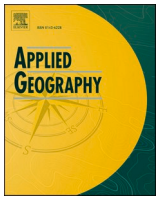
Applied Geography 172 (2024) 103416

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog



Disaggregating census data for population mapping using a Bayesian Additive Regression Tree model

Ortis Yankey *, Chigozie E. Utazi, Christopher C. Nnanatu, Assane N. Gadiaga, Thomas Abbot, Attila N. Lazar, Andrew J. Tatem

University of Southampton, Worldpop Research Group, Highfield, Southampton, SO17 1BJ, UK

ARTICLE INFO

Handling Editor: Dr. Y.D. Wei

Keywords:
Population modelling
Bayesian dasymetric population
Ghana
Population disaggregation
Bayesian Additive Regression Tree
Random forest
WorldPop

ABSTRACT

Population data is crucial for policy decisions, but fine-scale population numbers are often lacking due to the challenge of sharing sensitive data. Different approaches, such as the use of the Random Forest (RF) model, have been used to disaggregate census data from higher administrative units to small area scales. A major limitation of the RF model is its inability to quantify the uncertainties associated with the predicted populations, which can be important for policy decisions. In this study, we applied a Bayesian Additive Regression Tree (BART) model for population disaggregation and compared the result with a RF model using both simulated data and the 2021 census data for Ghana. The BART model consistently outperforms the RF model in out-of-sample predictions for all metrics, such as bias, mean squared error (MSE), and root mean squared error (RMSE). The BART model also addresses the limitations of the RF model by providing uncertainty estimates around the predicted population, which is often lacking with the RF model. Overall, the study demonstrates the superiority of the BART model over the RF model in disaggregating population data and highlights its potential for gridded population estimates.

1. Introduction

Population figures at small area scales are of great importance to policymakers since they offer crucial insights on the magnitude, structure, spatial distribution, and temporal changes of a nation's populace. The availability of spatially precise population data plays a crucial role in shaping policies pertaining to various domains such as humanitarian and disaster response, healthcare provisions, disease surveillance, resource allocation, monitoring development indicators, and evaluating economic growth (Leasure et al., 2023; Linard et al., 2010; Tatem, 2022; Tuholske et al., 2021; Utazi et al., 2018). For example, during natural disasters such as earthquakes, floods, or pandemics, accurate population data becomes critical for emergency response and disaster relief efforts. Population figures help identify the number of people affected, assess their needs, and plan rescue operations, evacuation measures, and the distribution of essential supplies like food, water, and medical aid. This data also aids in identifying vulnerable populations, such as the elderly, children, or people with disabilities, who may require specific assistance (Holt et al., 2019; Tenerelli et al., 2015; UN-SPIDER, 2023; UNFPA, 2020).

The most accurate, spatially detailed, and reliable source of population information is typically the national population and housing census, which are usually conducted at 10-year intervals within countries. Such exercises gather information on several aspects of a country's population characteristics, such as age, sex, race, and occupation that are useful for a country's decision-making processes. Nevertheless, despite the valuable insights that censuses offer regarding population characteristics, several challenges hinder the ability of certain low- and middle-income countries, particularly those in Sub-Saharan Africa (SSA), to carry out regular decennial censuses. These challenges include significant financial burdens, political instability, and conflict, as well as logistical and implementation deficiencies (Olorunfemi & Fashagba, 2021; Skinner, 2018).

Population projections have been one of the means to address data gaps during the intercensal period. However, population projections are saddled with a lot of uncertainties, particularly demographic changes arising from epidemics, conflict, and migration patterns during the intercensal period, which can lead to inaccuracies in the projected figures (O'Sullivan, 2023; Park & LaFrombois, 2019). Coupled with such constraints is the fact that projected estimates are produced at a coarse

* Corresponding author.
E-mail address: O.Yankey@soton.ac.uk (O. Yankey).

https://doi.org/10.1016/j.apgeog.2024.103416
Received 15 January 2024; Received in revised form 6 September 2024; Accepted 7 September 2024
Available online 14 September 2024

administrative level, such as the country or regional level, and are lacking at small area levels because of the challenge of sharing population data at sensitive small area scales (Skinner, 2018). Integrating data collected at coarser administrative levels with other forms of data, such as health facility or catchment area data, for small-scale estimations then becomes a challenge. Lack of population data at a granular level, such as enumeration areas, towns, and sub-districts, means that we are unable to make accurate and reliable population decisions at these levels.

To address population data needs at small geographic scales, there are now several global and continental gridded population datasets that are based on different modelling approaches and input data layers (Leyk et al., 2019). Some of these global population datasets include the WorldPop population datasets (Gaughan et al., 2016; McKeen et al., 2023; Sorichetta et al., 2015; Tatem, 2017) produced by the WorldPop Research Group at the University of Southampton and the Gridded Population of the World version 4 (GPW4) produced by the Center for International Earth Science Information Network (CIESIN) at the University of Columbia (CIESIN, 2018). The Global Human Settlement Layer-Population produced by the European Commission Joint Research Center and CIESIN (Florczyk et al., 2019), the LandScan Global Population Database (LandScan Global) produced by Oak Ridge National Laboratory (Sims et al., 2023) and the World Population Estimation (WPE) (Nordstrand & Frye, 2014) produced by the Environmental Systems Research Institution (ESRI), as well as many other human settlement products defining the density or distribution of human population globally.

The method for population mapping for producing these datasets involves the redistribution of population numbers from a spatially coarse administrative unit scale to target grid cells. This form of population redistribution has involved diverse methodological approaches to population redistribution. The most basic form of these methods is the areal weighting approach (Eicher & Brewer, 2001), in which populations are spread out evenly across grid cells without taking into account the need for geospatial covariates or other data that could help explain differences in how populations are redistributed. In response to this limitation, dasymetric population mapping approaches (Mennis, 2003; Mennis & Hultgren, 2006) have emerged. This approach utilizes ancillary geospatial covariates to inform the redistribution of the population to target grid cells. These geospatial covariates are used to create a weighting layer for population redistribution from coarse to fine spatial scales. There are different types of dasymetric mapping techniques used, ranging from the simple binary dasymetric approach (Eicher & Brewer, 2001) to more advanced ones like intelligent dasymetric mapping (Mennis & Hultgren, 2006) and even hybrid dasymetric approaches that use statistical and machine learning techniques like random forest to create a weighting layer that is used for population disaggregation (Stevens et al., 2015).

The global and gridded population datasets are based on some of these approaches. For example, GPW v4.0 uses an areal weighting approach, whereas the WPE uses a dasymetric approach using a likelihood surface derived from geospatial covariates (Nordstrand & Frye, 2014). In contrast, WorldPop global population data products use weighted dasymetric disaggregation using random forests (McKeen et al., 2023; Sorichetta et al., 2015). Recently, a new approach called POMELO (Population Mapping by Estimation of Local Occupancy Rates), a deep learning approach by Metzger et al. (2022), demonstrates efficacy in disaggregating population data and has been used to disaggregate population data in Tanzania, Zambia, and Mozambique.

While a wide variety of methods and products are available for population disaggregation at the small area level, a major limitation of these methods is their inability to quantify the uncertainty around the predicted population. These models offer only a single point estimate representing the population count for the target cell, without including the confidence interval associated with such estimates. Estimates of predicted population uncertainty are useful in addressing inherent biases and variability in population estimates that may arise because of biases in data input. Inherent biases and variability in the data used to fit a population model are not eliminated but they are propagated into the final prediction estimates and as such, the final predicted population estimates should ideally have a quantified level of uncertainty associated with the predicted population so that users of such data may become aware of the possible range of values for the expected population. For instance, in the event of a natural disaster, the absence of uncertainty intervals for the estimated population could hinder the ability to accurately determine the minimum or maximum number of people who may require resources.

Another issue in population modelling, particularly with top-down estimations, is the lack of validation of the predicted population at the grid cell level using external datasets, specifically precise ground data. Due to the lack of detailed population data to make such a comparison, countries lack a ground truthing assessment of whether population numbers at the grid cell level accurately reflect actual population numbers on the ground. Even in advanced economies, only a few high-income countries, such as Northern Ireland (Martin et al., 2011), Sweden (Archila Bustos et al., 2020), Poland and Portugal (Calka & Bielecka, 2020), as well as some countries in Mainland Southeast Asia (Yin et al., 2021) and Bioko Island in Equatorial Guinea (Fries et al., 2021) have had studies validating gridded population estimates with actual population numbers obtained from their national statistical offices. The absence of detailed data in the majority of low-income settings to enable validation means that we are unable to assess the level of biases introduced in the predictions.

In summary, two of the main limitations associated with top-down population disaggregation are: (1) our inability to quantify the uncertainties around the predictions; and (2) our inability to validate the gridded population numbers with actual observed ground data at the grid cell due to the absence of such data. These two issues have been unexplored in top-down gridded population modelling. The objectives of this work are therefore to:

1. Apply a Bayesian Additive Regression Tree (BART) approach to disaggregating population totals from a higher administrative unit to the gridcell level and to estimate the uncertainties (upper and lower credible intervals) associated with the predictions. This objective will involve a comparative assessment of the RF approach, which is used for top-down population disaggregation, and the BART approach.
2. Conduct a simulation study by simulating a “true” population at the pixel level, aggregating it to a higher administrative level, and then applying both BART and RF approaches to disaggregate the aggregated population back to the pixel level. We will compare the disaggregated gridcell population with the simulated “true” population to evaluate the relative performance of the RF Model and the BART approach. The purpose of the simulation study is to help understand the strengths and limitations of both models.

2. Methods

2.1. Data sources and processing

2.1.1. The 2021 national population census for Ghana

Data were obtained from recent national population census of Ghana, conducted in 2021 (Ghana Statistical Service, 2022). The primary objective of this census was to enumerate every individual, gather information on households, and compile the demographic and socio-economic characteristics of the country’s population. According to the Ghana Statistical Service (2022), the total population of Ghana in 2021 was 30,832,019 (30.8 million) people. For our analysis, we obtained population totals at the district level from the Ghana Statistical Service, which corresponds to administrative level 2. In 2021, Ghana was divided into 16 regions, further subdivided into 261 local districts for

administrative purposes. District boundary shapefiles was also provided by the same agency. In our analysis, we will seek to disaggregate the district-level population counts to 100m pixels (see Fig. 1).

