SUPPLEMENTARY APPENDIX

Supplement to: Thomson DR, Rhoda DA , Tatem AJ, Castro MC. Gridded population survey sampling: A systematic review and strategic research agenda.

Top-down gridded population datasets

A number of gridded population datasets are available across low- and middle-income countries (LMICs) (Table S1). All datasets available at the time of this writing are derived from "top-down" models which disaggregate census or other population counts into small grid cells. These models produce "pycnophylactic" estimates such that the cell counts re-aggregate to the counts of input administrative data.[1] Generally input population counts are adjusted to UN projections before modelling,[2] however this still means that countries with the greatest need for gridded population sampling have the least accurate top-down gridded population datasets. Additional factors influence the accuracy of top-down modelled population estimates, namely the aggregation scale of the input census data, modelling approach, and area of the output grid cell.

Scale of input data. The most important factor is the aggregation scale of the model input population data (e.g., census).[3] This is intuitive – the more detailed and accurate the input dataset, the more accurate the output estimates will be in small grid squares.

Modelling approach. The simplest top-down models assume that the population is spread evenly across grid cells within administrative units (e.g. GPWv4[4,5]) or landcover types (GHS-POP,[6,7] HRSL,[8] ESRI WPE[9]). These modelling techniques are more mechanical than statistical, and thus do not result in estimates of model error. These models only produce reasonably accurate cell-level estimates if a highly accurate dataset of built-up areas is used to mask unpopulated areas, and the input population data is disaggregated and recent,[3] all of which are rare in LMICs.

Complex modelling techniques using multiple spatial covariates (e.g., WorldPop,[10,11] WorldPop-Global,[10,11] LandScan-Global,[12] Demobase[13]) are employed to produce substantially more accurate gridded population estimates. WorldPop and WorldPop-Global are free, publicly available 100x100m datasets of the residential population based on a regression tree machine-learning method, and are accompanied by prediction errors.[10] Error estimates are derived at the geographic scale of the input population data by reserving a subset of the input data for comparison against the model output. Neither WorldPop nor WorldPop-Global mask built-up areas, thus they produce small, non-zero population predictions in deserts, savannahs, and forests (e.g., 0.0001 persons per cell). Differences between these datasets are that WorldPop is available for every fifth year in most LMICs based on all available spatial covariates, while WorldPop-Global has annual estimates for all countries and is modelled from a reduced set of covariates that are available globally.

Demobase is a free, publicly available 100x100m dataset of the residential population in three countries based on semi-automated classification of high- and medium-resolution satellite imagery, with prediction errors at the scale of the input population data.[13] LandScan-Global is a 1x1km dataset of the "ambient" population; a 24-hour average of daytime commuter population and night-time residential population.[12] Neither the source data nor the model code are released publicly, and most users pay a fee to access the data. This dataset is derived with a smart interpolation approach and model error estimates are not provided.[12]

A common issue across all top-down gridded population datasets is they allocate population to areas that show human activity according to satellite imagery and GIS datasets. This means that population estimates are sometimes allocated at airports, universities, factories, and government buildings. Misallocation of population in gridded population models can be reduced by including covariates that represent points of interest and infrastructure where people tend not to live.

Area of output grid cells. The geographic size of the output cells influences accuracy at the cell-level. Generally, estimates in smaller cells have greater uncertainty, and accuracy improves with cell size. For household survey sampling, however, cell-level accuracy must be balanced against feasibility of cell size for fieldwork; in dense urban contexts, a 100x100m grid cell might contain 1000s of people. Gridded population datasets with small cells are easy to aggregate into larger units, however, complex methods are required to disaggregate cells that are too populous.[14] WorldPop and WorldPop-Global offer the most flexibility in terms of small cell size, high model accuracy,[10] and full coverage in LMICs resulting in their use in numerous surveys.[15–19] The older LandScan-Global dataset was used in a number of early gridded population surveys.[20–23]

Bottom-up gridded population datasets

To generate gridded population estimates in countries without a recent or accurate census, "bottom-up" models are currently under development to estimate population counts based on recent micro-census samples rather than full censuses.[24] These models draw geostatistical relationships between population density in the micro-census units and settlement types and other spatial covariates to predict population counts in un-sampled areas of the country. These census-independent gridded population estimates are soon expected for multiple LMICs from the GRID3 and LandScan-HD projects.[25,26]

