GridSample

User Manual

FLOWMINDER.ORG

This manual was authored by Dana R Thomson from Flowminder Foundation.
Version 2.1 (Updated: October 2019)

Recommended citation: Thomson, Dana R. 2019. GridSample User Manual. Flowminder Foundation: Southampton UK. Available at: http://gridsample.org/tutorial

Contents

About 3

Why use GridSample? 3

Before using GridSample 5

Survey Design 5

Using GridSample 7

Sample selection 7

How GridSample works 13

GridSample output 14

Implementing your survey 15

Approaches & tools 15

After your survey 19

Sample weights 19

Data cleaning & analysis 23

Practice exercises 24

Exercise 1: 2014-15 Rwanda DHS 24

Exercise 2: 2017 Kathmandu SUE 27

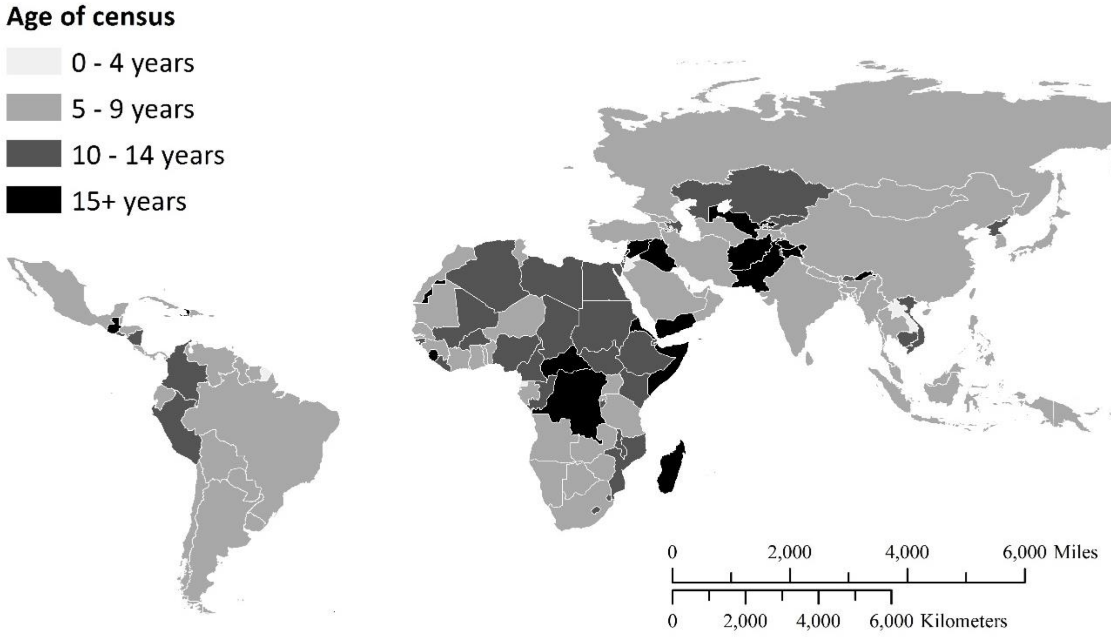
Additional resources 31

Exercise answers 33

About

Why use GridSample?

Outdated or inaccurate census sample frame. Standard “top-down” gridded population datasets, such as WorldPop-Global, reflect the population totals in census data. While census population totals may be inaccurate, the relative distribution of the population in WorldPop-Global may be more accurate than the original census data, and thus serve as a better sample frame. This is because gridded population datasets disaggregate population totals to areas with new population growth (reflected in new housing developments, roads, and changes in land cover) and in informal settlements not originally counted in the census.



(updated April 2019)

Spatial oversampling to improve small area estimates. In household surveys, it is common to collect latitude-longitude coordinates of clusters, and to use survey results to generate small area estimates in administrative areas smaller than strata. Small area estimates based on standard samples of the population distribution are prone to reduced accuracy in remote rural areas, and within highly heterogeneous urban populations. Spatial oversampling ensures spatial coverage of the sample and a base-level of accuracy in small area estimates across the study area.

Custom sizes clusters. In standard census-based sampling, cluster size is determined by the size of census enumeration areas, with each cluster containing approximately 500 people. When the sample frame is generated from a gridded population dataset such as WorldPop-Global, there is good reason to consider clusters of different sizes.

> Small clusters (~75 people each) can be generated to perform one-stage sampling. One-stage sampling in which all eligible households in the cluster are interviews, is an attractive option in complex settings where large numbers of vulnerable and mobile populations live. One-stage sampling allows for the household listing and interviews to be conducted on the same day, removing the months, or even years, long delay between listing and interviewers faced in standard multi-stage sampling.
>
> Medium clusters (~500 people each) are approximately the size of a census enumeration area and can be used to replicate standard household survey designs and methods with a gridded population sample frame.
>
> Large clusters (~1200 people each) are an attractive option in settings where the census is extremely outdated, for example more than 15 years old. WorldPop-Global estimates are generally more accurate as its 100m X 100m grid cells are aggregated into larger units. The use of larger areas in the sample frame helps to smooth out errors in the gridded population dataset.

Before using GridSample

Survey Design

Before using GridSample, have a clear research question, decide the survey design and calculate sample size.

Several resources are available to assist with survey planning including materials produced by the World Health Organization (WHO) Vaccination Coverage Cluster Surveys, Multiple Indicator Cluster Surveys (MICS), and Demographic and Health Surveys (DHS).

The primary question affects the survey design, sample size, and budget. Identify the type of question that the survey aims to address.

Calculating a sample size that both meets the inferential goals of the survey and the budget constraints is an iterative process that must be negotiated within a survey steering group. Use the WHO Vaccination Coverage Survey sample size + budget spreadsheet to generate multiple side-by-side scenarios.

Checklist before using GridSample

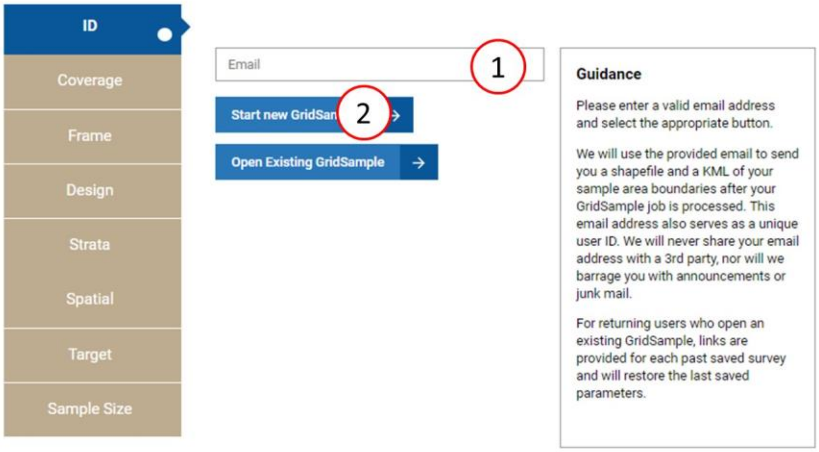
- I know the coverage area of the survey
- I know the year the survey will be implemented
- I know whether the survey will be stratified, and the geographic boundaries of those strata
- I know the multi-stage cluster sampling design
- I know the target cluster size
- I know the target sample size (number of clusters, and households per cluster)

Classification question	**Descriptive or estimation question**	**Comparative or hypothesis-test question**
A classification survey labels groups as “pass”, “marginal”, or “fail” to inform programmatic decisions	*An estimation survey results in quantitative estimates in one or more groups*	*A comparative survey make a quantitative estimate of difference between two groups, or of change between two time points*
Example question: Which health districts have low coverage of antenatal care for pregnant women?	Example question: What percent of pregnant women receive one or more antenatal care visits?	Example question: Has the percent of pregnant women receiving antenatal care increased since the last survey?
Inferential goal: • *Pass/fail threshold* *Uncertainty is reported as:* • *Misclassification of pass* • *Misclassification of fail*	*Inferential goal:* • *Coverage estimate* *Uncertainty is reported as:* • *Confidence interval* *Design parameter:* • α *(alpha) - the probability of Type 1 error - the hypothesis test declares the difference to be statistically significant when, in truth, it is not*	*Inferential goal:* • *Minimum detectable difference between 2 groups* *Uncertainty is reported as:* • *Confidence interval* *Design parameter:* • α *(alpha) - the probability of Type 1 error - the hypothesis test declares the difference to be statistically significant when, in truth, it is not* • *1-*β *(beta) - the power of the test - this is the probability that the hypothesis correctly identified a statistically significant difference*
Example parameters: • Pass/fail threshold: 90% • Misclassify passes: 5% • Misclassify fails: 10%	Example parameters: • α = 95% (95% confidence interval)	Example parameters: • difference = 10% or more • 2-sided hypothesis test • α = 95% (95% confidence interval) • power = 80% (β = 20%)
Generally requires 15+ PSUs per stratum	Generally requires 30+ PSUs per stratum	Sample size is highly dependent on the difference between groups and design parameters. 30-60 PSUs per stratum.

Using GridSample

Sample selection

Sample selection steps are demonstrated below with Exercise 1, reproducing the survey design of the 2015 Rwanda Demographic and Health Survey in GridSample.



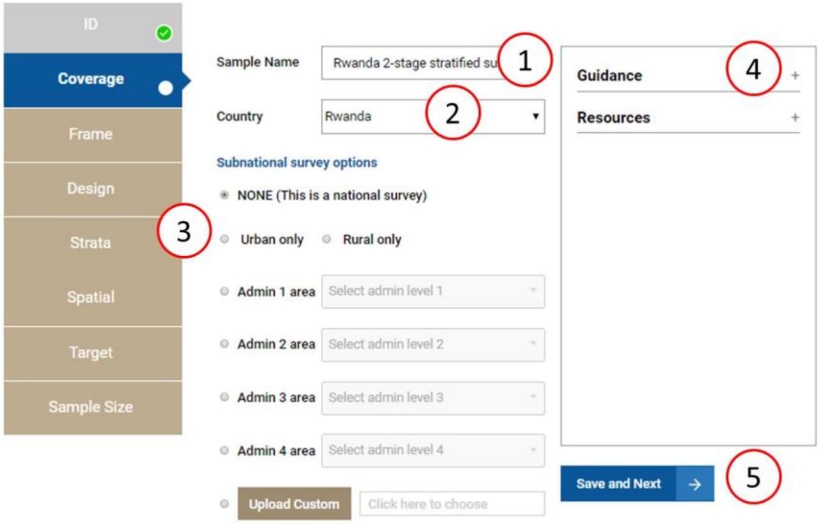
Guidance

Please enter a valid email address and select the appropriate button.

We will use the provided email to send you a shapefile and a KML of your sample area boundaries after your GridSample job is processed. This email address also serves as a unique user ID. We will never share your email address with a 3rd party, nor will we barrage you with announcements or junk mail.