2.2. Geospatial covariates data

We considered a diverse range of geospatial covariates, which are either directly or indirectly related to population distribution across the country. The geospatial covariates include land use and land cover data, climate variables such as temperature and rainfall, physical features and infrastructure such as roads and schools, and settlement data such as building footprints. We subjected all the potential geospatial covariates to covariate selection by assessing the relationship between log-population densities (people/area) and these geospatial covariates using a scatterplot and Pearson correlation coefficient. This process enabled us to retain 24 geospatial covariates with the strongest linear association with population densities. A detailed list of these 24 geospatial covariates and their sources is included in the supplementary materials. These geospatial covariates were also selected based on their availability across the entire study location and whether they can be mapped accurately as a high-resolution geospatial layer. In cases where the original data were not in raster format, we processed them into raster files at a resolution of 100m. These geospatial covariates have been previously linked to predicting population distribution in existing studies (Leyk et al., 2019; Lloyd et al., 2017, 2019; Nieves et al., 2017), thereby justifying their inclusion in our current research. To maintain consistency and alignment, we projected all geospatial covariates using the same spatial reference system. We then processed and stacked these covariates together into a single file using the R software (R Core Team, 2020). Subsequently, we conducted a zonal statistics operation by extracting observed geospatial covariate values for each district using their mean values. The schema in Fig. 2 illustrates the process of geospatial covariate processing and model fitting. The geospatial covariate used in this study have been mapped and can be found in the supplementary materials.

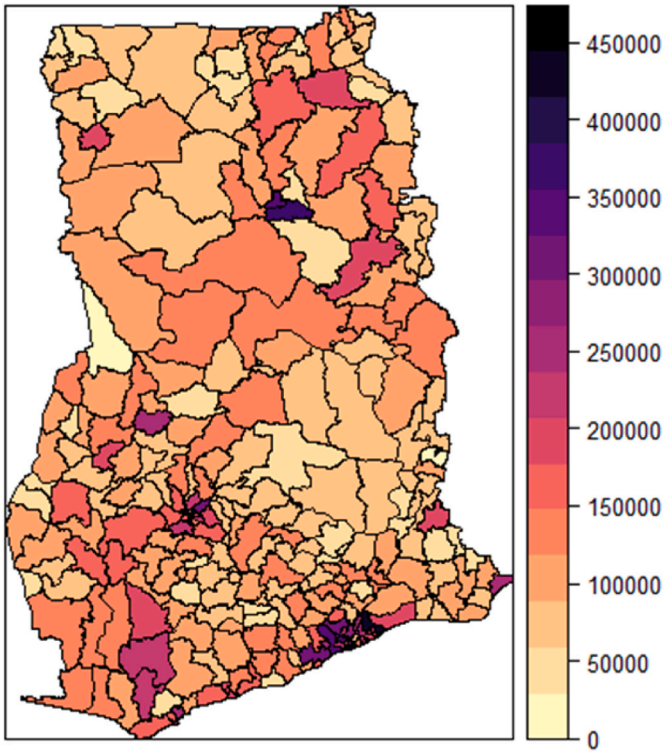


Fig. 1. The spatial distribution of the population at the district level in Ghana from the 2021 national population and housing census.

3. Method

3.1. Dasymetric population disaggregation

The dasymetric population disaggregation modelling approach involves the process of fitting a model to estimate predicted population density, which is used as a weighting layer. This weighting layer is then utilized to redistribute observed population data from a larger administrative unit to target gridcells or small area. Various weighting procedures have been employed for the purpose of disaggregating the overall population (Doxsey-Whitfield et al., 2015; Mennis, 2003; Mennis & Hultgren, 2006). One of the approaches to the population disaggregation has been the use of a Random Forest (RF) Model. The utilization of this approach was initially introduced by Stevens et al. (2015) and subsequently, numerous studies have embraced and expanded upon this strategy (Gaughan et al., 2016; McKeen et al., 2023; Stevens et al., 2020). The use of the RF model have been found to have a higher degree of accuracy in disaggregating population totals (Tatem, 2022)

The process of disaggregating the population into small areas involves a three-step methodology. The objective is to predict population density, which is used as a weighting layer. To achieve this, a random-forest model is employed, with population density serving as the response variable and the geospatial covariates acting as the predictors. The estimation of the response variable, population density, involves dividing the entire population within a specific administrative unit by the area of that unit. The response variable is then log-transformed and combined with district-level geospatial covariates to fit the model. In the second stage, the fitted model is subsequently used to make a prediction on a set of geospatial covariates at 100m to obtain a predicted population density. The predicted population density of a specific administrative unit is aggregated to obtain the total predicted density for that administrative unit. For a given grid cell, the predicted population density for that gridcell is divided by the aggregated (total) predicted density for the administrative unit in which the gridcell is located to obtain a weight for that grid cell given as

$$w_i = \frac{exp(\hat{y}_i)}{\sum_{i=1}^{n} exp\ (\hat{y}_i)} \tag{1}$$

where w_i is the weight for the ith gridcell, $exp(\hat{y}_i)$ is the exponentiated predicted density and n represent the total number of grid cells in each administrative unit.

In the final stage, the predicted population of a particular target gridcell or small area such as an enumeration area is estimated by multiplying the weight w_i by the observed total population for that administrative unit given as:

$$Population_i = total\ population_j \times w_i \tag{2}$$

where j represent the j th administrative unit. This ensures that the population is disaggregated proportionally to the population density within the admin unit. The population disaggregation is constraint to only settled pixels i.e., pixels with building count information which depict locations of human settlement across the landscape. In this study, the disaggregation of the population was done using both Random Forest Model and Bayesian Addictive Regression Tree Model in R.

3.2. Random-Forest (RF) model

A RF is an ensemble machine learning algorithm that can be used for both classification and regression models (Breiman, 2001). RF is built from a set of decision trees to form a "forest" and the predictions from the individual decision trees are combined to obtain a final prediction. By combining multiple decision trees to make the prediction, a RF model reduces the problem of overfitting and underfitting, which can be an issue in decision tree models (Hastie et al., 2009). A RF model works by

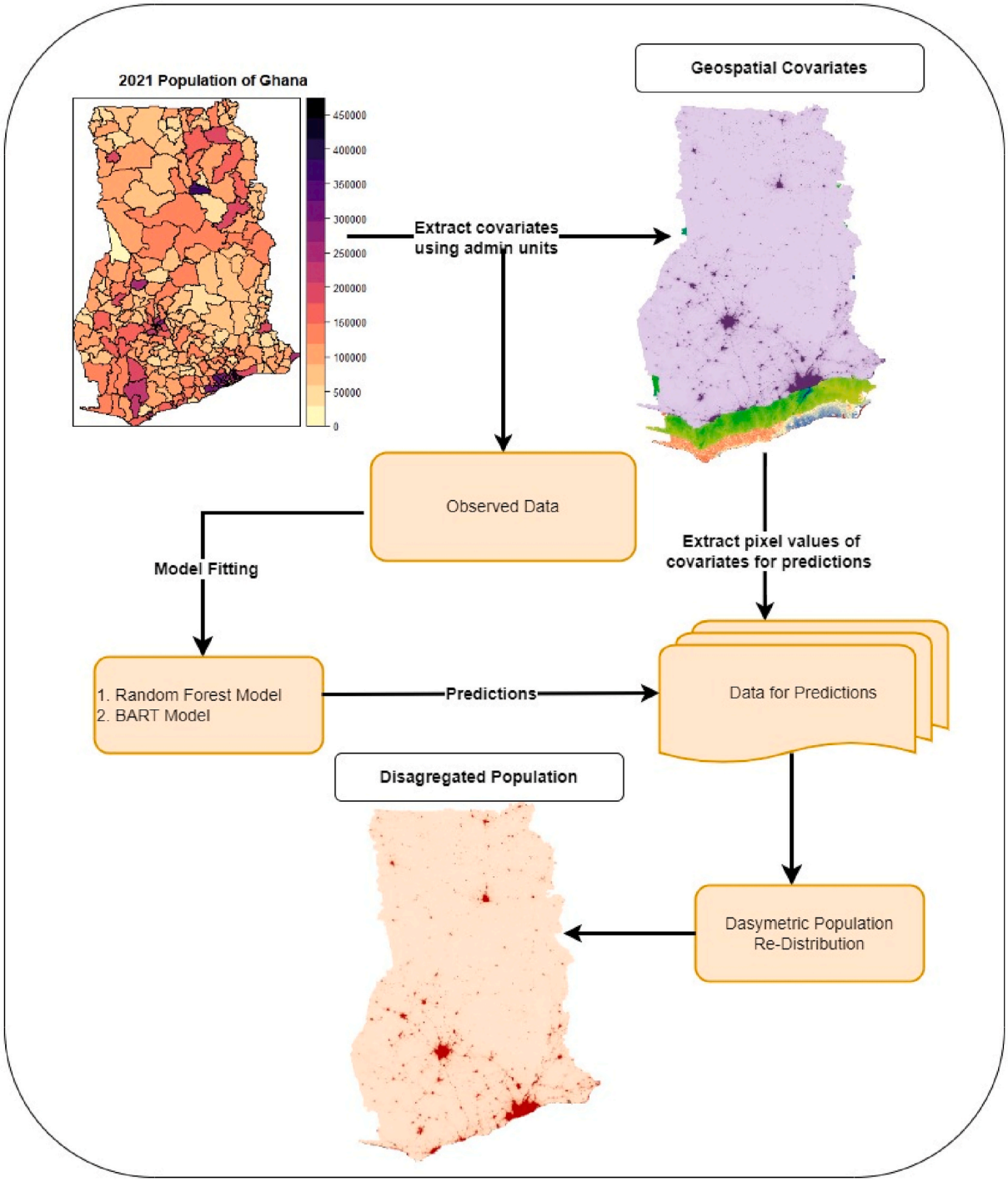


Fig. 2. Illustration of the process of geospatial covariate processing and model fitting.

creating a bootstrapped dataset, which involves selecting a random subset of the dataset (with replacement) to create multiple decision trees. Each tree is trained on a different subset of the data in a process known as bagging, with each tree predicting the outcome variable based on the selected features. This ensures diversity in the decision trees. The predicted value based on the decision trees is combined with the final predictions summarized using their median or mean value. The RF model can then be used to make predictions based on new data points. In this study, we used the randomForest package (Liaw & Wiener, 2002) in R for the population disaggregation. We fine-tuned the hyperparameters of the RF model using the tuneRF function from the randomForest package to select the optimal mtry parameter, which represents the number of features randomly sampled at each split. The tuneRF function employs a local search method to find the best value for mtry by iteratively adjusting it and evaluating the out-of-bag (OOB) error. Through this process, an mtry of 16 was found to yield the lowest OOB error and was consequently used in the final model.

