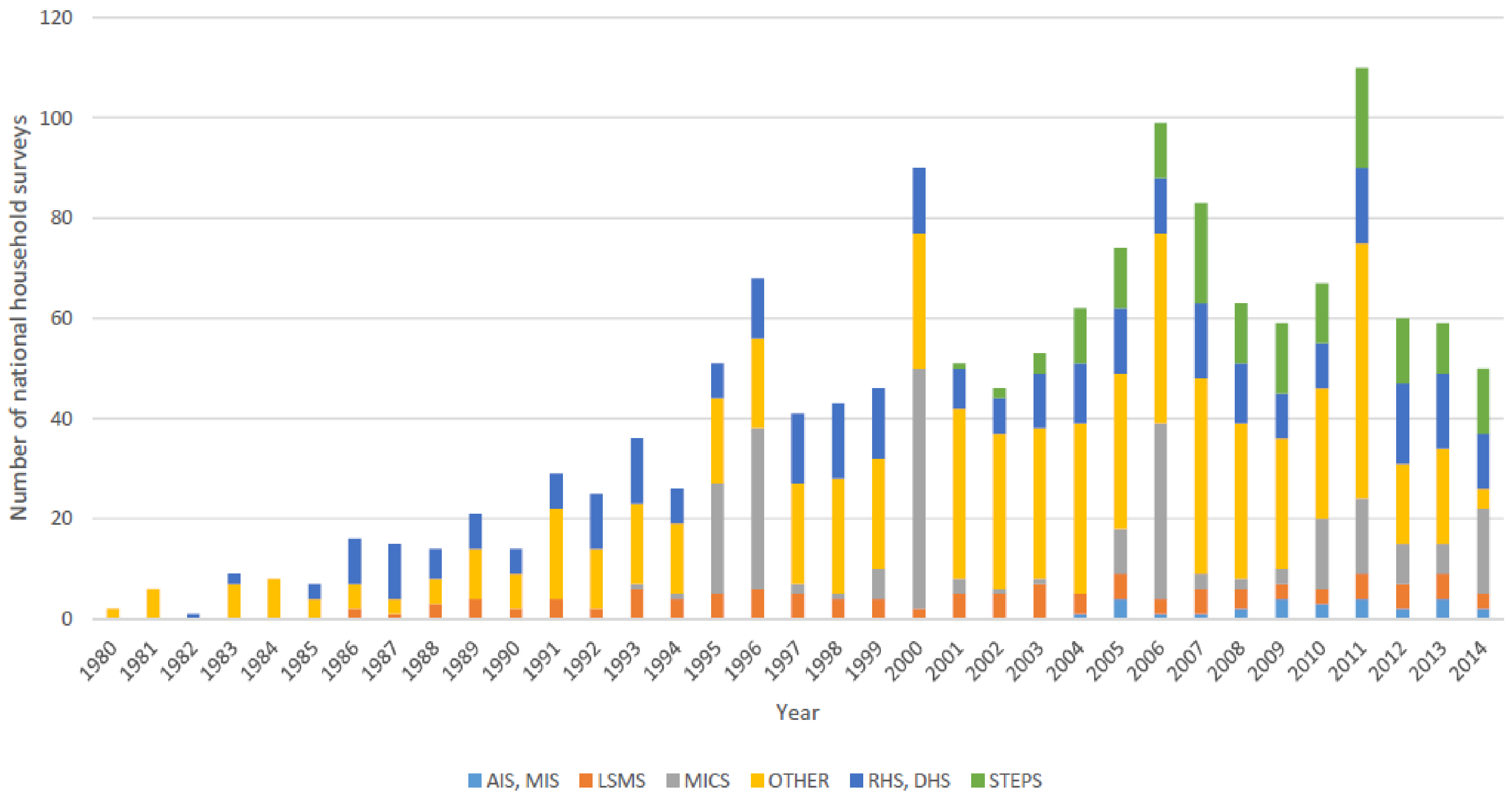
GridSample

Free, open tool to select accurate household surveys from gridded population data

