Disaggregating 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

GISRUK 2024

Summary

Fine-scale population census data are often lacking due to the challenge of sharing such sensitive data at granular scales. In this study, we compare the Random Forest (RF) model and the Bayesian Additive Regression Tree (BART) model for population disaggregation using both census data from Ghana and simulated data. The BART model outperforms the RF model in out-of-sample predictions for metrics like bias, mean squared error, and root mean squared error. It also provides uncertainty estimates around the predicted population, which is often lacking with the RF model. This study highlights the BART model's superiority in disaggregating population data.

Keywords: Population Modelling, Population Disaggregation, Bayesian Top-Down Population, Random Forest, Bayesian Additive Regression Tree

Introduction

Population figures at small area scales are crucial for policymakers as they offer insights on the magnitude, structure, spatial distribution, and temporal changes of a nation's population. Accurate population data is essential for emergency response and disaster relief efforts, especially during natural disasters (Tenerelli et al., 2015; UN-SPIDER, 2023; UNFPA, 2020). National population and housing censuses are the most accurate, spatially detailed, and reliable sources of population information. However, census data or projected population data are produced at a higher 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). Integration of data collected at higher administrative levels with other forms of data, such as health facilities or catchment areas 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.

Several global and continental gridded population datasets have been produced to fill in these gaps at the small-area level (Leyk et al., 2019) These include the LandScan Global Population Datasets (Sims et al., 2023), the Gridded Population of the World version 4 (CIESIN, 2018), the WorldPop population datasets (Tatem, 2017), and the Global Human Settlement Layer-Population (Florczyk et al., 2019). One of the methods for producing such data is dasymetric population mapping, where population numbers at higher administrative units are disaggregated at small area levels using ancillary geospatial covariates, which inform the model.

Two major limitations are associated with this method: (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 modeling. The objectives of this work were therefore to:

1. apply a Bayesian Additive Regression Tree (BART) approach to disaggregate population totals from a higher administrative unit to the grid cell level and to estimate the uncertainties associated with the predictions. This objective involved a comparative assessment of the BART approach and the RandomForest (RF) algorithm for dasymetric population mapping. The RF is one of the most popular algorithms for population mapping.
2. To validate the gridded population estimates with "true" simulated population numbers at the gridcell level in a simulation study. The predicted gridcell population was compared with the simulated "true" population to assess the comparative performance of both the RF model and the BART model.

Method

The study used the 2021 National Population Census of Ghana at the district level, which corresponds to administrative level 2 for the population disaggregation. We combined the observed population numbers with a wide range of geospatial covariates related to population distribution across the country for the modelling. The study used dasymetric population disaggregation modelling, which involves fitting a model to estimate predicted population density, which is used as a weighting layer to redistribute observed population data from a larger administrative unit to target gridcells or small areas (Stevens et al., 2020). The model was fitted using both the BART (Kapelner and Bleich, 2013) and the RF models. Model metrics were calculated to compare the performance of the two models, including bias, imprecision, mean square error (MSE), root mean square error (RMSE), Pearson correlation, pseudo-R-squared, and 95% coverage.

We also simulated gridcell population counts and aggregated them to a higher administrative level. We then used both the RF and BART models to disaggregate the total simulated population from the higher administrative level to obtain pixel-level population estimates. We compared the true simulated pixel population count and the disaggregated pixel population estimates on a pixel-by-pixel level. We also calculated model metrics to see how well both models worked with simulated data.

Results

Simulation Study

The BART model showed superior performance across all model metrics, including in-sample and out-of-sample predictions (Table 1). The BART model achieved a nearly perfect percentage of variance explained by geospatial covariates, reaching 100%, while the RF model achieved 96% in in-sample prediction. The RF model exhibited a slight tendency to overfit the data, while the BART model demonstrated optimal performance. There was an 81% correlation between the BART predicted estimates and the true simulated pixel-level estimates for the disaggregated population numbers at the grid cell level. This was in contrast to a 66%

correlation between the RF model estimates and the simulated true estimates. The BART model also had lower values for imprecision, MSE, and RMSE compared to the RF model, indicating that the BART model provides a better approach to disaggregating population totals at small area levels compared to the RF model.

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		

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 superior 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.

2021 National Population Census Disaggregation

The superiority of the BART model in disaggregating census data compared to the RF model was also observed when both models were used to disaggregate the 2021 National Population Census for Ghana. Notably, the BART model outperformed the RF model in both in-sample and out-sample predictions (Table 2). 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. Overall, the model evaluation metrics from the out-of-sample prediction show that the BART model performs better than the RF model, which is similar to what we found in the simulation study.

Table 2. Goodness of fit metrics of 2021 Population Census Disaggregation

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

The BART model also addresses the limitation of the RF model, which is its inability to quantify the uncertainty around the predictions. With the BART model, we did posterior simulations from the parameter estimates, calculated credible intervals around the the predictions and used these credible intervals to quantify the uncertainty around the predictions. From the BART model in Fig 1, values for the upper credible interval ranges between 0.24 to 568.25 and the lower credible interval ranges between 0.08 to 287.85 people per pixel. The uncertainty around the predictions ranges from 0.50 to 2.42, and the coefficient of variation for most of the gridcells are less than 0.2, indicating less variability around the mean predicted population count.