3.3. *Bayesian Addictive Regression Tree (BART)*

A BART (Chipman et al., 2010) combines Bayesian statistical techniques and ensemble learning to create a flexible and more powerful predictive model. The underlying assumption of a BART model is based primarily on a Bayesian probability model, unlike a RF model which is

largely driven by the algorithm. Because the BART model is Bayesian, the BART model consists of a set of priors for the structure and the leaf parameters and a likelihood for the data in the terminal nodes. The aim of the priors is to provide regularization, preventing any single regression tree from dominating the total fit (Kapelner & Bleich, 2013). Because BART is Bayesian, predicted values from each tree, known as posteriors, is not a point estimate but rather a probability distribution for each prediction, which represents the uncertainty estimate from the predictions. In this work, we used the bartMachine R package (Kapelner & Bleich, 2013) and used the default priors in this package, which provide good performance (Kapelner and Bleich, 2013). We used the bartMachineCV function for hyperparameter tuning, specifically for the number of trees, burn-in iterations, and post-burn-in iterations, to select the optimal values for the final modeling. The bartMachineCV function employs a cross-validation approach to determine the best hyperparameters. BART is based on a Markov Chain Monte Carlo (MCMC) simulation where each posterior prediction is based on a previous prediction in a sequence. To derive the predicted population at the pixel-level and account for uncertainty, we obtained 1000 simulated posteriors, representing predicted population densities after the model had converged. These posteriors were used to generate a weighting layer, and each weight was subsequently multiplied by the observed population, as specified in equation (2). The predicted population for each pixel was summarized using their mean, while measures of uncertainty were obtained based on the lower (2.5th quantile) and upper (97.5th quantile) credible intervals.

3.4. Model validation

Both the RF model and the BART model were used in separate analyses to predict population density (weighting layer) based on the geospatial covariate. We partitioned our district into training and test data. Out of 261 districts in Ghana, 70% of the districts, comprising 183 districts were used as training set, while the remaining 30% (78 districts) were reserved for testing purposes. We have included in the supplementary materials a map showing the districts that were used as training and testing datasets (see Fig. 2 of supplementary materials). Model metrics were calculated to compare the performance of the two models. We computed in-sample (train dataset) and out-of-sample (test dataset) predictions and used the observed and predicted values to compute the following metrics:

1. Bias = mean of the residuals (prediction -observed)
2. Imprecision = standard deviation of the residuals
3. Mean Square Error (MSE) = mean of the square residuals.
4. Root Mean Square Error (RMSE) = square root of the mean square error
5. Pearson Correlation(r) = Estimate the correlation between the observed and predicted values. The range of value is between 0 and 1. A higher correlation indicates a better data fit.
6. Pseudo R-Squared (R^2) = proportion of variance in population density explained by the geospatial covariates.
7. 95 % Coverage = This was estimated only for the BART model. It calculates the percentage of observations falling between the upper and lower credible intervals.

The formula for each of the metrics can be found in the supplementary.

3.5. Simulation study

We conducted a simulation study to investigate the predictive performance of the RF model and the BART model. The simulation study is separate and independent from the actual census disaggregation described above. This simulation study was to ascertain the strengths and limitations of both the RF and the BART models at a granular level. Because both RF and BART are tree-based machine learning approaches, we decided to use a regression tree model, which is a much simpler model, to mimic the model structure of both the RF and BART models in generating a dataset that captures the structure of both models. To obtain our simulated dataset, we followed the following steps:

1. Using the district shapefile of Ghana as our study domain and the geospatial covariates extracted, we fitted a regression tree model to the observed district dataset to gain insights into the model's structure and the covariates used for its construction. The observed response used in fitting the model is population density, whereby population density is the observed population divided by the total area of a given district. The regression tree model selected three covariates for building the tree: residential total area (referred to as x_1), residential total length (referred to as x_2), and slope (referred to as x_3). In the supplementary material, we have provided the structure of the regression tree construction (Fig. 3 of supplementary materials).
2. We stacked the raster files of these three covariates together to extract their respective pixel values. We used the extracted pixel values and the true parameter values from the regression tree model to simulate gridcell level population density values. For the simulation process, we employed the following true parameter values: alpha (α) = 1.9, beta (β) = (-6.04, -4.74, -4.13, -3.39, -1.97, -0.26), with each beta value corresponding to one of the six nodes in the decision tree model.
3. After getting the simulated population density, we derived the simulated population for the *ith* pixel by multiplying the population density (μ_i) by building count for the *i*th pixel. Subsequently, we generated a simulated total population count for the pixel by sampling from a Poisson distribution. This approach was employed to mimic a bottom-up population modelling technique (Boo et al., 2022; Darin et al., 2022; Leasure et al., 2020) where the predicted population density is multiplied by the building count within a Poisson likelihood. The model is of the form:

$$Simulated\ Pixel\ Population_i \sim Poisson(\mu_i * B_i) \quad (3)$$

Where μ_i is the simulated population density and B_i is building count which is an additional geospatial covariate introduced into the modelling. This methodology ensured that our simulated pixel-level population closely resembled the bottom-up population modelling process, where the predicted population is adjusted by the building count with a Poisson likelihood, enhancing the fidelity of our simulated population estimation.

4. We then aggregated the simulated pixel population to the district level to obtain the simulated district population. Both RF and the BART models were used to disaggregate the simulated district population totals back to the pixel level using the geospatial covariates.
5. Finally, we compare the disaggregated pixel-level population from both the RF model and the BART model with the "true" simulated pixel-level population to investigate the predictive performance of both models. Fig. 3 provides a flow diagram for the simulation study.

4. Results simulation study

4.1. Simulated population data characteristics

Fig. 4 illustrates the spatial distribution of the simulated population at a pixel level (map A) and at the district level (map B). To create map B, the simulated population data at the pixel level was aggregated to the district level, where the population within each district was calculated by summing the populations of all pixels contained within it. Upon closer inspection of map, A, distinct population clusters become apparent, with noticeably higher concentrations along the south-eastern

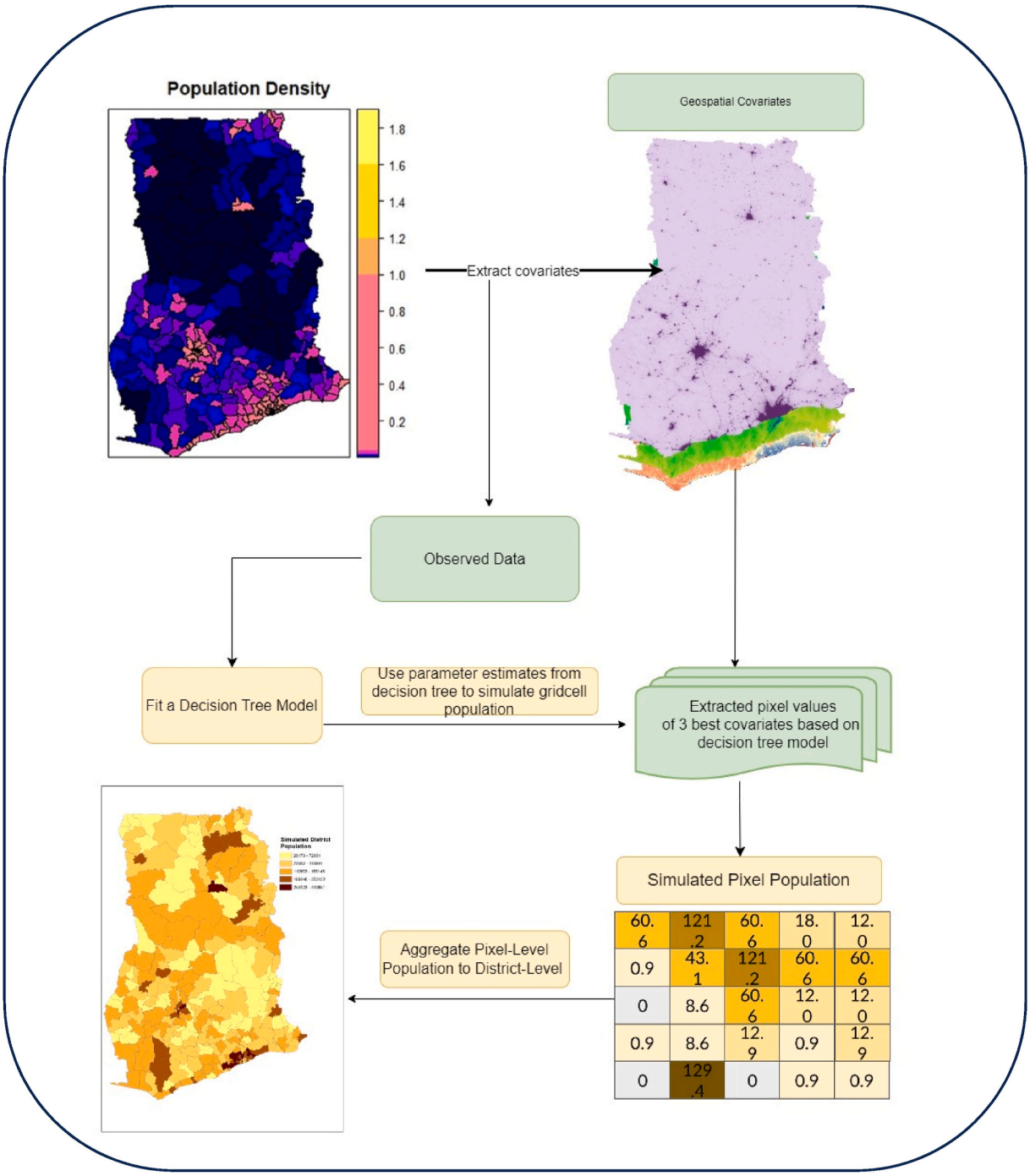


Fig. 3. A summarized workflow of how the simulation study was conducted.

coast as well as in the mid-central region, along with a smaller concentration in a northern-central area. These higher population clusters are the main cities in Ghana. In addition to the maps in Fig. 4, Table 1 gives important descriptive statistics for both the simulated population at the pixel level and the aggregated population at the district level. The median population per pixel is 9 people, with the lowest and highest population counts per pixel being 1 and 753, respectively. On the other hand, when considering the aggregated district-level population, the mean population per district amounts to 125,889 individuals, with the minimum and maximum population counts per district being 20,671 and 580,866, respectively.