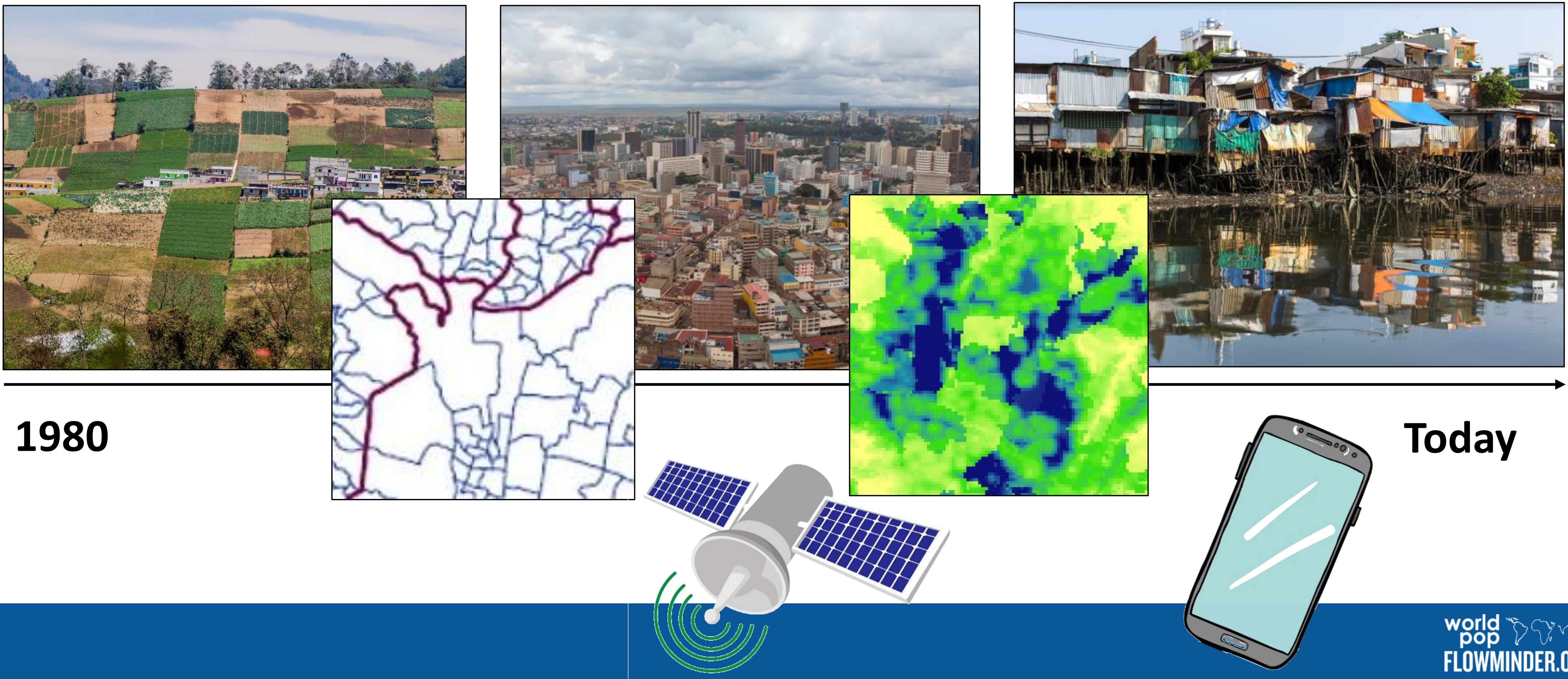
BigSurv18

Dana R. Thomson
WorldPop University of Southampton
Flowminder Foundation

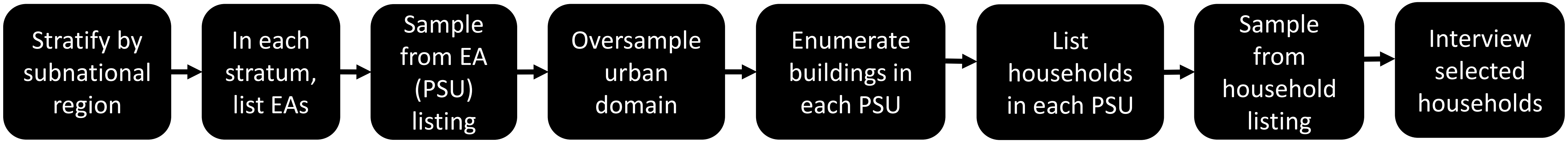
Household surveys are a (the?) main source of health and demographic data in LMICs



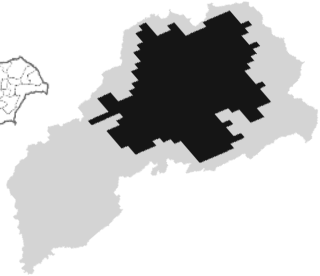
Since 1980, contexts in LMICs have changed, but survey methods and tools have not

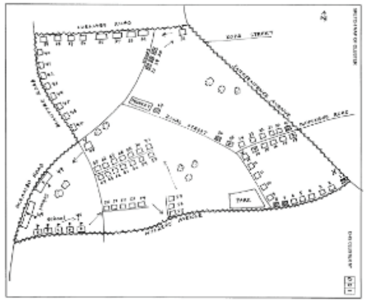


Typical household survey workflow



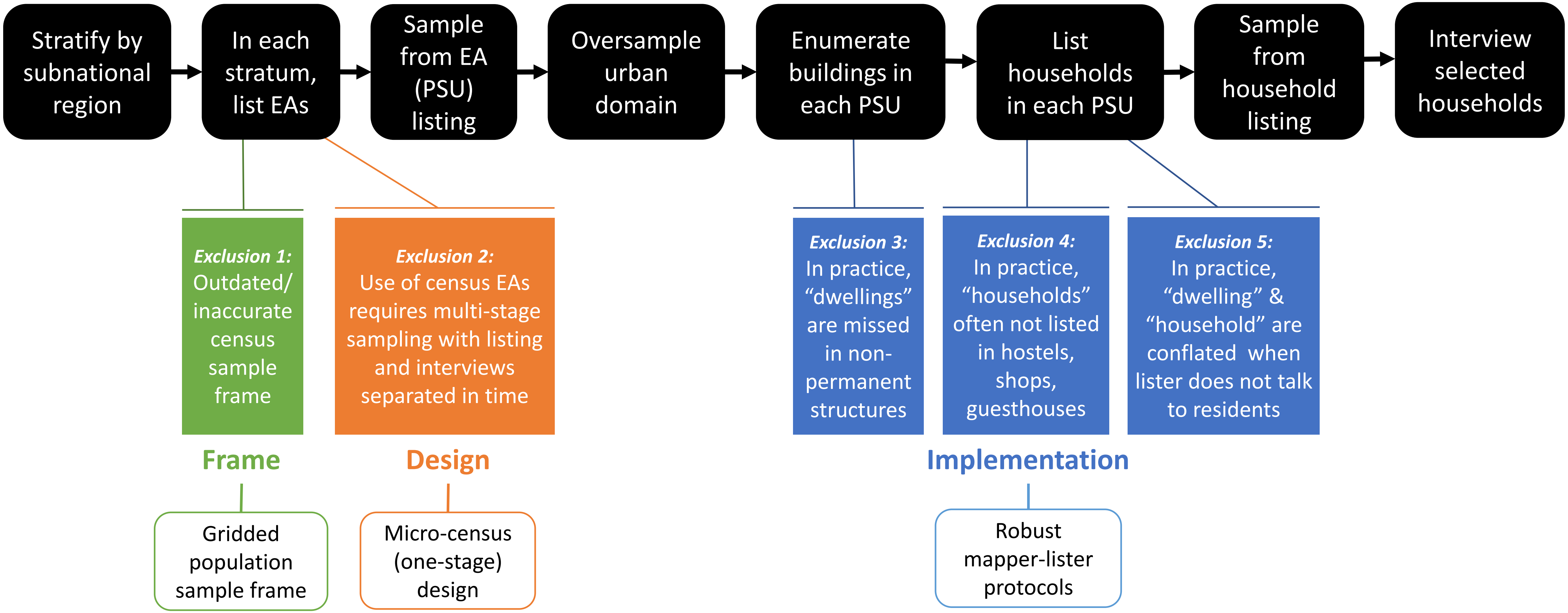
Stratify by subnational region → In each stratum, list EAs → Sample from EA (PSU) listing → Oversample urban domain → Enumerate buildings in each PSU → List households in each PSU → Sample from household listing → Interview selected households





Building	Dwelling	Household
1	1	1
1	1	2
1	2	1
1	3	1
2	1	1
3	1	1
4	1	1
…..	……	……

Unintentional exclusion of the poorest & vulnerable



Stratify by subnational region → In each stratum, list EAs → Sample from EA (PSU) listing → Oversample urban domain → Enumerate buildings in each PSU → List households in each PSU → Sample from household listing → Interview selected households

Exclusion 1: Outdated/ inaccurate census sample frame

Exclusion 2: Use of census EAs requires multi-stage sampling with listing and interviews separated in time

Exclusion 3: In practice, "dwellings" are missed in non-permanent structures

Exclusion 4: In practice, "households" often not listed in hostels, shops, guesthouses

