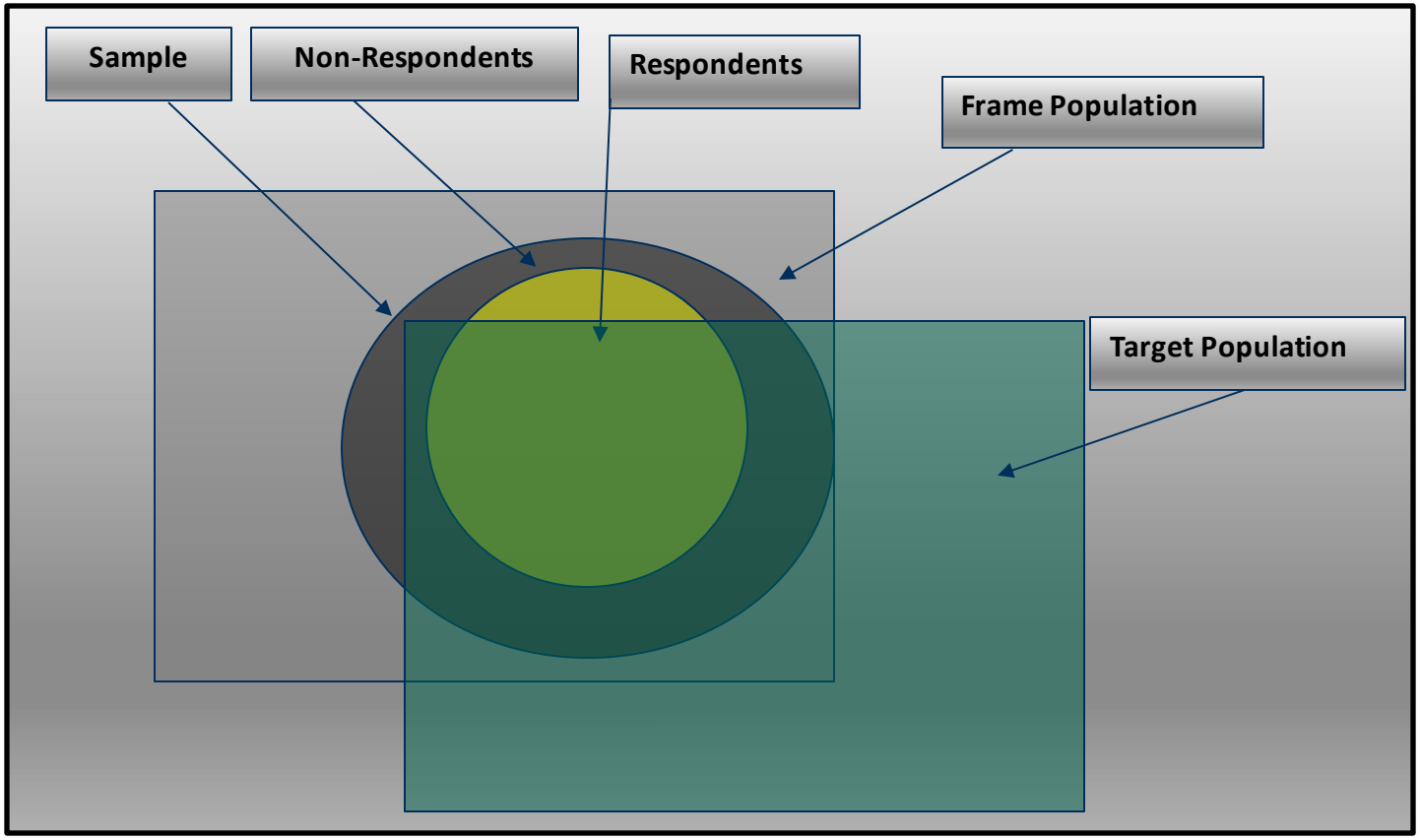
Sampling Frame Requirements

- It must be complete, i.e. cover each sampling unit in the target population once and only once.
- It must be current, i.e. the frame data must cover the target population as present during the survey period.
- It must be reasonably informative, i.e. contain information which can be used to make the sampling design more efficient.