4.2. *Model fitting metrics*

The simulated district population was disaggregated using both RF model and a BART model. The models were rigorously trained using 70% of the dataset, comprising 183 districts, while the remaining 30% (78 districts) were reserved for testing purposes. Both RF and BART methodologies were applied to both the training and test sets of simulated data. Upon comprehensive evaluation, the BART model demonstrated significantly superior performance across all evaluated metrics. For instance, the BART model achieved a nearly perfect percentage of variance (R^2) explained by geospatial covariates, nearly reaching 100%, while the RF model attained 96% in in-sample prediction. Furthermore, test data prediction accuracy favoured the BART model as compared to

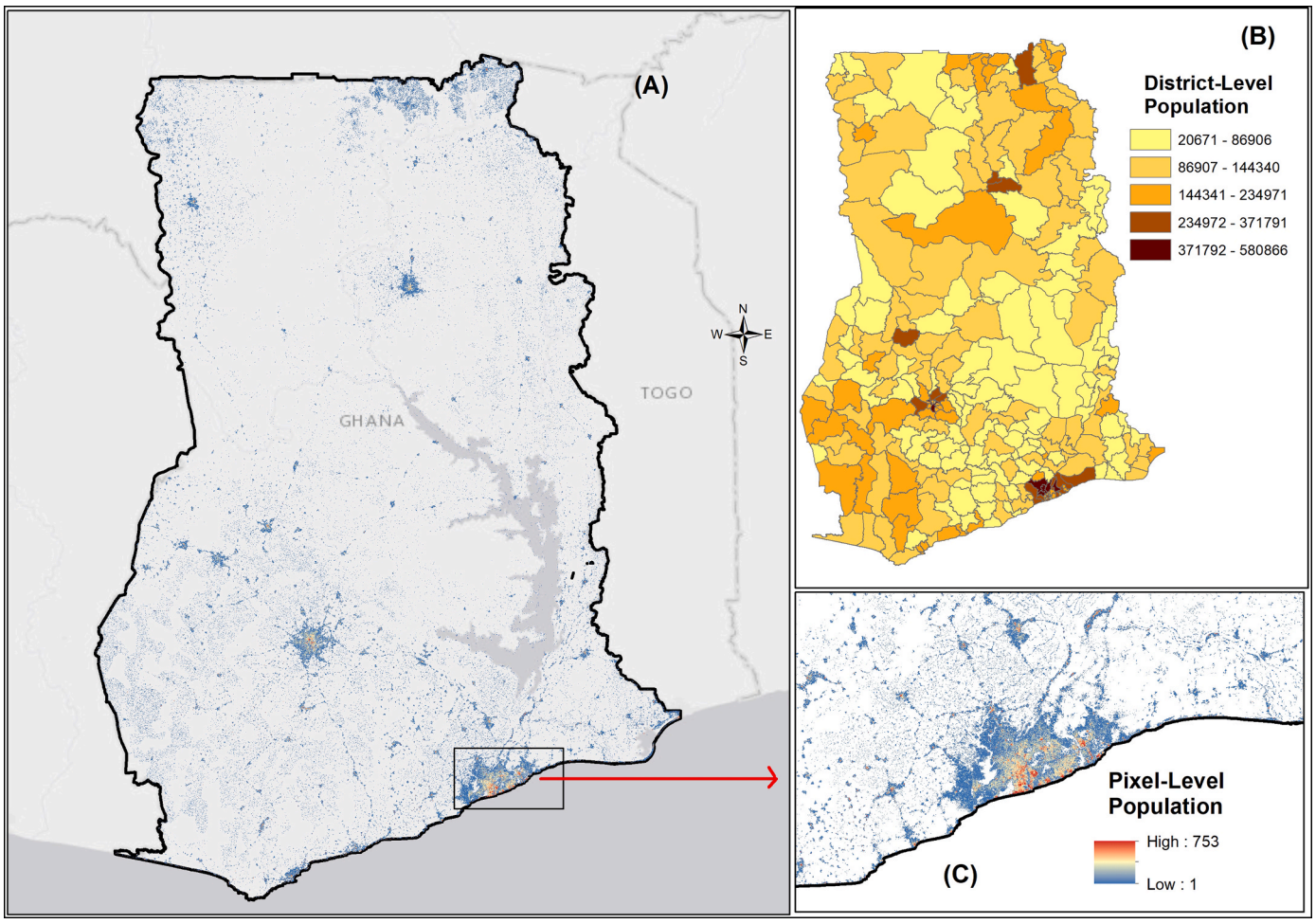


Fig. 4. Simulated population counts at the pixel level (map A) and corresponding aggregated population counts at the district levels (map B). Panel C offers a detailed view of a densely populated area (Greater Accra, the capital city of Ghana).

Table 1
Descriptive statistics of the simulated data.

	Pixel-Level Population	District-Level Population
Minimum	1	20671
1st Quantile	5	71546
Median	9.00	103120
Mean	21.07	125889
3rd Quantile	18	150895
Maximum	753	580866
Std Deviation	38.29	86429.18

the RF model (Table 2). It is noteworthy that the RF model exhibited a slight tendency to overfit the data, whereas the BART model demonstrated optimal performance.

In addition, we leveraged the complete simulated district level dataset encompassing 261 districts to fit a model, and subsequently employed this model to make pixel-level predictions using the geospatial covariates. The predicted population density served as a weighting factor to disaggregate the simulated district-level population, yielding predicted pixel-level populations. To gauge the accuracy of these predictions, we compared them to the simulated pixel-level population data. Results from the RF model indicated a 66% correlation (Pearson – r) between the predicted pixel-level population and the "true" simulated pixel population estimates, as detailed in Table 2. Conversely, the BART model showed a notably higher 81% correlation between predicted pixel-level populations and the "true" simulated pixel-level estimates. Additionally, various other evaluation metrics, such as the imprecision, MSE, and RMSE, exhibited substantially lower values in the BART model

Table 2
Goodness of fit metrics of the simulated data.

Models	Predictions	Bias	Imprecision	MSE	RMSE	Pearson r	R^2	% Coverage
RF	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		

Note: Model metrics were computed using residuals (predicted – observed values). A lower value for bias, imprecision, mean squared error (MSE), and root mean square error (RMSE) signifies a better fit of the model. Conversely, a higher value for correlation and the percentage of variance explained by the geospatial covariates indicates a more accurate and robust model fit.

in comparison to the RF model. For example, the MSE and the RMSE for the RF model was 826 and 28.7 respectively compared to the BART model which had a MSE and RMSE of 503.42 and 22.44 respectively. This underscores the superior capability of the BART model in accurately disaggregating simulated populations, aligning them closely with the simulated pixel level population in contrast to the RF model.

4.3. Distribution of disaggregated simulated population

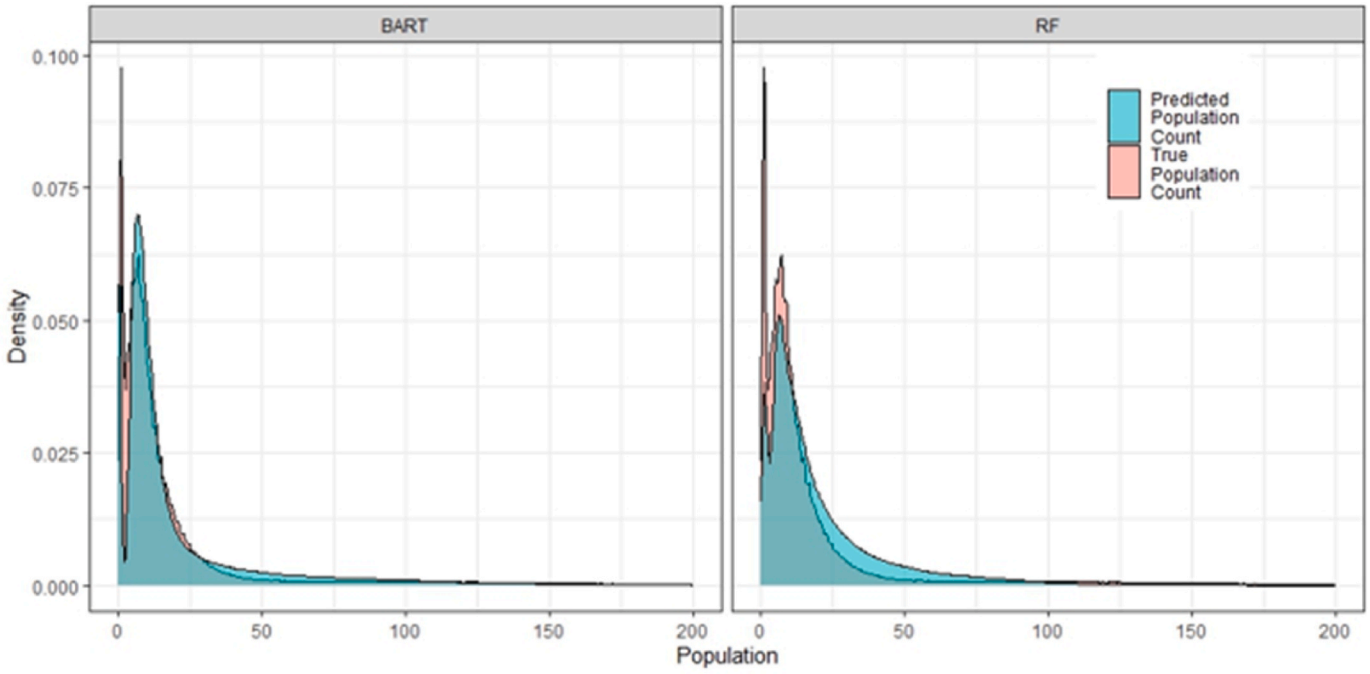
We further examined the distribution of the disaggregated simulated population at the pixel level with the "true" simulated population also at the pixel level using a density plot and provided descriptive statistics of the population count in Fig. 5. The distribution of the predicted population from the BART model closely mimics the distribution of the true simulated population count compared to the RF model. This is also evident in the table, in which the summarized distribution from the BART model is closer to the true simulated population compared to the RF model. However, the BART model yielded a much wider range, as indicated by the minimum value of 0.13 and a higher value of 431.61.

