USAID
FROM THE AMERICAN PEOPLE

DHS WORKING PAPERS

Gridded Population Sampling and the Demographic and Health Surveys

Mahmoud Elkasabi
Ruilin Ren

2020 No. 176

DEMOGRAPHIC AND HEALTH SURVEYS

September 2020

DHS Working Papers No. 176

Gridded Population Sampling and the Demographic and Health Surveys

Mahmoud Elkasabi
Ruilin Ren

The DHS Program
ICF
Rockville, Maryland, USA

September 2020

Corresponding author: Mahmoud Elkasabi, The DHS Program, ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA; phone: +1 301-572-0484; fax: +1 301-407-6501; email: Mahmoud.Elkasabi@icf.com

Editor: Diane Stoy
Document Production: Chris Gramer

This study was conducted with support from the United States Agency for International Development (USAID) through The DHS Program (#720-OAA-18C-00083). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

Recommended citation:

Elkasabi, Mahmoud and Ren, Ruilin. 2020. *Gridded population sampling and the Demographic and Health Surveys*. DHS Working Papers No. 176. Rockville, Maryland, USA: ICF.

CONTENTS

TABLES

ABSTRACT

This paper addresses gridded population sampling as opposed to the traditional procedures of sampling for population-based household surveys. The paper explores the possibility that gridded population datasets could replace the traditional census frames for sampling design. To evaluate the strengths and weaknesses of gridded population sampling designs, this paper contrasts the major sample selection steps and design features of gridded population sampling to traditional area probability sampling, and compares a sampling design for a Demographic and Health Survey (DHS) survey as an example of a traditional area sample design, with a gridded population sampling design for the same survey.

Key words: area sampling, DHS, gridded population sampling, multistage sampling design

ACRONYMS AND ABBREVIATIONS

DHS	Demographic and Health Surveys
EA	enumeration area
GIS	geographic information system
GPS	global positioning system
MOS	measure of size
PPES	probability proportionate to estimated population size
PPS	proportional to size
PSU	primary sampling unit
SSU	secondary selection unit

1 BACKGROUND

Demographic and Health Surveys (DHS) are nationally representative, population-based household surveys designed to produce a wide range of data on demographic and health topics. In all DHS surveys, multistage stratified probability samples are selected. In a typical DHS survey, a two-stage stratified sampling design is used. Census enumeration areas (EAs) are selected in the first stage as primary sampling units (PSUs) with probability proportional to size (PPS). In the second stage, residential households are randomly selected from each selected PSU after a household listing operation has generated an updated, complete list of residential households in the selected PSU (ICF International 2012). In the past few years, *gridded population sampling* methods have been used for population-based household surveys in settings where census frames were either outdated or not available (Elsey et al. 2016; Galway et al. 2012; Muñoz and Langeraar 2013; Thomson et al. 2012). Advancements in gridded population sampling have increased interest in adopting these methods in DHS surveys. The objective of this paper is to compare gridded population sampling methods with the traditional area sampling method, which is the standard DHS sampling approach. The paper will also address the DHS Program's position on adopting methods of gridded population sampling in DHS surveys.

Gridded population sampling is not a sampling approach that is distinct from area-based probability sampling. In fact, gridded population sampling designs are area probability sampling designs that are GIS/GPS-assisted, especially in the construction of the gridded sampling frames. The designs might also use GIS/GPS to operationalize the procedures required for developing a probability sample of the target population (Heeringa 2018). This paper addresses gridded population sampling through a comparison of GIS/GPS-assisted and traditional procedures. We explore the possibility that GIS/GPS-assisted gridded population designs could replace the traditional census frames for sampling design. To evaluate the strengths and weaknesses of gridded population sampling designs, this paper contrasts the major sample selection steps and design features of gridded population sampling to traditional area probability sampling.

2 SAMPLING DESIGN IN DHS SURVEYS

2.1 Sampling Frame

A *sampling frame* is a complete list of all sampling units that entirely cover the target population. A sampling frame allows for a probability selection of sampling units. The availability of a suitable sampling frame determines the feasibility of conducting a DHS. In a typical DHS survey, the list of *enumeration areas* (EAs)[1] from the most recently completed population census is the sampling frame. In some surveys, an existing master sample is used instead[2]. The frame usually covers the entire geographic area of a country, without omission or overlap. Maps usually exist for each EA with clearly defined boundaries. Such maps are usually digitalized and supported by GIS technology. Each EA typically has a unique identification code or a series of codes, and at least one measure of size (MOS) estimate (population and/or number of households), based on the most recent census or updates from the field. Before using the census frames for the sampling design, the quality of the frame is examined by checking attributes such as frame coverage and population distribution, identification and coding, MOS and consistency (ICF International 2012).

