GridSample2.0 Algorithm

Technical Manual

Updated: 07 July 2019

By:

Dana R Thomson, GridSample Manager
Flowminder Foundation
www.flowminder.org

FLOWMINDER.ORG

Table of Contents:

About

In countries where census data are outdated, inaccurate or too geographically course to be used effectively as a survey sample frame, gridded population data are being used to select representative, complex household surveys.

Top-down gridded population data have been available for decades and represent disaggregated census population counts in small grid cells. Disaggregation is based on spatial datasets such as land cover type and road locations. Two such datasets in particular - WorldPop and Landscan - have been used in dozens of published and unpublished gridded population surveys in low- and middle-income country (LMIC) surveys. WorldPop has the advantage of being free, having small ~100m X 100m cells (roughly the size of a city block), and representing residential population counts. Alternatively, many Landscan users must pay for data access, cells are ~1km X 1km, and the dataset represents a 24-hour ambient population averaging counts of nighttime residents and daytime commuters. Census-independent bottom-up gridded population datasets are expected to become available for many countries in 2019 based on satellite imagery and micro-census counts. GridSample2.0 assume use of WorldPop estimates.

Until recently, gridded population surveys were generated with *adhoc* tools and methods. However, gridded population sampling is becoming easier with tools such as GridSample.org, a free click-and-point web tool designed for survey planners in LMICs. GridSample.org includes pre-loaded gridded population data and boundary datasets, and supports standard complex survey designs used in the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Living Standard Measurement Surveys (LSMS), and other routine surveys.

GridSample.org is created by Flowminder, a non-profit organisation whose mission is to improve public health and wellbeing in LMICs using data from mobile operators, satellites, and geo-located household surveys.

This document provides technical details of the GridSample2.0 algorithm which is used by GridSample.org and can facilitate gridded population sampling offline.

License

Vocabulary

Cell - Refers to a unit in a gridded population dataset. Also called a grid square, raster cell, or pixel.

EA (Enumeration Area) - Usually refers to census enumeration areas, though might be used more generally to refer to any small area unit that is mutually exclusive and entirely covers the survey coverage area. In typical household survey, EAs are used as the first-stage, or primary, sampling units.

PSU (Primary sampling unit) - The selection of small areas (often EAs or cells) made during the first-stage of sampling.

Cluster - A less technical word for PSU.

Domain - A sub-region/sub-population classification, usually referring to urban/rural.

Coverage - The extent of the survey area.

Stratification - Mutually exclusive areas that entirely cover the survey coverage area, in which independent samples are selected. Administrative boundaries are often used as strata boundaries. Ideally, populations within a strata are more similar than between strata. Stratification is used in household surveys to make independent estimates for sub-populations.

PPS (probability proportionate to size) sampling - A systematic sample selection technique that ensures each member of the population has an equal, or known, probability of selection. PPS sampling is implemented by listing each small area and its population total, determining a population increment that will result in the correct number of sampling units (clusters/PSUs), and a random starting value. A cumulative sum of population is calculated for the list of sampling areas (EAs/cells). Clusters/PSUs are selected by starting from the random value, and adding the increment value, and selecting clusters which contain the increment population value. This systematic sampling approach results in a selection of clusters/PSUs that is proportional to their population size. See the MICS guidance on Designing and Selecting the Sample for more detail.

Serpentine sampling - A technique in which EAs/cells or ordered from west-to-east, north-to-south before PPS sampling. This ensures maximum spatial coverage of clusters/PSUs in the PPS sample.

Implicit stratification - A technique in which EAs/cells are ordered by domain before sampling. This ensures representative coverage across across domains. Most household surveys use PPS

serpentine sampling with implicit stratification by urban/rural domain. This design is implemented by ordering EAs/cells by urban and rural first, then from west-to-east and north-to-south second, and finally sampling with PPS. In GridSample2.0 the GHS-SMOD datasets is assumed for implicit stratification such that the sample is representative of high density urban, low density urban, rural, and (optionally) unsettled domains.

Oversampling - After PPS sampling is performed, adding clusters/PSUs to ensure a minimum sample size in an important sub-group (which is not already represented by strata). In DHs surveys in majority rural countries, for example, the capital city is often oversampled to achieve a minimum sample size of urban households. Oversampling of a sub-group results in unequal chance of selection of population members. Thus, sampling weights must be employed to ensure that oversampled populations contribute less per member to mean, percent, and total calculations. Without getting into details, sampling probabilities are easier to conceive and calculate when oversampling is performed without replacement.

With replacement - Oversampled clusters/PSUs are traded out of the sample, and thus could theoretically be resampled. This is the method used in the GridSample R algorithm available on R CRAN in order to preserve the total desired sample size and restrict resources spent.

Without replacement - Oversampled clusters/PSUs are added to the sample. This is the method used in GridSample2.0 to be consistent with the majority of DHS and similar survey designs, ensuring accurate calculation of sampling weights.

Parallel processing - A way of processing a computer operation by splitting it into parts and executing each part simultaneously on a different processor on the same computer. Parallel processing is used to split and run each GridSample2.0 job by strata.

Resample - Refers to changing the size of a grid cell through aggregation or disaggregation. In GridSample2.0, the smallest cell is assumed to be WorldPop ~100m X 100m cells. We only resample to larger cells comprised of whole original cells (e.g. ~200m X 200m cells, or ~700m X 700m cells).