Exclusion 5: In practice, "dwelling" & "household" are conflated when lister does not talk to residents

Frame

Gridded population sample frame

Design

Micro-census (one-stage) design

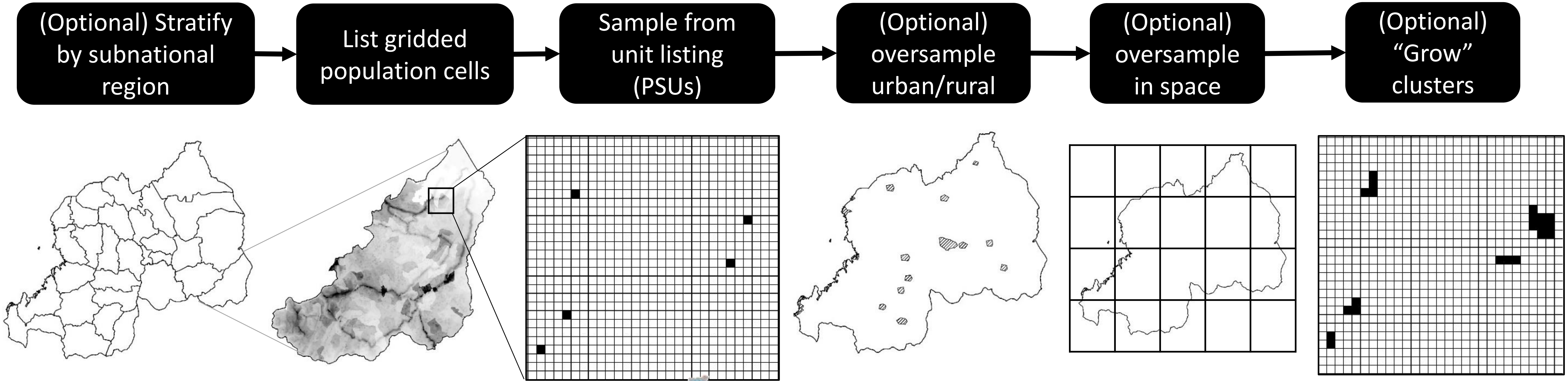
Implementation

Robust mapper-lister protocols

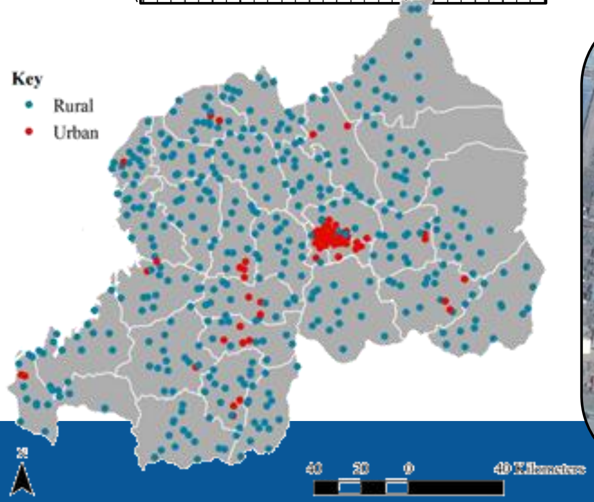
Free gridded population sampling tools

Features	GridSample R package (2016)	GridSample2.0 GridSample.org (2019)
Allocation of clusters to strata	Equal	Equal, Custom, Proportional
Oversampling options	U/R, Spatial	Custom, Spatial
Geographic boundaries (coverage, strata)	Own shapefiles	Own shapefiles, Pre-defined
Definition of clusters	Any grid cell size, optionally “grow” cluster after selection	Single grid cells, Mutli-cell units, Own shapefile
Publicly available, free	Yes	Yes

GridSample R algorithm



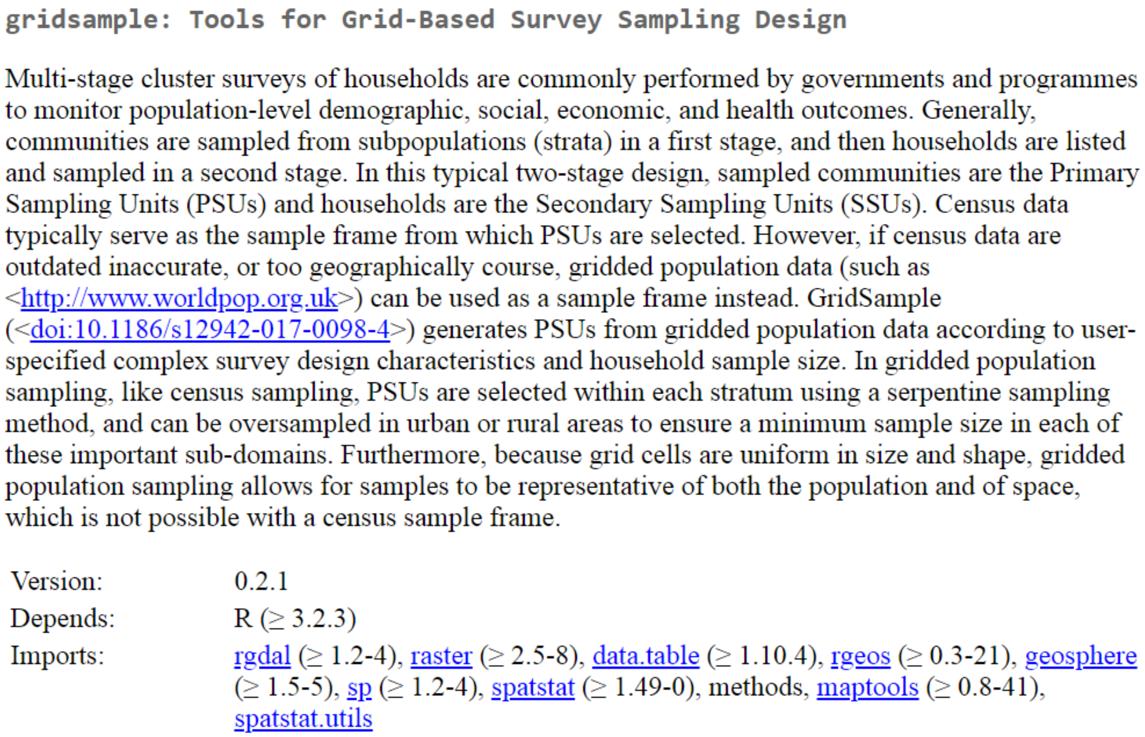
(Optional) Stratify by subnational region → List gridded population cells → Sample from unit listing (PSUs) → (Optional) oversample urban/rural → (Optional) oversample in space → (Optional) "Grow" clusters