When a census frame or a master sample is not available for DHS surveys, alternative frames with full coverage for the target population are potential options.[3] The options are evaluated by a number of factors such as (1) Does the frame completely cover the target population? (2) Are the PSUs too large or too small for a DHS survey? (3) Are there cartographic materials that allow for the boundaries of PSUs to be precisely identified in the field? and (4) Are geographic/administrative variables available that can be used in sample stratification?

2.2 Sampling Stages

Most DHS surveys use a two-stage stratified sampling design, in which EAs are selected in the first stage with PPS selection. In the second stage, residential households are selected with equal probability from an updated household listing of each selected EA. The DHS samples are typically stratified by regions and urban/rural residence. In the first stage of sampling, EAs are selected according to the allocation of the EA sample over the sample strata. After the selection of the EAs, a household listing operation is conducted in every selected EA and a complete list of households is constructed. In the second stage, a fixed number of households is selected from each EA. From the selected households, all women age 15–49 who are usual members of the household or who slept in the household the night before the survey are eligible for a woman's individual survey. In all households or in a subsample of households, all men of reproductive age (typically age 15–49, 15–54, or 15–59) are eligible for a man's individual survey.

[1] An EA is a geographical area that serves as a counting unit for the population census, and is usually a city block in urban areas or a village in rural areas with 200-300 households.

[2] The master sample is a representative random sample of EAs selected from the most recent completed population census.

[3] In the Egypt DHS 2014, a complete list of all Shiakhas (urban administrative unit) and villages served as the sampling frame. In Congo (DRC), an electoral frame was used for the 2013-14 Congo (DRC) DHS.

2.3 Fieldwork

Fieldwork begins with the household listing process that takes place between the first and second stages of sampling. This process is essential to the construction of a list of households in each selected EA, from which a sample of households can be selected at random in the second stage. Since census maps typically exist for each EA with clearly defined boundaries (natural boundaries such as rivers, streets, and roads), field enumerators use these maps to locate the selected EAs in the field. After they tour the EA and match the actual boundaries on the ground with the map boundaries, the enumerators begin the extensive process of household listing. In situations where matching the boundaries on the ground is not possible due to low quality maps or real geographic changes on the ground, the field workers seek help from local authorities.

2.4 Survey Weights

All probability sampling surveys use survey weights to guarantee the actual representativeness of the sample. The DHS sampling stages and procedures allow for a straightforward calculation of survey weights that account for all selection probabilities of the sampling units. The calculations benefit from the available measure of size for each EA in the census frame, and the number of listed households in each EA based on the household listing.

3 THE GRIDDED POPULATION SAMPLING

3.1 Sampling Frame

The gridded population datasets contain model-produced grid cells that are uniform in areal size and shape, but variable in population totals.[4] Models can either disaggregate census data and administrative population counts or predict population density based on small area surveys and other spatial covariates, such as road infrastructure and nightlight intensity. Gridded population datasets also vary in the geographic scale of the grid cells. For example, WorldPop offers 100 × 100 m grid cells, whereas LandScan offers 1 × 1 km grid cells (Quader et al. 2020; Thomson et al. 2017) Therefore, the choice of the gridded population datasets definitely influences the subsequent sampling stages, where small grid cells can be grouped into bigger sampling units or large grid cells can be segmented into smaller sampling units. To be used as a sampling frame, the gridded population datasets might be supplemented with other cartographic materials that support the sample design. For example, to clearly define the sample strata, a shapefile[5] of boundaries of the sample strata should be overlaid with the gridded population dataset (Thomson et al. 2017).

3.2 Sampling Stages

Similar to the traditional sampling method, gridded population sampling is a multistage sampling design, but with grid cells or groups of grid cells that are selected in the first stage. There are several approaches to first-stage sample selection and the subsequent stages. With the *segmentation approach*, grid cells (such as 1 km × 1 km) are selected with the probability proportionate to estimated population size (PPES) as PSUs, and then segmented into secondary selection units (SSUs) that are selected in the second stage of selection (Thomson et al. 2012). The segmentation process can be done by smaller grid cells or by manual delineation that uses satellite imagery. With the *point approach*, randomly generated points are located within grid cells where the number of points is proportional to the estimated population. A random sample of points are selected as PSUs centroids, and the PSU boundaries are then manually delineated based on satellite imagery (Muñoz and Langeraar 2013). With the *growth approach*, sample grid cells (such as 100 m × 100 m) are randomly selected with PPES. In a second step, PSUs are developed by including neighboring areas around selected seed cells until a minimum population threshold is achieved (Elsey et al. 2016).