Coverage by frame and final respondents



Groves et al., 2011, p 55

Potential Remedies

- Administrative Data
- Demographic Surveys
- Dual/Multi frame survey
- Satellite Data
- Call Data Records

Building a Spatial Synthetic Population (I)

Ingredients:

1 Raster of build area 3by3 m preprocessed by Facebook labs

1 PSU Boundary file (i.e. shape)

1 Population Census file

1 Survey Data file base on the same census frame

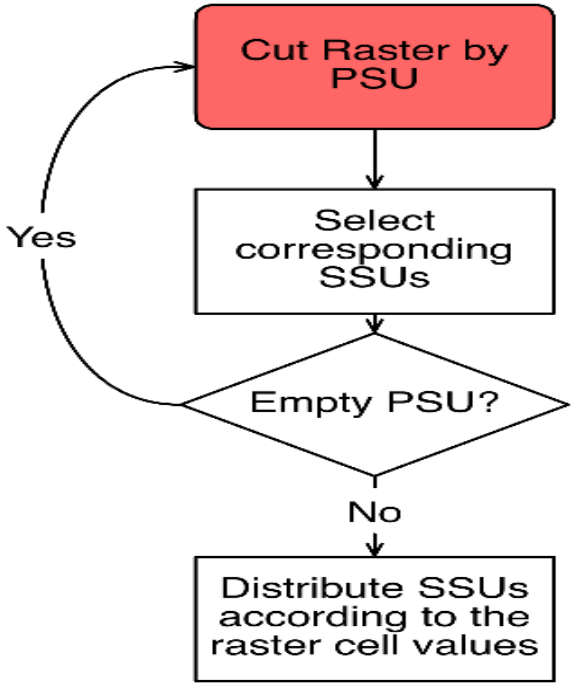
1) Generate Synthetic Population on True Census Data (R's simpop) → Can be skipped if privacy is not an issue

- Population Values are consistent down to PSU level, i.e. Age, Employment, Relationship to household head etc.
- Preserves HH structure and regional structure

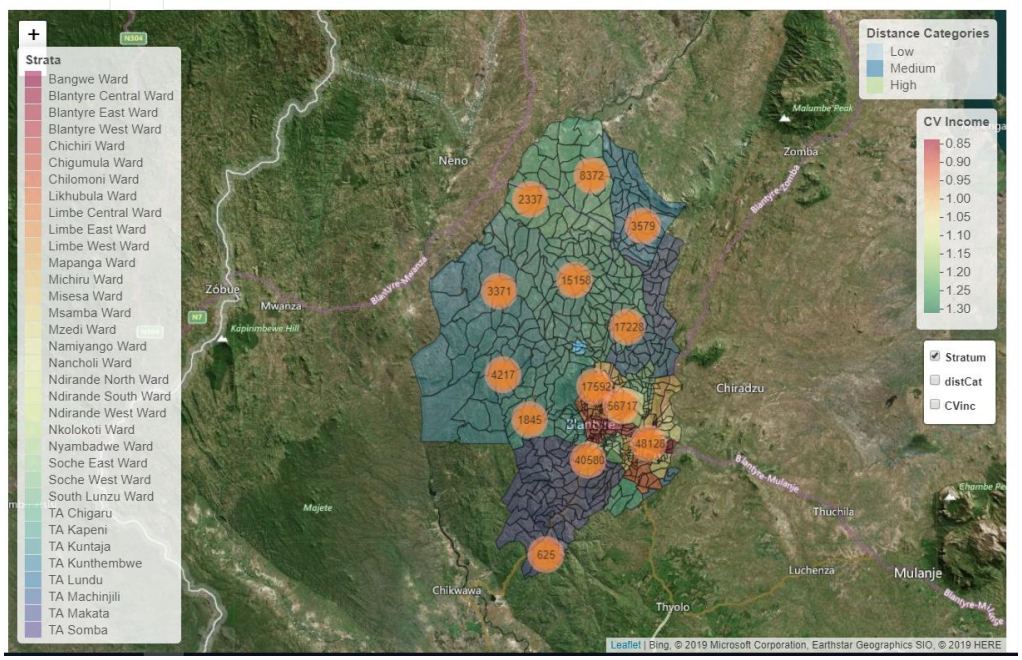
2) Simulate variables of interest not in census, i.e aggregate household consumption through 2-level random effects model estimated from survey data

Building a Spatial Synthetic Population (II)

3) Allocate synthetic population across space, according to the following algorithm, preserving the household counts at the corresponding census segment (i.e. PSU)