Output:

Gridded PSU boundaries with specified population and maximum area

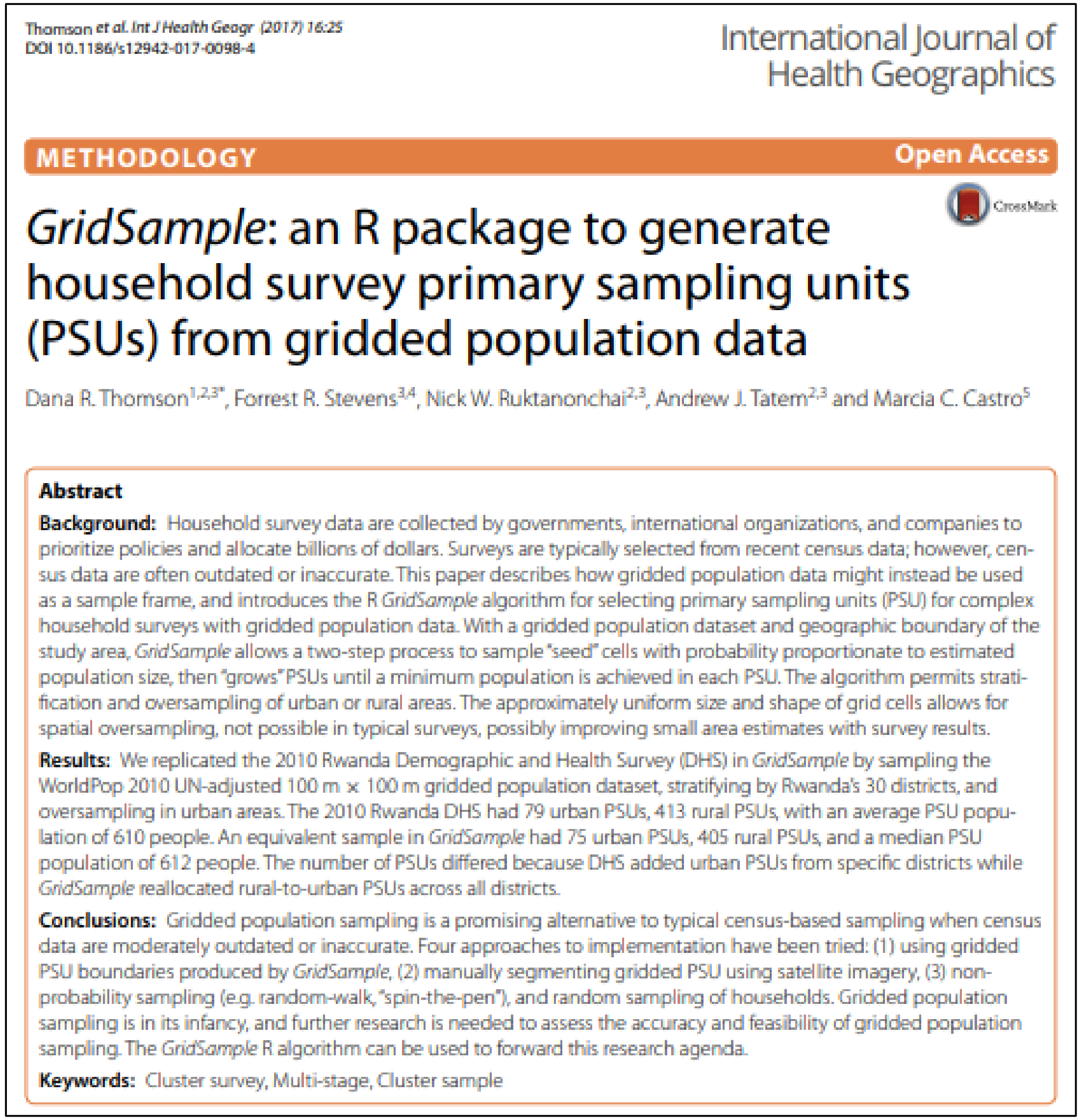
GridSample R algorithm



gridsample: Tools for Grid-Based Survey Sampling Design

Multi-stage cluster surveys of households are commonly performed by governments and programmes to monitor population-level demographic, social, economic, and health outcomes. Generally, communities are sampled from subpopulations (strata) in a first stage, and then households are listed and sampled in a second stage. In this typical two-stage design, sampled communities are the Primary Sampling Units (PSUs) and households are the Secondary Sampling Units (SSUs). Census data typically serve as the sample frame from which PSUs are selected. However, if census data are outdated inaccurate, or too geographically course, gridded population data (such as <http://www.worldpop.org.uk>) can be used as a sample frame instead. GridSample (<doi:10.1186/s12942-017-0098-4>) generates PSUs from gridded population data according to user-specified complex survey design characteristics and household sample size. In gridded population sampling, like census sampling, PSUs are selected within each stratum using a serpentine sampling method, and can be oversampled in urban or rural areas to ensure a minimum sample size in each of these important sub-domains. Furthermore, because grid cells are uniform in size and shape, gridded population sampling allows for samples to be representative of both the population and of space, which is not possible with a census sample frame.

Version:	0.2.1
Depends:	R (≥ 3.2.3)
Imports:	rgdal (≥ 1.2-4), raster (≥ 2.5-8), data.table (≥ 1.10.4), rgeos (≥ 0.3-21), geosphere (≥ 1.5-5), sp (≥ 1.2-4), spatstat (≥ 1.49-0), methods, maptools (≥ 0.8-41), spatstat.utils



Thomson *et al. Int J Health Geogr* (2017) 16:25
DOI 10.1186/s12942-017-0098-4

International Journal of Health Geographics

METHODOLOGY Open Access

GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data

Dana R. Thomson[1,2,3*], Forrest R. Stevens[3,4], Nick W. Ruktanonchai[2,3], Andrew J. Tatem[2,3] and Marcia C. Castro[5]

Abstract

Background: Household survey data are collected by governments, international organizations, and companies to prioritize policies and allocate billions of dollars. Surveys are typically selected from recent census data; however, census data are often outdated or inaccurate. This paper describes how gridded population data might instead be used as a sample frame, and introduces the R *GridSample* algorithm for selecting primary sampling units (PSU) for complex household surveys with gridded population data. With a gridded population dataset and geographic boundary of the study area, *GridSample* allows a two-step process to sample "seed" cells with probability proportionate to estimated population size, then "grows" PSUs until a minimum population is achieved in each PSU. The algorithm permits stratification and oversampling of urban or rural areas. The approximately uniform size and shape of grid cells allows for spatial oversampling, not possible in typical surveys, possibly improving small area estimates with survey results.