Parameters

Description	Dataset/parameter name	Value
The desired action and output	action	"create_histogram" produces a histogram of sample frame unit population counts "get_strata_pop_values" produces a csv with population total and proportion per stratum "Create_sample" selects PSUs from the sample frame with PPS and implicit stratification by urbanicity based on the parameters provided
The ID for this sample	uniqueID	Positive integer.
The folder location to save output	output_dir	String. A file pathway Example: "c:/project/output/"
The name of the output shp and kml files to be created after PSUs are selected	PSU_filename	String. A file name Example: "Kathmandu.shp"
The gridded population dataset on which the sample frame is based	pop_raster	String. A file pathway and projected raster file Example: "c:/project/data/WorldPop/ NPL_ppp_v2c_2020_UNadj.tif"
The gridded dataset used for urban/rural implicit stratification, definition of multicell sample frame units with gridEZ algorithm, or defining an urbanOnly or ruralOnly coverage area	ghs_mod_raster	String. A file pathway and projected raster file Example: "c:/project/data/GHSL/ GHS_SMOD_POP2015_GLOBE_R2016A.tif"
The coverage area of the survey *(If coverage = shapefile)*	coverage_poly_filename	String. A file pathway and projected shapefile or "None" Example: "c:/project/data/GADM/NPL_0.shp"
The coverage area of the survey *(If coverage = GHS-SMOD)*	coverage_ur_option	"urbanOnly" means that GHS-SMOD = 3 "ruralOnly" means that GHS-SMOD = 1 or 2 "None" means this option is not used
The sample frame definition	cfg_frame_type	"single" means that one grid cell (of any dimension) is equal to one sample frame unit "multi" means that one or more grouped grid cells (e.g. gridEZ algorithm) equals one sample frame unit "own" means that a shapefile is used to define the boundaries of each sample frame unit
The size of each grid cell sample frame unit *(If frame type = single)*	cfg_resample_size	Positive integer. Number of original cell units per edge of new (resampled) grid cell units or "None" Example: 1 = ~100mX100m cells, 10 = ~1kmX1km cells
Classifies grid cell sample	cfg_exclude_pop_per_cell	Positive decimal value or "None"

frame unit to exclude because the estimated population is below a specified threshold *(If frame type = single)*		
Classifies sample frame units to exclude based on GHS-SMOD "unsettled" areas *(If frame type = single or multi)*	cfg_exclude_ghssmod0_bool	"True" or "False"
The size of each gridEZ sample frame unit *(If frame type = multi)*	cfg_multi_cell_cluster_size	"small" means each unit has a target of 75 people (15-20 households) and max area of 1kmX1km "medium" means each unit has a target of 500 people (~125 households) and max area of 3kmX3km "large" means that each unit has a target of 1200 people (~300 households) and max area of 5kmX5km "None" when this parameter is not used
An administrative or other sub-national boundary to aid the gridEZ algorithm to define sample frames *(If frame type = multi)*	grid_ea_strata_file	String. A pathway and raster file name or "None" ONLY USED with gridEZ algorithm, not with GridSample
The field name that uniquely identifies administrative area in grid_ea_strata_file *(If frame type = multi)*	grid_ea_strata_file_id_field	String. A field name or "None" ONLY USED with gridEZ algorithm, not with GridSample
An administrative or other boundary to use directly as the sample frame *(If frame type = own)*	cfg_own_frame_file	String. A pathway and raster file name or "None"
The field name that uniquely identifies areas in cfg_own_frame_file *(If frame type = own)*	cfg_own_frame_id	String. A field name or "None"
Whether stratification will be used, and if so, how strata will be defined	cfg_stratification_method	"None" means there are no strata within coverage area "urbanRural" means that ghs_mod_raster will be used to create an urban stratum where GHS-SMOD=3, and a rural stratum where GHS-SMOD=2, 1, or 0 "adminArea" is only used with GridSample.org "custom_upload" means the user defines strata boundaries with strata_poly_filename
A file indicating boundaries of strata within the coverage area *(If strata method = custom)*	strata_poly_filename	String. A file pathway and projected shapefile or "None" Example: "c:/project/data/GADM/NPL_2.shp"
The field name that uniquely identifies strata IDs in	strata_ID_field	String. A field name or "None"