4.4. Results of 2021 population census for Ghana

Both the RF and the BART model were employed to disaggregate the 2021 population census data for Ghana. Table 3, presented below, offers insights into the performance of these two models in disaggregating the 2021 population data. Notably, the BART model outperformed the RF model in both in-sample and out-sample predictions. The BART model exhibited a substantially higher percentage of variance explained, nearing 100%, as opposed to the RF model's 96%. Out-of-sample metrics showed BART's strength in population disaggregation, which added to the case for its better performance. For instance, the out-of-sample RMSE for the RF model stood at 0.15, while the BART model demonstrated a significantly lower RMSE of 0.05. Collectively, model evaluation metrics from the out-of-sample prediction affirm the better model performance of BART in contrast to the RF model, which is similar to what we found in the simulation study.

4.5. Variable importance

We also conducted an evaluation of the key covariates selected for



Comparison of Disaggregated Gridcell Population with "true" Simulated Population

	True Population	BART	RF
Minimum	1	0.13	0.30
1st Quantile	5	6.21	6.86
Median	9.00	10.11	13.10
Mean	21.07	21.07	21.07
3rd Quantile	18	20.22	26.20
Maximum	753	431.61	342.84
Std Deviation	38.29	29.60	23.02
Correlation(r)		0.81	0.66

Fig. 5. Display a density plot and descriptive statistics for the disaggregated population in comparison to the true population. The density plot has been intentionally truncated to show a population count of fewer than 200 people per pixel to enhance visual clarity.

Table 3
Goodness of fit metrics of 2021 population census disaggregation.

Models	Predictions	Bias	Imprecision	MSE	RMSE	Pearson r	R^2	% in Credible Interval
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

partitioning the feature space and predicting population density within both the RF and the BART models. These models generate measures of variable importance by assessing how frequently a feature is employed to split the data within the regression tree. In our study, we used 24 geospatial covariates, and Fig. 6 shows the geospatial covariate selection process. We saw that the choice of covariates for partitioning the feature space when predicting population density was not always the same. In the RF model, which is represented in the upper plot, residential total length emerged as the most influential covariate, contributing to 12.75% of the model's predictive power. Following closely were residential total area (12.49%) and residential count (10.64%). Conversely, the BART model identified residential density (14.36%) as its most

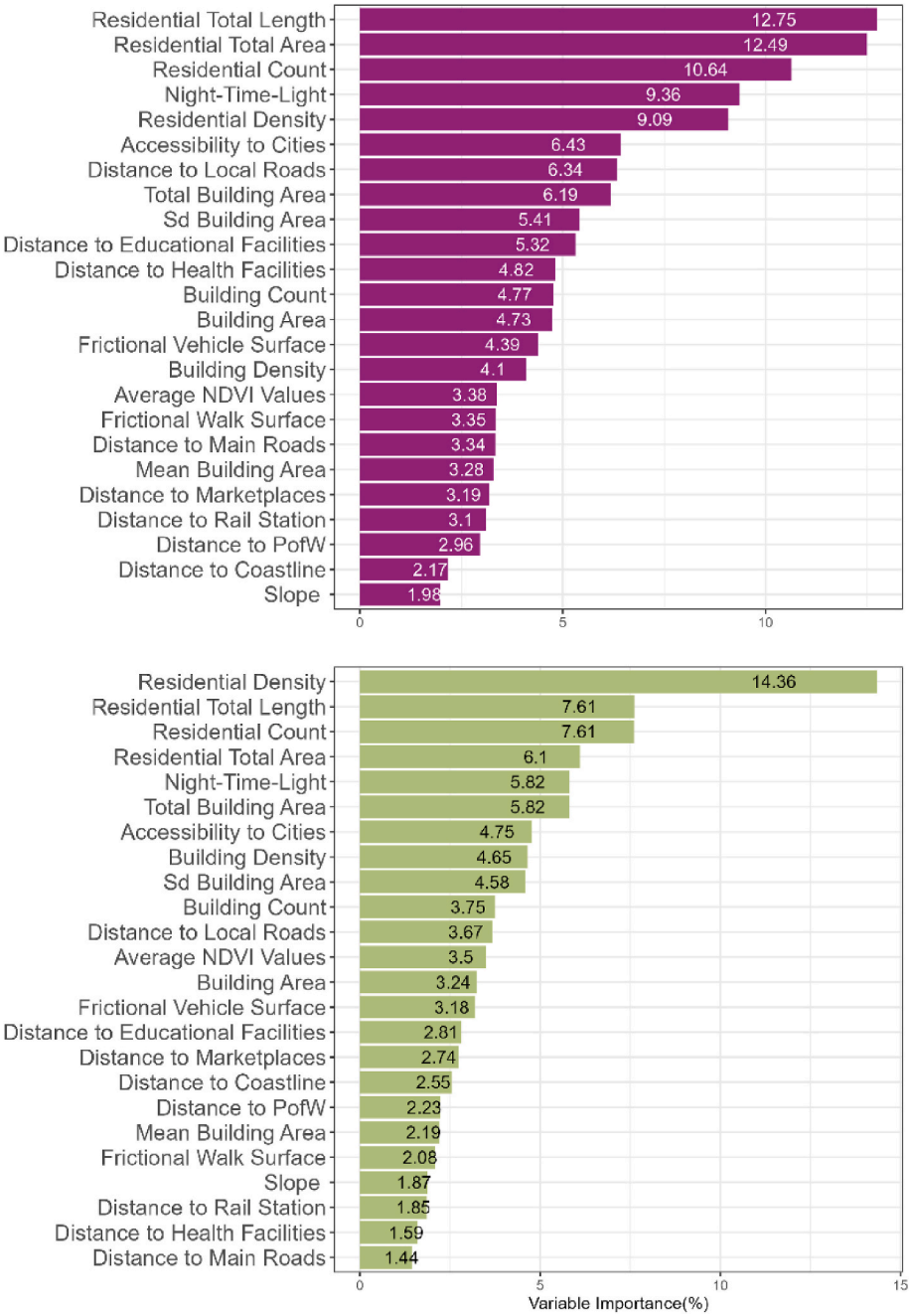


Fig. 6. Shows the geospatial covariate selection process in the BART and RF models. The upper plot shaded in purple is a covariate ranking from the RF model, and the lower plot coloured in light green is a covariate ranking from the BART model.

crucial covariate, with residential total length i.e., perimeter of residential buildings (7.61%) and residential count (7.61%) also featuring prominently among the top three. Comparatively, the least influential covariates in the RF model included slope (1.98%), distance to the coastline (2.17%), and distance to place of worship (2.96%). Meanwhile, in the BART model, the three least significant covariates were distance to main roads (1.44%), distance to health facilities (1.59%), and distance to rail stations (1.85%). In summary, our analysis underscores notable disparities in how the RF and BART models prioritize covariates for partitioning the feature space in the context of population disaggregation.

4.6. Mapping disaggregated population

The fitted model, utilizing the entire dataset, was employed for predicting population figures based on pixel-level covariates. The resultant population distributions at grid cell level have been mapped in Fig. 7. We also provide a density plot and summary statistics for the predicted population. Visually, it becomes apparent that both the disaggregated population from the RF and BART models exhibit strikingly similar spatial patterns. Regions with a higher concentration of people per pixel are predominantly situated along the southeastern coast, in the south-central region, and in the north-central areas of the country. The density plots from both models are closely aligned with each other, except that the BART model has a much higher peak compared to the RF model. Similar to the simulation study, the BART model has a wide range, with the predicted population ranging from 0.14 to 390.38. Across the majority of pixels, the population count is less than 30 individuals and the mean population per pixel is 20 people. However, there are notable exceptions in the form of highly dense population centers where the total number of people exceeds 50, particularly evident along the southeastern coast of the map. Fig. 8 compares the spatial distribution of the disaggregated population for Accra, the capital city of Ghana, superimposed on a recent high-resolution satellite image. Both the BART and RF models show similar spatial patterns of population distribution. However, the pixel-level population counts tend to vary between the two models, with the BART model producing slightly higher counts compared to the RF model.

4.7. BART model uncertainty

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. Unlike a single-point estimate, the BART model generates a posterior distribution, allowing us to derive credible intervals that give a range of plausible values.

In Fig. 9 below, we present the lower and upper credible intervals, the uncertainty, and the coefficient of variation for both Ghana and its capital city, Accra. The upper credible interval values range from 0.24 to 568.25 people per pixel, while the lower credible interval values range from 0.08 to 287.85 people per pixel. This range provides a clear picture of the potential variability in the population estimates.

The uncertainty around these predictions ranges from 0.50 to 2.42, indicating the degree of confidence in the model's predictions. Additionally, 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. This low variability indicates that the BART model provides robust and reliable population estimates. The credible intervals produced by the BART model capture the inherent uncertainty in population estimates at a granular level, which is very important for key policy and decision-making processes.

5. Discussions

Top-down population disaggregation has been applied in many countries to disaggregate the national census or projected population from a higher administrative unit to a lower one (Gaughan et al., 2016; McKeen et al., 2023; Sorichetta et al., 2015). Most of these studies have involved the use of a RF model for dasymetric population disaggregation (Leyk et al., 2019; Stevens et al., 2015). In this study, we examine a new Bayesian approach to population disaggregation using a BART model and compare this novel approach to the RF model. We compared the performance of the BART model and the RF model using both simulated data and the 2021 national population census of Ghana.

