A novel geospatial data augmentation strategy for producing high-resolution population estimates using sparse health campaign data

Chibuzor Christopher Nnanatu[1*] Assane Gadiaga[1], Heather Chamberlain[1], Thomas Abbott[1], Attila Lazar[1], Andrew Tatem[1]

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

Background:

Demographic datasets containing counts of people per household are often collected opportunistically during healthcare campaigns. These datasets can be integrated with satellite observations (settlements) and other geospatial covariates within a robust statistical modelling framework, to produce comparable small area estimates of population within countries where census estimates are either outdated or incomplete [1]. However, population datasets based on healthcare campaigns are often susceptible to either under- or over-count and may often be only scarcely available for population estimation. Here, datasets from the National Malaria Elimination Programme (NMEP) campaigns conducted in 2022 in Nigeria, were augmented to produce small area population estimates (along with the uncertainties) across the entire 774 local governments in the country. The data augmentation strategy which took advantage of the large spatial coverage of the 2021 Multiple Indicator Cluster Surveys (MICS) data, was implemented using advanced Bayesian hierarchical regression population modelling framework [2,3]. Our methodology highlights the usefulness of health campaign datasets in population estimation and how a sample-based nationally representative household survey data can be used to fill data gaps within a population input dataset with sparse coverage and could easily be adapted to other contexts.

Data:

- Nationally representative 2021 MICS data:
  - Anonymised household (HH) listing for sampled clusters (2079 clusters)
  - XY coordinates of the clusters available
  - HH totals were multiplied by HH survey weights
  - Weighted HH were aggregated at the Local Government Area (LGA) level
- State-level 2022 NMEP data:
  - Anonymised HH listing for 6 states with complete spatial coverage (Delta, Kaduna, Katsina, Kano, Niger, Taraba)
  - XY coordinates of the households available
  - HH totals were aggregated at the LGA level
- Rasterised settlement extent at 100m resolution [4]

Methods:

The large spatial coverage of the 2021 MICS data provided an opportunity to create a scale factor between the NMEP and MICS datasets. The estimated scale factor which was predicted at locations with no NMEP/MICS overlaps, was used to augment the NMEP data and provide country-wide full counts, which were then used to train model parameters for grid-cell predictions using the integrated nested Laplace approximation in conjunction with the stochastic partial differential equations (INLA-SDPE,[5]) frameworks. Data cleaning, model implementation and grid cell predictions relied on R programming software [6].