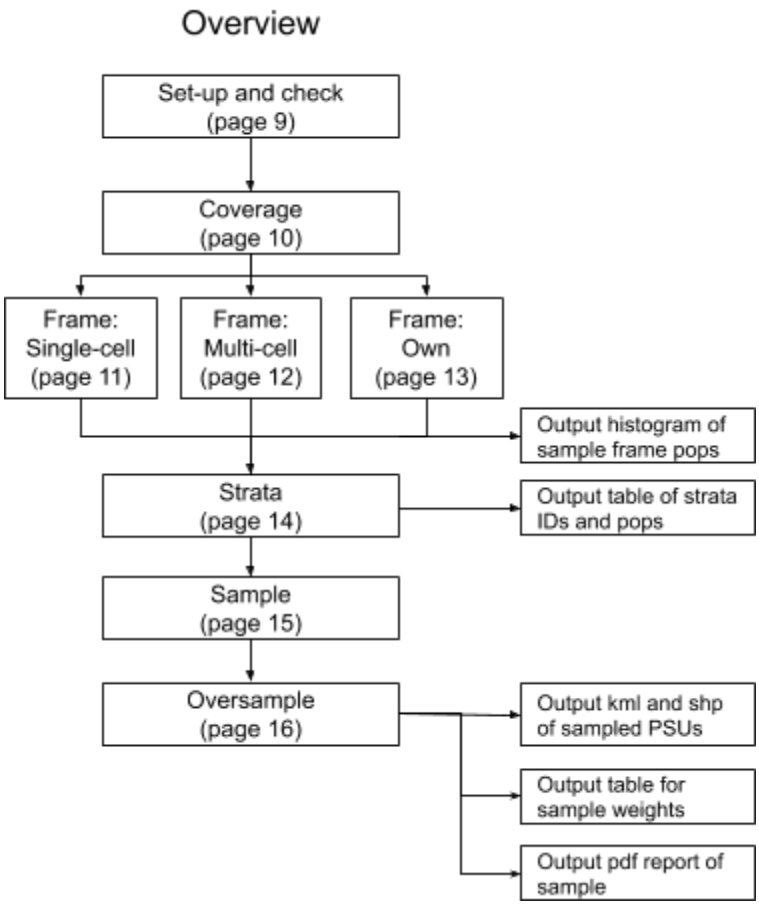
strata_poly_filename *(If strata method = custom)*		
The field name that uniquely identifies strata names in strata_poly_filename *(If strata method = custom)*	strata_name_field	String. A field name or “None”
Whether spatial oversampling will be used, and if so, the size of spatial oversampling units	cfg_oversample_grid_spatial_scale	Positive integer and multiple of cfg_resample_size or “None”. This is the number of original cell units (eg ~100m X 100m) per edge of the spatial oversample units. Example: 1000 = ~100kmX100km cells
The average household size in the coverage area, or by strata, to estimate number of households per sample frame unit based on the population per sample frame unit	cfg_hh_size	Positive decimal. Can be defined by strata in curly brackets. Example for strata 1, 2, and 3: {u'1': 3.5, u'2': 4.0, u'3': 4.5}
The number of PSUs to select in the coverage area, or by strata is strata are specified. The user should consider the expected target population per household; average household size(s); whether the sample will follow a one-stage or two-stage design; whether PSU allocation to strata is equal, proportional, optimal, or other design; and whether different numbers of households will be sampled in urban versus rural PSUs.	cfg_psu_per_strata	Positive integer. Can be defined by strata in curly brackets. Example for strata 1, 2, and 3: {u'1': 100, u'2': 110, u'3': 240}
A “seed” value for the random number generator so the exact sample can be replicated	cfg_random_number	An integer.

Additional notes:

- The gridEA algorithm was renamed gridEZ after GridSample2.0 development had occurred, thus our code refers to this algorithm as gridEA. gridEZ was released in March 2019 on GitHub.
- Some terms in the code could be improved.
    - multicell_psu_raster might have been more aptly named multicell_unit_raster as it defines the units that comprise the sample frame.
    - cells_allocated might have been more aptly named units_allocated as it refers to any sample frame unit (cells, gridEZ units, or own units)
    - GHS-SMOD is sometime written incorrectly as GHS-MOD.

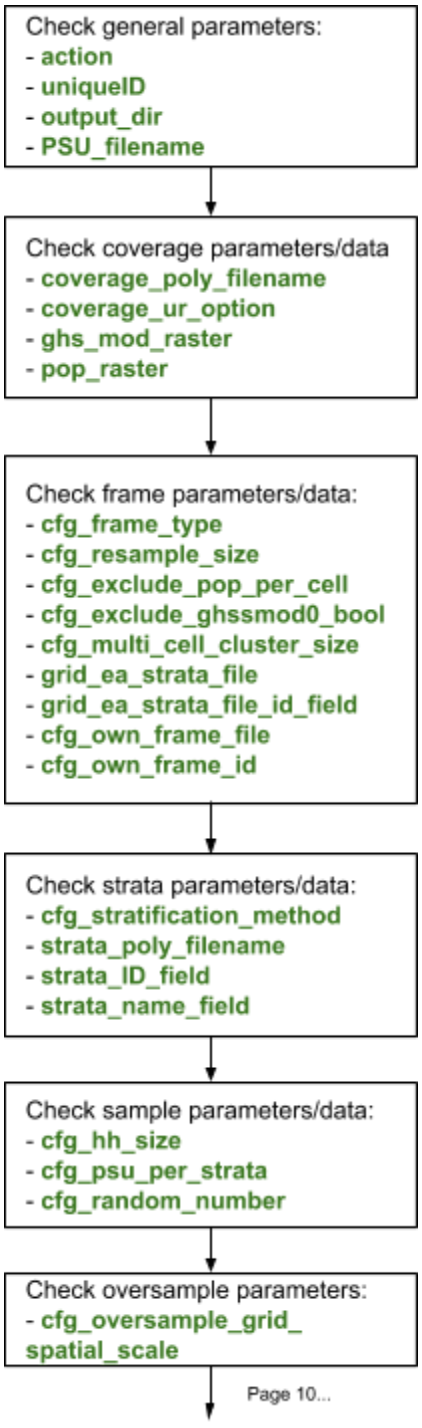
Functionality

The GridSample2.0 algorithm is available in [Flowminder GitHub] account.



Set-up and Check

This section loads the datasets and parameters, and flags any datasets or parameters that are missing.