Results: We replicated the 2010 Rwanda Demographic and Health Survey (DHS) in *GridSample* by sampling the WorldPop 2010 UN-adjusted 100 m × 100 m gridded population dataset, stratifying by Rwanda's 30 districts, and oversampling in urban areas. The 2010 Rwanda DHS had 79 urban PSUs, 413 rural PSUs, with an average PSU population of 610 people. An equivalent sample in *GridSample* had 75 urban PSUs, 405 rural PSUs, and a median PSU population of 612 people. The number of PSUs differed because DHS added urban PSUs from specific districts while *GridSample* reallocated rural-to-urban PSUs across all districts.

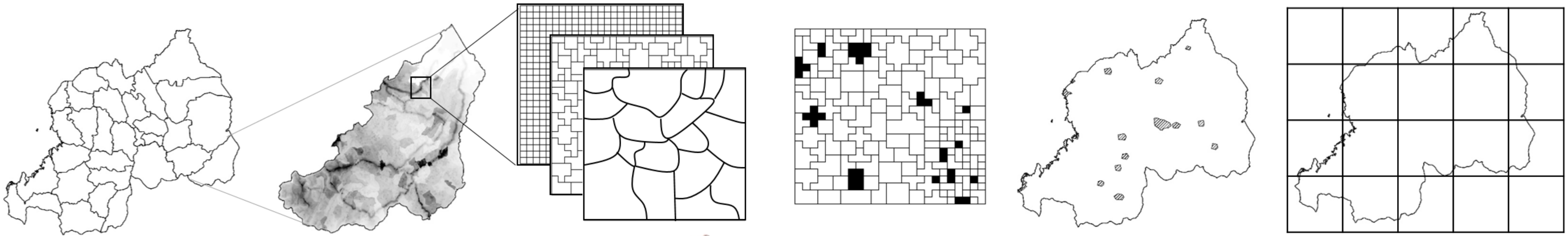
Conclusions: Gridded population sampling is a promising alternative to typical census-based sampling when census data are moderately outdated or inaccurate. Four approaches to implementation have been tried: (1) using gridded PSU boundaries produced by *GridSample*, (2) manually segmenting gridded PSU using satellite imagery, (3) non-probability sampling (e.g. random-walk, "spin-the-pen"), and random sampling of households. Gridded population sampling is in its infancy, and further research is needed to assess the accuracy and feasibility of gridded population sampling. The *GridSample* R algorithm can be used to forward this research agenda.

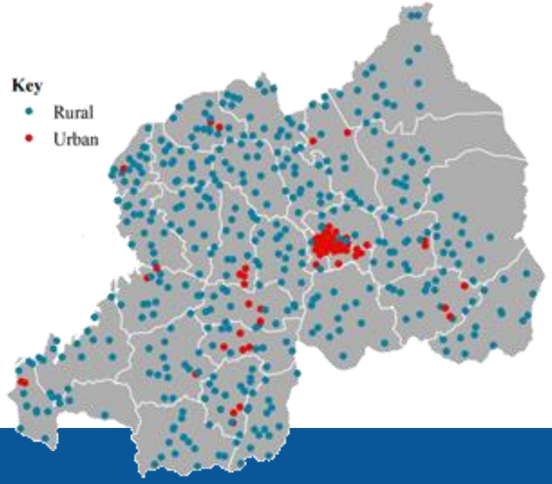
Keywords: Cluster survey, Multi-stage, Cluster sample

	GridSample **R package (2016)**	**GridSample2.0** **GridSample.org (2019)**
Pros	Free, public code Use on a PC, offline Customizable – own gridded population, own shp files	Free, public code which can be used and customized offline Free point-n-click website Preloaded or custom files online Sample probabilities are exact
Cons	Max out RAM on a PC when sample covers large area “Growth” algorithm sample probabilities are a proxy, not exact Requires intermediate R skills	Not yet available

GridSample2.0

(Optional) Stratify by subnational region → Create and list single-cell, multi-cell, or polygon units from gridded population data → Sample from unit listing (PSUs) → (Optional) oversample urban/rural → (Optional) oversample in space

