Gridded population survey sampling: A review of the field and strategic research agenda

Dana R Thomson[1,2,*], Dale A Rhoda[3], Andrew J Tatem[1,2], Marcia C Castro[4]

1. Flowminder Foundation, 15-17 Cumberland Place, Southampton SO15 2BG, UK
2. WorldPop, Department of Geography and Environmental Science, University of Southampton, University Road, Southampton SO17 1BJ, UK
3. Biostat Global Consulting, 870 High Street, Worthington OH 43085, USA
4. Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston MA 02115, USA

* Corresponding author

ABSTRACT

Objective: In low- and middle-income countries (LMICs), household survey data are a main source of information for planning, evaluation, and decision-making. Standard surveys are based on censuses, however, for many LMICs it has been more than ten years since their last census and they face high urban growth rates. Over the last decade, survey designers have begun to use modelled gridded population estimates as sample frames. We summarize the state of the emerging field of gridded population survey sampling, focussing on LMICs.

Methods: We performed a systematic review and identified 23 national and sub-national gridded population-based household surveys implemented across 18 LMICs.

Findings: Gridded population surveys used automated and manual approaches to derive clusters from WorldPop and LandScan gridded population estimates. After sampling, many surveys interviewed all households in each cluster or segment, though some sampled households from larger clusters. Tools to select gridded population survey clusters include the GridSample R package, Geo-sampling tool, and GridSample.org. In the field, gridded population surveys generally relied on geographically accurate maps based on satellite imagery or OpenStreetMap, and a tablet or GPS technology for navigation.

Conclusions: For gridded population survey sampling to be adopted more widely, several strategic questions need answering regarding cell-level accuracy and uncertainty of gridded population estimates, the methods used to group/split cells into sample frame units, design effects of new sample designs, and feasibility of tools and methods to implement surveys across diverse settings.

KEY WORDS
census, survey design, household survey, LMIC, WorldPop, LandScan

INTRODUCTION

Household surveys provide insight into the distribution of health, demographics, economics, and behaviours of populations, and are a primary resource for decision-making across low- and middle-income countries (LMICs). Household survey data are used to estimate more than a quarter of the Sustainable Development Goal (SDG) indicators, to generate small area estimates (SAEs) of indicators that support decision-making in decentralized health systems,[1] and inform the distribution development funding to, and within, LMICs. Nevertheless, as the use of household surveys has increased over the last 40 years, data accuracy have likely decayed because survey methods have not changed while population characteristics and behaviours have – drastically.

Survey sampling methods have been mature for decades. The Demographic and Health Surveys (DHS),[2] Multiple Indicator Cluster Surveys (MICS),[3] and Living Standards Measurement Surveys (LSMS)[4] have collectively supported hundreds of surveys in over 130 countries since 1980 using essentially the same methods. They follow a stratified two-stage cluster design. In stage one, census enumeration areas (EAs) are selected with probability proportionate to population size (PPS) from administrative units (e.g., provinces), and a field-based household mapping-listing activity is conducted in each selected cluster. In stage two, households are sampled from the household listing by an impartial central team, and interviewers return to selected households to administer questionnaires.

The last 40 years have seen increases in mobility of LMIC populations, urbanisation, and socioeconomic disparities within cities.[5] The urban poorest include climate and political refugees, seasonal migrants, and rural migrants, as well as multi-generation slum dwellers, street-sleepers, and marginalized minorities.[5] Concurrently, availability of technologies (e.g., mobile phones) and new data (e.g., high-resolution satellite imagery) has rapidly increased, though few new technologies and datasets were incorporated into standard survey practice. This mismatch has resulted in challenges to sample frame and field protocol accuracy.[6,7] Furthermore, the SDGs have increased emphasis on disaggregated indicators,[8] raising concerns about survey designs for accurate SAEs, which we highlight below.

First, countries with the greatest need for household survey data are the least likely to have an appropriate census sample frame. One in four LMICs has not had a census in the last 10 years.[9] High rates of urban growth and mobility in LMICs mean that megacities in Asia, and soon Africa, grow by 1,000 people per day.[10] Since 2000, the average household survey sample frame in LMICs is seven years old, with some surveys using 15 (Pakistan) and 30 (DR Congo) year old sample frames.[11] Vulnerable populations are most likely to be excluded from surveys due to an outdated sample frame because population growth is greater among lower-income households, and they are more likely to be undercounted in censuses.[7]

Second, standard survey methods, largely developed for rural settings 40 years ago, struggle to sample mobile and vulnerable households accurately.[12] Delays between the household mapping-listing and interviews means that mobile and vulnerable households are more likely to be counted as non-responders or be under-listed. Furthermore, the mappers-listers – responsible for generating the final household sample frame – have short interactions with residents, and may make assumptions about the number of households in dwellings.[11] In LMICs that do not have geocoded census EA boundaries, mapping-listing activities rely on hand-sketched maps and subjective descriptions of EA boundaries by local leaders, leading to potential bias.

Third, in recent years, funders and decision-makers have pushed for important outcomes to be measured at smaller administrative scales (e.g., district) for policy planning and evaluation.[1,8]

Increased availability of satellite imagery has enabled outcomes to be modelled at fine-scale using geostatistical SAE techniques.[13] However, SAEs based on the stratified two-stage PPS design tend to have large uncertainty in sparsely-sampled rural areas and in heterogeneous urban settings.[13]

Gridded population estimates can provide more up-to-date and detailed population counts than outdated census frames, permit new survey designs such as one-stage sampling to eliminate time between listing and interviews, and facilitate spatial oversampling to improve survey-based SAEs. Gridded population datasets represent estimates of the population in grid cells as small as a city block. “Top-down” datasets disaggregate census counts to grid cells, while “bottom-up” estimates are based on micro-census population counts.[14] Currently, nine sources of “top-down” estimates are available across multiple LMICs from WorldPop,[15–18] LandScan,[19] Columbia-CIESIN,[20,21] European Commission,[22,23] ESRI,[24] Facebook,[25] and US Census,[26] and are compared elsewhere.[27] Two sources of “bottom-up” estimates are in production by LandScan[28] and GRID3.[29] In gridded population sampling, grid cells are aggregated into clusters of a desired population size, and used in place of census EAs.