For returning users who open an existing GridSample, links are provided for each past saved survey and will restore the last saved parameters.

1. Enter an email address to receive your sample output
2. Select “Stare new GridSample” or open an existing GridSample job



Sample Name: Rwanda 2-stage stratified su

Country: Rwanda

Subnational survey options

● NONE (This is a national survey)

○ Urban only ○ Rural only

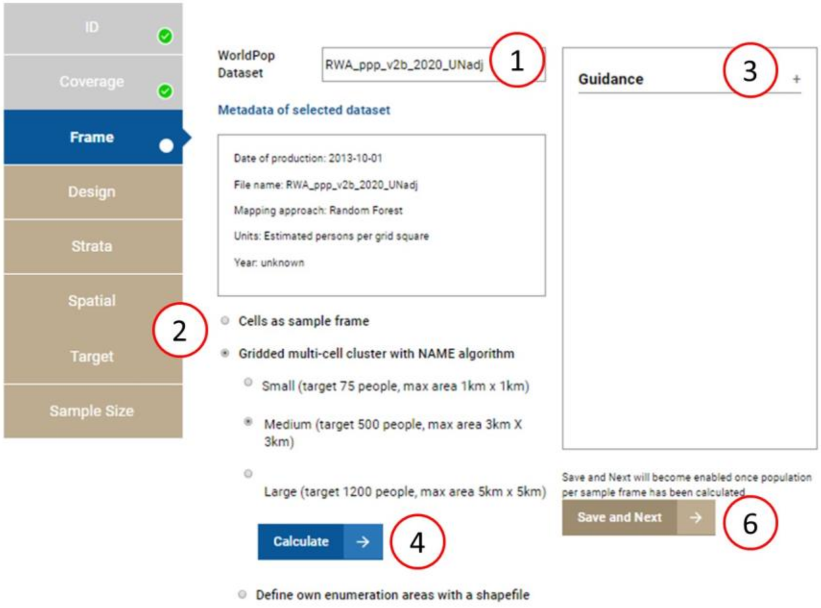
○ Admin 1 area: Select admin level 1

○ Admin 2 area: Select admin level 2

○ Admin 3 area: Select admin level 3

○ Admin 4 area: Select admin level 4

1. Give your sample a name
2. Select the country
3. Define the survey coverage area
4. See the guidance text if needed
5. Select “Save and Next” when done defining coverage



WorldPop Dataset: RWA_ppp_v2b_2020_UNadj

Metadata of selected dataset

Date of production: 2013-10-01
File name: RWA_ppp_v2b_2020_UNadj
Mapping approach: Random Forest
Units: Estimated persons per grid square
Year: unknown

○ Cells as sample frame

● Gridded multi-cell cluster with NAME algorithm

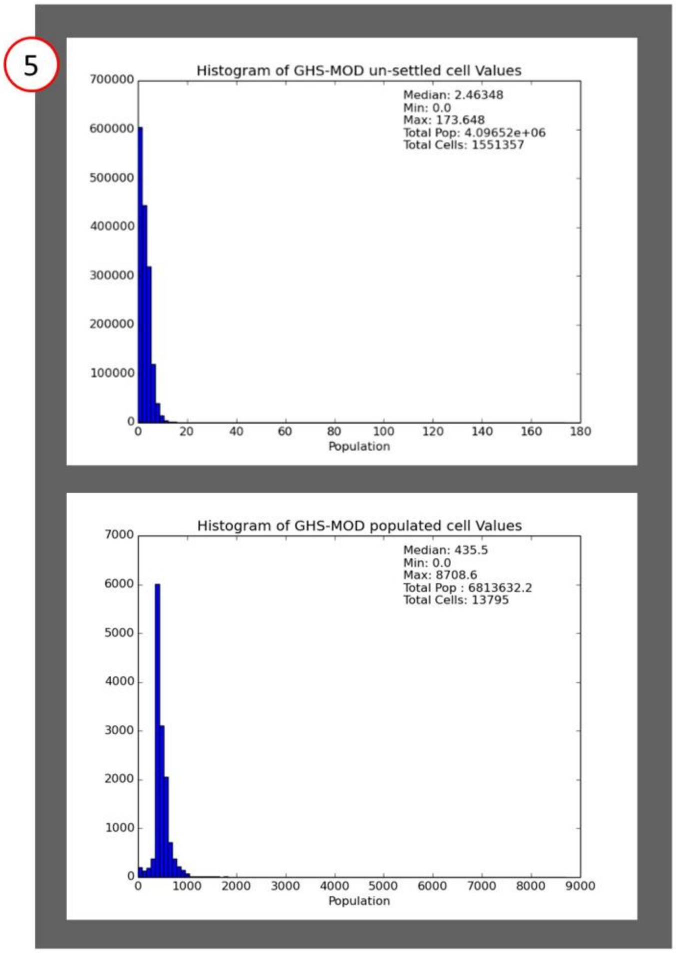
○ Small (target 75 people, max area 1km x 1km)

● Medium (target 500 people, max area 3km X 3km)

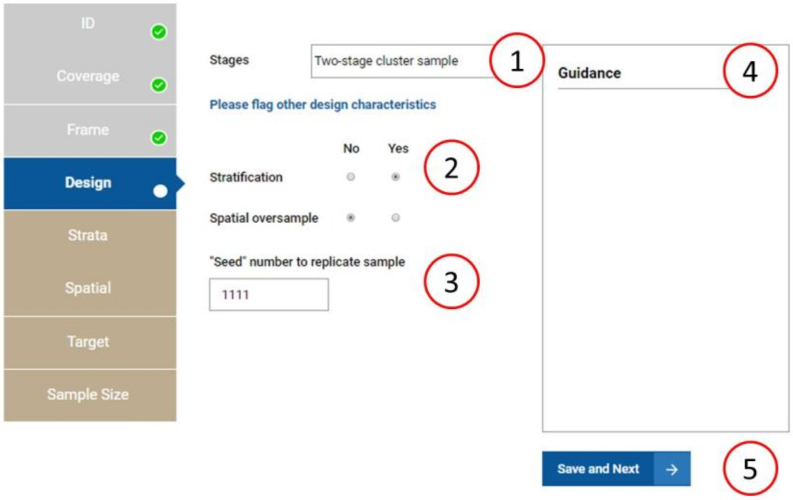
○ Large (target 1200 people, max area 5km x 5km)

○ Define own enumeration areas with a shapefile

Save and Next will become enabled once population per sample frame has been calculated.



1. Select the WorldPop dataset to use as the foundation of the sample frame
2. Define the sample frame unit
3. See the guidance text if needed
4. Calculate population per sample frame unit
5. Review the histogram(s) of sample frame unit population estimates
6. When you are satisfied with the gridded population-based sample frame, select “Save and Next”



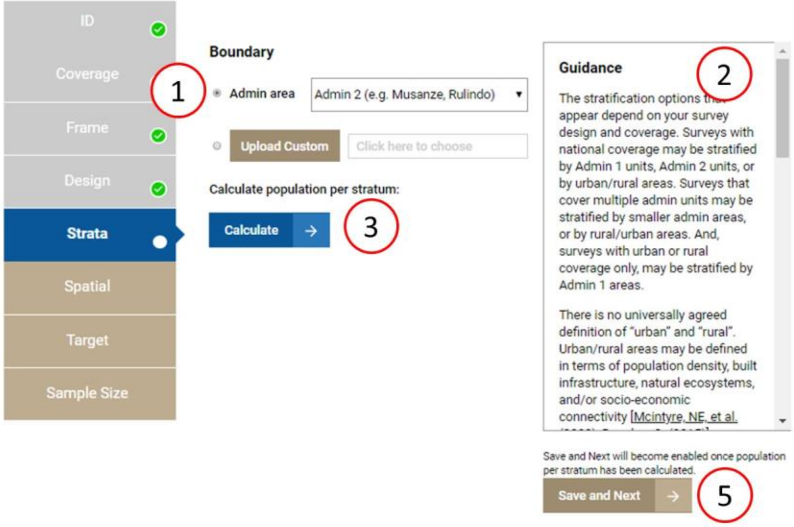
Stages: Two-stage cluster sample

Please flag other design characteristics

	No	Yes
Stratification	○	●
Spatial oversample	●	○

"Seed" number to replicate sample: 1111

1. Define whether the survey will have one or two stages of sampling
2. Select whether you will use stratification or sample oversampling
3. Optionally specify a "seed" number to replicate the sample
4. See the guidance text if needed
5. Select "Save and Next" when done flagging the sample design features



Boundary

● Admin area: Admin 2 (e.g. Musanze, Rulindo) ▾

○ Upload Custom: Click here to choose

Calculate population per stratum:

Calculate →

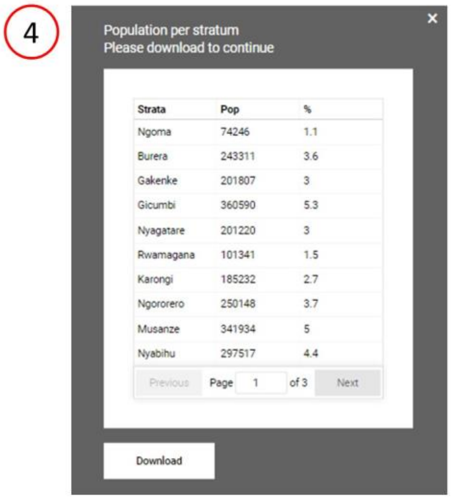
Guidance

The stratification options that appear depend on your survey design and coverage. Surveys with national coverage may be stratified by Admin 1 units, Admin 2 units, or by urban/rural areas. Surveys that cover multiple admin units may be stratified by smaller admin areas, or by rural/urban areas. And, surveys with urban or rural coverage only, may be stratified by Admin 1 areas.

There is no universally agreed definition of "urban" and "rural". Urban/rural areas may be defined in terms of population density, built infrastructure, natural ecosystems, and/or socio-economic connectivity [Mcintyre, NE, et al.

Save and Next will become enabled once population per stratum has been calculated.