Key:

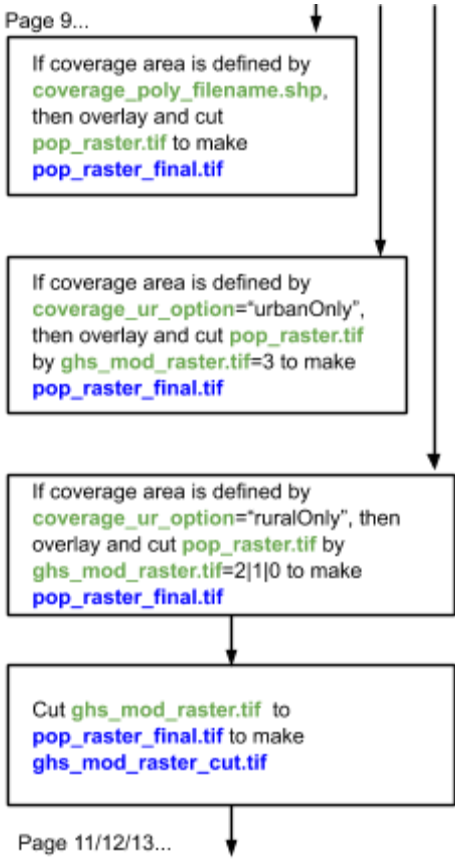
Green = parameter/data specified by user

Blue = generated dataset

Red = calculated value

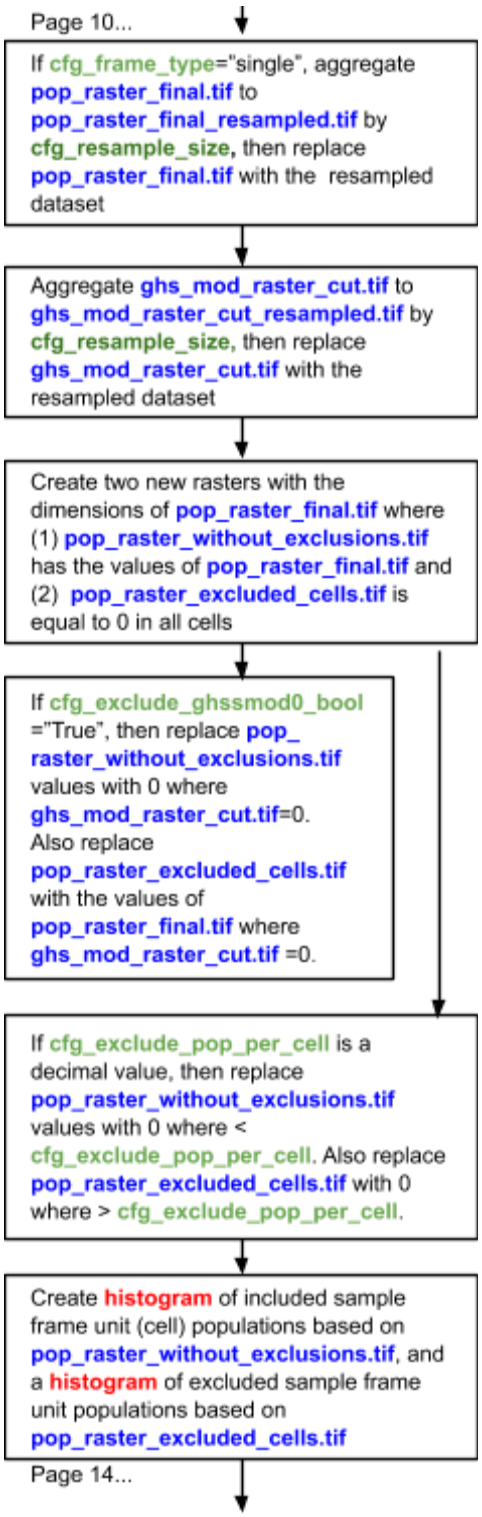
Coverage

Define the population raster and GHS-SMOD dataset by the coverage area.