3.3 Fieldwork

Several approaches have been used to identify the selected PSUs or SSUs in the field and to select households. When boundaries of grid cells are used as boundaries for PSUs or SSUs, identifying the selected sampling units in the field can be challenging (Elsey et al. 2016). Whereas, when PSUs or SSUs are *manually delineated* using satellite imagery to follow sensible boundaries such as rivers, streets, and

[4] Different datasets are available based on the models used to construct the datasets' grid cells and to estimate the population size per each cell. These include WorldPop, GPWv4, GHS-POP, GRUMP, UNEP, and Landscan.

[5] A type of file used to store the geometric location and attribute information of geographic features represented by points, lines, or polygons (areas).

roads, it is easy for the field teams to find the selected PSUs/SSUs and perform the household enumeration by using sketch or digital maps, or satellite imagery. However, such manual delineation is subject to human bias (Muñoz and Langernaar 2013; Thomson et al. 2012).

Although a traditional household listing can be done, many non-traditional approaches are used with gridded population sampling. These include non-probability sampling approaches that are prone to selection bias, such as the random-walk, spin-the-pen, or other semi-probability sampling approaches such as selecting every tenth household (Chen et al. 2019; Elsey et al. 2016; Galway et al. 2016**;** Sollum et al. 2011). The traditional household listing has also been replaced by digitizing dwelling point locations in satellite imagery followed by selecting sampling points at random (Escamilla et al. 2014; Lin and Kuwayama 2016; Pearson et al. 2015; Siri et al. 2008; Wampler et al. 2013).

3.4 Survey Weights

Since gridded population sampling is a multistage sampling design, survey weights can be calculated by considering selection probabilities from different stages (Heeringa 2018). However, the inclusion probabilities of the grid cells or the aggregated cells are sensitive to the size of these cells, where the quality of these sizes might be questionable in some gridded population datasets. The approach to calculation is also not clear when techniques like the growth approach are used (Thomson et al. 2017; Qader et al. 2020).

4 COMPARISON BETWEEN THE GRIDDED POPULATION SAMPLING AND THE TRADITIONAL SAMPLING DESIGN OF DHS SURVEYS

Table 1 compares a sampling design for a DHS survey as an example of a traditional area sample design, with a gridded population sampling design for the same survey. The following criteria guide the comparison: (1) sampling frame; (2) sampling stages; (3) fieldwork; and (4) survey weights. With each criteria, we address the pros, cons, and potential challenges.

Table 1 Comparison of a traditional sampling design with a gridded population sampling design for a DHS survey

	Traditional approach	Gridded population sampling
1. Sampling frame	Frame of EAs from the most recent census.	A gridded population dataset that covers the whole area land of the country.
Coverage	The sampling frame provides complete coverage for the whole population.	The sampling frame provides complete coverage for the whole population.
Accessibility and summary tables of the sampling frame	In most of the cases, the sampler must have access to the census frame for checking, study, and production of the summary tables, which include necessary information for the sample design.	Checking the quality of the gridded population datasets and producing summary tables require special GIS technical expertise.
Accessibility and flexibility of the frame for sample selection	The sampler can sort the frame properly to achieve the designed and implicit stratifications. The sampler can decide which software will be used for sample selection.	Selecting the sample PSUs might require using several GIS tools and packages.
Size of PSUs	MOS from the last census (number of households) is provided for each EA, and used for the PPS selection of EAs as PSUs. Based on the actual size of the EA, large EAs might be segmented during the household listing, and small EAs might be combined with adjacent EAs to form bigger PSUs.	Disaggregated administrative population counts are used as a measure of size for each grid cell. The last census is a key source of these counts. Grid cells might be selected as PSUs or aggregated in larger PSUs with appropriate size.
Boundaries of PSUs[T]	Most use natural landmarks and roads as boundaries. A census map is available for each EA with clearly identifiable boundaries and text descriptions.	Boundaries of grid cells are mostly imaginary boundaries. When grid cells are aggregated into a larger PSU, manual delineation based on satellite imagery can be used to form PSUs with natural boundaries. Such manual delineation is subject to human bias.
Urban-rural definition[T]	The urban-rural definition in a census frame reflects a country's official definition of urban and rural areas. The DHS surveys use the country's officially defined urban and rural areas.	The gridded sampling frames use different model based urban-rural simulations. These may not correspond to the country's official definition of urban and rural areas.
Variables for stratification[T]	Stratification variables (regions by urban/rural) are available for each EA in the census frame, and/or updated regions that reflect the recent administrative changes.	To define stratification variables, gridded population datasets might need to be supplemented by the boundary shapefiles of the strata. If accessible, such shapefiles might not reflect recent administrative changes.
Special population coverage[G]	The census frame may not be up to date for identification of special target populations such as a displaced population, a nomadic population, and a population that lives in slum areas.	Based on current high resolution imagery and pattern recognition, the gridded population frame may be used to identify special target populations such as displaced populations, nomadic populations, and populations living in slum areas.