Results from applying the BART model and the RF model indicate the superior performance of the BART model compared to the RF model for population disaggregation across all model metrics (bias, imprecision, MSE, RMSE, and correlation). For example, based on the simulation study, the BART model more accurately recovered the "true" simulated population at the pixel level. The correlation between the simulated disaggregated population and the "true" simulated population at the pixel level was 81% for the BART model, compared to 66% for the RF model. This suggests that the RF model may systematically underestimate or overestimate the population within a grid cell compared to the BART model. The higher performance of the BART model over the RF model can be attributed to the Bayesian framework of the BART model, which allows for better regularization and adaptability to the handling of high-dimensional data. For example, the BART model provides a set of priors for the various parameters in the model, such as the tree structure, the leave parameters, and the error variance component, which tunes the model to optimum performance, allowing it to capture complex interactions and non-linear relationships (Kapelner & Bleich, 2013).

Recently, there have been new methodological developments, such as applying deep learning approaches (Metzger et al., 2022, 2023) and stack ensemble machine learning approaches that combine multiple algorithms like RF, XGBoost, and LightGBM as base models for population disaggregation (Chen et al., 2024; Tu et al., 2022; Zhang et al., 2024). These new approaches improve model predictions by leveraging individual models' strengths. However, a major limitation is the lack of uncertainty quantification in the disaggregated population estimates.

Uncertainty in population estimates can result from different sources such as discrepancies in input data sources and the estimation methodology (Leyk et al., 2019; Tatem, 2022). Official population censuses, or projected populations, typically provide observable population totals for dasymetric population disaggregation. It is well known that population counts from censuses can have problems like net undercounting, omissions and hard-to-count populations (O'Hare, 2019; Robinson, 2011) propagating errors in the total population count. Similarly, subnational scale projected population numbers can be highly uncertain, especially because of population changes caused by epidemics, conflict, and migration patterns during the intercensal period (O'Sullivan, 2023; Park & LaFrombois, 2019). Additional ancillary geospatial covariates used as input datasets can be an additional source of errors, coupled with differences in methodological methods and limitations associated with these dasymetric mapping models (such as overfitting and erroneous model assumptions), as well as the inability to capture spatial heterogeneity in the input population, all of which combine to produce uncertainty in gridded population output.

Methods for producing gridded population estimates should be viewed as a stochastic process where multiple levels of uncertainty arise, and quantifying such levels of uncertainty in the final gridded output enables users of such products to be aware of the variability of the predicted population estimates. However, this is not the case with many dasymetric population mapping approaches, such as the RF model, that treat the relational problem between the observed population data and the ancillary geospatial covariates as a deterministic rather than a stochastic process. These approaches generate a single point estimate as the predicted population for a given grid cell and do not incorporate a quantitative mechanism for measuring uncertainty around such estimates. The BART model, on the other hand, is a stochastic model that combines decision trees with Bayesian modelling to capture both

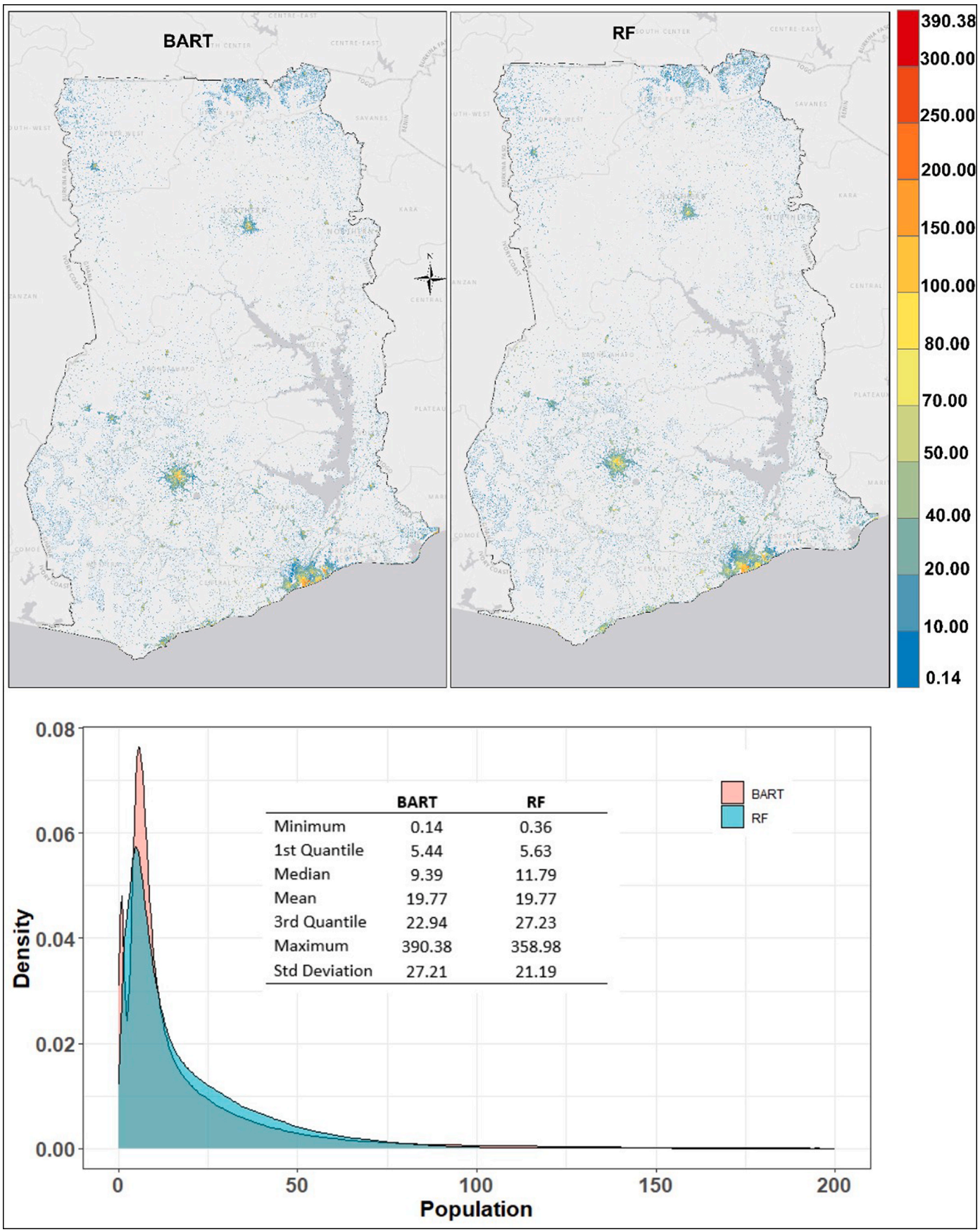


Fig. 7. Map of the disaggregated population at 100m resolution using the two modelling approaches. The disaggregation is constrained to only settled pixels. The figure also provides a density plot showing the distribution of the predicted population.

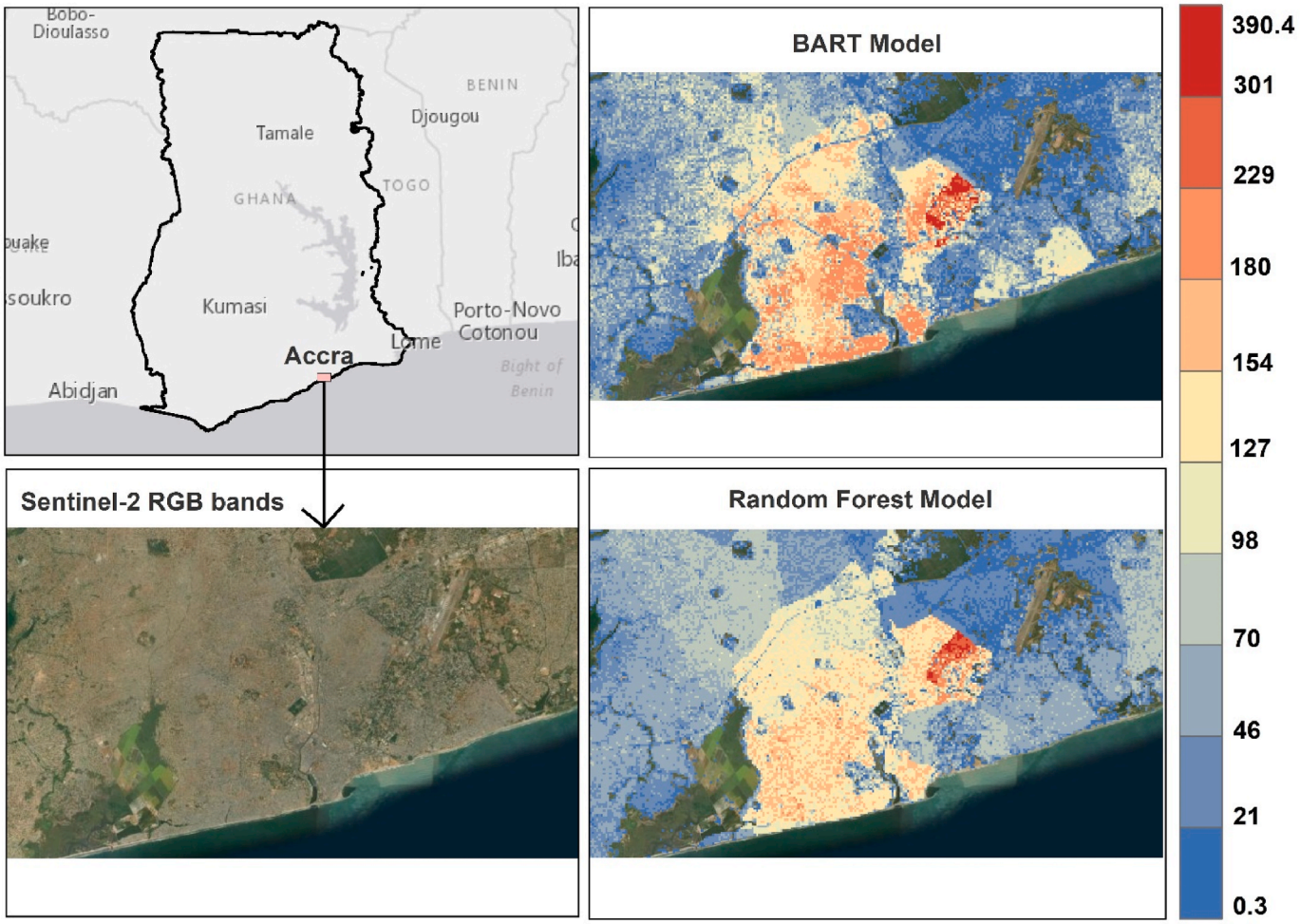


Fig. 8. Illustrates the disaggregated predicted population distribution at 100m spatial resolution for Accra the capital city of Ghana superimposed on high-resolution satellite The imagery was downloaded from Copernicus on 13th May 2024.

systematic patterns and random variations in the data, thereby quantifying uncertainty in the disaggregated population distribution. This gives the BART model a distinct advantage over other dasymetric population mapping methods.