In this paper, we review the datasets, tools, and methods used in gridded population surveys, and outline a research agenda that would equip survey designers to decide when gridded population sampling can be viable and preferable to census-based sampling. The aim is to encourage new research and practices that improve the accuracy of survey data and, ultimately, improve targeting of resources toward mobile and vulnerable populations.

METHODS

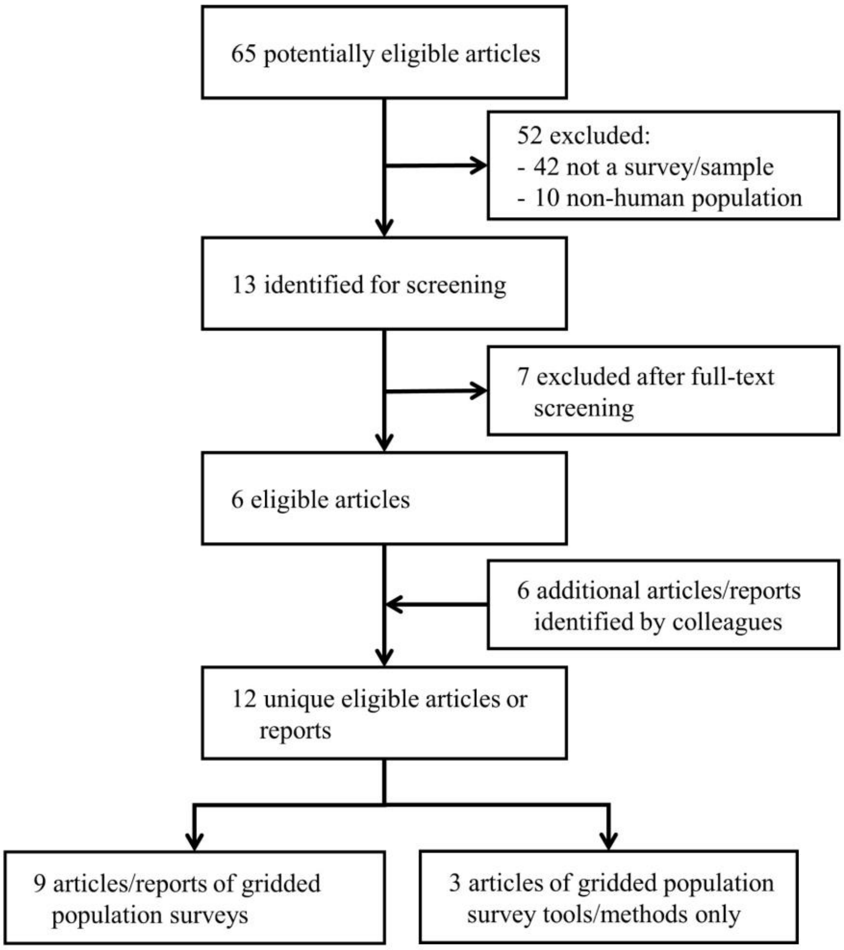
We conducted a systematic review in Scopus using the terms: (“gridded” OR “landscan” OR “worldpop” OR “gpw” OR “ghs-pop” OR “hrsl” OR “wpe” OR “demobase”) AND (“population” OR “household”) AND “survey”. Article abstracts were independently screened by co-authors DRT and DAR and retained if they referred to sampling of human populations. We additionally solicited reports, websites, and articles from colleagues. DRT performed a full-text review of all screened articles and reports, and retained those that described a method, tool, or survey based on gridded population data. Retained publications were reviewed for sample frame, sample design, tools, and protocols used. A strategic research agenda was iteratively developed among co-authors with feedback from survey experts.

FINDINGS

The systematic review resulted in 65 potentially eligible and six eligible articles describing a gridded population survey, tool, or method. Solicitation of documents from colleagues resulted in six additional resources (Figure 1). Although we did not restrict our search by geography, all identified gridded population surveys were located in LMICs and were motivated by an outdated or unavailable census. This literature review resulted in 23 surveys across 18 LMICs: Bangladesh,[11,30,31] Brazil,[31] Colombia,[31] DR Congo,[32,33] Guatemala,[31] Ghana,[31] India,[31] Iraq,[34] Kenya,[31] Mozambique,[33] Myanmar,[35] Nepal,[11,12,30,33] Nigeria,[31] Rwanda,[31] Somalia,[36,37] Thailand,[31] Uganda,[31] and Vietnam[11,30] (Table 1). Three resources described tools or methods for selecting gridded population survey clusters.[38–40]

 doi:10.20944/preprints201911.0072.v1

Figure 1. Systematic review selection criteria



Most sample frames in early surveys were derived from LandScan-Global 1x1km estimates,[31,32,34,35] while most recent surveys derived sample frames from WorldPop 100x100m estimates (Table 1).[11,12,30,33,36] The final selection of households followed two approaches. First, all eligible households in a cluster or segment were interviewed (called one-stage hereafter). Second, households were sampled within clusters or segments before interviewing (called two-stage hereafter). We noted whether household sampling was conducted with a robust probability method (complete mapping-listing of households before sampling households), or a non-probability method (random-walk or spin-the-pen).[41]