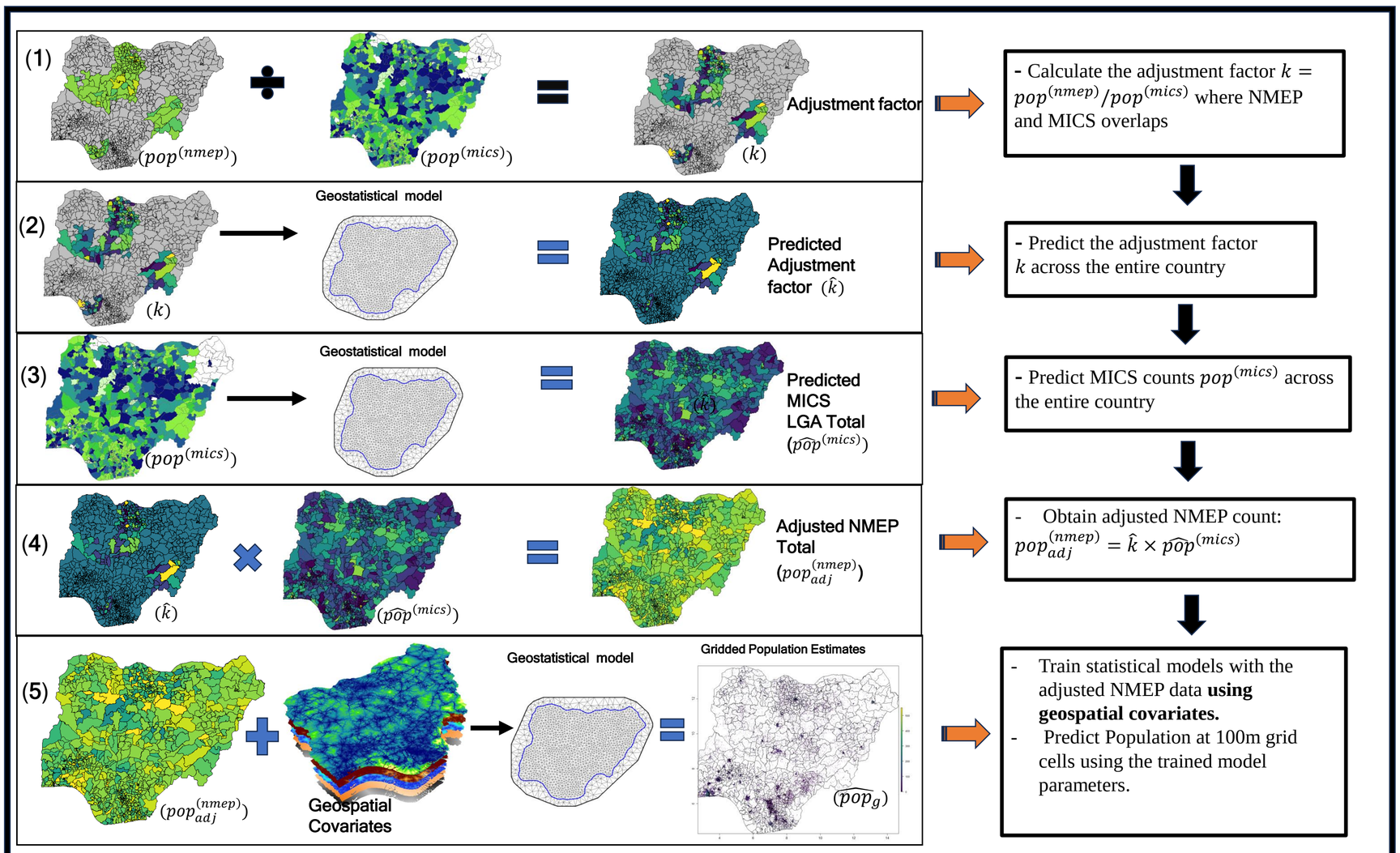

Figure 1. A schematic representation of the data augmentation steps

RESULTS:

- A model-based data augmentation scheme was implemented for NMEP data taking advantage of the large spatial coverage of the MICS data.
- Cross-validated model fit metrics (in-sample and out-sample) of the best fit model show evidence of high predictive ability (Table 1) with a predicted counts vs observed counts correlation value of not less than 78%. .
- Estimates of population were then obtained across all the grid cells in the country (Figure 3)

Table 1. Cross-validated model fit metrics

Cross-Validation	MAE	RMSE	BIAS	COR
In-sample	167897.7	289569.6	15938.6	0.78
Out-sample	166541.2	275143.6	19435.6	0.79