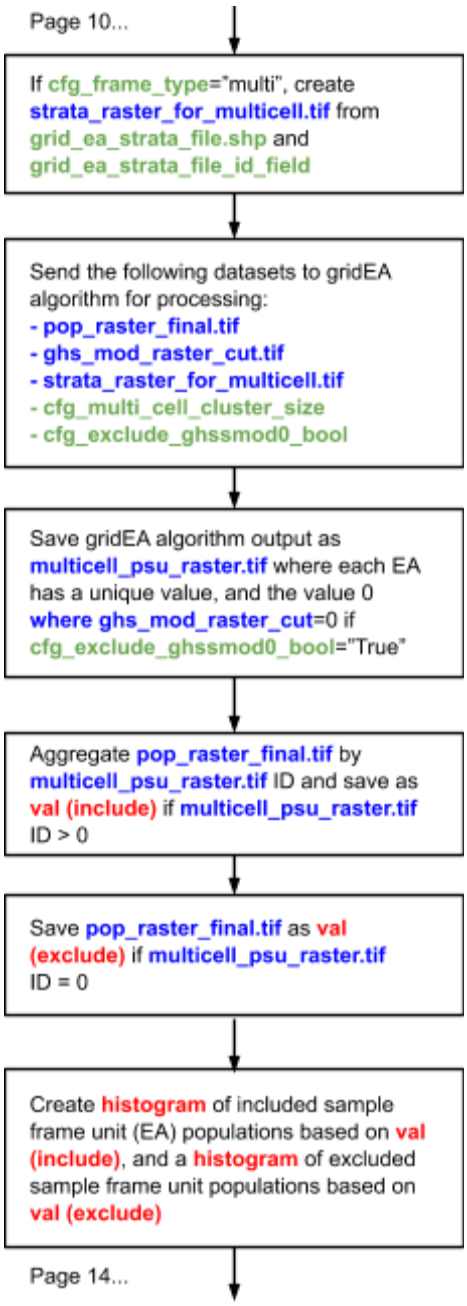
Frame: Single-cell

Create a sample frame of grid cells



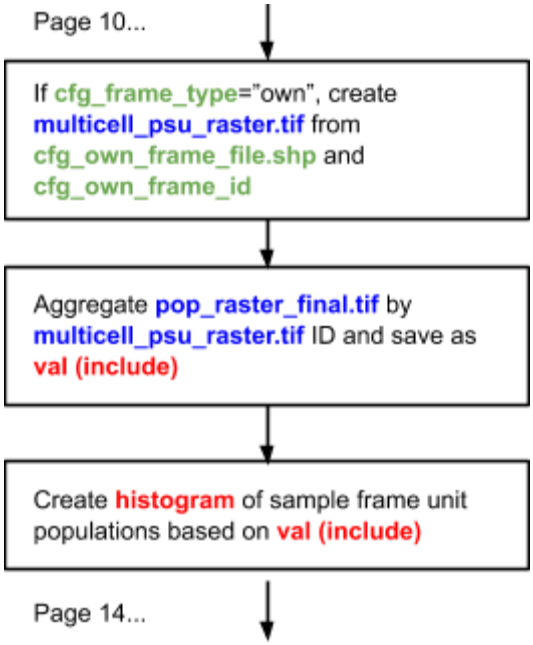
Frame: gridEZ

Create a sample frame of multi-cell enumeration-area-like units using the gridEZ algorithm