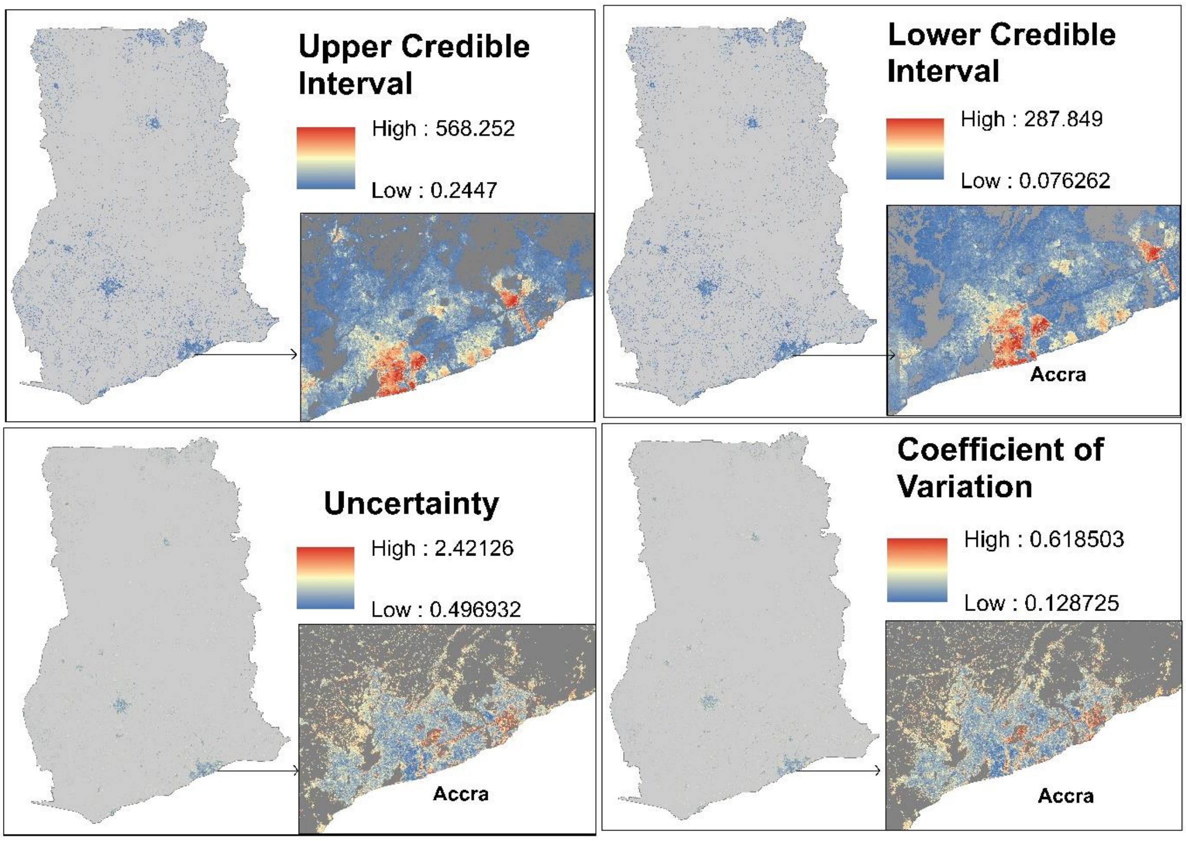

Fig. 1 shows the uncertainty surrounding the predictions. The uncertainty was calculated using the formula (upper credible interval – lower credible interval)/mean population. The coefficient of variation was calculated by dividing the mean population by the standard deviation. A low coefficient of variation 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.

Discussion

The study compares a Bayesian approach to population disaggregation using a BART model and a Random Forest (RF) model for predicting population figures based on pixel-level covariates. The study found that the BART model was able to recover the "true" simulated population much better than the RF model across all model metrics (bias, imprecision, MSE, RMSE, and correlation). This suggests that the RF model may systematically underestimate or overestimate the number of people within a grid cell compared to the BART model. This underestimation or overestimation may have national policy implications, particularly for healthcare campaigns and disaster relief efforts in LMICs where regular population data at granular scales is lacking. The study highlights the need for more accurate and reliable population mapping methods for top-down population disaggregation. The BART approach also quantifies uncertainty around the predictions, which is good for policy decisions.

Our study stands as a significant 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.

Acknowledgements

We would like to acknowledge the Ghana Statistical Service for providing the 2021 census data for this study. We also want to acknowledge GRID3 for funding this project through the Bill and Melinda Gates Foundation.

References

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

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), 290498.

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

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.

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.

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.

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, *4*(1), 1-4.

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.

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. https://www.unfpa.org/resources/value-modelled-population-estimates-census-planning-and-preparation