Nineteen of the 23 surveys followed a one-stage design for one of four reasons. First, one-stage sampling saved time and costs by eliminating, or reducing, the mapping-listing activity.[30,33] Second, it restricted fieldwork to one visit in insecure or hard-to-reach areas.[33,34] Third, it provided a simple field protocol and required less training of interviewers which was assumed to ensure higher data quality.[32] Fourth, in complex, dynamic urban environments, it removed the time lag between mapping-listing and interviewing, guarding against under-listing of mobile or vulnerable households.[12,30] One survey compared one- and two-stage gridded population sampling in Kathmandu, Nepal, and found that when interviewers (one-stage) rather than the mapper-listers (two-stage) performed the household listing, non-family and single-adult households were more likely to be identified, possibly because interviewers spent more time building rapport with residents.[11] This study also found lower response rates in one-stage samples.[11]

Table 1. Summary of gridded population surveys including their designs

Country & Year (if reported)	Design: Coverage, Strata, Stages	Cluster & Household Sample Size	Gridded population dataset	Target Population, Main topic(s)
DR Congo 2010[32]	Idjwi Island, no strata, one-stage	50 clusters, 2078 HHs	2001 LandScan-Global	All women age 18-50, Maternal and child health
Myanmar 2010[35]	Chin state, urban/rural strata, multi-stage (spin-the-pen)	90 clusters, 720 HHs	2005 LandScan-Global (rural only)	Household head age 18+, Health, human rights
Iraq 2011[34]	National, governorates strata, multi-stage (random-walk)	100 clusters, 1960 HHs	2008 LandScan-Global	Household head age 18+, Mortality
Bangladesh[31]	National, division X urbanicity strata, one-stage	148 clusters, 3296 HHs	2012-2016 LandScan-Global	Adult age 18+, Topics not reported
Brazil[31]	National, region X poverty strata, one-stage	149 clusters, 3652 HHs		
Colombia[31]	National, region X poverty strata, one-stage	152 clusters, 2706 HHs		
Colombia[31]	National, region X poverty strata, one-stage	152 clusters, 3037 HHs		
Ghana[31]	National, region X poverty X urbanicity strata, one-stage	151 clusters, 3113 HHs		
Guatemala[31]	National, departament X urbanicity strata, one-stage	211 clusters, 3057 HHs		
India[31]	Three states, district X urbanicity strata, one-stage	467 clusters, 10,824 HHs		
Kenya[31]	National, province X poverty strata, one-stage	143 clusters, 3364 HHs		
Nigeria[31]	National, region X poverty strata, one-stage	147 clusters, 3042 HHs		
Rwanda[31]	National, province X poverty strata, one-stage	150 clusters, 3096 HHs		
Thailand[31]	National, region X poverty strata, one-stage	150 clusters, 3136 HHs		
Thailand[31]	National, region X poverty strata, one-stage	150 clusters, 3275 HHs		
Uganda[31]	National, region strata, one-stage	146 clusters, 3075 HHs		
Nepal 2015[12]	Kathmandu Valley, no strata, multi-stage	90 clusters, 1,310 HHs (planned)	2014 WorldPop-Random Forest	Woman age 18+, Maternal and child health
Mozambique 2017[33]	Six districts, district strata, one-stage	234 clusters, 4998 HHs	2017 WorldPop-Random Forest	Caregiver of child age 12-18, Child health
DR Congo 2017[33]	Kinshasa, communes strata, one-stage	210 clusters, 1,850 HHs	Bespoke derived from administrative records	Household head, Food insecurity
Somalia 2017[36,37]	National, region X urbanicity, multi-stage	405 clusters, 6,284 HHs	Modified 2015 WorldPop-Land Cover	Household head, Economic
Nepal 2017[11,30,33]	Kathmandu valley, no geographic strata, one-stage & multi-stage	60 clusters, 1200 HHs	2017 WorldPop-Random Forest	Adult age 18+, Economic, non-communicable disease
Bangladesh 2018[11,30]	Two communities, community strata, one-stage	20 clusters, 400 HHs	2020 WorldPop-Random Forest	
Vietnam 2018[11,30]	Long Bien District, no strata, one-stage	20 clusters, 400 HHs		

Four tools and numerous ad-hoc geographic information system (GIS) approaches were described to select gridded population survey clusters (Table 2), and resulted in various forms of a gridded population sample frame, visualized in Figure 2. First, the open-source GridSample R package was released by Thomson and colleagues in 2016[38] and has been used in six sub-national surveys.[12,30,33] It treats the gridded population dataset as the sample frame and selects grid cells with PPS allowing for stratification, oversampling in urban/rural domains, and spatial oversampling.[38] The algorithm runs on a personal computer and is limited by the computer's memory. All datasets must be pre-processed and specified by the user, allowing use of any gridded population but also requiring GIS and/or R programming skills. The algorithm enables optional "growth" of clusters to a minimum population size or maximum area by randomly adding neighbouring cells after selection of "seed" cells with PPS. While this process results in clusters with roughly consistent population size for improved fieldwork, the population counts in the "grown" clusters do not reflect the population counts used for sample selection, and thus may skew sample weights.[38] The output is a shapefile of cluster boundaries, with attributes of estimated population counts.

Table 2. Comparison of tools for gridded population sampling

Feature	GridSample R	Geo-sampling	Ad-hoc GIS	GridSample2.0	GridSample.org
Public	Yes	No	Yes	Yes	Yes
Free	Yes	No	Some	Yes	Yes
Skill level required	Advanced	Advanced	Advanced	Advanced	Basic
User selects the sample	Yes	No	Yes	Yes	Yes
Gridded pop	Any	LandScan-Global	Any	Any	WorldPop-Global
Preloaded/ provided data	No	Yes	Some	No	Yes
Pre-forms clusters	No	Yes	Some	Yes	Yes
Citations	11,12,30,33,38	31,39	32,34–36,40	github.com/Flowminder/ GridSample2.0	GridSample.org

Second, the Geo-sampling survey tool was created by RTI and used in 13 national and sub-national surveys.[31] It is designed for use with 1x1km grid cells, and supports a multi-stage stratified sampling approach. After administrative units are sampled with PPS, 1x1km cells are sampled with PPS. To improve fieldwork, 1x1km cells with fewer than 250 persons are excluded, potentially biasing the sample toward higher-density populations. The sampled 1x1km cells are partitioned into 150m, 100m or 50m grid cells depending on population density. Next, a deep-learning residential scene classification model is used to identify and exclude small cells without settlement, and disaggregate the 1x1km population to remaining small cells. Finally, three of the small cells are selected at random for a one-stage sample.[39] Clients are provided a shapefile of the final cluster boundaries and population counts.