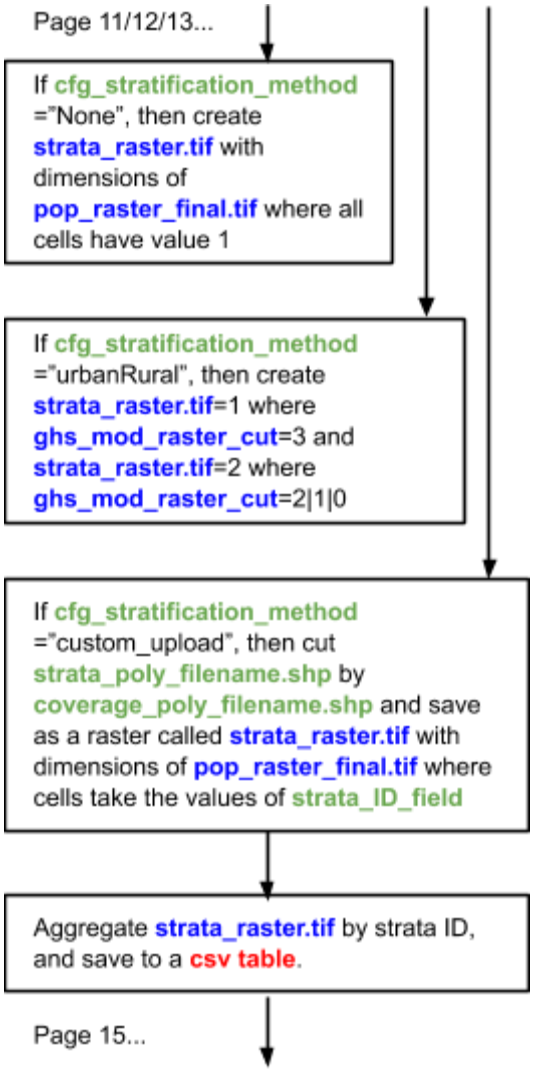
Frame: Own

Create a sample frame from census EAs or similar vector shapefile



Stratification

Stratify the sample and calculate population per stratum

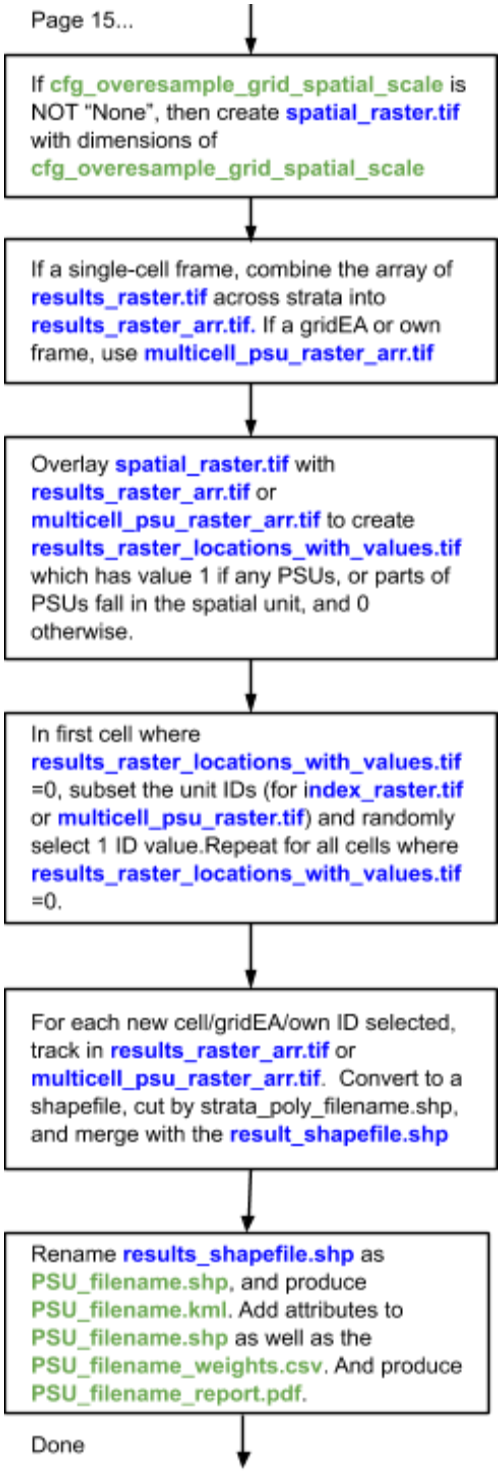


Sampling

Within each geographic strata, implicitly stratify by degree of urbanicity and select a sample of units (cells, gridEZ, or own) with PPS using serpentine sampling

Oversampling

Identify areas without a PSU and select a unit (cell, gridEZ, or own) at random



Example code and output

City sample: Surveys for Urban Equity (SUE) - Kathmandu

This example replicates a survey described by the the Surveys for Urban Equity project. This is an example of a one-stage household survey sample in Kathmandu Valley, Nepal in which each cluster has 15-20 households and does not exceed 1km X 1km in area. The sample frame is built using the gridEZ tool (referred to as gridEA tool in the code) with WorldPop, GHS-SMOD, and GADM 4th-level admin data. To select this sample in GridSample2.0, we define the coverage area as all urban areas in Nepal, as defined by GHS-SMOD = high dense urban (value 3). We then stratify by 2nd-level GADM administrative units, and set the number of sampled clusters in each strata to zero except in Bagmati (unit 1107) where Kathmandu Valley is located. In urban Bagmati (Kathmandu Valley), we select 60 clusters. Note that additional back-up clusters were selected in practice.

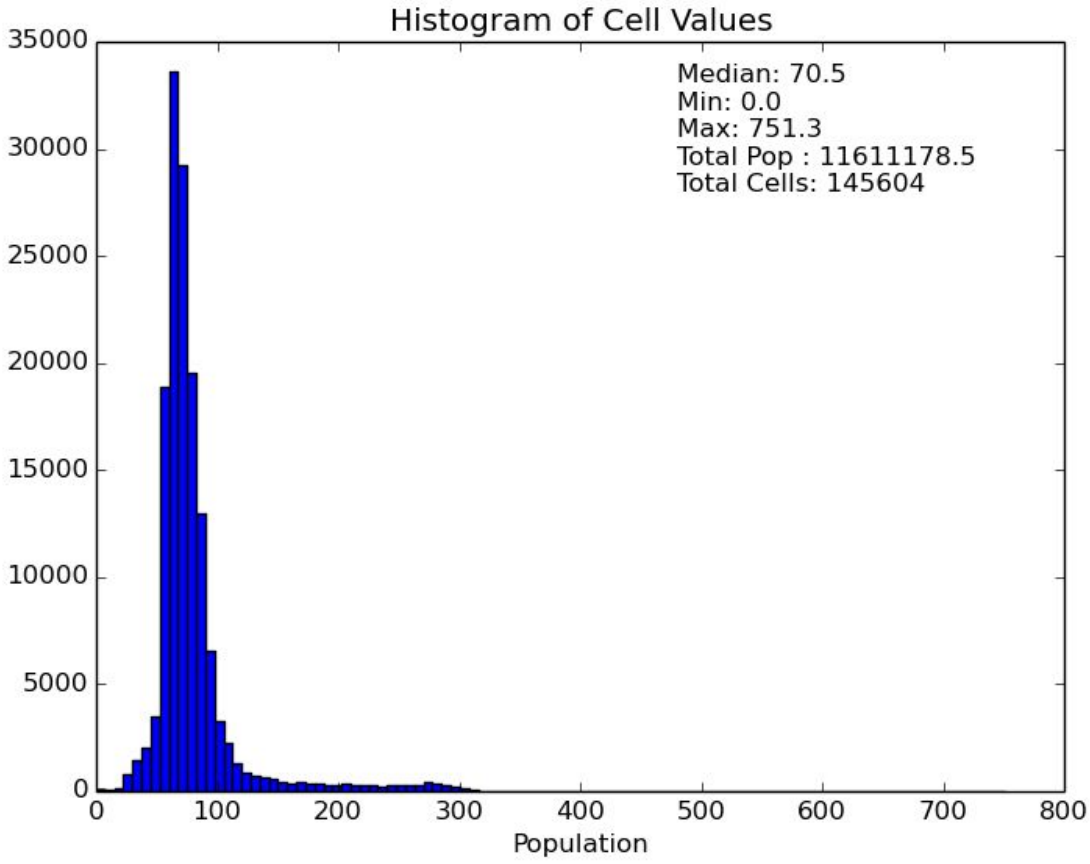
Generate a gridded enumeration area sample frame