Save and Next →



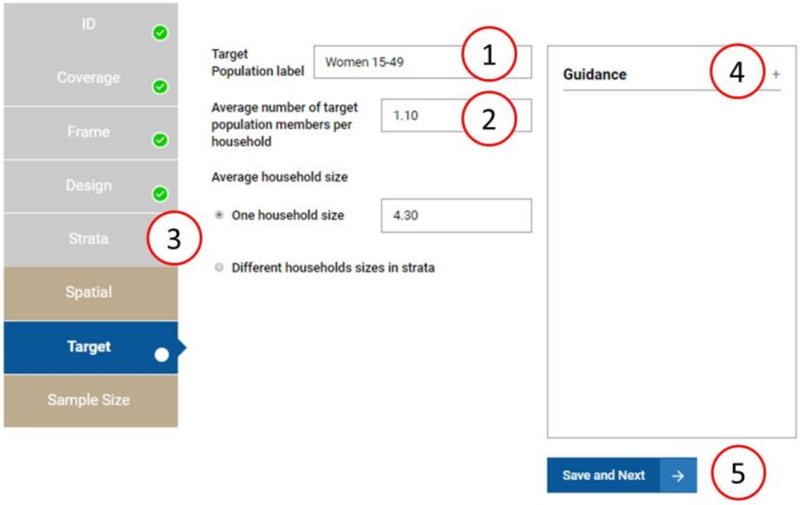
Population per stratum
Please download to continue

Strata	Pop	%
Ngoma	74246	1.1
Burera	243311	3.6
Gakenke	201807	3
Gicumbi	360590	5.3
Nyagatare	201220	3
Rwamagana	101341	1.5
Karongi	185232	2.7
Ngororero	250148	3.7
Musanze	341934	5
Nyabihu	297517	4.4

Previous Page 1 of 3 Next

Download

1. Define which boundaries will define strata
2. See the guidance text if needed
3. Calculate population per stratum
4. Review the population per stratum
5. If satisfied with the strata, select "Save and Next"



Target Population label: Women 15-49

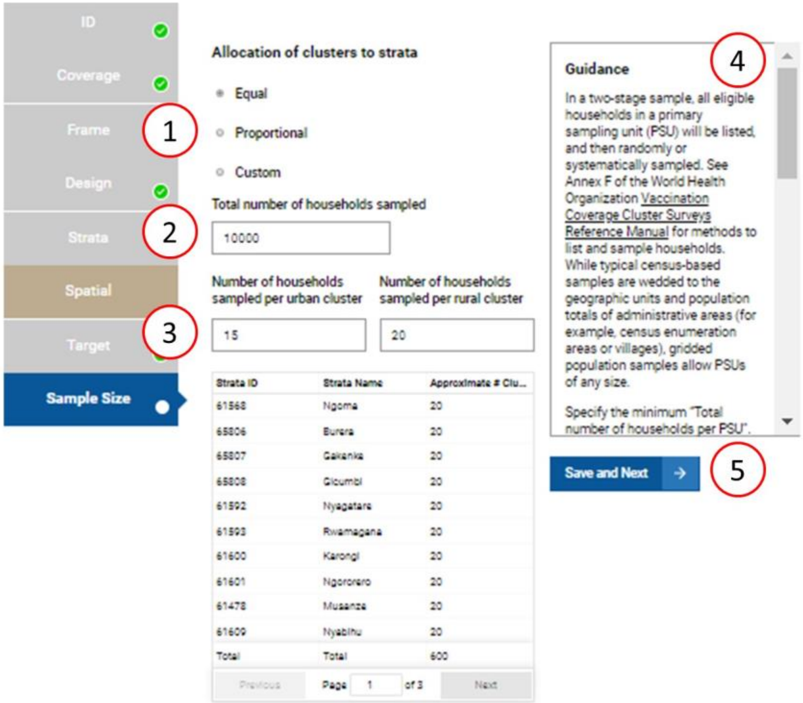
Average number of target population members per household: 1.10

Average household size

● One household size: 4.30

○ Different households sizes in strata

1. Give a descriptive label to your target population
2. Enter the average number of target population members per household (eg from DHS report)
3. Enter the average household size
4. See the guidance text if needed
5. When the target population is defined, select “Save and Next”



Allocation of clusters to strata

● Equal

○ Proportional

○ Custom

Total number of households sampled: 10000

Number of households sampled per urban cluster: 15

Number of households sampled per rural cluster: 20

Strata ID	Strata Name	Approximate # Clu...
61568	Ngoma	20
65806	Burera	20
65807	Gakenke	20
65808	Gicumbi	20
61592	Nyagatare	20
61593	Rwamagana	20
61600	Karongi	20
61601	Ngororero	20
61478	Musanze	20
61609	Nyabihu	20
Total	Total	600

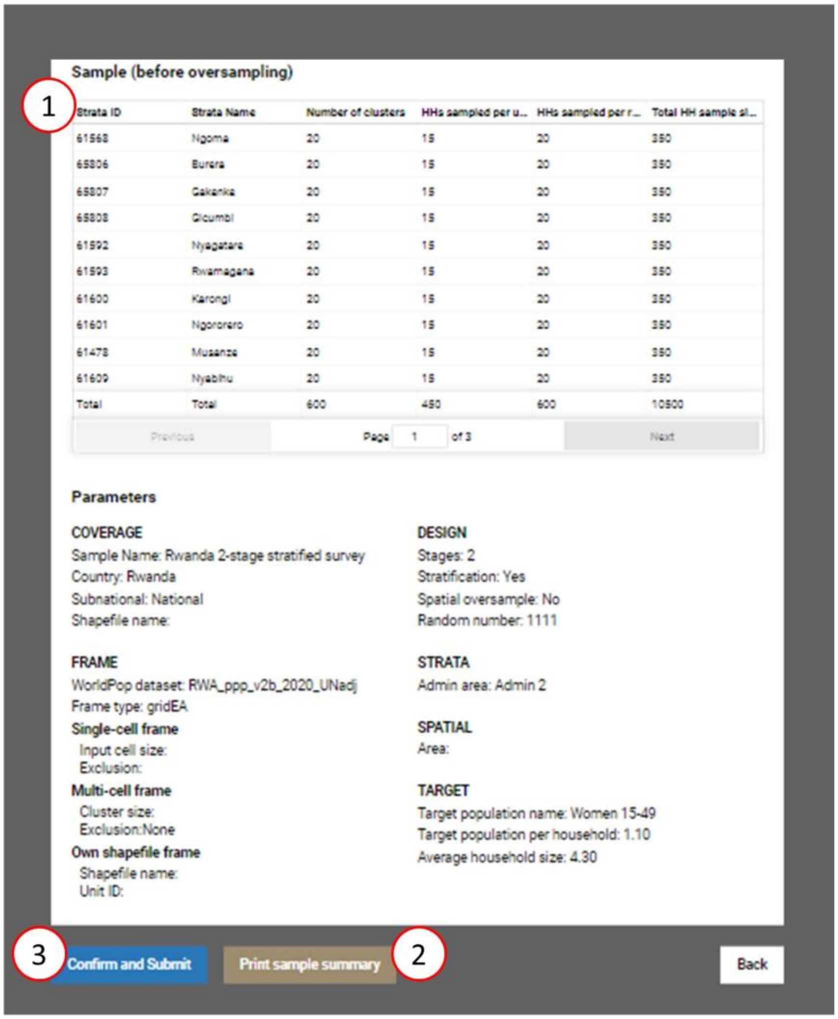
Page 1 of 3

Guidance

In a two-stage sample, all eligible households in a primary sampling unit (PSU) will be listed, and then randomly or systematically sampled. See Annex F of the World Health Organization Vaccination Coverage Cluster Surveys Reference Manual for methods to list and sample households. While typical census-based samples are wedded to the geographic units and population totals of administrative areas (for example, census enumeration areas or villages), gridded population samples allow PSUs of any size.

Specify the minimum "Total number of households per PSU".

1. If sample is stratified, define how to allocate clusters to strata
2. Enter total number of household to be sampled
3. If a two-stage sample, specify the number of households to be sampled per cluster
4. See the guidance text if needed
5. When the sample size is entered, select “Save and Next”



Sample (before oversampling)

Strata ID	Strata Name	Number of clusters	HHs sampled per u...	HHs sampled per r...	Total HH sample si...
61568	Ngoma	20	15	20	350
65806	Burera	20	15	20	350
65807	Gakenke	20	15	20	350
65808	Gicumbi	20	15	20	350
61592	Nyagatare	20	15	20	350
61593	Rwamagana	20	15	20	350
61600	Karongi	20	15	20	350
61601	Ngororero	20	15	20	350
61478	Musanze	20	15	20	350
61609	Nyabihu	20	15	20	350
Total	Total	600	450	600	10500

Parameters

COVERAGE
Sample Name: Rwanda 2-stage stratified survey
Country: Rwanda
Subnational: National
Shapefile name:

FRAME
WorldPop dataset: RWA_ppp_v2b_2020_UNadj
Frame type: gridEA
Single-cell frame
Input cell size:
Exclusion:
Multi-cell frame
Cluster size:
Exclusion:None
Own shapefile frame
Shapefile name:
Unit ID:

DESIGN
Stages: 2
Stratification: Yes
Spatial oversample: No
Random number: 1111

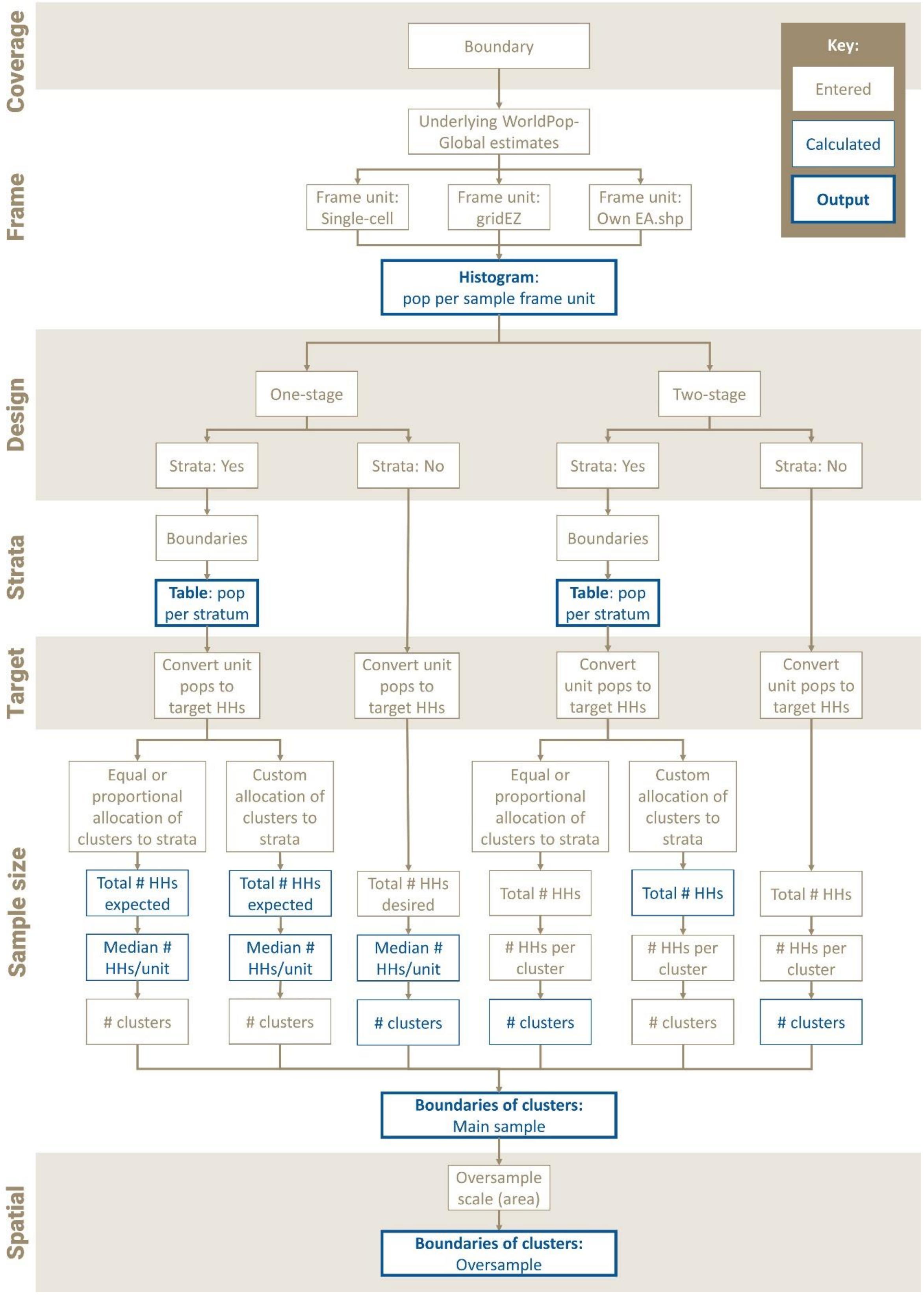
STRATA
Admin area: Admin 2

SPATIAL
Area:

TARGET
Target population name: Women 15-49
Target population per household: 1.10
Average household size: 4.30

1. Review the sample summary
2. Optionally print the summary to store for your records
3. When ready, select "Confirm and Submit"

How GridSample works



GridSample output

After a GridSample job is complete, the user receives an email with a link to download the following as a 7zip file:

Shapefile and **KML** file of the selected cluster boundaries

- Shapefile can be opened with ArcGIS or QGIS
- KML file can be opened with Google Earth

Excel spreadsheet with shapefile attributes

- Use attribute values to calculate sample weights

PDF report summarising the input files and parameters

- Summary of survey parameters
- Map of each input dataset, and links to original input data sources

Key to shapefile and Excel attributes

Attribute	Description
cl_id:	Unique numeric ID of cluster
str_id:	Unique number ID of strata
cl_type:	Whether cluster was selected during "main" PPS sampling or spatial "oversample"
s_cl_st:	Number of selected clusters in stratum
s_cl_tot:	Number of selected clusters in coverage area
s_cl_pop:	Estimated population in selected cluster
s_cl_hh:	Estimated households in selected cluster = s_cl_pop / st_hhsiz
urb_rur:	GHS-SMOD classification where 3=high dense urban, 2=low dense urban, 1=rural, 0=unsettled
u_n:	Number of sample frame units in the survey coverage area
u_medpop:	Median population per sample frame unit
st_hhsiz:	Average household size specified in this stratum
st_pop_n:	Estimated population in stratum
st_pop_p:	Percent of population in the coverage area falling in this stratum
st_hh_n:	Estimated households in stratum = st_pop_n / st_hhsiz
st_hh_p:	Percent of households in the coverage area falling in this stratum
st_name:	Unique alphanumeric name of stratum
ori_id	Unit (cluster) ID from the original sample frame

Implementing your survey

Approaches & tools

Four approaches and various tools are available to implement your gridded population survey.

300m X 300m cell

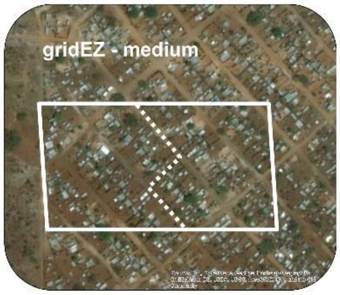
Cell

Grid cells of any size can be generated and sampled with GridSample. Cells will always be a multiple of 100m X 100m grid cells. This approach results in a sample frame of units that are uniform in size, but not in population. This approach is generally not used on its own for household survey sampling, but might be useful for other applications.

gridEZ - medium

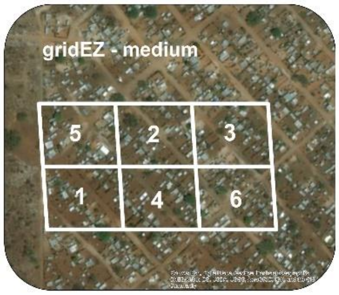
gridEZ

A sample frame of multi-cell units with approximately the same population in each unit and maximum area can be generated with the gridEZ algorithm. Multi-cell units can be sampled and used directly, segmented manually along natural features, or segmented automatically along 100m X 100m cell boundaries for field work.



gridEZ original boundaries. This approach was used in Nepal by Elsey et al (2016). It was also used in Mozambique by World Vision International (2018) and is a featured case study.

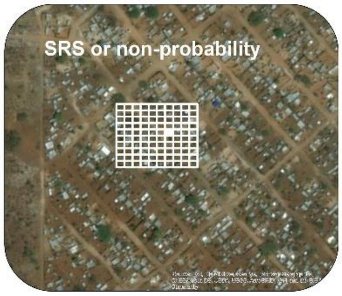
gridEZ, segment along natural boundaries. This approach was used in Myanmar by Munoz and Langeraar (2013) and in Somalia by Pape and Wollberg (2019). It was also used in Nepal by Elsey et al (2018) and is a featured case study.



gridEZ, segment along sub-cell boundaries. This approach was used by the World Food Programme (2018) in DR Congo and is a featured case study. Sub-cells were randomized and fully enumerated until the target number of households per cluster was achieved, allowing for one field visit and calculation of sample probability weights as a one-stage segmented survey design.

Own EA.shp boundaries

GridSample allows users to define custom sample frame boundaries by uploading a zipped shapefile. This approach is suitable if census enumeration areas or other boundaries are available, and population estimates need to be updated. The output from GridSample and fieldwork approach is identical to a standard household survey.



Simple random sample (SRS) or non-probability

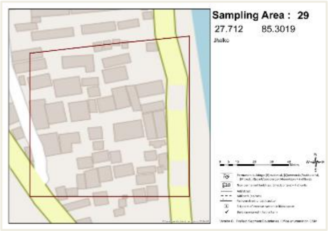
Some researchers, particularly in urban areas or other densely settled areas (e.g. IDP camps) survey a random sample of households. A simple random sample of buildings can be selected via GridSample by first sampling 100m X 100m WorldPop cells, then using a technique such as random point placement, or random selection of mini grid cells in a GIS to identify one building at random.

This approach can also be used to identify a random starting point for non-probability sampling techniques such as “random walk” or “spin-the-pen”. Galway et al. (2012) used this approach in Iraq.

Higher-tech tools

Navigation: In urban areas where most roads are mapped in OpenStreetMap, MAPS.ME allows for offline navigation over very long distances. In areas where OpenStreetMap data are sparse, navigation based on GPS coordinates and place names may be necessary.

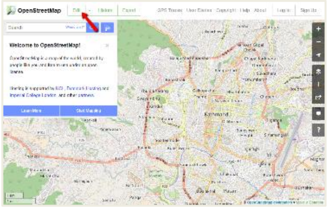
Within clusters, an app such as OSMAnd app can be used for offline navigation based on preloaded MBTiles, displaying a blue dot at the device location.



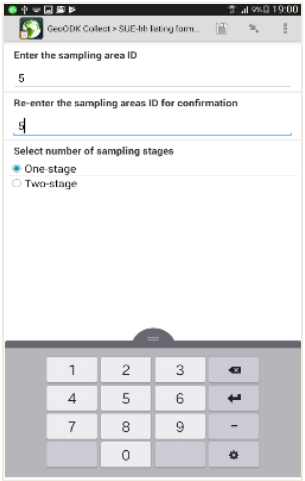
Mapping: Applications such as Vespucci allow for tablet-based updates to OpenStreetMap in the field. While this is a seemingly efficient way to reduce steps during field work, we have found that field-based tablet editing of OpenStreetMap is time-intensive and frustrating for staff working on small screens often in adverse weather conditions.

In our experience, survey teams almost unanimously prefer printed geographic maps in the field with OpenStreetMap or satellite imagery as a base layer. This is because paper maps can be marked up quickly in the field with a pencil, and edits to

OpenStreetMap can be made quickly after field work in the comfort of an office. Furthermore, field teams that have compared paper and tablet mapping say that paper maps produced in ArcGIS or QGIS helped to facilitate positive conversations with residents about the survey while editing maps on tablets in the field fuelled suspicion.



Whichever mapping approach you choose, it is wise to update roads and building footprints in OpenStreetMap using iDeditor or similar tool before visiting the field. GridSample cluster boundaries can be visualized on top of OpenStreetMap as a GPX trace file, keeping the dataset private. QGIS and a number of free apps can be used to transform the GridSample shapefile or KML file of cluster boundaries to a GPX file.


GeoODK Collect > SUE-hh listing form...

Enter the sampling area ID

5

Re-enter the sampling areas ID for confirmation

5

Select number of sampling stages

● One-stage
○ Two-stage

Listing: A number of apps are available to collect household listing data. Many of these apps, including OpenMapKit, GeoODK, and KoBoToolbox, also allow for collection of spatial data. Other tablet-based apps include the World Bank's Survey Solutions tool which enables monitoring of field workers.

Questionnaire: The same apps used for listing - OpenMapKit, GeoODK, Kobo Collect, Survey Solutions - can be used to administer questionnaires.

Tablet-based data collection requires a server, someone to design and configure the data collection form, and someone to set up, secure, and maintain all of the devices. All of the linked apps and programs are free, and most are open source.

The Surveys for Urban Equity guides for survey planners and field teams provide guidance to implement higher-tech field tools and methods for a gridded population survey.

Lower-tech tools

Navigation: If you are working in a context without power, in a team with limited technical skills, or your field staff face high security risks, you will likely opt to use lower-tech tools, and possibly avoiding tablets or GPS units in the field. Lower-tech navigation to clusters can be done with a travel map and asking for directions based on place names.

Mapping: Going lower-tech does not mean that field staff need to sacrifice geographically accurate maps. Two simple options

Cluster 82
Legend
rwanda_survey.kml
Google Earth
N
300 m

are available to produce field maps to navigate and update building footprints in the field.

In rural contexts, satellite imagery from Google Earth is generally a suitable base map. Simply double-click the KML file of cluster boundaries provided in the GridSample output to visualize cluster boundaries in Google Earth. Then zoom to each cluster and print, recording the Cluster ID on each map.

In dense urban contexts where buildings are attached, the OpenStreetMap base layer may be needed to distinguish buildings and walking paths. The Field Papers website can be used to generate a map with the OpenStreetMap or Bing imager base layer for each cluster.