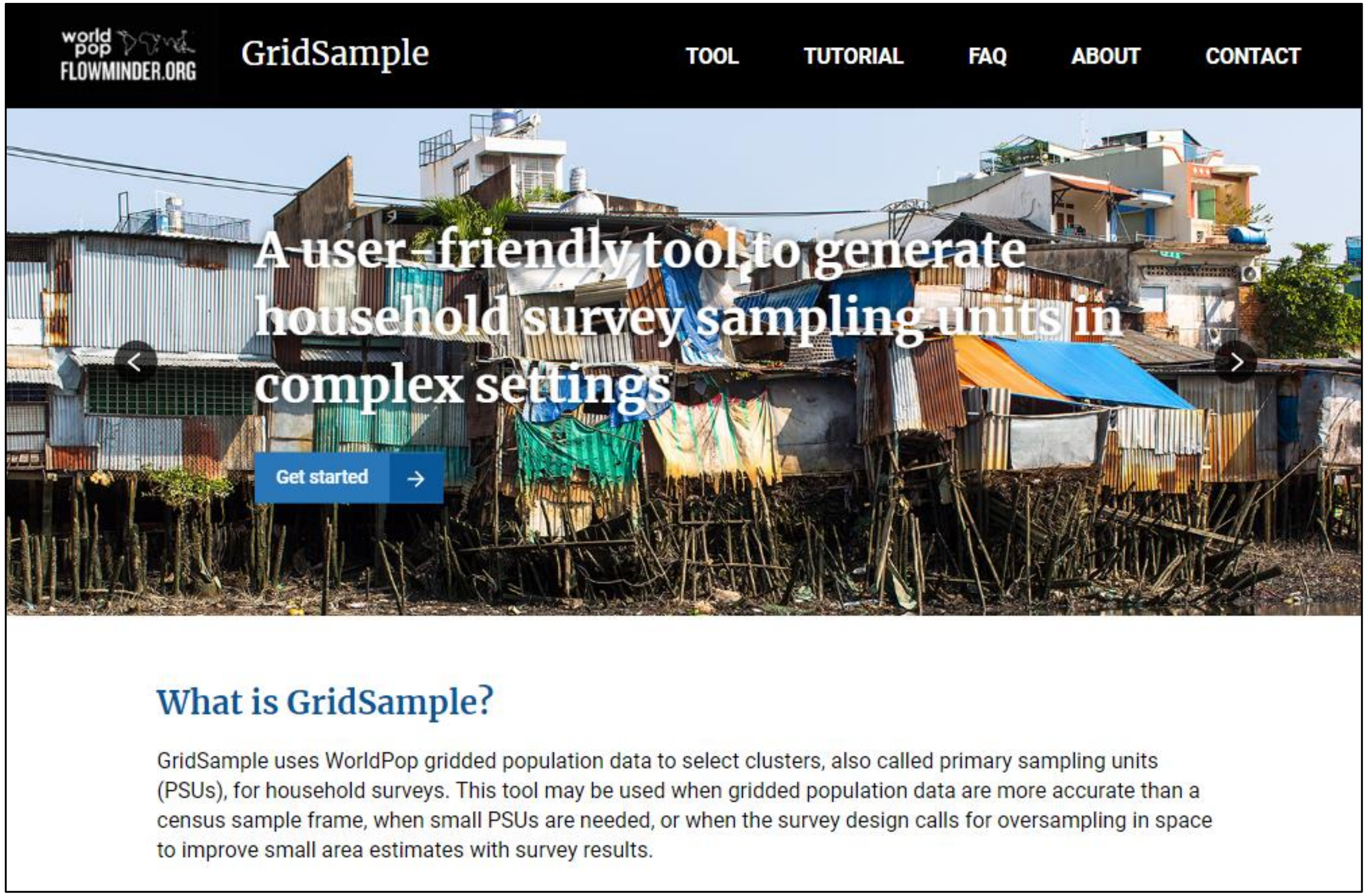




Output:

Gridded PSU boundaries with specified population and maximum area

GridSample2.0 (BETA)



A user-friendly tool to generate household survey sampling units in complex settings

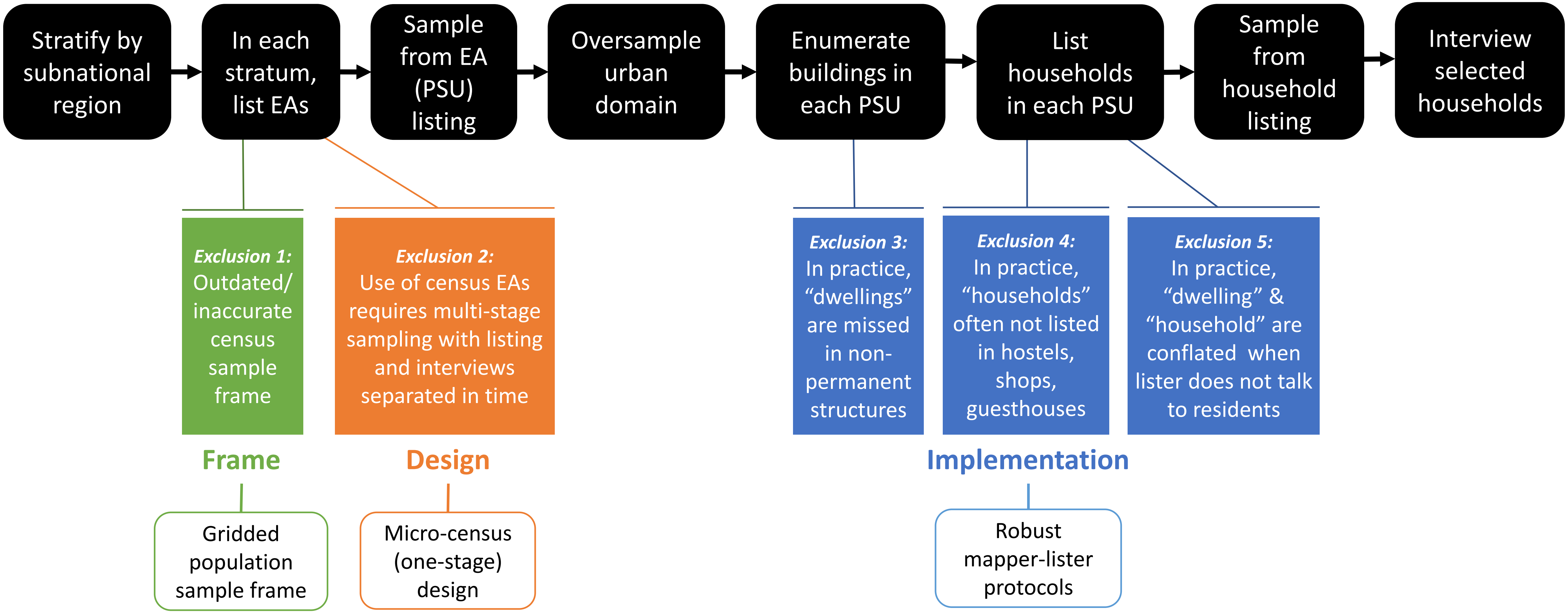
What is GridSample?

GridSample uses WorldPop gridded population data to select clusters, also called primary sampling units (PSUs), for household surveys. This tool may be used when gridded population data are more accurate than a census sample frame, when small PSUs are needed, or when the survey design calls for oversampling in space to improve small area estimates with survey results.

Surveys using GridSample R algorithm

Year	Country	Coverage	Frame	Implementer	More information
2010	DR Congo	2 sub-districts	LandScan Global	Harvard	Thomson, et al. (2012)
2015	Nepal	Metro area	WorldPop	Leeds / HERD	Elsey, et al. (2016)
2017	Somalia	National	WorldPop	World Bank	
2017	Nepal Bangladesh Vietnam	Metro area 2 wards 1 district	WorldPop	Surveys for Urban Equity consortium	SUE project website (Leeds)
2017	Mozambique	6 districts	WorldPop	World Vision International	
2017	DR Congo Uganda	Kinshasa Kampala	own 50m grid	World Food Programme	

Evaluating methods to improve survey accuracy



Stratify by subnational region → In each stratum, list EAs → Sample from EA (PSU) listing → Oversample urban domain → Enumerate buildings in each PSU → List households in each PSU → Sample from household listing → Interview selected households

Exclusion 1: Outdated/ inaccurate census sample frame

Exclusion 2: Use of census EAs requires multi-stage sampling with listing and interviews separated in time

Exclusion 3: In practice, “dwellings” are missed in non-permanent structures

Exclusion 4: In practice, “households” often not listed in hostels, shops, guesthouses

Exclusion 5: In practice, “dwelling” & “household” are conflated when lister does not talk to residents