Third, many gridded population surveys developed ad-hoc approaches to sampling using GIS software, such as ArcGIS. Galway et al. (2012) sampled 1x1km cells with PPS, then randomly selected one household in one building and performed a random walk.[34] Thomson et al. (2012) converted 1x1km population counts to random points, selected points at random, manually delineated clusters within cells around selected points, and performed a one-stage sample.[32] Muñoz and Langeraar (2013) proposed an approach for 1x1km cells, though it is unclear if a survey followed.[40] In this approach, 1x1km cells are aggregated to 3x3km grid cells and sampled with PPS.

Figure 2. Visualisation of approaches to derive a gridded population sample frame

Then 1x1km grid cells are combined within selected 3x3km cells to achieve a minimum population and sampled with PPS. Next, they select a 1x1km (or larger) area and manually delineate segments of approximately 100 households each. One segment is randomly selected, households are listed via a mapping-listing activity, and finally a sample of households is selected.[40] Sollom et al. (2011) joined 1x1km gridded population estimates to rural village point locations and sampled points with PPS, and then used spin-the-pen to sample households in the field.[35] Finally, Qader et al. (2019) used gridded population estimates to update census EA counts in urban areas where EA boundaries were available, and used a quadtree method to create different sized grid cells with approximately the same population each in rural areas.[36] The combined frame was sampled with PPS before manually segmenting and randomly selecting one household per segment.[36]

Fourth, GridSample.org is a free web-based tool released in 2019 that runs the open-source GridSample2.0 algorithm developed by Flowminder Foundation. It provides a point-and-click interface, preloaded datasets, and guidance to enter parameters and select clusters for a gridded population survey. It also leverages gridEZ, a publicly-available algorithm, to group cells into clusters before sampling. Preloaded datasets include WorldPop-Global 100x100m gridded population estimates, GADM administrative boundaries, and GHS-SMOD urban/rural boundaries. All surveys are implicitly stratified by level of urbanicity; stratification and spatial oversampling are supported; and custom coverage, strata, or sample frame boundaries can be uploaded by users. GridSample.org is designed for low-bandwidth settings, running sample selection remotely on a super-computer. The user is emailed a shapefile of cluster boundaries, population estimates to calculate sample weights, and a report.

A range of simple-to-advanced tools has been used to implement gridded population surveys. Lower-tech field tools include use of paper maps displaying cluster boundaries over satellite imagery in Google Earth, and paper listing forms and questionnaires.[32–34] Higher-tech field tools include tablet-based applications for navigation,[11,31] paper field maps designed in GIS,[11,12,33,34,36] and tablet-based household listing and/or questionnaires.[11,12,31,33] Satellite imagery was essential to all gridded populations surveys to manually segment along roads, rivers, and other features,[30,32,40] and as a field map base layer.[31–34,36] In some surveys, satellite imagery was used to digitize building footprints and roads in OpenStreetMap which was then displayed as a field map base layer.[12,30] Many teams included points of interest from OpenStreetMap or GPS coordinates of recognizable intersections/structures on field maps to aid navigation.[12,30,32,36]

DISCUSSION

The successful implementation of gridded population sample surveys across a variety of settings bodes well for this emerging field. However, a survey statistician considering whether to recommend an outdated census-based frame or a gridded population frame is faced with questions about sample frame accuracy, methods to form and select sample frame units, and optimal survey designs. Next, we outline a research agenda to equip survey designers to identify situations where gridded population sampling can be a feasible and trustworthy option. The agenda shows key stages of a gridded population survey and available options (Figure 3).

