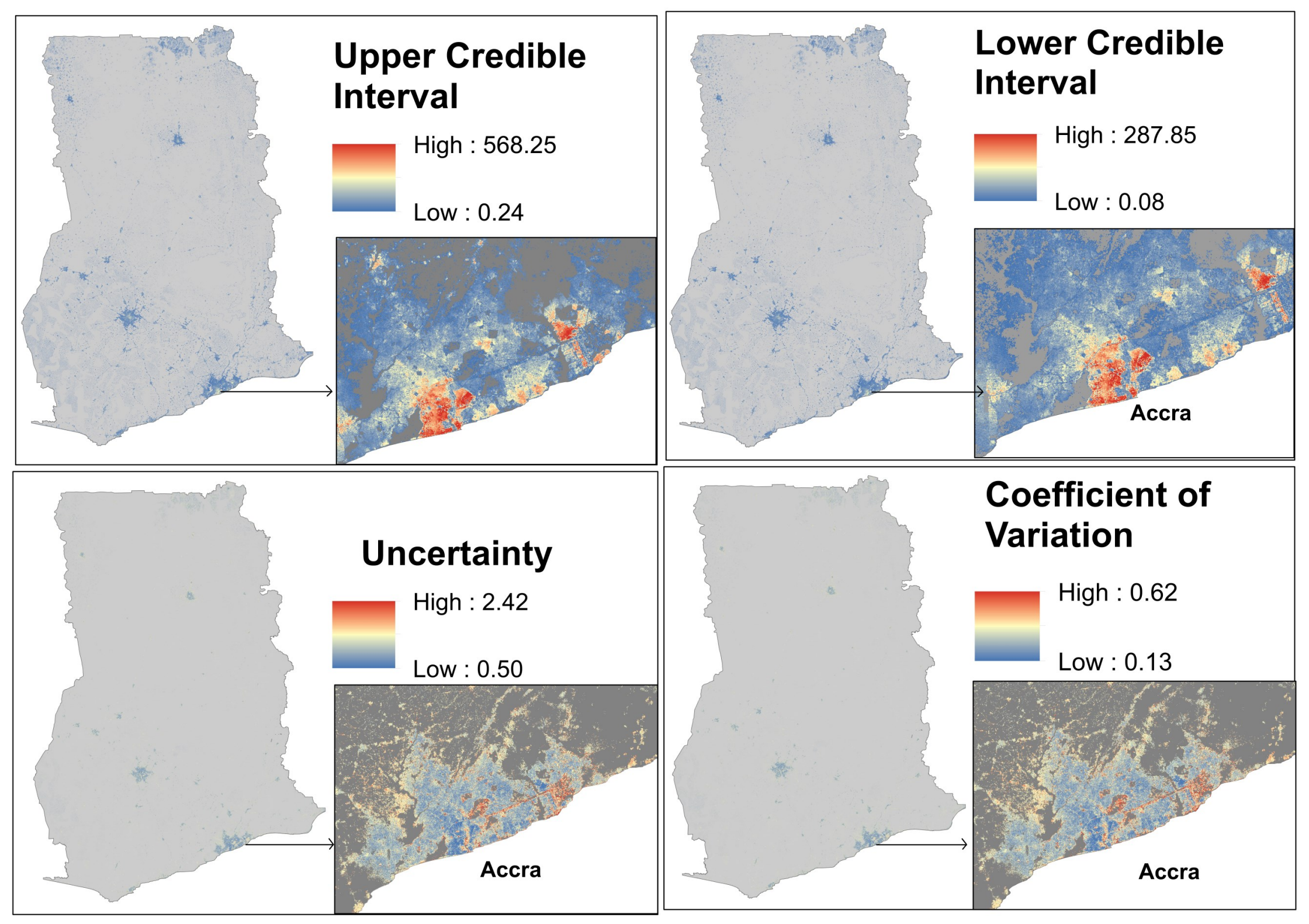


Figure 3: Uncertainty estimate from the BART model

- Figure 2 depicts the disaggregated population's spatial distribution based on the BART and RF models.
- Both models have similar spatial patterns; however, the BART model has a wide range of values compared to the RF model.
- One significant advantage of the BART model over the RF model is its ability to predict credible intervals (CIs), which provide estimates for both the lower and upper bounds of population counts from the disaggregation process.
- In Figure 3, we present the lower and upper credible intervals, the uncertainty, and the coefficient of variation for both Ghana and its capital city, Accra.
- The coefficient of variation for most of the grid cells is less than 0.2, suggesting that there is relatively low variability around the mean predicted population count.
- In conclusion, this study compares the relative performance of a new approach to population disaggregation using a BART model with an already existing approach, the RF model. Model performance was better in the BART model compared to the RF model.

To read the full paper, kindly scan this QR code

REFERENCES

Tatem, A. (2022). Small area population denominators for improved disease surveillance and response. Epidemics, 41, 100641

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS One*, *10*(2), e0107042

McKeen, T., Bondarenko, M., Kerr, D., Esch, T., Marconcini, M., Palacios-Lopez, D., Zeidler, J., Valle, R.C., Juran, S., Tatem, A.J. and Sorichetta, A., 2023. High-resolution gridded population datasets for Latin America and the Caribbean using official statistics. *Scientific Data*, *10*(1), p.436.