Many current surveys require field staff to hand-sketch maps of roads, buildings, and points of interest on a blank piece of paper. This approach is decades old but does not result in a geographically accurate map, and it is time intensive for field staff.

Listing and questionnaire: Printed paper forms provide a low-tech solution to conduct the household listing and administer questionnaires.

The linked tools are free and open source.

After your survey

Sample weights

Sample weights are necessary to make accurate estimates about the population from a survey with a complex design (e.g. cluster sampling). Sample weight calculations are described below, and an Excel template file is provided at GridSample.org/tutorial to calculate sample weights. Inputs come from three sources of information:

- GridSample: Cluster ID (`cl_id`), strata name (`st_name`), strata ID (`str_id`), number of households in strata (`st_hh_n`), and number of households per cluster (`s_cl_hh`)
- Mapping-listing: Number of clusters visited, number of segments created in each cluster during pre-field and post-field (one-stage only) enumeration, number of households listed in each (segmented) cluster, and number of households selected for sampling in each (segmented) cluster (two-stage only)
- Interview: Number of responded households per cluster (household questionnaire), and number of non-responding individuals per cluster (individual questionnaire)

Household sample (design) weight. The formulas use 2 indices: 1…*k* strata (or entire coverage area) and 1…*i* cluster. The household sample (design) weight – the probability that cluster *i* is selected – is given by:

$$w_{hh.d} = \frac{G_k/g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times b_{ik}$$

Where:

n_k is the number of selected clusters in stratum *k*

G_k is the estimated total population in stratum *k* from GridSample

g_{ik} is the estimated population in cluster *i* in stratum *k* from GridSample

m_{ik} is the number of households sampled in cluster *i* and stratum *k* during fieldwork

M_{ik} is the number of total households enumerated in cluster *i* and stratum *k* during fieldwork

b_{ik} is the number of segments (if segmentation was performed before and after enumeration, then $b_{ik} = b_{ik.before} \times b_{ik.after}$)

Household response weight. Interviewers will list households and record household and individual response rates during fieldwork. After interviews are completed, calculate household response weight - the probability that cluster *i* is found and sampled, and households are found and respond – is given by:

$$w_{hh.r} = \frac{n_k}{n_{k*}} \times \frac{m_k}{m_{k*}}$$

Where:

n_k is the number of selected clusters in stratum *k*

m_k is the number of households sampled in stratum *k* during fieldwork

n_{k*} is the number of found and sampled clusters in stratum *k*

m_{k*} is the number of found and responded households in stratum *k*

Household sample weight. To calculate the raw household sample weight, multiply the sample design weight and household response weight like this:

$$w_{hh} = \frac{G_k/g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times b_{ik} \times \frac{n_k}{n_{k*}} \times \frac{m_k}{m_{k*}}$$

Note, in one-stage samples, $\frac{M_{ik}}{m_{ik}}$ is equal to 1, and in two-stage samples, b_{ik} is usually equal to 1.

Individual sample weight. The individual sample weight includes four additional terms to account for the sampling of one respondent among all eligible respondents in the household, and the response rate of those respondents. The individual sample weight is given by:

$$w_{ind.s} = \frac{G_k/g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times b_{ik} \times \frac{n_k}{n_{k*}} \times \frac{m_k}{m_{k*}} \times \frac{U_{ik}}{u_{ik}} \times \frac{u_k}{u_{k*}}$$

Where:

U_{ik} is the number of eligible individuals in cluster *i* and stratum *k*

u_{ik} is the number of sampled individuals in cluster *i* and stratum *k*

u_k is the number of sampled individuals in stratum *k*

u_{k*} is the number of responded individuals in stratum *k*

Normalizing sample weights. Household surveys are often “normalized” or “standardized” such that the sum of weighted respondents equals the sum of respondents. Conceptually, each observation in the sample represents slightly more or slightly less than 1 household or person. To normalize sample weights, apply the below formulas:

$$w_{hh\_norm} = w_{hh} \times \frac{\sum(m_{ik*})}{\sum(w_{hh} \times m_{ik*})}$$

Where:

m_{ik*} is the number of respondents with a completed interview in cluster *i* in stratum *k*

w_{hh} is the raw household sample weight in cluster *i* in stratum *k*

$$w_{ind\_norm} = w_{ind} \times \frac{\sum(u_{ik*})}{\sum(w_{ind} \times u_{ik*})}$$

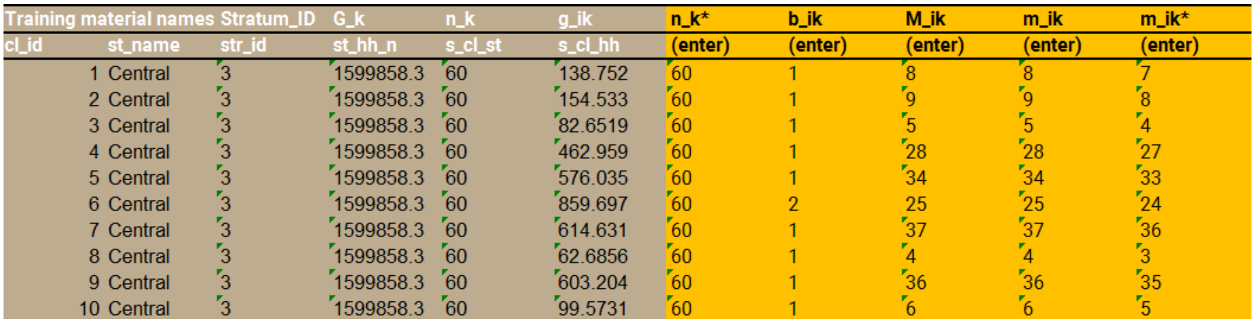
Where:

u_{ik*} is the number of individuals with a completed interview in cluster *i* in stratum *k*

w_{ind} is the raw individual sample weight in cluster *i* in stratum *k*

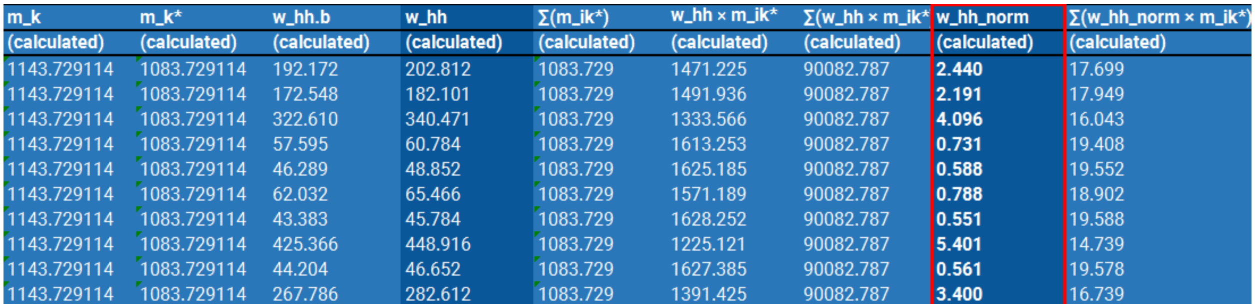
Calculating sample weights. Calculate sample weights in the provided Excel template, or in a statistical software programme such as SPSS or Stata. The provided Excel template is available at GridSample.org/tutorial.

To calculate household sample weights in the template, copy corresponding GridSample values into the grey columns of the “hh_survey_weights” tab. Strata and cluster values are included in an Excel file along with other downloaded outputs from GridSample. After performing the household listing and interviews, enter field-generated information into the template’s orange columns.



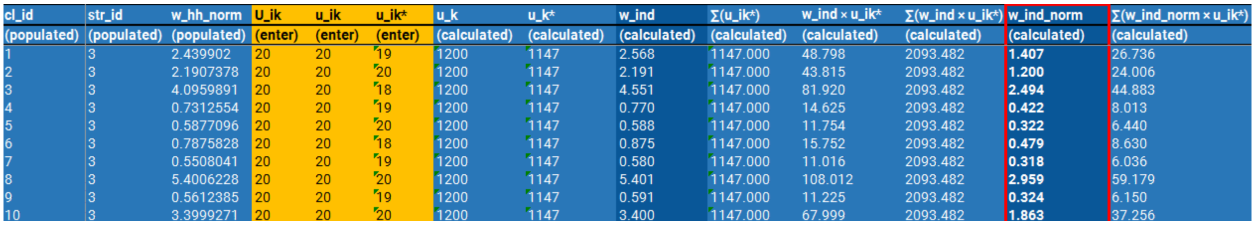
Training material names		Stratum_ID	G_k	n_k	g_ik	n_k*	b_ik	M_ik	m_ik	m_ik*
cl_id	st_name	str_id	st_hh_n	s_cl_st	s_cl_hh	(enter)	(enter)	(enter)	(enter)	(enter)
1	Central	3	1599858.3	60	138.752	60	1	8	8	7
2	Central	3	1599858.3	60	154.533	60	1	9	9	8
3	Central	3	1599858.3	60	82.6519	60	1	5	5	4
4	Central	3	1599858.3	60	462.959	60	1	28	28	27
5	Central	3	1599858.3	60	576.035	60	1	34	34	33
6	Central	3	1599858.3	60	859.697	60	2	25	25	24
7	Central	3	1599858.3	60	614.631	60	1	37	37	36
8	Central	3	1599858.3	60	62.6856	60	1	4	4	3
9	Central	3	1599858.3	60	603.204	60	1	36	36	35
10	Central	3	1599858.3	60	99.5731	60	1	6	6	5

The template's blue columns with populate; the dark blue columns show the household sample weight and normalized household sample weight for each cluster. Analysts generally prefer to use the normalized weights, outlined in red in the template.



m_k	m_k*	w_hh.b	w_hh	Σ(m_ik*)	w_hh × m_ik*	Σ(w_hh × m_ik*	w_hh_norm	Σ(w_hh_norm × m_ik*)
(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)
1143.729114	1083.729114	192.172	202.812	1083.729	1471.225	90082.787	**2.440**	17.699
1143.729114	1083.729114	172.548	182.101	1083.729	1491.936	90082.787	**2.191**	17.949
1143.729114	1083.729114	322.610	340.471	1083.729	1333.566	90082.787	**4.096**	16.043
1143.729114	1083.729114	57.595	60.784	1083.729	1613.253	90082.787	**0.731**	19.408
1143.729114	1083.729114	46.289	48.852	1083.729	1625.185	90082.787	**0.588**	19.552
1143.729114	1083.729114	62.032	65.466	1083.729	1571.189	90082.787	**0.788**	18.902
1143.729114	1083.729114	43.383	45.784	1083.729	1628.252	90082.787	**0.551**	19.588
1143.729114	1083.729114	425.366	448.916	1083.729	1225.121	90082.787	**5.401**	14.739
1143.729114	1083.729114	44.204	46.652	1083.729	1627.385	90082.787	**0.561**	19.578
1143.729114	1083.729114	267.786	282.612	1083.729	1391.425	90082.787	**3.400**	16.739