Based on our findings, we can see in Fig. 9 that the BART model does not just give us a single point estimate for the predicted population. Instead, it gives us a posterior distribution that shows the possible range of values for the disaggregated population. This considers the inherent biases and uncertainty in the modelling process. We can summarize the posterior distribution to obtain the mean predicted value and the credible intervals surrounding the estimates. This capability provides robust uncertainty estimates, invaluable for guiding evidence-based policy decisions on critical national issues such as epidemiological surveillance, disaster preparedness, and resource allocation.

For instance, in low-income settings, some research has shown that vaccination campaigns often fail to achieve coverage goals due to uncertainty about target population size and distribution (Bharti et al., 2016) as well as the use of outdated population numbers (Tatem, 2022). Simply providing the total number of expected children within a given health catchment area without factoring into account the uncertainty in such estimates could be misleading. Health care workers may visit the community and find the total number of children undercounted, potentially leading to inefficiencies in vaccine delivery. However, with the BART model, we are able to estimate the plausible range of expected child populations within specific communities and make provision for the required number of vaccines, which is crucial for optimizing logistics and vaccine distribution.

Both the BART and the RF models selected different covariates in partitioning the feature space for the modeling. However, a common trend emerged from the analysis. In both the BART and RF models, settlement, or built-up data, such as residential count, residential total area, building count, etc., was always ranked higher as a covariate for the modeling, while geospatial variables such as slope, distance to health facilities, and distance to hospitals were less important. This underscores the pivotal role of settlement and built-up area datasets in accurately modeling human population distribution (Wardrop et al., 2018). The increasing availability of such settlement data derived from remote sensing imagery further underscores the need to be circumspect in selecting different settlement products for population modeling. In a recent study by Chamberlain et al. (2023) on building footprint comparison across Africa, they found that the different building footprints are spatially heterogeneous and are not interchangeable, and that these footprint datasets have clear differences in various metrics derived from them, such as building count and building area. Stevens et al. (2020) highlight the need to integrate multiple sources of settlement data in top-down modeling to leverage the strengths of each dataset while mitigating individual misclassifications.

Both BART and RF models are valuable methodologies for producing gridded population estimates, which have become essential in research and decision-making on critical global issues such as urbanization (Santillan & Heipke, 2024), climate change (Hanberry, 2022), conflict (Sbarra et al., 2023; Zheng et al., 2022), public health (Florio et al.,

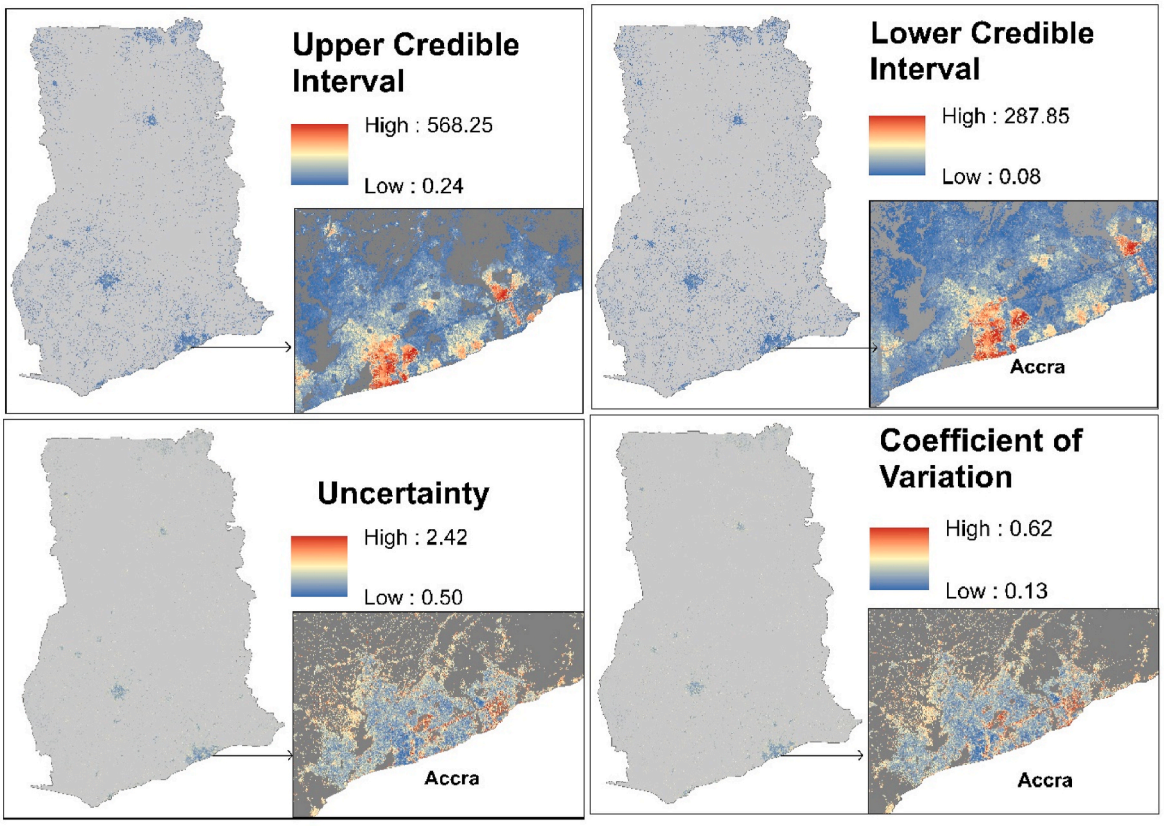


Fig. 9. The uncertainty surrounding the BART model predictions. The uncertainty was calculated using the formula (upper credible interval – lower credible interval)/mean population. The coefficient of variation was calculated by dividing the mean population by the standard deviation. A low coefficient of variation indicates that the predicted population is tightly clustered around the mean, while a high coefficient of variation indicates a wider variability around the predicted population.

2023; WHO, 2023) and planning for sustainable development (Tuholske et al., 2021). Unlike population projections or census data, which are typically produced at coarser spatial scales (e.g., administrative units or national boundaries) and updated infrequently, gridded population estimates are provided at a high spatial resolution at granular levels. This allows governments, organizations, and researchers to better understand human distribution across geographic areas. The fine-grained nature of these estimates makes them crucial for addressing real-world challenges that demand detailed and timely information about population distribution. For instance, gridded population estimates are highly valuable in disaster and conflict response planning. Natural disasters such as hurricanes, floods, earthquakes, and wildfires often affect vast areas with varying population densities. In such scenarios, knowing precisely where people live can significantly enhance emergency response efforts. By overlaying hazard maps with population grids, emergency planners can prioritize evacuation routes, allocate resources to the most affected areas, and minimize the loss of life. Similarly, in the case of pandemics or disease outbreaks, gridded population data can help track transmission patterns and guide the deployment of healthcare services.

Our study stands as a milestone in the field of top-down dasymetric population modelling, being the first to apply a Bayesian approach to the modelling. While the RF model has been the de facto choice in previous research, its inability to provide uncertainty estimates around predictions has been a notable limitation. The BART model, as a pioneering Bayesian approach in this context, not only outperforms the RF model but also provides a means to quantify prediction uncertainties. This innovation has the potential to transform top-down population disaggregation, offering a powerful tool for researchers and policymakers alike. By adopting this Bayesian top-down model, future researchers can harness its capabilities to improve the accuracy and precision of population distribution estimates, ultimately advancing our understanding of human demographics at local scales and informing critical decision-making processes.

A future research project we intend to perform is to introduce Bayesian hierarchical Modelling approaches for top-down disaggregation. The model we present here is “tree-based” and does not consider the hierarchical nature of different administrative levels and categorical covariate types that may influence population count. Future research will focus on using popular Bayesian algorithms such as INLA (Lindgren & Rue, 2015), which is computationally faster for top-down population disaggregation. We also seek to extend such analysis to multiple time periods to come up with computationally efficient ways of disaggregating the population and make possible projections into the future.

Despite the significance of this study, a limitation of the BART approach is the computational cost and time associated with the modelling. BART is a Bayesian approach that is based on Markov Chain Monte Carlo (MCMC) simulation. In MCMC, the posterior distribution is simulated using a series of chains until convergence is reached. Each chain of posterior iterations is based on the one before it. This imposes a substantial computational cost and time on fitting a model and making predictions. For example, the BART model took about 3 h to fit and make predictions using the 2021 census data for Ghana, whereas the RF model took less than an hour to disaggregate the same dataset. The RF model therefore offers a faster approach to top-down disaggregation. However, it is imperative to note that despite the huge computation cost of the BART model, computational cost should not be sacrificed for efficiency in disaggregating populations.

In conclusion, we consider this work a useful addition to the growing

body of methodology for dasymetric top-down population disaggregation. The BART approach to gridded population estimation can be used in addition to the RF model that is already in place, and it can give unique information about how to measure uncertainty in the predictions. Predictions from the RF model and the BART approach can be compared, and informed decisions can be made based on the model's metrics.

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

The 2021 census population data used in this study was provided by the Ghana Statistical Service and this dataset is available upon reasonable request at the Ghana Statistical Service website https://www.statsghana.gov.gh/.

Geospatial covariate data can also be downloaded from the WorldPop Research Group at the University of Southampton's website https://www.worldpop.org/datacatalog/

Funding

This work was part of the GRID3 – Phase 2 Scaling project with funding from the Bill and Melinda Gates Foundation [grant number: INV-044979]. GRID3 project partners include GRID3 Inc., WorldPop at the University of Southampton, and the Center for International Earth Science Information Network in the Columbia Climate School at Columbia University.