Continued…

Table 1—*Continued*

2. Sampling stages[T]	When using a census EA frame, a typical two-stage design sampling procedure is straightforward.	Since grid cells are uniform in shape and areal size, the multistage design might involve more than two stages. The sampling procedure can be sophisticated, and might be labor intensive. It might also be difficult to control the work quality since some designs depend heavily on the surveyor's subjective judgments.
3. Fieldwork	Fieldwork begins with the household listing process.	Fieldwork begins with the household listing process.
Identifying the boundaries[T]	Boundaries are identifiable especially when sensible boundaries are used such as natural boundaries—rivers, streets, and roads.	When imaginary boundaries of grid cells are used, identifying the boundaries can be challenging. Boundaries are identifiable if sensible boundaries are used.
Household listing	Household listing is a necessary if a random sample of households is to be selected. DHS uses a standard process of listing where all structures are listed, and a random sample of the listed households is selected.	Gridded population sampling also needs a household listing because the satellite photos are not sufficiently detailed to allow a direct selection of households. Many non-traditional approaches are used with gridded population sampling. However, a complete household listing can be used and a truly random sample of households can be selected.
4. Survey weights[T]	The sampling stages and procedures used in the DHS allow for a straightforward calculation for survey weights that account for all selection probabilities for all sampling stages and all sampling units.	The sampling stages and selection procedures can be executed in a way that allows for a proper calculation of selection probabilities in each sampling stage and for each sampling unit. With the *point approach* and *growth approach*, the selection probability can be challenging because it depends on how the sample point grows and the surveyor's subjective judgements.

[T] Traditional sampling is potentially superior to gridded population sampling.

[G] Gridded population sampling is potentially superior to traditional sampling.

5 DISCUSSION

Gridded population sampling methods are multistage sampling designs that are GIS/GPS assisted in the construction of the sampling frame. With procedures constructed for multistage sampling that are similar to the traditional approach, gridded population sampling is a promising approach for population surveys, especially when access to a census frame is limited. This paper compared the traditional method and the gridded sampling approach by several criteria, including the frame coverage, sampling stages, fieldwork, and survey weights.

Several issues should be considered when using gridded population sampling. The following points need to be addressed:

1. ***The accuracy of the gridded population datasets.*** Several gridded population datasets that were used as sampling frames in different circumstances are available. The accuracy of these datasets varies by the scale of the output grid, the complexity of models used to construct the grid cells, and the geographic scale and accuracy of the input census data and covariates used in the model. To prepare a sampling frame for a household survey, measuring and improving accuracy by considering all these factors might be challenging for some survey practitioners.

2. ***The stratification variables.*** To construct a complete sampling frame with all variables needed for a sampling design such as the stratification variables, the gridded population datasets may need to be supplemented with other datasets from different sources, such as a shapefile of boundaries of regions or urban/rural areas. This can be a challenge for survey practitioners because these shapefiles might not publicly accessible, especially for the urban/rural areas.

3. ***The sampling stages and the size of PSUs.*** The use of gridded PSUs or larger units influences the sampling inclusion probabilities, and also the feasibility and the protocols used in the field. In addition, other alternatives, such as the *point approach* and the *growth approach*, are prone to bias because they involve considerable subjective judgment.

4. ***The boundaries of the final sampling units.*** When gridded PSU boundaries are used, field protocols should be developed and used in conjunction with high-resolution satellite imagery to guide fieldworkers in accurately locating the selected grids in the field. In contrast, when boundaries are manually delineated to follow sensible boundaries related to features in the real world, fieldwork becomes more feasible but is more labor intensive.

5. ***Resources and technical expertise and skills.*** Designing and implementing a gridded population sampling design require access to technical expertise, implementation skills, and technology. These might affect the cost and timeline of the survey, and should be considered and properly evaluated before proceeding with the design.

6. ***Survey weights.*** A multistage sampling design must be constructed in a way that would allow for a straightforward calculation of the overall selection probabilities and survey weights. An area gridded population sampling, together with a standard household listing, will outperform the *point* and *growth approaches*.

Gridded population sampling is a potential alternative for a DHS survey when a census frame is not available in a country. In such situations, the gridded population sampling should be assessed in conjunction with other alternative sampling frames available in the country. In this assessment, in addition to the six points mentioned above, *legality and justifiability*[6] should be considered.