This code calls the gridEZ algorithm and generates a sample frame of units with 15-20 households each in areas less than 1km X 1km.

```
def gridsample(

        # General parameters
        action = "create_histogram",
        uniqueID = 1,
        output_dir = c:/project/output/,
        ghs_mod_raster = c:/project/data/GHSL/GHS_SMOD_POP2015_GLOBE_R2016A.tif,
        pop_raster = c:/project/data/WorldPop/NPL_ppp_v2c_2020_UNadj.tif,
        coverage_poly_filename = c:/project/data/GADM/NPL_0.shp,
        coverage_ur_option = urbanOnly,

        # Sample frame parameters
        cfg_frame_type = 'multi',
        cfg_multi_cell_cluster_size = None,
        cfg_exclude_ghssmod0_bool = False,
        grid_ea_strata_file = c:/project/data/GADM/NPL_4.shp,
        grid_ea_strata_file_id_field = ID_4,

        # Below parameters not used for this step or for this survey design
        strata_poly_filename = ,
        strata_ID_field = None,
        strata_name_field = None,
```

```
cfg_stratification_method = None,
cfg_psu_per_strata,
cfg_hh_size = None,
cfg_exclude_pop_per_cell = None,
cfg_random_number = None,
cfg_resample_size = None,
cfg_oversample_grid_spatial_scale = None,
cfg_own_frame_file = None,
cfg_own_frame_id = None,
PSU_filename =

):
```

This output is a histogram of sample frame unit population counts within the coverage area.



Calculate population per strata

This code calculates the total population and proportion of population in each strata of the survey coverage area.

```
def gridsample(

       # General parameters
       action = "get_strata_pop_values",
       uniqueID = 1,
       output_dir = c:/project/output/,
       ghs_mod_raster = c:/project/data/GHSL/GHS_SMOD_POP2015_GLOBE_R2016A.tif,
       pop_raster = c:/project/data/WorldPop/NPL_ppp_v2c_2020_UNadj.tif,
       coverage_poly_filename = c:/project/data/GADM/NPL_0.shp,
       coverage_ur_option = urbanOnly,

       # Sample frame parameters
       cfg_frame_type = 'multi',
       cfg_multi_cell_cluster_size = None,
       cfg_exclude_ghssmod0_bool = False,
       grid_ea_strata_file = c:/project/data/GADM/NPL_4.shp,
       grid_ea_strata_file_id_field = ID_4,

       # Strata parameters
       cfg_stratification_method = True,
       strata_poly_filename = c:/project/data/GADM/NPL_2.shp,
       strata_ID_field = ID_2,
       strata_name_field = NAME_2,

       # Below parameters not used for this step or for this survey design
       cfg_psu_per_strata,
       cfg_hh_size = None,
       cfg_exclude_pop_per_cell = None,
       cfg_random_number = None,
       cfg_resample_size = None,
       cfg_oversample_grid_spatial_scale = None,
       cfg_own_frame_file = None,
       cfg_own_frame_id = None,
       PSU_filename =

):
```

This output is a table of population counts and proportions by strata.

```
Strata       Pop
Koshi        998609
Mechi        541247
Sagarmatha   921618
Mahakali     218662
Seti         373121
Bheri        603799
Karnali      1447
Rapti        112539
Dhaualagiri  78623
Gandaki      429469
Lumbini      1328009
Bagmati      3277721
Janakpur     1177830
Narayani     1536909
```

Select a gridded population sample

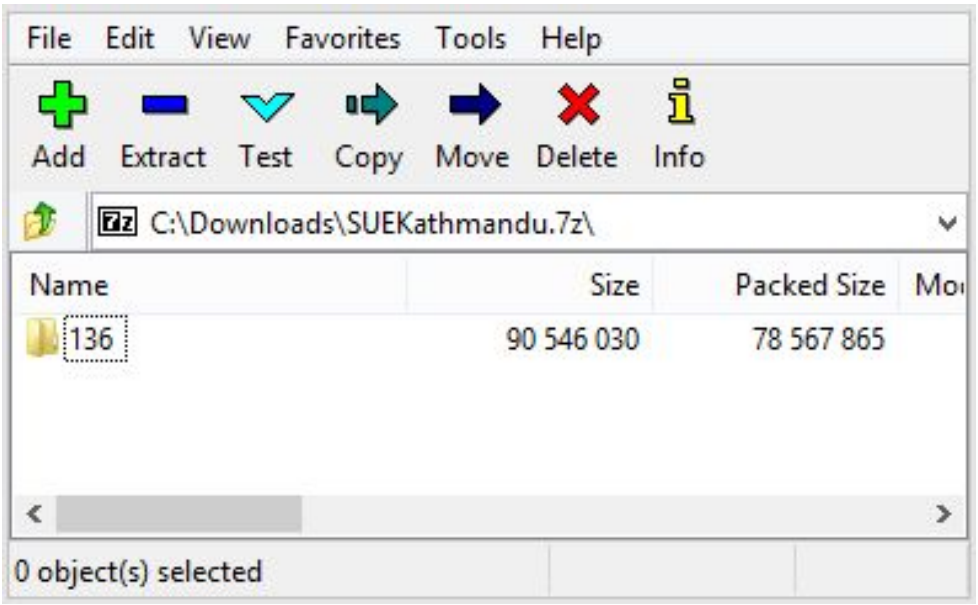
This code uses the sample frame and strata population counts to select a sample of units to serve as survey clusters.

```
def gridsample(

       # General parameters
       action = "get_strata_pop_values",
       uniqueID = 1,
       output_dir = c:/project/output/,
       ghs_mod_raster = c:/project/data/GHSL/GHS_SMOD_POP2015_GLOBE_R2016A.tif,
       pop_raster = c:/project/data/WorldPop/NPL_ppp_v2c_2020_UNadj.tif,
       coverage_poly_filename = c:/project/data/GADM/NPL_0.shp,
       coverage_ur_option = urbanOnly,

       # Sample frame parameters
       cfg_frame_type = 'multi',
       cfg_multi_cell_cluster_size = None,
       cfg_exclude_ghssmod0_bool = False,
       grid_ea_strata_file = c:/project/data/GADM/NPL_4.shp,
       grid_ea_strata_file_id_field = ID_4,
```