If individuals were listed, sampled, and interviewed within each selected household, then use the "ind_survey_weights" tab in the Excel template to calculate individual sample weights. Household sample weights will be pre-populated; simply enter field-generated information about individuals in the template's orange columns.



cl_id	str_id	w_hh_norm	U_ik	u_ik	u_ik*	u_k	u_k*	w_ind	Σ(u_ik*)	w_ind × u_ik*	Σ(w_ind × u_ik*)	w_ind_norm	Σ(w_ind_norm × u_ik*)
(populated)	(populated)	(populated)	(enter)	(enter)	(enter)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)	(calculated)
1	3	2.439902	20	20	19	1200	1147	2.568	1147.000	48.798	2093.482	**1.407**	26.736
2	3	2.1907378	20	20	20	1200	1147	2.191	1147.000	43.815	2093.482	**1.200**	24.006
3	3	4.0959891	20	20	18	1200	1147	4.551	1147.000	81.920	2093.482	**2.494**	44.883
4	3	0.7312554	20	20	19	1200	1147	0.770	1147.000	14.625	2093.482	**0.422**	8.013
5	3	0.5877096	20	20	20	1200	1147	0.588	1147.000	11.754	2093.482	**0.322**	6.440
6	3	0.7875828	20	20	18	1200	1147	0.875	1147.000	15.752	2093.482	**0.479**	8.630
7	3	0.5508041	20	20	19	1200	1147	0.580	1147.000	11.016	2093.482	**0.318**	6.036
8	3	5.4006228	20	20	20	1200	1147	5.401	1147.000	108.012	2093.482	**2.959**	59.179
9	3	0.5612385	20	20	19	1200	1147	0.591	1147.000	11.225	2093.482	**0.324**	6.150
10	3	3.3999271	20	20	20	1200	1147	3.400	1147.000	67.999	2093.482	**1.863**	37.256

Data cleaning & analysis

Household listing data collected via paper forms should be entered into Excel or OpenOffice Calc. Counts from the household listing and household response rates are needed for sample weight calculations.

Questionnaire data collected via paper forms should be double entered and cleaned using a software such as CSPro.

Household listing and questionnaire data collected via tablet should be reviewed and quality checked throughout the data collection process. Data should then be imported into a statistical software programme, such as SPSS or Stata, for further processing.

All summary statistics generated from complex household surveys should account for unequal probability of selection by applying sample weights, clustering of observations, and if applicable, stratification. Here are several resources to conduct survey data analysis in SPSS, Stata, and other statistical software programmes:

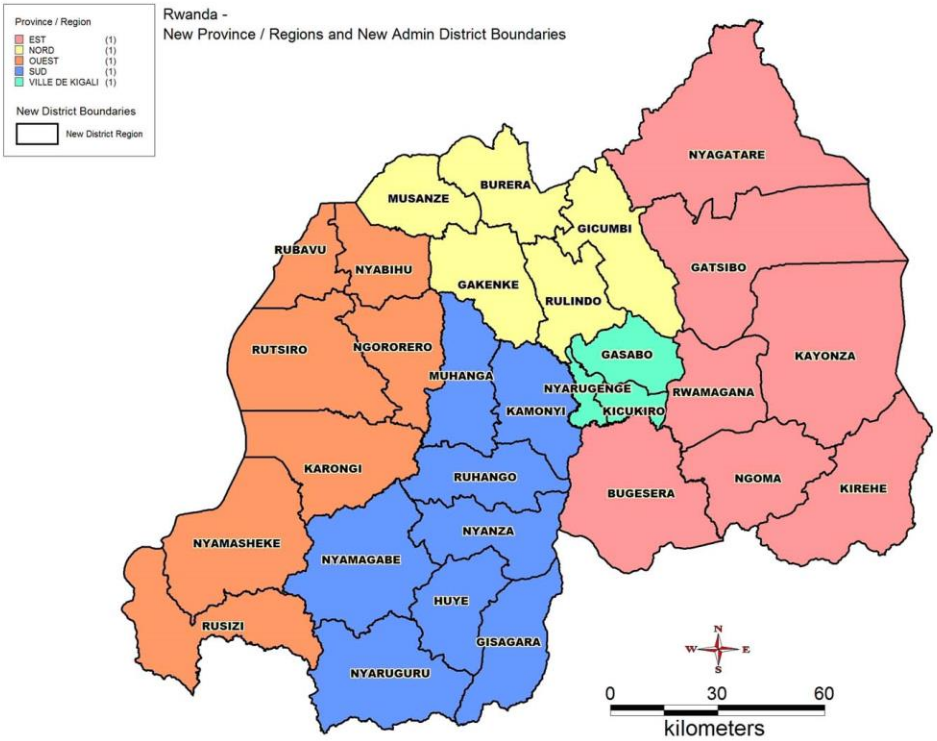
- Population Survey Analysis
- UNC - Carolina Population Center
- UCLA Institute for Digital Research & Education

Practice exercises

Exercise 1: 2014-15 Rwanda DHS

The 2014-15 Rwanda Demographic and Health Survey interviewed all women age 15-49 in selected households about their own health and wellbeing and the health of their children. The final report describes the survey design on pages 7 and 8 as follows:

> *The 2014-15 RDHS followed a two-stage sample design and was intended to allow estimates of key indicators at the national level as well as for urban and rural areas, five provinces, and each of Rwanda's 30 districts (for some limited indicators). The first stage involved selecting sample points (clusters) consisting of EAs delineated for the 2012 RPHC. A total of 492 clusters were selected, 113 in urban areas and 379 in rural areas.*
>
> *The second stage involved systematic sampling of households. A household listing operation was undertaken in all of the selected EAs from July 7 to September 6, 2014, and households to be included in the survey were randomly selected from these lists. Twenty-six households were selected from each sample point, for a total sample size of 12,792 households. However, during data collection, one of the households was found to actually be two households, which increased the total sample to 12,793. Because of the approximately equal sample sizes in each district, the sample is not self-weighting at the national level, and weighting factors have been added to the data file so that the results will be proportional at the national level.*



These tables from the 2014-15 RDHS final report provide additional information to reproduce this survey design in GridSample.

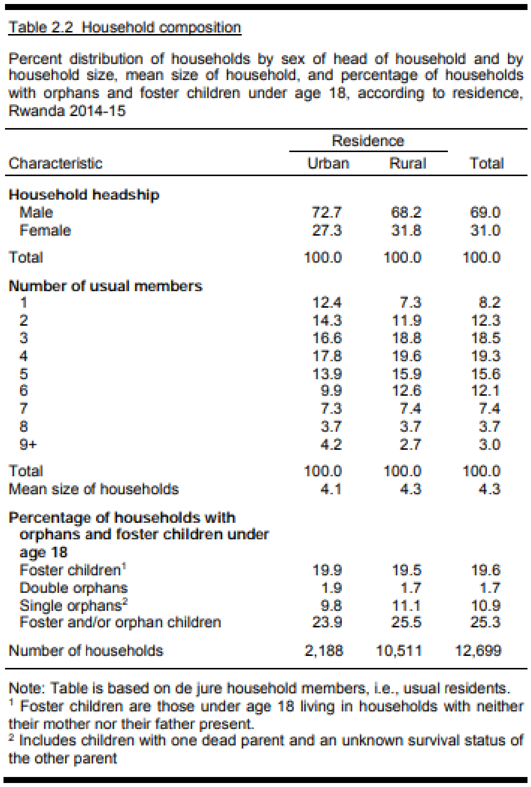


Table 2.2 Household composition

Percent distribution of households by sex of head of household and by household size, mean size of household, and percentage of households with orphans and foster children under age 18, according to residence, Rwanda 2014-15

Characteristic	Residence		Total
	Urban	Rural	
Household headship			
Male	72.7	68.2	69.0
Female	27.3	31.8	31.0
Total	100.0	100.0	100.0
Number of usual members			
1	12.4	7.3	8.2
2	14.3	11.9	12.3
3	16.6	18.8	18.5
4	17.8	19.6	19.3
5	13.9	15.9	15.6
6	9.9	12.6	12.1
7	7.3	7.4	7.4
8	3.7	3.7	3.7
9+	4.2	2.7	3.0
Total	100.0	100.0	100.0
Mean size of households	4.1	4.3	4.3
Percentage of households with orphans and foster children under age 18			
Foster children[1]	19.9	19.5	19.6
Double orphans	1.9	1.7	1.7
Single orphans[2]	9.8	11.1	10.9
Foster and/or orphan children	23.9	25.5	25.3
Number of households	2,188	10,511	12,699

Note: Table is based on de jure household members, i.e., usual residents.

[1] Foster children are those under age 18 living in households with neither their mother nor their father present.

[2] Includes children with one dead parent and an unknown survival status of the other parent

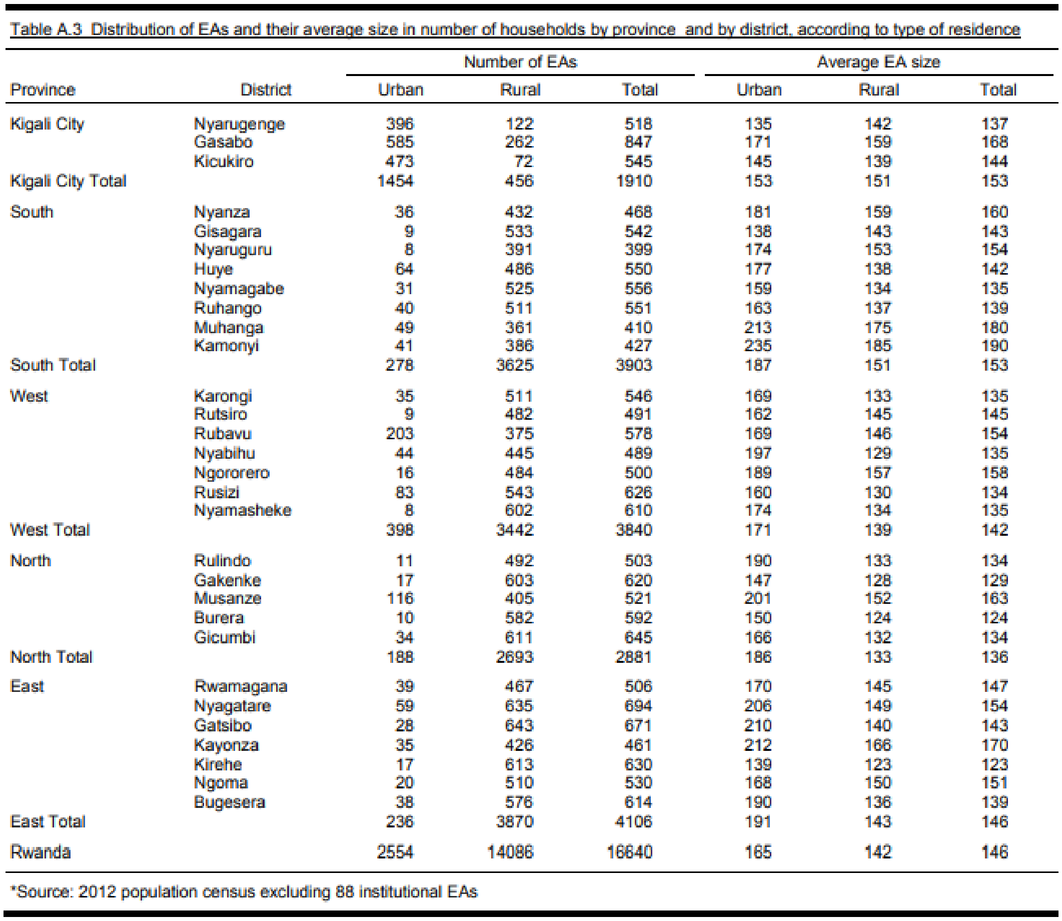


Table A.3 Distribution of EAs and their average size in number of households by province and by district, according to type of residence