[6] The DHS Program does not decide which sampling frame to use for a survey. It is the host country that decides which sampling frame is to be used. If a census frame is available, the host country may not accept another sampling frame other than the last population census frame, even if that census frame is old.

REFERENCES

Chen, X., H. Hu, X. Xu, J. Gong, Y. Yan, and F. Li. 2018. "Probability Sampling by Connecting Space with Households Using GIS/GPS Technologies." *Journal of Survey Statistics and Methodology* 6 (2): 149-168. https://doi.org/10.1093/jssam/smx032.

Elsey, H., D. R. Thomson, R. Y. Lin, U. Maharjan, S. Agarwal, and J. Newell. 2016. "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." *Journal of Urban Health* 93(3): 526–537. https://doi.org/10.1007/s11524-016-0046-9.

Galway, L., N. Bell, A. S. Sae, A. Hagopian, G. Burnham, A. Flaxman, W. M. Weiss, J. Rajaratnum, and T. K. Takaro. 2012. "A Two-Stage Cluster Sampling Method Using Gridded Population Data, a GIS, and Google Earth™ Imagery in a Population-Based Mortality Survey in Iraq." *International Journal of Health Geographics* 11:12. https://doi.org/10.1186/1476-072X-11-12.

Heeringa, S. G. 2018. "Discussion of 'Probability Sampling by Connecting Space with Households Using GIS/GPS Technologies' by Chen, X.; Xu, X.; Gong, J.; Yan, Y.; Fang, L." *Journal of Survey Statistics and Methodology* 6 (2):169–181. https://doi.org/10.1093/jssam/smy004.

ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual*. Calverton, MD, USA: ICF International. https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

Lin, Y., and D. P. Kuwayama. 2016. "Using Satellite Imagery and GPS Technology to Create Random Sampling Frames in High Risk Environments." *International Journal of Surgery* 32: 123-128. https://doi.org/10.1016/j.ijsu.2016.06.044.

Muñoz, J., and W. Langeraar. 2013. "A census-independent sampling strategy for a household survey in Myanmar." https://www.winegis.com/images/census-independent-GIS-based-sampling-strategy-for-househld-surveys-plan-of-action%20removed.pdf.

Pearson, A. L., A. Rzotkiewicz, and A. Zwickle. 2015. "Using Remote, Spatial Techniques to Select A Random Household Sample in a Dispersed, Semi-Nomadic Pastoral Community: Utility for A Longitudinal Health and Demographic Surveillance System." *International Journal of Health Geographics* 14: 33. https://doi.org/10.1186/s12942-015-0026-4.

Qader, S. H., V. Lefebvre, A. J. Tatem, U. Pape, W. Jochem, K. Himelein, A. Ninneman, et al. 2020. "Using Gridded Population and Quadtree Sampling Units to Support Survey Sample Design in Low-Income Settings." *International Journal of Health Geographics* 19 (1): 1-16. https://doi.org/10.1186/s12942-020-00205-5.

Siri, J. G., K. A. Lindblade, D. H. Rosen, B. Onyango, J. M. Vulule, L. Slutsker, and M. L. Wilson. 2008. "A Census-Weighted, Spatially-Stratified Household Sampling Strategy for Urban Malaria Epidemiology." *Malaria Journal* 7 (39): 1-10. https://doi.org/10.1186/1475-2875-7-39.

Sollom, R, A. K. Richards, P. Parmar, L. C. Mullany, S.B. Lian, V. Iacopino, and C. Beyrer. 2011. "Health and Human Rights in Chin State, Western Burma: A Population-Based Assessment Using Multistaged Household Cluster Sampling. *PLoS Medicine* 8 (2): e1001007. https://doi.org/10.1371/journal.pmed.1001007.

Thomson, D. R., M. B. Hadley, P. G. Greenough, and M. C. Castro. 2012. "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* 12 (1): 959. https://doi.org/10.1186/1471-2458-12-959.

Thomson, D. R., F. R. Stevens, N. W. Ruktanonchai, A. J. Tatem, M. C. Castro. 2017. "*GridSample*: An R Package to Generate Household Survey Primary Sampling Units (PSUs) from Gridded Population Data." *International Journal of Health Geographics* 16 (1): 25. https://doi.org/10.1186/s12942-017-0098-4.

Wampler, P. J., R. R. Rediske, and A. R. Molla. 2013. "Using Arcmap, Google Earth, and Global Positioning Systems to Select and Locate Random Households." *International Journal of Health Geographics* 12: 3. https://doi.org/10.1186/1476-072X-12-3.