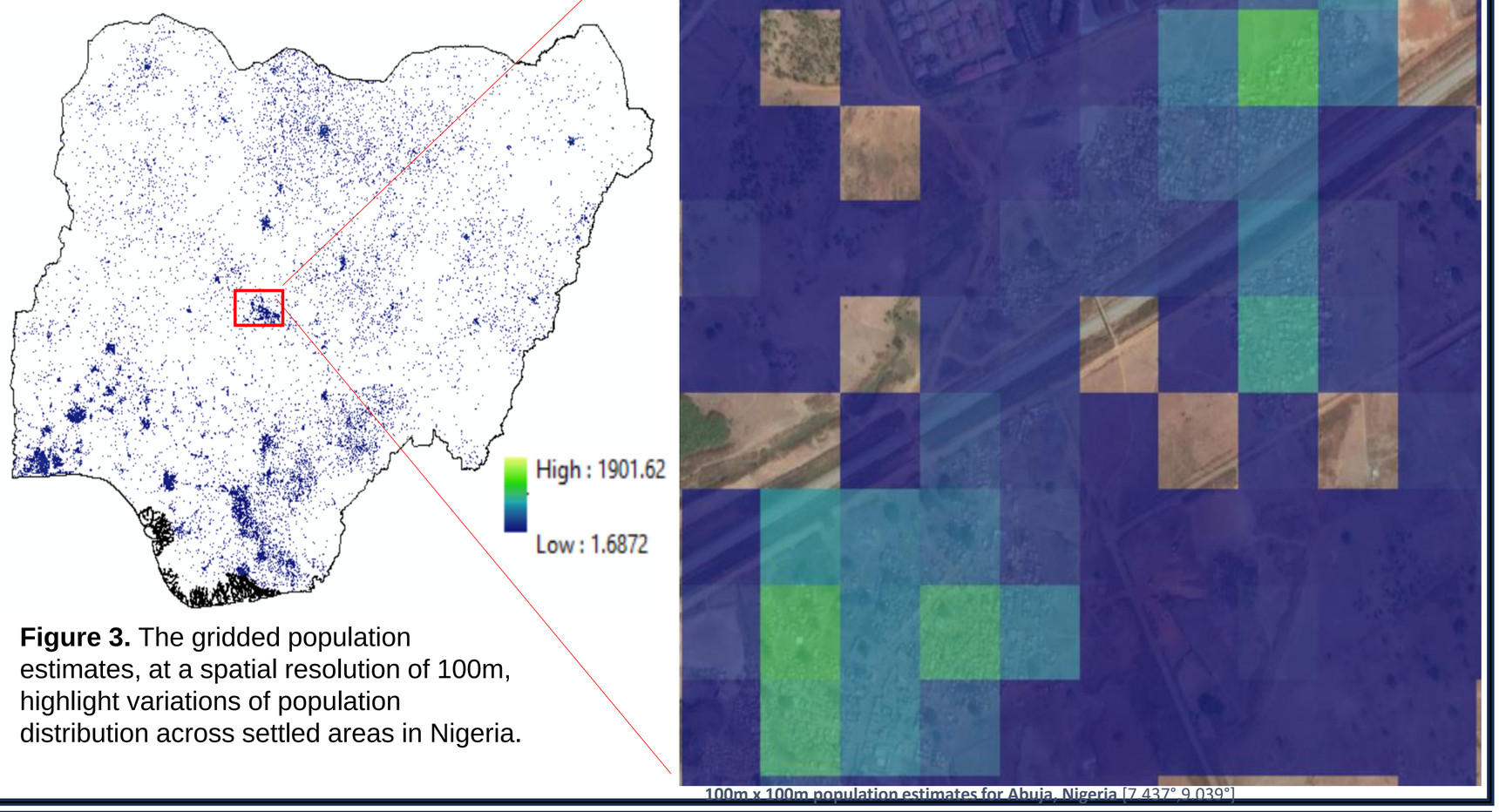

Figure 3. The gridded population estimates, at a spatial resolution of 100m, highlight variations of population distribution across settled areas in Nigeria.

BIBLIOGRAPHY:

[1] UNFPA, The Value of Modelled Population Estimates for Census Planning and Preparation. 2020b, UNFPA: New York, USA. https://www.unfpa.org/sites/default/files/resource-pdf/Technical-Guidance-Note_Vaue_of_Modeled_Pop_Estimates_in_Census_FINAL.pdf

[2] Leasure, D. R., W. C. Jochem, E. M. Weber, V. Seaman and A. J. Tatem (2020). "National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty."Proceedings of the National Academy of Sciences: 201913050. DOI: 10.1073/pnas.1913050117. https://www.pnas.org/doi/pdf/10.1073/pnas.1913050117

[3] Nnanatu C., Yankey O., Abbott T. J., Lazar A. N., Darin E., Tatem A. J. (2022) Bottom-up gridded population estimates for Cameroon (2022), version 1.0. https://dx.doi.org/10.5258/SOTON/WP00662

[4] Center for International Earth Science Information Network (CIESIN), Columbia University. 2024. GRID3 NGA - Settlement Extents v3.1. New York: GRID3. https://doi.org/10.7916/x9xg-e262

[5] Rue, Havard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. " Journal of the Royal Statistical Society, Series B 71 (2): 319–92. https://doi.org/10.1111/j.1467-9868.2008.00700.x

[6] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Acknowledgements

This work was part of the GRID3 – Phase 2 Scaling project, with funding from the Bill & Melinda Gates Foundation (INV-044979). Project partners included GRID3, the Center for International Earth Science Information Network (CIESIN) in the Earth Institute at Columbia University and WorldPop at the University of Southampton. We thank NMEP and UNICEF for sharing their survey data.