Province	District	Number of EAs			Average EA size		
		Urban	Rural	Total	Urban	Rural	Total
Kigali City	Nyarugenge	396	122	518	135	142	137
	Gasabo	585	262	847	171	159	168
	Kicukiro	473	72	545	145	139	144
Kigali City Total		1454	456	1910	153	151	153
South	Nyanza	36	432	468	181	159	160
	Gisagara	9	533	542	138	143	143
	Nyaruguru	8	391	399	174	153	154
	Huye	64	486	550	177	138	142
	Nyamagabe	31	525	556	159	134	135
	Ruhango	40	511	551	163	137	139
	Muhanga	49	361	410	213	175	180
	Kamonyi	41	386	427	235	185	190
South Total		278	3625	3903	187	151	153
West	Karongi	35	511	546	169	133	135
	Rutsiro	9	482	491	162	145	145
	Rubavu	203	375	578	169	146	154
	Nyabihu	44	445	489	197	129	135
	Ngororero	16	484	500	189	157	158
	Rusizi	83	543	626	160	130	134
	Nyamasheke	8	602	610	174	134	135
West Total		398	3442	3840	171	139	142
North	Rulindo	11	492	503	190	133	134
	Gakenke	17	603	620	147	128	129
	Musanze	116	405	521	201	152	163
	Burera	10	582	592	150	124	124
	Gicumbi	34	611	645	166	132	134
North Total		188	2693	2881	186	133	136
East	Rwamagana	39	467	506	170	145	147
	Nyagatare	59	635	694	206	149	154
	Gatsibo	28	643	671	210	140	143
	Kayonza	35	426	461	212	166	170
	Kirehe	17	613	630	139	123	123
	Ngoma	20	510	530	168	150	151
	Bugesera	38	576	614	190	136	139
East Total		236	3870	4106	191	143	146
Rwanda		2554	14086	16640	165	142	146

*Source: 2012 population census excluding 88 institutional EAs

How would you enter parameters into GridSample to reproduce the 2014-14 RDHS?

Coverage

Subnational option (pick one):

- □ None (National survey)
- □ Urban only
- □ Rural only
- □ Admin 1 area(s):______________________________
- □ Admin 2 area(s):______________________________
- □ Admin 3 area(s):______________________________
- □ Admin 4 area(s):______________________________

Frame

WorldPop-Global Dataset (year):______________________________

Gridded multi-cell cluster with gridEZ algorithm (select one):

- □ Small (target 75 people, max areas 1km X 1km
- □ Medium (target 500 people, max area 3km X 3km)
- □ Large (target 1200 people, max area 5km X 5km)

Design

Stages (select one):

- □ One-stage cluster sample
- □ Two-stage cluster sample

Stratification (select one):

- □ No
- □ Yes

Spatial oversample (select one):

- □ No
- □ Yes

Strata

Admin area boundaries (select one):

- □ Admin 1 (e.g. Umujyi wa Kigali, Amajyaruguru)
- □ Admin 2 (e.g. Musanze, Rulindo)
- □ Admin 3 (e.g. Busogo, Cyuve, Gacaca)
- □ Admin 4 (e.g. Gisesero, Kavumu, Nyagisozi)

Target

Target Population Label:______________________________

Average number of target population members per household: ____________

Average household size: ______________________________

Sample size

Allocation of clusters to strata (select one):

- □ Equal
- □ Proportional
- □ Custom (describe ______________________________)

Total number of households sampled:______________________________

Number of households sampled per urban cluster:____________________

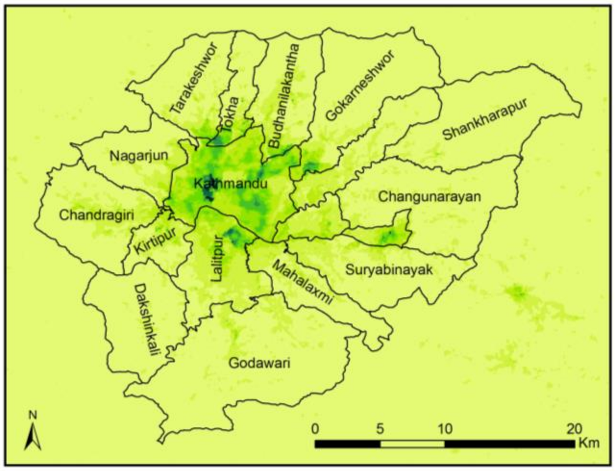
Number of households sampled per rural cluster:____________________

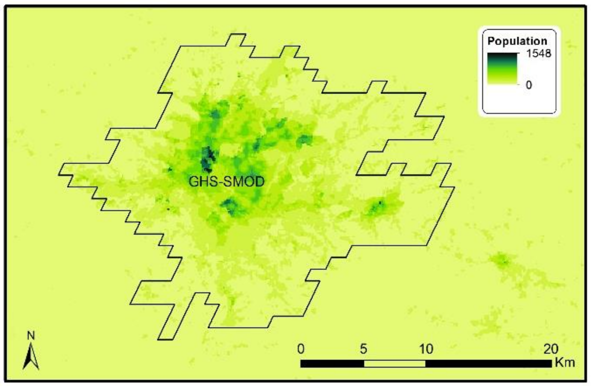
Exercise 2: 2017 Kathmandu SUE

The 2017 Surveys for Urban Equity in Kathmandu Valley aimed to overcome several challenges that lead to exclusion of the urban poorest in standard household surveys as described on page 2 of the SUE survey protocol.

> *First, census data, which is used to select first-stage samples, is often outdated and undercounts informally settled households.*
>
> *Second, by design, surveys typically exclude the homeless and institutional populations. Use of two-stage cluster sampling methods requires two visits to households over several months or years, resulting in underlisting or higher non-response by mobile and fragile households.*
>
> *Third, underlisting and undersampling of poorer households can occur if standardised, detailed protocols are not used by enumerators to interact with residents during the household listing process. For example, multihousehold dwellings will be underlisted if the enumerator assumes one dwelling to be occupied by one household or poorer members of households, such as guards and servants may be excluded.*
>
> *Furthermore, periurban communities frequently home to urban migrants and slum areas, maybe classified as rural.*

The protocol includes the following figures demonstrating that population in the Kathmandu Valley extend well beyond the official city administrative boundaries, and are located within the "high-dense urban" area defined by GHS-SMOD. Thus GHS-SMOD is used to define the Kathmandu Valley survey coverage boundary.





The survey protocol goes on to describe the survey design as follows on page 4:

> *To compare the effectiveness of one-stage sampling compared with two-stage sampling, in the Kathmandu survey we will randomly allocate half of the clusters to each approach. [For the purposes of this exercise, let us assume that one-stage sampling will be conducted in all clusters.]*
>
> *One-stage sampling of approximately 20 households in each sampling area will be facilitated by the use of WorldPop 100 m×100m grid cells rather than much larger census enumeration areas as the sampling frame.*
>
> *We will interview adults, 18 years and above…*
>
> *We will aim for a sample size of 1200 in the Kathmandu survey…. This assumes... one eligible individual per household.*
>
> *This sample population will be distributed across 60 clusters...*

Additional information about average household size is needed to replicate this survey design in GridSample. For this information, we use the 2016 Nepal Demographic and Health Survey final report.

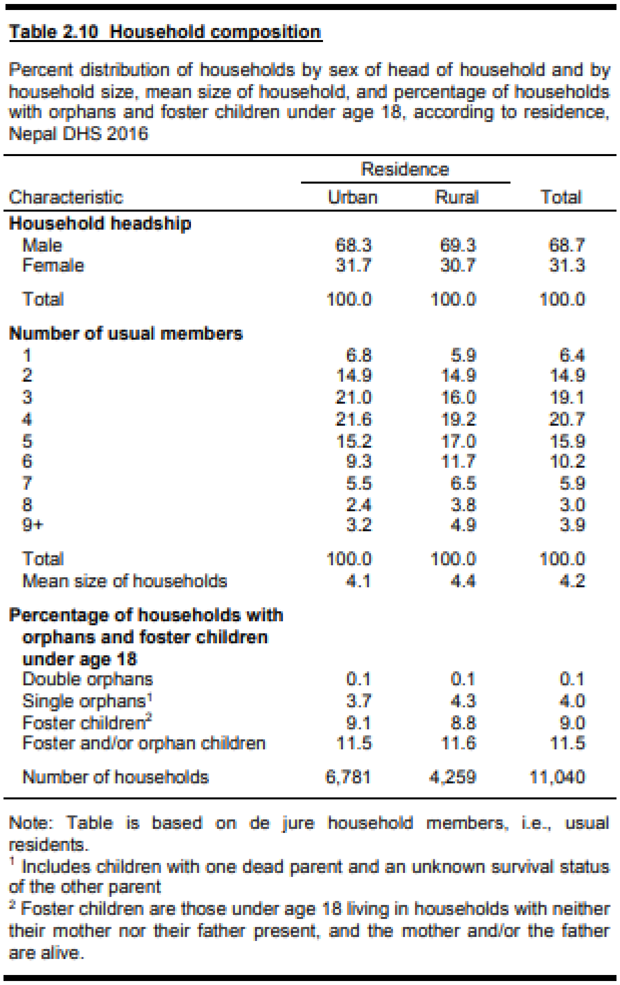


Table 2.10 Household composition

Percent distribution of households by sex of head of household and by household size, mean size of household, and percentage of households with orphans and foster children under age 18, according to residence, Nepal DHS 2016