Frame

Gridded population sample frame

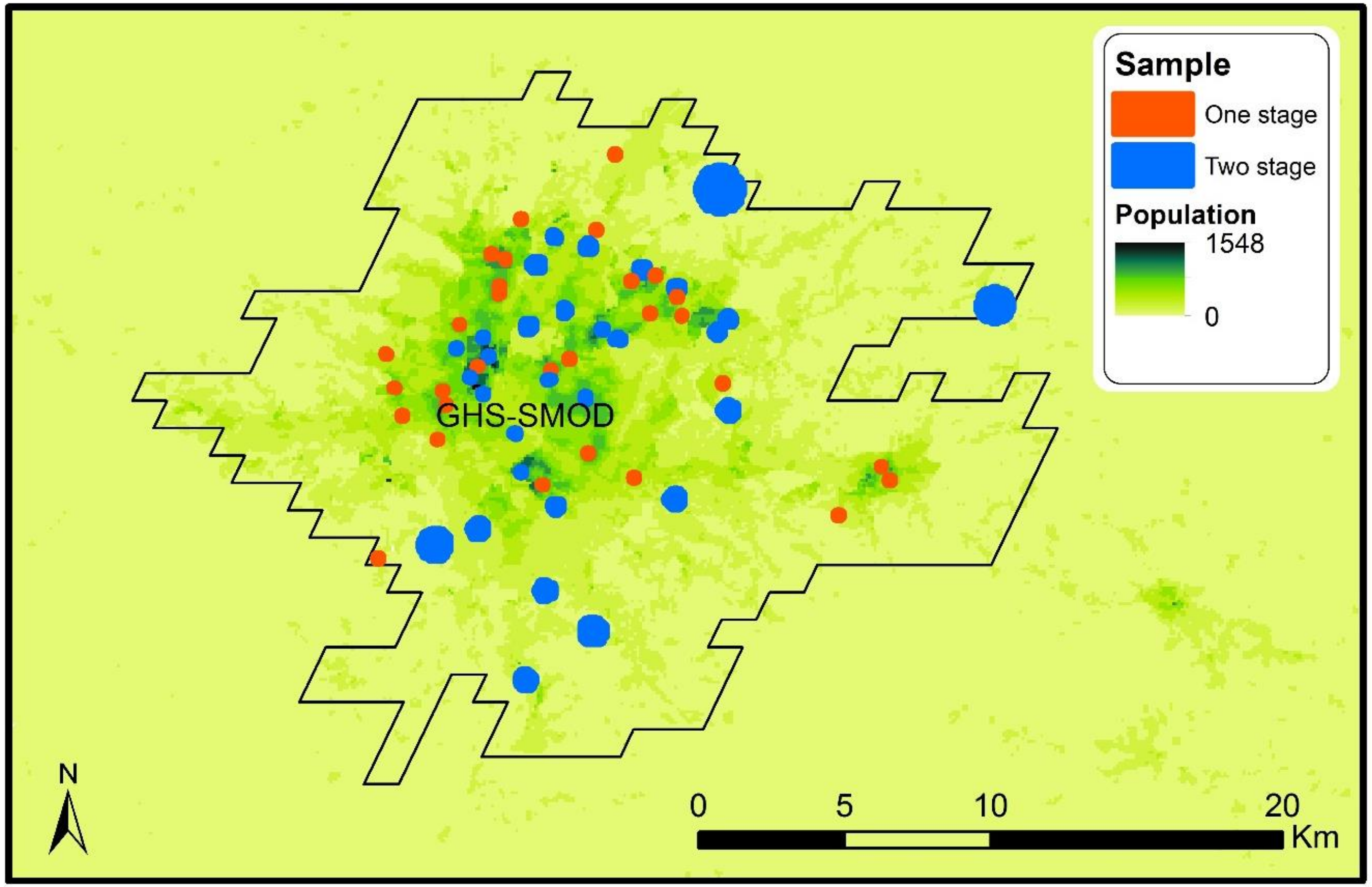
Design

Micro-census (one-stage) design

Implementation

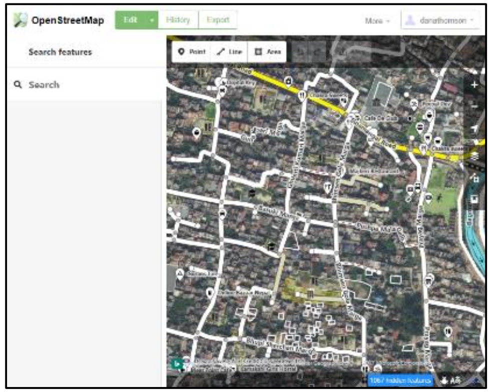
Robust mapper-lister protocols

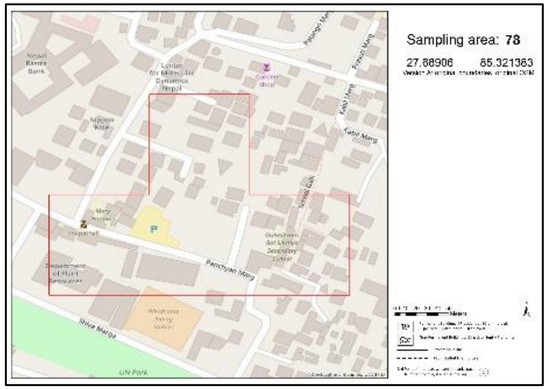
Case study: Surveys for Urban Equity



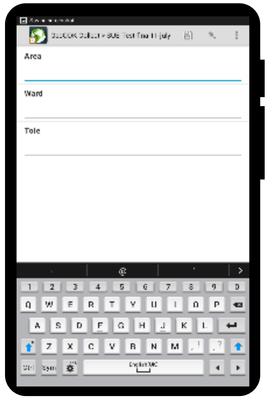
Implementation: Surveys for Urban Equity

Pre-field enumeration in OpenStreetMap → Field enumeration (paper or digital) → List households (paper or digital) → Interview selected households




Sampling area: 78

27.88906 85.321383


Area

Ward

Tole

Output:

Geographically accurate digital map of each PSU, and a digital listing of households

Implementation: Surveys for Urban Equity

Listing quality	Micro-census (one-stage)	Two-stage
Target households	20	200
Listed households - median	19	132
- mean	18	160
- range	6-37	37-483

Case study: Surveys for Urban Equity

- Technical Feasibility
  - Geographically accurate maps essential in complex urban settling
  - Vertical listing: Building → Level → Dwelling → Household
  - Paper map used in field
    - Communicate with community members
    - Paper record for quality assurance
    - Record residential shacks/tents offline to prevent harm
- Time / Cost Savings
  - Two-stage – same cost and time as typical census-based survey
  - Micro-census – cheaper and faster than typical census-based survey
- Improved Accuracy
  - Micro-census sample: more adult men from non-family, lower-income households