Declaration

The authors have no conflict of interest to declare.

CRediT authorship contribution statement

Ortis Yankey: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Chigozie E. Utazi:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Christopher C. Nnanatu:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Assane N. Gadiaga:** Software, Methodology, Data curation. **Thomas Abbot:** Software, Resources, Data curation. **Attila N. Lazar:** Writing – review & editing, Supervision. **Andrew J. Tatem:** Writing – review & editing, Funding acquisition.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable reviews in shaping this manuscript, as well as the Ghana Statistical Service for providing the 2021 census data for this study. We would also like to thank GRID3 for funding this project through the Bill and Melinda Gates Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2024.103416.

References

Archila Bustos, M. F., Hall, O., Niedomysl, T., & Ernstson, U. (2020). A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015. *Population and Environment, 42*, 255–277.

Bharti, N., Djibo, A., Tatem, A. J., Grenfell, B. T., & Ferrari, M. J. (2016). Measuring populations to improve vaccination coverage. *Scientific Reports, 6*(1), Article 34541.

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., & Fuller, T. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications, 13*(1), 1330.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Calka, B., & Bielecka, E. (2020). GHS-POP accuracy assessment: Poland and Portugal case study. *Remote Sensing, 12*(7), 1105.

Chen, Y., Xu, C., Ge, Y., Zhang, X., & Zhou, Y.n. (2024). A 100-m gridded population dataset of China's seventh census using ensemble learning and geospatial big data. *Earth System Science Data Discussions, 2024*, 1–19.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). *Bart: Bayesian additive regression trees*.

CIESIN. (2018). *Gridded population of the world, version 4 (GPWv4): Population count adjusted to match 2015 revision of UN WPP country totals, revision 11*. NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4PN93PB

Darin, E., Kuépié, M., Bassinga, H., Boo, G., Tatem, A. J., & Reeve, P. (2022). The population seen from space: When satellite images come to the rescue of the census. *Population, 77*(3), 437–464.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography, 1*(3), 226–234.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science, 28*(2), 125–138.

Florczyk, A. J., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., Melchiorri, M., Pesaresi, M., Politis, P., & Schiavina, M. (2019). GHSL data package 2019. *Luxembourg, eur, 29788*(10.2760), Article 290498.

Florio, P., Freire, S., & Melchiorri, M. (2023). Estimating geographic access to healthcare facilities in sub-saharan Africa by degree of urbanisation. *Applied Geography, 160*, Article 103118.

Fries, B., Guerra, C. A., García, G. A., Wu, S. L., Smith, J. M., Oyono, J. N. M., Donfack, O. T., Nfumu, J. O. O., Hay, S. I., & Smith, D. L. (2021). Measuring the accuracy of gridded human population density surfaces: A case study in Bioko Island, Equatorial Guinea. *PLoS One, 16*(9), Article e0248646.

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., Ye, X., Linard, C., Hornby, G. M., & Hay, S. I. (2016). Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data, 3*(1), 1–11.

Ghana Statistical Service. (2022). 2021 population and housing census (2021 PHC), version 1.0 of the public use dataset (august 2022),provided by the national data archive. www.statsghana.gov.gh.

Hanberry, B. B. (2022). Global population densities, climate change, and the maximum monthly temperature threshold as a potential tipping point for high urban densities. *Ecological Indicators, 135*, Article 108512.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Additive models, trees, and related methods. *The elements of statistical learning: Data mining, inference, and prediction* (pp. 295–336).

Holt, J. B., Matthews, K. A., Lu, H., Wang, Y., LeClercq, J. M., Greenlund, K. J., & Thomas, C. W. (2019). Small area estimates of populations with chronic conditions for community preparedness for public health emergencies. *American journal of public health, 109*(S4), S325–S331.

Kapelner, A., & Bleich, J. (2013). *Bartmachine: Machine learning with bayesian additive regression trees*. arXiv preprint arXiv:1312.2171.

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, 117*(39), 24173–24179. https://doi.org/10.1073/pnas.1913050117

Leasure, D. R., Kashyap, R., Rampazzo, F., Dooley, C. A., Elbers, B., Bondarenko, M., … Akimova, E. T. (2023). Nowcasting daily population displacement in Ukraine through social media advertising data. *Population and Development Review, 49*(2), 231–254.

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., Rose, A., Stevens, F. R., Blankespoor, B., & Frye, C. (2019). The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data, 11*(3), 1385–1409.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Linard, C., Alegana, V. A., Noor, A. M., Snow, R. W., & Tatem, A. J. (2010). A high resolution spatial population database of Somalia for disease risk mapping. *International Journal of Health Geographics, 9*(1), 1–13.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software, 63*(19).

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., & MacManus, K. (2019). Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big earth data, 3*(2), 108–139.

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data, 4*(1), 1–17.

Martin, D., Lloyd, C., & Shuttleworth, I. (2011). Evaluation of gridded population models using 2001 Northern Ireland Census data. *Environment and Planning A, 43*(8), 1965–1980.

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

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42.
Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194.
Metzger, N., Daudt, R. C., Tuia, D., & Schindler, K. (2023). *High-resolution population maps derived from sentinel-1 and sentinel-2*. arXiv preprint arXiv:2311.14006.
Metzger, N., Vargas-Muñoz, J. E., Daudt, R. C., Kellenberger, B., Whelan, T. T.-T., Ofli, F., Imran, M., Schindler, K., & Tuia, D. (2022). Fine-grained population mapping from coarse census counts and open geodata. *Scientific Reports, 12*(1), Article 20085.
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. *Journal of The Royal Society Interface, 14*(137), Article 20170401.
Nordstrand, E., & Frye, C. (2014). World population estimate. https://doi.org/10.13140/RG.2.2.18213.14565.
O'Hare, W. P. (2019). *Differential undercounts in the US census: Who is missed?* Springer Nature.
Olorunfemi, J., & Fashagba, I. (2021). Population census administration in Nigeria. *Nigerian Politics*, 353–367.
O'Sullivan, J. N. (2023). Demographic delusions: World population growth is exceeding most projections and jeopardising scenarios for sustainable futures. *World, 4*(3), 545–568.
Park, Y., & LaFrombois, M. E. H. (2019). Planning for growth in depopulating cities: An analysis of population projections and population change in depopulating and populating US cities. *Cities, 90*, 237–248.
R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing *(No Title)*.
Robinson, J. G. (2011). *Coverage of population in census 2000 based on demographic analysis: The history behind the numbers*. US Census Bureau, Population Division.
Santillan, J. R., & Heipke, C. (2024). Assessing patterns and trends in urbanization and land use efficiency across the Philippines: A comprehensive analysis using global Earth observation data and SDG 11.3. 1 indicators. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 1–24.
Sbarra, A. N., Rolfe, S., Haeuser, E., Nguyen, J. Q., Adamu, A., Adeyinka, D., Ajumobi, O., Akunna, C., Amusa, G., & Dahiru, T. (2023). Estimating vaccine coverage in conflict settings using geospatial methods: A case study in borno state, Nigeria. *Scientific Reports, 13*(1), Article 11085.
Sims, K., Reith, A., Bright, E., Kaufman, J., Pyle, J., Epting, J., Gonzales, J., Adams, D., Powell, E., Urban, M., & Rose, A. (2023). *LandScan Global 2022 Version 2022) [raster digital data]*. Oak Ridge National Laboratory. https://doi.org/10.48690/1529167
Skinner, C. (2018). Issues and challenges in census taking. *Annual Review of Statistics and its Application, 5*, 49–63.
Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data, 2*(1), 1–12.
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), Article e0107042.
Stevens, F. R., Gaughan, A. E., Nieves, J. J., King, A., Sorichetta, A., Linard, C., & Tatem, A. J. (2020). Comparisons of two global built area land cover datasets in methods to disaggregate human population in eleven countries from the global South. *International Journal of Digital Earth, 13*(1), 78–100.
Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data, 4*(1), 1–4.
Tatem, A. (2022). Small area population denominators for improved disease surveillance and response. *Epidemics, 41*, Article 100641.
Tenerelli, P., Gallego, J. F., & Ehrlich, D. (2015). Population density modelling in support of disaster risk assessment. *International Journal of Disaster Risk Reduction, 13*, 334–341.
Tu, W., Liu, Z., Du, Y., Yi, J., Liang, F., Wang, N., Qian, J., Huang, S., & Wang, H. (2022). An ensemble method to generate high-resolution gridded population data for China from digital footprint and ancillary geospatial data. *International Journal of Applied Earth Observation and Geoinformation, 107*, Article 102709.
Tuholske, C., Gaughan, A. E., Sorichetta, A., de Sherbinin, A., Bucherie, A., Hultquist, C., Stevens, F., Kruczkiewicz, A., Huyck, C., & Yetman, G. (2021). Implications for tracking SDG indicator metrics with gridded population data. *Sustainability, 13*(13), 7329.
UN-SPIDER. (2023). *How are population and settlement data used in disaster risk reduction and response efforts?* UN-SPIDER. Retrieved 08/01/2024 from https://www.un-spider.org/links-and-resources/daotm/daotm-populationandsettlementdata.
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.
Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., & Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine, 36*(12), 1583–1591.
Wardrop, N., Jochem, W., Bird, T., Chamberlain, H., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., & Tatem, A. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences, 115*(14), 3529–3537.
WHO. (2023). *World malaria report 2023*. World Health Organization.
Yin, X., Li, P., Feng, Z., Yang, Y., You, Z., & Xiao, C. (2021). Which gridded population data product is better? Evidences from Mainland Southeast Asia (MSEA). *ISPRS International Journal of Geo-Information, 10*(10), 681.
Zhang, H., Fu, J., Li, F., Chen, Q., Ye, T., Zhang, Y., & Yang, X. (2024). Fine-scale population mapping on Tibetan plateau using the ensemble machine learning methods and multisource data. *Ecological Indicators, 166*(112307).
Zheng, Z., Wu, Z., Cao, Z., Zhang, Q., Chen, Y., Guo, G., Yang, Z., Guo, C., Wang, X., & Marinello, F. (2022). Estimates of power shortages and affected populations during the initial period of the Ukrainian-Russian conflict. *Remote Sensing, 14*(19), 4793.