Table S1. Summary of gridded population datasets available for LMICs

Approach	Name	Coverage	Resolution	Years	Available
Top-down	GPWv4 [4,5]	Global	~1x1km	2000-2020	Yes – free
	GHS-POP [7,28]	Global	250x250m	1975-2015	Yes – free
	HRSL (Facebook) [8]	33 countries	~30x30m	2015	Yes – free
	ESRI WPE [9]	Global	150x150m	2016	Yes – paid
	LandScan-Global [12]	Global	~1x1km	2000-2017	Yes – paid
	Demobase [13]	3 countries	~100x100m	2003-2013	Yes – free
	WorldPop-Land Cover[32,33]	57 countries	~100x100m	2010-2015	Yes – free
	WorldPop-Random Forest [10,11]	69 countries	~100x100m	2010-2020	Yes – free
	WorldPop-Global [10,11]	Global	~100x100m	2000-2020	Yes – free
Bottom-up	LandScan HD	21 countries	~100x100m	varying	No (2019)
	GRID3	5 countries	~100x100m	varying	No (2019)

Gridded population sample frame attributes

Gridded population datasets are not provided with urban/rural classes, administrative unit names, or estimates of sub-populations because they are designed to be aggregated into any desired spatial unit. Publicly available datasets can be used to classify a gridded population dataset with a geographic information system (GIS) (e.g., ArcGIS, QGIS) or statistical program (e.g., R, Python). Urban/rural datasets include the Global Urban Footprint (GUF)[27] dataset of 85x85m grid cells classified as built-up or not built-up, and the Global Human Settlement GHS-SMOD[28] dataset of 30x30m grid cells classified as high-dense urban, low-dense urban, rural, and unsettled based on the GHS-POP population density and GHS-BUILT-UP built areas datasets. Administrative boundaries are available as shapefiles through a number of initiatives including GADM,[29] UN-SALB,[30] and MapLibrary.[31]

References

1. Tobler WR. Smooth Pycnophylactic Interpolation for Geographical Regions. *J Am Stat Assoc*. 1979;74:519-536.
2. United Nations Department of Economic and Social Affairs. World Population Prosects. Population Division. https://population.un.org/wpp/Download/Standard/Population/. Published 2017. Accessed May 20, 2019.
3. Hay S, Noor A, Nelson A, Tatem A. The accuracy of human population maps for public health application. *Trop Med Int Heal*. 2005;10(10):1073-1086. doi:10.1111/j.1365-3156.2005.01487.x
4. Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4). Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4F47M2C. Published 2016. Accessed February 19, 2017.
5. Doxsey-Whitfield E, MacManus K, Adamo SB, et al. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. *Pap Appl Geogr*. 2015;1(3):226-234. doi:10.1080/23754931.2015.1014272
6. European Commission. GHS Population Grid. http://ghsl.jrc.ec.europa.eu/ghs_pop.php. Published 2017. Accessed May 18, 2017.
7. Pesaresi M, Ehrlich D, Florczyk AJ, et al. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014*. Ispra, Italy: European Union; 2016. doi:10.2788/253582
8. Facebook Connectivity Lab and Center for International Earth Science Information Network—CIESEN—Columbia University. High Resolution Settlement Layer (HRSL). Source imagery for HRSL 2016 DigitalGlobe. https://ciesin.columbia.edu/data/hrsl/. Published 2016. Accessed March 10, 2017.
9. Frye C, Nordstrand E, Wright DJ, Terborgh C, Foust J. Using Classified and Unclassified Land

Cover Data to Estimate the Footprint of Human Settlement. *Data Sci J*. 2018;17:1-12. doi:10.5334/dsj-2018-020

10. Stevens FR, Gaughan AE, Linard C, Tatem AJ. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS One*. 2015;10(2):e0107042. doi:10.1371/journal.pone.0107042

11. WorldPop. Data Availability. http://www.worldpop.org.uk/data/data_sources. Published 2017. Accessed February 2, 2017.

12. Dobson JE, Brlght EA, Coleman PR, Worley BA. LandScan: A Global Population Database for Estimating Populations at Risk. *Photogramm Eng Remote Sensing*. 2000;66(7):849-857.