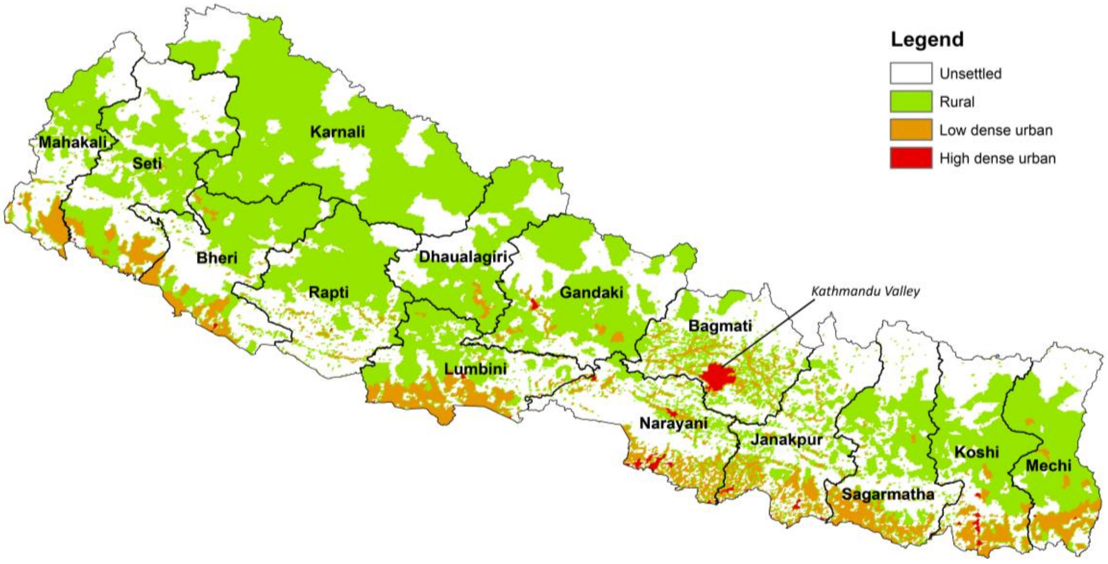
	Residence		
Characteristic	Urban	Rural	Total
Household headship			
Male	68.3	69.3	68.7
Female	31.7	30.7	31.3
Total	100.0	100.0	100.0
Number of usual members			
1	6.8	5.9	6.4
2	14.9	14.9	14.9
3	21.0	16.0	19.1
4	21.6	19.2	20.7
5	15.2	17.0	15.9
6	9.3	11.7	10.2
7	5.5	6.5	5.9
8	2.4	3.8	3.0
9+	3.2	4.9	3.9
Total	100.0	100.0	100.0
Mean size of households	4.1	4.4	4.2
Percentage of households with orphans and foster children under age 18			
Double orphans	0.1	0.1	0.1
Single orphans[1]	3.7	4.3	4.0
Foster children[2]	9.1	8.8	9.0
Foster and/or orphan children	11.5	11.6	11.5
Number of households	6,781	4,259	11,040

Note: Table is based on de jure household members, i.e., usual residents.

[1] Includes children with one dead parent and an unknown survival status of the other parent

[2] Foster children are those under age 18 living in households with neither their mother nor their father present, and the mother and/or the father are alive.

Although GridSample does not provide individual metropolitan boundaries for specific cities, it does provide boundaries of all urban areas (defined as GHS-SMOD=high dense urban) and multiple levels of sub-national administrative units. A clever solution is possible in GridSample using the provided parameter options to limit the coverage area to urban area(s) in a sub-region of the country. Consider the below map, and see if you can figure it out.



How would you enter parameters into GridSample to reproduce the 2017 Kathmandu SUE?

Coverage Subnational option (pick one):

- □ None (National survey)
- □ Urban only
- □ Rural only
- □ Admin 1 area(s):____________________
- □ Admin 2 area(s):____________________
- □ Admin 3 area(s):____________________
- □ Admin 4 area(s):____________________

Frame WorldPop-Global Dataset (year):____________________

Gridded multi-cell cluster with gridEZ algorithm (select one):

- □ Small (target 75 people, max areas 1km X 1km
- □ Medium (target 500 people, max area 3km X 3km)
- □ Large (target 1200 people, max area 5km X 5km)

Design Stages (select one):

- □ One-stage cluster sample
- □ Two-stage cluster sample

Stratification (select one):

- □ No
- □ Yes

Spatial oversample (select one):

- □ No
- □ Yes

Strata Admin area boundaries (select one):

- □ Admin 1 (e.g. Central, East)
- □ Admin 2 (e.g. Koshi, Mechi)
- □ Admin 3 (e.g. Bhojpur, Dhankuta, Morang)
- □ Admin 4 (e.g. Aamtep, Annapurna, Baikunthe)

Target Target Population Label:____________________

Average number of target population members per household: __________

Average household size: ____________________

Sample size Allocation of clusters to strata (select one):

- □ Equal
- □ Proportional
- □ Custom

Number of clusters per stratum:

Stratum Name	Number of clusters

Additional resources

Gridded population sampling manual

- Thomson DR, Bhattarai R, Dhungel R, Gajurel S, Singh S, Manandhar S, Khanal S. 2018. Surveys for Urban Equity (SUE) Project: Planning Team Guide. Leeds: Leeds University. 147 p. Available at: https://medicinehealth.leeds.ac.uk/downloads/download/95/planning_team_guide.

Gridded population survey implementations

- Pape UJ and Wollburg PR. 2019. Estimation of Poverty in Somalia Using Innovative Methodologies. World Bank: Washington DC USA. Available at: http://documents.worldbank.org/curated/en/509221549985694077/Estimation-of-Poverty-in-Somalia-Using-Innovative-Methodologies.
- Cajka J, Amer S, Ridenhour J, Allpress, J. 2018. Geo-Sampling in Developing Nations. Int J Soc Res Methodol, 21(6): doi:10.1080/13645579.2018.1484989.
- Elsey H, Poudel AN, Ensor T, et al. 2018. 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, 8 doi: 10.1136/bmjopen-2018-024182.
- Vulnerability Analysis and Mapping team. 2018. Urban Essential Needs Assessment in the five communes of Kimbanseke, Kinsenso, Makala, N'sele and Selembao (Kinshasa) World Food Programme: Kinshasa DRC and Rome Italy. Available at: https://docs.wfp.org/api/documents/WFP-0000099888/download/.
- Thomson DR and Hesse JB. 2018. GridSample: Household surveys with gridded population data to overcome outdated/inaccurate census frame whilst saving time and cost. World Vision International: Mozambique. Presentation at World Data Forum. Available at: https://undataforum.org/WorldDataForum/sessions/ta2-23-innovate-or-perish-household-surveys-in-a-changing-data-landscape/.
- Elsey H, Thomson DR, Lin RY, et al. 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. J Urban Health, 93(3): doi:10.1007/s11524-016-0046-9.
- Muñoz J, Langeraar W. 2013. A census-independent sampling strategy for a household survey in Myanmar. Sistemas Integrales: Santiago Chile. Available at: http://winegis.com/images/census-independent-GIS-based-sampling-strategy-for-household-surveys-plan-of-action%20removed.pdf.

- Galway L, Bell N, Shatari SAE, et al. 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. Int J Health Geogr, 11: doi:10.1186/1476-072X-11-12.
- Thomson DR, Hadley MB, Greenough PG, et al. 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: doi:10.1186/1471-2458-12-959.
- Sollom R, Richards AK, Parmar P, et al. 2011. Health and human rights in Chin State, Western Burma: A population-based assessment using multistaged household cluster sampling. PLoS Med, 8(2): doi:10.1371/journal.pmed.1001007.

Exercise answers

Coverage

Subnational option (pick one):

- ✓ None (National survey)
- □ Urban only
- □ Rural only
- □ Admin 1 area(s):______________________________
- □ Admin 2 area(s):______________________________
- □ Admin 3 area(s):______________________________
- □ Admin 4 area(s):______________________________

Frame

WorldPop-Global Dataset (year):__2019____________________

Gridded multi-cell cluster with gridEZ algorithm (select one):

- □ Small (target 75 people, max areas 1km X 1km
- ✓ Medium (target 500 people, max area 3km X 3km)
- □ Large (target 1200 people, max area 5km X 5km)

Average HHs per EA: 146
Average HH size: 4.3
146 × 4.3 = 628 pop per EA

Design

Stages (select one):

- □ One-stage cluster sample
- ✓ Two-stage cluster sample

Stratification (select one):

- □ No
- ✓ Yes

Spatial oversample (select one):

- ✓ No
- □ Yes

Strata

Admin area boundaries (select one):

- □ Admin 1 (e.g. Umujyi wa Kigali, Amajyaruguru)
- ✓ Admin 2 (e.g. Musanze, Rulindo)
- □ Admin 3 (e.g. Busogo, Cyuve, Gacaca)
- □ Admin 4 (e.g. Gisesero, Kavumu, Nyagisozi)

Target

Target Population Label:__Women 15-49____________________

Average number of target population members per household: _1.1_________

Average household size: __4.3___________________________

Sample size

Allocation of clusters to strata (select one):

- □ Equal
- □ Proportional
- ✓ Custom (describe: 20 clusters in each of 3 Kigali districts, 16 clusters in each of other districts)

Total number of households sampled:__12,792____________________

Number of households sampled per urban cluster:__26_________________

Number of households sampled per rural cluster:__26_________________

Coverage

Subnational option (pick one):

- □ None (National survey)
- □ Urban only
- □ Rural only
- □ Admin 1 area(s):________________________________
- ✓ Admin 2 area(s):_Bagmati__________________________
- □ Admin 3 area(s):________________________________
- □ Admin 4 area(s):________________________________

Frame

WorldPop-Global Dataset (year):__2019______________________

Gridded multi-cell cluster with gridEZ algorithm (select one):

- ✓ Small (target 75 people, max areas 1km X 1km
- □ Medium (target 500 people, max area 3km X 3km)
- □ Large (target 1200 people, max area 5km X 5km)

Target population: 75
Average HH size: 4.1
75 × 4.1 = 18 HHs per unit

Design

Stages (select one):

- ✓ One-stage cluster sample
- □ Two-stage cluster sample

Stratification (select one):

- □ No
- ✓ Yes

Spatial oversample (select one):

- ✓ No
- □ Yes

Strata

Admin area boundaries (select one):

- ✓ Urban / rural
- □ Admin 1 (e.g. Central, East)
- □ Admin 2 (e.g. Koshi, Mechi)
- □ Admin 3 (e.g. Bhojpur, Dhankuta, Morang)
- □ Admin 4 (e.g. Aamtep, Annapurna, Baikunthe)

Target

Target Population Label:_Adults 18+________________________

Average number of target population members per household: _1________

Average household size: _4.1__________________________

Sample size

Allocation of clusters to strata (select one):

- □ Equal
- □ Proportional
- ✓ Custom

Number of clusters per stratum:

Stratum Name	Number of clusters
Urban	60
Rural	0