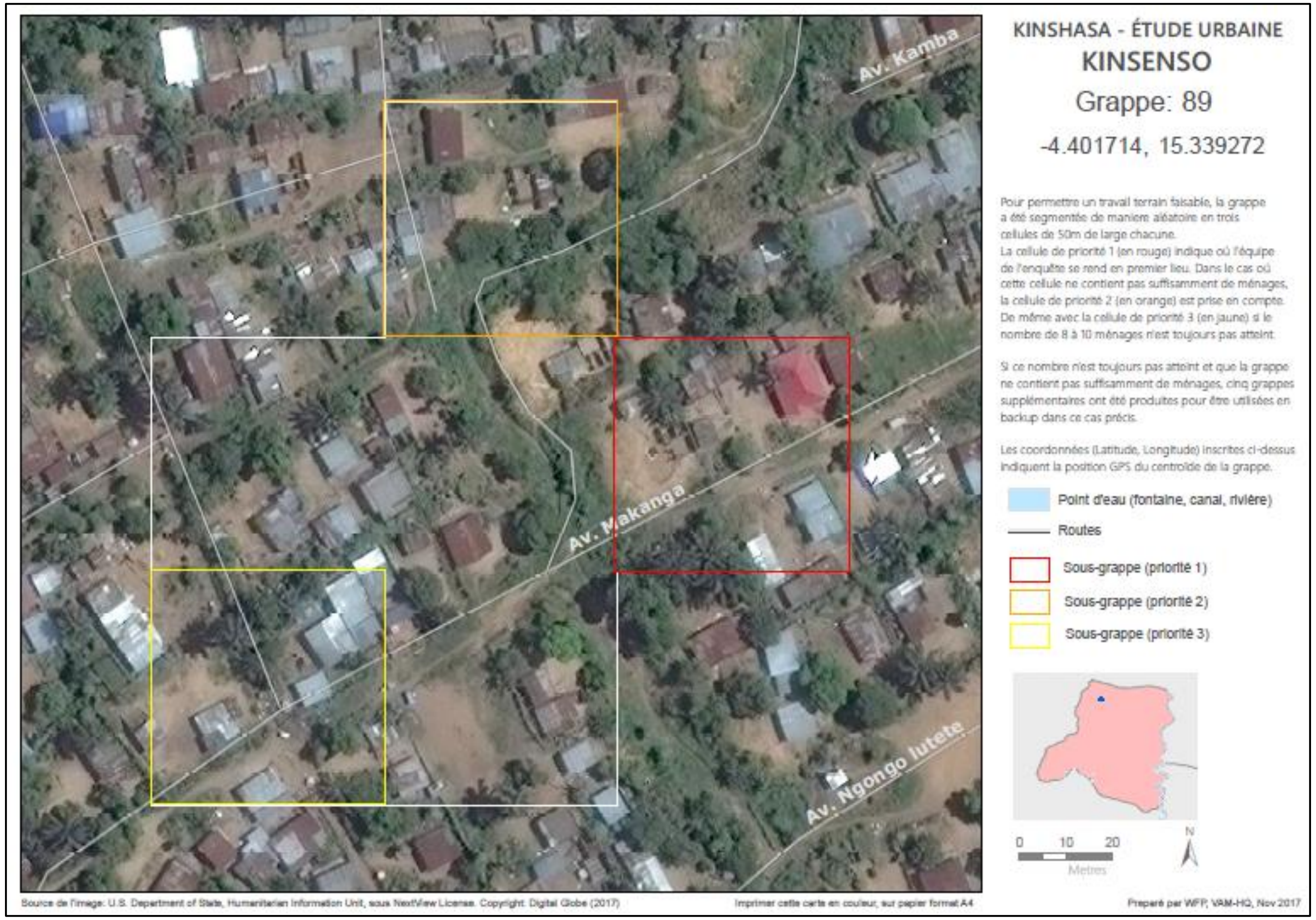
Different household composition in one-stage

Household (DHS/MICS def.)	One-stage	Two-stage
Size (mean)	4.4	4.7
Configuration		
Single or non-family	10.3%	6.7%
Family	86.6%	91.4%
Family + servant	1.5%	1.4%
Other	1.5%	0.6%
N (weighted)	510	590

Household composition affects outcome estimates

Outcome (DHS/MICS def.)	One-stage	Two-stage
Unimproved water	3.5%	1.9%
Unimproved sanitation	1.6%	1.1%
Overcrowding (>3/room)	10.0%	12.6%
Non-cement-bonded walls	1.8%	9.2%
Insecure tenure	6.5%	4.6%
Monthly income (rupees)	40,785	41,106
N (weighted)	510	590

Case study: World Food Programme



KINSHASA - ÉTUDE URBAINE

KINSENSO

Grappe: 89

-4.401714, 15.339272

Pour permettre un travail terrain faisable, la grappe a été segmentée de maniere aléatoire en trois cellules de 50m de large chacune.
La cellule de priorité 1 (en rouge) indique où l'équipe de l'enquête se rend en premier lieu. Dans le cas où cette cellule ne contient pas suffisamment de ménages, la cellule de priorité 2 (en orange) est prise en compte. De même avec la cellule de priorité 3 (en jaune) si le nombre de 8 à 10 ménages n'est toujours pas atteint.

Si ce nombre n'est toujours pas atteint et que la grappe ne contient pas suffisamment de ménages, cinq grappes supplémentaires ont été produites pour être utilisées en backup dans ce cas précis.

Les coordonnées (Latitude, Longitude) inscrites ci-dessus indiquent la position GPS du centroïde de la grappe.

Source de l'image: U.S. Department of State, Humanitarian Information Unit, sous NextView License. Copyright: Digital Globe (2017)

Imprimer cette carte en couleur, sur papier format A4

Preparé par WFP, VAM-HQ, Nov 2017

Applications

- Who
  - Program evaluations, research studies (Government, academic, NGO)
  - Rapid assessments, M&E (NGOs)
  - Municipal and district governments (Official statistics, priority setting)
  - National surveys (Official statistics, priority setting)
- Why
  - Overcome outdated/inaccurate census sample frame
  - Define smaller (or custom-sized) clusters
  - Sample population with PPS + oversample in space for improved SAE with results
  - Prevent non-sampling error during fieldwork
  - Costs and time savings
  - Leverage existing OSM maps, satellite imagery, and population estimates

Thank you!

worldpop

FLOWMINDER.ORG

dana.r.thomson@gmail.com

References

- GridSample (2019-expected). www.gridsample.org
- Surveys for Urban Equity. (2018) https://medhealth.leeds.ac.uk/info/691/research/2388/sue
- Thomson, et al. (2017) https://doi.org/10.1186/s12942-017-0098-4
- Elsey, et al. (2016) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4899330/
- Thomson, et al. (2012) https://doi.org/10.1186/1471-2458-12-959