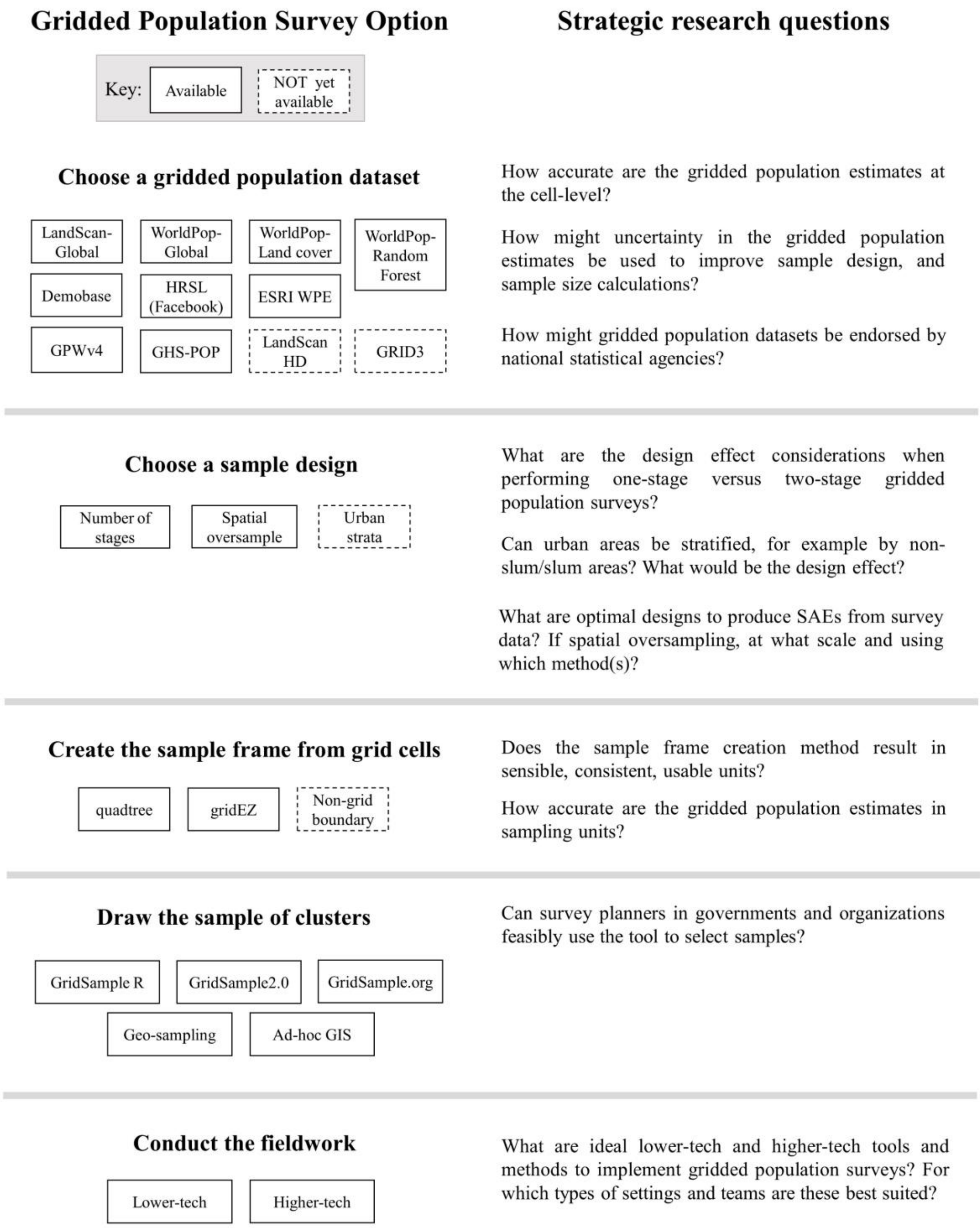
Choose gridded population

In Supplement 1, we summarize available gridded population datasets and what is known about their accuracy. “Top-down” datasets that restrict estimates to settled areas are likely to underestimate rural, and overestimate urban, populations because small settlements are often undetected in the settlement layer. Conversely, datasets that estimate population in all landmasses likely overestimate rural, and underestimate urban, population because fractions of the population are allocated to unsettled cells. Factors that affect accuracy include the model, accuracy, and aggregation of the input population dataset, whether residential or “ambient” population is modelled, accuracy and type of covariates, and area of the cell in which population is estimated.[27]

A major gap is that cell-level accuracy is not known for any “top-down” gridded population datasets. To assess accuracy, a recent census disaggregated to household locations would be needed, though this is rarely, if ever, available. The next best option is comparison of modelled gridded population estimates with micro-census counts from a sample of areas. Aggregated household listings from a recent geo-located household survey might serve this purpose, but to our knowledge, data sharing agreements have not been investigated or defined. Simulated household-level datasets are a third option.[42]

Presently, some “top-down” datasets include model errors at the scale of the input population dataset based on internal validation. New “bottom-up” datasets are expected to include cell-level uncertainty estimates. When those become available, survey designers will want to consider how uncertainty estimates might be used to improve sample designs or sample size calculations. In addition, DHS, MICS, LSMS, and other surveys are distributed via national statistical offices, and thus their sample frames hale from official sources. Processes are needed for national statistical agencies to engage with gridded population dataset production so that official endorsements might be made.

Figure 3. Strategic research agenda to determine when to use gridded population sampling



Choose sample design

One-stage sample designs may prove to be faster and cheaper than two-stage designs, and more accurately sample vulnerable urban populations, however, there can be a counter-balancing detriment of higher survey design effects due to more variable numbers of respondents per cluster, greater within-cluster homogeneity, and lower response rates. For survey designers to assess these trade-offs and to select a sample size that will meet stakeholders' goals for budget, timeline, and statistical precision, they need reliable projections of likely design effects in one-stage samples. A simulation study found that nearly twice as many one-stage clusters are needed to achieve the same precision as a two-stage survey, holding constant the number of respondents per cluster.[43]

Also, as urban settlement classification becomes increasingly possible,[44] survey designers need to understand how within-urban stratification affects the various sample designs used in gridded population, and other, surveys. With no way to stratify urban populations, all surveys are at risk of under-sampling or omitting slums and other vulnerable populations. In addition, research is needed to balance survey designs that can support both precise design-based estimation of outcomes and precise SAEs of indicators at fine geographic scales.[45]

Create sample frame

Existing sample frame approaches do not result in boundaries that are recognizable on the ground. Improved methods are needed to use natural features such as rivers and roads to delineate cluster boundaries from gridded population data. Survey designers need to be confident that clusters will yield the right number of eligible respondents and have a geographic area that can be canvassed by a field team in the time budgeted for fieldwork.

Draw sample

Several gridded population sampling tools and approaches are available, and their feasibility is influenced by cost, transparency of the methods, clarity of documentation, and usability by survey design professionals in government agencies and organizations who may not have advanced programming and GIS skills.

Conduct fieldwork

The emerging field of gridded population survey sampling should recommend tools and protocols for both lower- and higher-tech settings. Uniquely, gridded population surveys rely on access to up-to-date high-resolution satellite imagery (0.5m) for fieldwork. This is less of a challenge in urban areas worldwide thanks to Google Earth, Bing, and other free sites. However, imagery resolution in rural areas of LMICs is quite variable, with images sometimes being several years old. As a result, it would be difficult to implement gridded population surveys in areas of heavy forest or cloud cover.