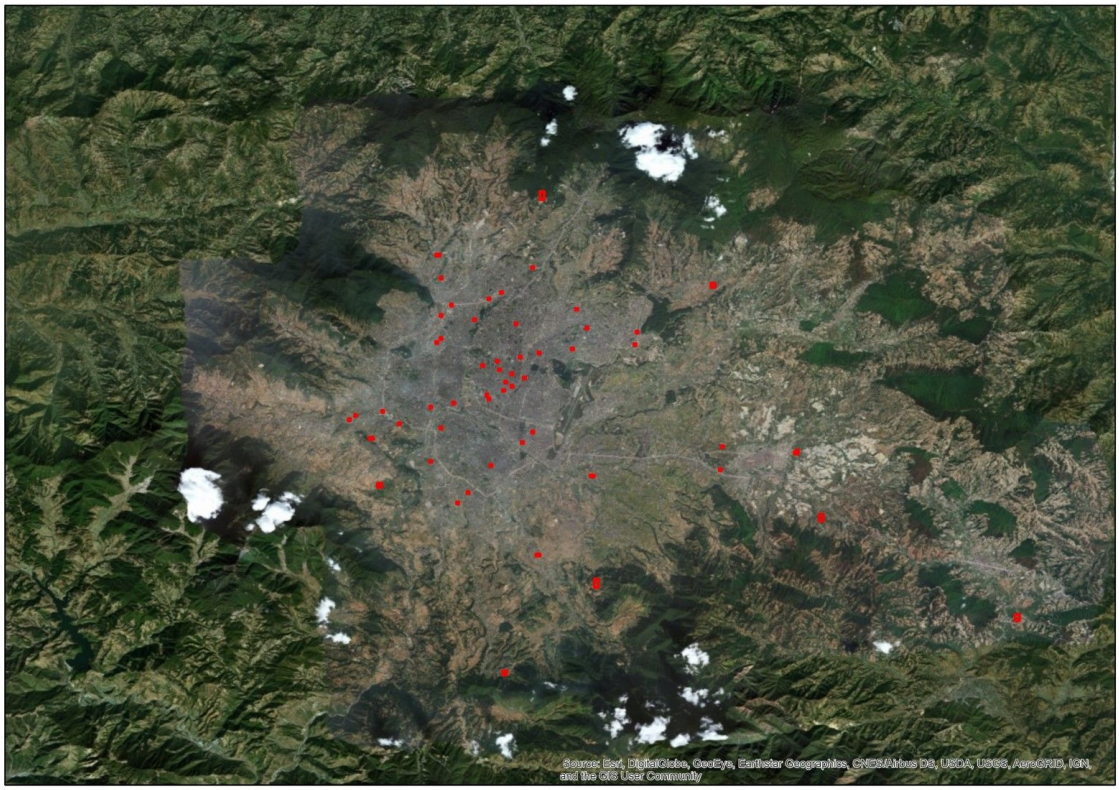
```
        # Strata parameters
        cfg_stratification_method = True,
        strata_poly_filename = c:/project/data/GADM/NPL_2.shp,
        strata_ID_field = ID_2,
        strata_name_field = NAME_2,

        # Sample selection parameters
        cfg_psu_per_strata ={u'1095': 0, u'1096': 0, u'1097': 0, u'1098': 0, u'1099':
        0, u'1107': 60, u'1104': 0, u'1105': 0, u'1102': 0, u'1103': 0, u'1100': 0,
        u'1101': 0, u'1108': 0, u'1109': 0},
        cfg_hh_size = 4.00,
        cfg_random_number = 111,
        PSU_filename = "Kathmandu.shp",

        # Below parameters not used for this survey design
        cfg_exclude_pop_per_cell = None,
        cfg_resample_size = None,
        cfg_oversample_grid_spatial_scale = None,
        cfg_own_frame_file = None,
        cfg_own_frame_id = None

):
```

This output is a shapefile and KML of cluster boundaries, a report of datasets and parameters used by GridSample2.0 to select the sample, and a tabular file with population estimates to calculate sample weights after response rates are known following the survey.

The output will be provided in a zipped .7z file



The user can visualise the enclosed KML in Google Earth, or the enclosed shapefile in ArcGIS or QGIS.



The user can open the enclosed sample weight file in Excel or OpenOffice to calculate sample weights after household/individual response rates are recorded in the field.

The enclosed PDF report visualizes all datasets and summarizes all parameters used to select the sample.