13. Azar D, Engstrom R, Graesser J, Comenetz J. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sens Environ*. 2013;130:219-232. doi:10.1016/j.rse.2012.11.022

14. Chew RF, Amer S, Jones K, et al. Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery. *Int J Health Geogr*. 2018;17(1):1-17. doi:10.1186/s12942-018-0132-1

15. Elsey H, Thomson D, Lin R, Maharjan U, Agarwal S, Newell J. Addressing Inequities in Urban Health: Do Decision-Makers Have the Data They Need? Report from the Urban Health Data Special Session at International Conference on Urban Health Dhaka 2015. *J Urban Heal*. 2016;93(3):526-537. doi:10.1007/s11524-016-0046-9

16. Elsey H, Poudel AN, Ensor T, et al. Improving household surveys and use of data to address health inequities in three Asian cities: protocol for the Surveys for Urban Equity (SUE) mixed methods and feasibility study. *BMJ Open*. 2018;8(11):e024182. doi:10.1136/bmjopen-2018-024182

17. Vulnerability Analysis and Mapping (VAM) unit. *Urban Essential Needs Assessment in the Five Communes of Kimbanseke, Kinsenso, Makala, N'sele and Selembao (Kinshasa)*. Rome Italy; 2018.

18. Pape U, Wollburg P. *Estimation of Poverty in Somalia Using Innovative Methodologies*. Washington DC USA; 2019. http://documents.worldbank.org/curated/en/509221549985694077/Estimation-of-Poverty-in-Somalia-Using-Innovative-Methodologies.

19. GridSample. World Vision International. Case Studies. http://gridsample.org/world-vision-international. Published 2019. Accessed May 12, 2019.

20. Thomson DR, Hadley MB, Greenough PG, Castro MC. Modelling strategic interventions in a population with a total fertility rate of 8.3: a cross-sectional study of Idjwi Island, DRC. *BMC Public Health*. 2012;12(1):959. doi:10.1186/1471-2458-12-959

21. Galway L, Bell N, Sae AS, et al. A two-stage cluster sampling method using gridded population data, a GIS, and Google EarthTM imagery in a population-based mortality survey in Iraq. *Int J Health Geogr*. 2012;11:12.

22. Cajka J, Amer S, Ridenhour J, Allpress J. Geo-sampling in developing nations. *Int J Soc Res Methodol*. 2018;21(6):729-746. doi:10.1080/13645579.2018.1484989

23. Sollom R, Richards AK, Parmar P, et al. Health and human rights in Chin State, Western Burma: A population-based assessment using multistaged household cluster sampling. *PLoS Med*. 2011;8(2):e1001007. doi:10.1371/journal.pmed.1001007

24. Wardrop NA, Jochem WC, Bird TJ, et al. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc Natl Acad Sci*. 2018;0:201715305. doi:10.1073/pnas.1715305115

25. UNFPA, WorldPop, Flowminder, CIESIN. Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). http://www.grid3.org/. Published 2018. Accessed November 21, 2018.

26. Oak Ridge National Laboratories. *LandScan HD: Human Settlement Mapping at Global Scale*. USA; 2015. https://www.youtube.com/watch?v=P84vxTT9Vos.

27. DLR Earth Observation Center. Global Urban Footprint (GUF). http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11725/20508_read-47944/. Published 2017. Accessed February 6, 2017.

28. European Commission. Global Human Settlement City Model (GHS-SMOD). http://ghsl.jrc.ec.europa.eu/faq.php. Published 2017. Accessed February 6, 2017.

29. GADM. Global administrative areas version 2.8. http://www.gadm.org/problems. Published 2015. Accessed March 3, 2017.

30. Geospatial Information Section and Statistics Division. United Nations Second Administrative Level Boundaries (UN-SALB). https://www.unsalb.org/data. Published 2019. Accessed April 23, 2019.

31. Map Maker Ltd. Map Library. http://www.maplibrary.org/library/stacks/Africa/index.htm. Published 2007. Accessed February 6, 2017.

32. Linard C, Gilbert M, Tatem AJ. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*. 2011;76(5):525-538. doi:10.1007/s10708-010-9364-8

33. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS One*. 2013;8(2):e55882. doi:10.1371/journal.pone.0055882