CONCLUSION

Organizations with skills in GIS and digital tools can successfully implement surveys with gridded population sample frames, which have the potential to yield samples that are more representative of mobile and vulnerable respondents than outdated census-based frames. However, census-based frames are likely to be considered a safe choice by many survey designers because censuses have long been the standard and their limitations are commonly accepted. To recommend a gridded population frame would involve risks and rewards that are currently difficult to quantify. New tools are needed to evaluate gridded population datasets and frames in specific country contexts, and to facilitate low-burden survey implementation. There are opportunities to develop tools for nearly every stage of survey planning and implementation, which ultimately will improve the accuracy of survey data.

REFERENCES

1. Victora CG, Black RE, Boerma JT, Bryce J. Measuring impact in the Millennium development Goal era and beyond: A new approach to large-scale effectiveness evaluations. Lancet. 2011;377:85-95.
2. ICF International. Demographic and Health Survey sampling and household listing manual. Calverton, MD USA; ICF International: 2012. https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf. Accessed 15 Oct 2019.
3. Multiple Indicator Cluster Surveys. MICS5 tools: manual for mapping and household listing. New York; UNICEF: 2013. http://mics.unicef.org/tools?round=mics5. Accessed 15 Oct 2019.
4. Grosh ME, Munoz J. A Manual for planning and implementing the Living Standards Measurement Study survey. LSMS Working Paper No. 126. Washington, DC USA: The World Bank; 1996.
5. UN-Habitat. Urbanization and development: emerging Futures. World Cities Report 2016. Nairobi, Kenya: UN-Habitat; 2016. http://wcr.unhabitat.org/wp-content/uploads/2017/02/WCR-2016-Full-Report.pdf. Accessed 15 Oct 2019.
6. Cobham A. Uncounted: Power, inequalities and the post-2015 data revolution. Development. 2014;57:320-337.
7. Lucci P, Bhatkal T, Khan A. Are we underestimating urban poverty? World Dev. 2018;103:297-310.
8. Golding N, Burstein R, Longbottom J, Browne AJ, Fullman N, Osgood-Zimmerman A, et al. Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. Lancet. 2017;390:2171-2182.
9. UN Statistics Division (UNSD). 2020 world population and housing census programme. New York, NY USA: UNSD; 2019. https://unstats.un.org/unsd/demographic-social/census/censusdates/. Accessed 15 Oct 2019.
10. UN Department of Economic and Social Affairs (UN-DESA). World urbanization prospects: the 2018 revision. https://population.un.org/wup/DataQuery/. New York, NY USA: UN-DESA; 2018. Accessed 15 Oct 2019.
11. Thomson DR, Bhattarai R, Khanal S, Manandhar S, Dhungel R, Gajurel S, et al. Feasibility of innovative tools and methods to improve household survey accuracy in complex urban settings: Multiple methods analysis of the Surveys for Urban Equity (SUE) project in Kathmandu, Dhaka, and Hanoi. [Preprint]. 2019:[21 p.] Available from: www.preprints.org/manuscript/201910.0320/v1. Accessed 28 Oct 2019.
12. 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;93:526-37.
13. Burgert-Brucker CR, Dontamsetti T, Gething PW. The DHS Program's modeled surfaces spatial datasets. Stud Fam Plann. 2018;49:87-92.
14. Wardrop NA, Jochem WC, Bird TJ, Chamberlain HR, Clarke D, Kerr D, et al. Spatially disaggregated population estimates in the absence of national population and housing census data. Proc Natl Acad Sci. 2018;115(14):3529-37.
15. Linard C, Gilbert M, Tatem AJ. Assessing the use of global land cover data for guiding large area population distribution modelling. GeoJournal. 2011;76:525-38.
16. 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:e55882.
17. 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:e0107042.

18. WorldPop. Data availability. http://www.worldpop.org.uk/data/data_sources. Published 2019. Accessed 15 Oct 2019.
19. Dobson JE, Brlght EA, Coleman PR, Worley BA. LandScan: a global population database for estimating populations at risk. Photogramm Eng Remote Sensing. 2000;66:849-857.
20. Doxsey-Whitfield E, MacManus K, Adamo SB, Pistolesi L, Squires J, Borkovska O, 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;1:226-34.
21. Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4). Socioeconomic Data and Applications Center (SEDAC). New York, NY USA: CIESIN; 2016. http://dx.doi.org/10.7927/H4F47M2C. Accessed 15 Oct 2017.
22. European Commission – Joint Research Centre (EC-JRC). Global Human Settlement Population (GHS-POP). https://ghsl.jrc.ec.europa.eu/ghs_pop2019.php. Published 2019. Accessed 15 Oct 2019.
23. Pesaresi M, Ehrlich D, Ferri S, Florczyk AJ, Freire S, Halkia M, 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 Commission; 2016.
24. 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.
25. Facebook Connectivity Lab and Center for International Earth Science Information Network—CIESIN—Columbia University. High Resolution Settlement Layer (HRSL). Source imagery for HRSL 2016 DigitalGlobe. https://ciesin.columbia.edu/data/hrsl/. Published 2019. Accessed 15 Oct 2019.
26. 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;130:219-32.
27. Leyk S, Gaughan AE, Adamo SB, Sherbinin A, Balk D, Freire S, et al. Allocating people to pixels: a review of large-scale gridded population data products and their fitness for use. Earth Syst Sci Data Discuss. 2019;11:1385-409.
28. Oak Ridge National Laboratories (ORNL). LandScan HD: Human settlement mapping at global scale [video]. https://www.youtube.com/watch?v=P84vxTT9Vos. Published 2015. Accessed 15 Oct 2019.
29. UNFPA, WorldPop, Flowminder, CIESIN. Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). http://www.grid3.org/. Published 2018. Accessed 15 Oct 2019.
30. Elsey H, Poudel AN, Ensor T, Mirzoev T, Newell JN, Hicks JP, 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:e024182.
31. Cajka J, Amer S, Ridenhour J, Allpress J. Geo-sampling in developing nations. Int J Soc Res Methodol. 2018;21:729-46.
32. 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;12:959.
33. GridSample. Case Studies. http://gridsample.org/case-studies. Published 2019. Accessed May 25, 2019.
34. Galway L, Bell N, Al Shatari SAE, Hagopian A, Burnham G, Flaxman A, 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.
35. Sollom R, Richards AK, Parmar P, Mullany LC, Lian SB, Iacopino V, et al. Health and human rights in Chin State, Western Burma: A population-based assessment using multistaged household cluster sampling. PLoS Med. 2011;8:e1001007.

36. Qader SH, Lefebvre V, Ninneman A, Himelein K, Pape UJ, Bengtsson L, et al. A novel approach to the automatic designation of pre-defined census enumeration areas and population sampling frames: a case study in Somalia. Washington, DC USA: The World Bank; 2019. http://documents.worldbank.org/curated/en/810771565268760366/A-Novel-Approach-to-the-Automatic-Designation-of-Predefined-Census-Enumeration-Areas-and-Population-Sampling-Frames-A-Case-Study-in-Somalia. Accessed 15 Oct 2019.
37. Pape U, Wollburg P. Estimation of poverty in Somalia using innovative methodologies. Washington, DC USA; The World Bank: 2019. http://documents.worldbank.org/curated/en/509221549985694077/Estimation-of-Poverty-in-Somalia-Using-Innovative-Methodologies. Accessed 15 Oct 2019.
38. Thomson DR, Stevens FR, Ruktanonchai NW, Tatem AJ, Castro MC. GridSample: An R package to generate household survey primary sampling units (PSUs) from gridded population data. Int J Health Geogr. 2017;16:25.
39. Chew RF, Amer S, Jones K, Unangst J, Cajka J, Allpress J, 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;17:12.
40. Muñoz J, Langeraar W. A Census-Independent Sampling Strategy for a Household Survey in Myanmar. Santiago Chile; 2013. http://winegis.com/images/census-independent-GIS-based-sampling-strategy-for-household-surveys-plan-of-action removed.pdf.
41. Grais RF, Rose AMC, Guthmann J. Don’t spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. Emerg Themes Epidemiol. 2007;4:8.
42. Thomson DR, Kools L, Jochem WC. Linking synthetic populations to household geolocations: a demonstration in Namibia. Data. 2018;3:30.
43. Kools L. One-stage versus two-stage cluster sampling, a simulation study. In: Essays on wealth, health, and data collection [dissertation]. Leiden, The Netherlands: Leiden University; 2018. https://openaccess.leidenuniv.nl/bitstream/handle/1887/67120/fulltext.pdf?sequence=1.
44. Mahabir R, Croitoru A, Crooks A, Agouris P, Stefanidis A. A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: trends, challenges and emerging opportunities. Urban Sci. 2018;2:8.
45. Sinha P, Gaughan AE, Stevens FR, Nieves JJ, Sorichetta A, Tatem AJ. Assessing the spatial sensitivity of a random forest model: application in gridded population modeling. Comput Environ Urban Syst. 2019;75:132-145.

ACKNOWLEDGEMENTS

Many thanks to Madeline Fabric, Clara Burgert-Brucker, and Sarchil Qader for their review and valuable feedback on an early draft. Also thanks to the participants at the Gates Foundation 2018 Gridded Population Sampling meeting for discussions and synthesis of gridded population survey evidence: Vincent Seaman, Brian Blankespoor, Hil Lyons, Julien Nobili, Kathleen Wannemuehler, Laina Mercer, Laura A Dwyer-Lindgren, Linnea Zimmerman, Michael Wild, Benjamin L. Nygren, Sabhanaz Rashid Diya, Tove Ryman, Eric M Weber, Wenfeng Gong, Win Brown, Sophia Magalona, Fiona Lau, and Jon Wakefield.

FUNDING

Dana R Thomson received funding from the UK Economic and Social Research Council (ESRC) grant ID: ES/5500161/1.

COMPETING INTERESTS

DRT is the creator of the GridSample R and GridSample2.0 python algorithms, and manager of GridSample.org. AJT is the director of the WorldPop team at University of Southampton. DAR and MCC declare no competing interests.

ROLES

DRT and DAR performed the literature review, and drafted the figures and text. MC and AT provided data interpretation and edits. All co-authors reviewed and approved the final manuscript.

 doi:10.20944/preprints201911.0072.v1

SUPPLEMENT 1

Supplement to: Thomson DR, Rhoda DA , Tatem AJ, Castro MC. Gridded population survey sampling: A review of the field 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 are weighted by land cover types (GHS-POP,[6,7] HRSL,[8] ESRI WPE,[9] WorldPop-Land Cover[10,11]). These modelling techniques are more mechanical than statistical, and thus do not result in estimates of model error. These models 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 both disaggregated and recent,[3] all of which are rare in LMICs.

Complex modelling techniques using multiple spatial covariates (e.g., WorldPop-Random Forest,[12,13] WorldPop-Global,[12,13] LandScan-Global,[14] Demobase[15]) are employed to produce substantially more accurate gridded population estimates. WorldPop-Random Forest 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.[12] 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-Random Forest 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-Random Forest is available for every fifth year in many LMICs based on all available spatial covariates, while WorldPop-Global has annual estimates for all countries, incorporates changes to urban extents over time, was more recently updated for most 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.[15]

LandScan-Global is a 1x1km dataset of the "ambient" population; a 24-hour average of daytime commuter population and night-time residential population.[14] 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.[14]

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, particularly effecting cell-level accuracy in urban areas. Misallocation of population in gridded population models may 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.[16] WorldPop-Random Forest and WorldPop-Global offer the most flexibility in terms of small cell size, high model accuracy,[12] and full coverage in LMICs resulting in their use in numerous surveys.[17–21] The older LandScan-Global dataset was used in a number of early gridded population surveys.[22–25] However datasets which restrict population estimates to settled areas may be more attractive to survey practitioners concerned with feasibility of fieldwork.

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.[26] 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, and will have the benefit of being constrained to settled areas.[27,28]

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)[29] dataset of 85x85m grid cells classified as built-up or not built-up, and the Global Human Settlement GHS-SMOD[6] 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,[30] UN-SALB,[31] and MapLibrary.[32]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 November 2019 doi:10.20944/preprints201911.0072.v1

Table S1. Summary of gridded population datasets available for LMICs

Approach	Name	Coverage	Producer	Resolution	Years	Available
Top-down	Gridded Population of the World v4 (GPWv4) [4,5]	Global	Columbia University - Center for International Earth Science Information Network (CIESIN)	~1x1km	2000-2020	Yes – free
	Global Human Settlement Population (GHS-POP) [6,7]	Global	Europe Commission - Joint Research Centre (JRC)	250x250m	1975-2015	Yes – free
	High Resolution Settlement Layer (HRSL) [8]	140 countries	Facebook & CIESIN	~30x30m	2015	Yes – free
	World Population Estimate (WPE) [9]	Global	ESRI	150x150m	2016	Yes – paid
	LandScan-Global [14]	Global	Oak Ridge National Laboratory	~1x1km	2000-2017	Yes – paid
	Demobase [15]	3 countries	United States Census Bureau	~100x100m	2003-2013	Yes – free
	WorldPop-Land Cover[10,11]	57 countries	WorldPop Project	~100x100m	2010- 2015	Yes – free
	WorldPop-Random Forest [12,13]	69 countries	WorldPop Project	~100x100m	2010-2020	Yes – free
	WorldPop-Global [12,13]	Global	WorldPop Project	~100x100m	2000-2020	Yes – free
Bottom-up	LandScan HD [27]	21 countries	Oak Ridge National Laboratory	~100x100m	varying	No (2020)
	GRID3 [28]	5 countries	WorldPop Project, Flowminder Foundation, CIESIN, UN Population Fund (UNFPA)	~100x100m	varying	No (2020)

References

1. Tobler WR. Smooth Pycnophylactic Interpolation for Geographical Regions. *J Am Stat Assoc* 1979;**74**:519—36.
2. UN Department of Economic and Social Affairs (UN-DESA). World population prospects. Population Division. New York, NY USA: UN-DESA; 2017. https://population.un.org/wpp/Download/Standard/Population/. Accessed 15 Oct 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**:1073—86.
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). New York, NY USA: CIESIN; 2016. http://dx.doi.org/10.7927/H4F47M2C. Accessed 15 Oct 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**:226—34.
6. European Commission – Joint Research Centre (EC-JRC). Global Human Settlement Population (GHS-POP). https://ghsl.jrc.ec.europa.eu/ghs_pop2019.php. Published 2019. Accessed 15 Oct 2019.
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 Commission; 2016.
8. Facebook Connectivity Lab and Center for International Earth Science Information Network—CIESIN—Columbia University. High Resolution Settlement Layer (HRSL). Source imagery for HRSL 2016 DigitalGlobe. https://ciesin.columbia.edu/data/hrsl/. Published 2019. Accessed 15 Oct 2019.
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.
10. Linard C, Gilbert M, Tatem AJ. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* 2011;**76**:525—38.
11. 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**:e55882.
12. 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**:e0107042.
13. WorldPop. Data availability. http://www.worldpop.org.uk/data/data_sources. Published 2019. Accessed 15 Oct 2019.
14. Dobson JE, Brlght EA, Coleman PR, Worley BA. LandScan: a global population database for estimating populations at risk. *Photogramm Eng Remote Sensing* 2000;**66**:849-857.
15. 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—32.
16. 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—17.
17. 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**:526—37.
18. 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**:e024182.
19. World Food Programme - Vulnerability Analysis and Mapping unit (WPF-VAM). Urban essential needs assessment in the five communes of Kimbanseke, Kinsenso, Makala, N'sele and Selembao (Kinshasa). Rome, Italy: WFP-VAM; 2018.
20. Pape U, Wollburg P. Estimation of poverty in Somalia using innovative methodologies. Washington, DC USA; The World Bank: 2019. http://documents.worldbank.org/curated/en/509221549985694077/Estimation-of-Poverty-in-Somalia-Using-Innovative-Methodologies. Accessed 15 Oct 2019.
21. GridSample. Case Studies. http://gridsample.org/case-studies. Published 2019. Accessed May 25, 2019.
22. 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**:959.
23. 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.
24. Cajka J, Amer S, Ridenhour J, Allpress J. Geo-sampling in developing nations. *Int J Soc Res Methodol* 2018;**21**:729—46.
25. 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**:e1001007.
26. 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;**115**:3529—37.
27. Oak Ridge National Laboratories (ORNL). LandScan HD: Human settlement mapping at global scale [video]. https://www.youtube.com/watch?v=P84vxTT9Vos. Published 2015. Accessed 15 Oct 2019.
28. UNFPA, WorldPop, Flowminder, CIESIN. Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). http://www.grid3.org/. Published 2018. Accessed 15 Oct 2019.
29. DLR Earth Observation Center (DLR-EOC). Global Urban Footprint (GUF). https://www.dlr.de/eoc/. Published 2017. Accessed 6 Feb 2017.
30. Global Administrative boundaries project (GADM). Global administrative areas version 3.6. https://gadm.org/data.html. Published 2018. Accessed 15 Oct 2019.
31. Geospatial Information Section and Statistics Division. United Nations Second Administrative Level Boundaries (UN-SALB). https://www.unsalb.org/data. Published 2019. Accessed 15 Oct 2019.
32. Map Maker Ltd. Map Library. http://www.maplibrary.org/library/stacks/Africa/index.htm. Published 2007. Accessed February 6, 2017.