The Result



ESA Landcover 15m

NB_LAB	LCCOwnLabel	R	G	B
0	No data	0	0	0
1	Tree cover areas	0	160	0
2	Shrubs cover areas	150	100	0
3	Grassland	255	180	0
4	Cropland	255	255	100
5	Vegetation aquatic or regularly flooded	0	220	130
6	Lichens Mosses / Sparse vegetation	255	235	175
7	Bare areas	255	245	215
8	Built up areas	195	20	0
9	Snow and/or Ice	255	255	255
10	Open Water	0	70	200

ESA Climate Change Initiative, 2017

Population Value	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
AGE	951,375	21.1	16.76	0	8	30	99
P20_ECONACTIVE	555,371	1.54	0.5	1	1	2	2
P26_EMPLOYMENT_STATUS	727,472	3.11	1.45	1	2	4	6
P18B_SCHOOL_GRADE	755,696	3.84	2.34	0	2	5	8
EMPL	555,371	0.46	0.5	0	0	1	1
Total HH consumption (local currency)	219,749	563,910.30	139,770.10	401,014.00	484,899.10	644,746.70	3,917,976.00

A quick note on the math

Consumption

$$consumption_{ij} = \beta_0 + \beta_1 size_{ij} + \sigma_j + \epsilon_{ij}$$

Design Weights

$$p_{design} = p_1 * p_2 = \frac{m}{M} * \frac{n}{N_M}$$

MSE (Cochran, 1977)

$$MSE(\hat{Y}) = E(\hat{Y} - Y)^2 = E[(\hat{Y} - \tilde{Y}) + (\tilde{Y} - Y)]^2$$

$$= E(\hat{Y} - \tilde{Y})^2 + 2E(\hat{Y} - \tilde{Y})(\tilde{Y} - Y) + (\tilde{Y} - Y)^2$$

$$Var(\hat{Y}) + Bias(\hat{Y})$$

Calibration weights (Saerndal & Lundstgroem, 2005)

$$w_i^c = w_i + w_i \boldsymbol{\lambda_r} \boldsymbol{x_k} \ with \sum_r w_i^c \boldsymbol{x_k} = \boldsymbol{X}$$

A quick note on the design

1) All designs are at least 2 stage designs, with the first stage being AREAS, commonly referred to as census districts, and in the context of Household Surveys called Primary Sampling Units (PSU). The second stage units (SSU) and final sampling units are households. Theoretically there's a third stage, all persons within the household, but usually not mentioned further.
2) In some designs PSUs are sampled at random, in others proportional to number of households in the area.
3) Some designs use stratification.

Results: Standard Frame

CENSUS

Target Value & Design	MSE	Est. Pop. Mean	CV%
Age PPS	1.23	21.14	1.43
Age PPS (wrong size)	2.21	21.16	1.92
Age Random	1.56	21.11	1.6
Age STR	1.56	21.13	1.52
Age STRPPS	1.34	21.11	1.37
Consumption PPS	0.91	563329.11	0.91
Consumption PPS (wrong size)	1.45	563365.35	1.19
Consumption Random	1.06	563828.89	1.05
Consumption STR	0.62	563819.86	1.01
Consumption STRPPS	0.52	564101.1	0.9
Employment Ratio PPS	3.57	0.45	4.26
Employment Ratio PPS (wrong size)	4.88	0.45	5.16
Employment Ratio Random	4.37	0.45	4.38
Employment STR	4.02	0.46	4.09
Employment STRPPS	3.36	0.45	3.91
Population Count PPS	1.38	NA	1.5
Population Count PPS (wrong size)	11.99	NA	8.93
Population Count Random	5.78	NA	6.14
Population Count STR	5.01	NA	5.54
Population Count STRPPS	1.44	NA	1.47

SAMPLE SIZES

n_psu	n_ssu
80	12
71	12
80	12
80	12
80	12
71	12
71	12
71	12
71	12
71	12
71	12
71	12
71	12
71	12
71	12
120	12
120	12
120	12
120	12
120	12

HYBRID

Target Value & Design	MSE	Est. Pop. Mean	CV%
Age PPS	1.87	21.13	1.91
Age PPS (calibrated)	0.33	21.1	0.32
Age STRPPS	1.67	21.16	1.71
Age STRPPS (calibrated)	0.29	21.1	0.28
Consumption PPS	1.31	563728.96	1.24
Consumption STRPPS 1	1.13	563774.77	1.1
Employment PPS	5.37	0.45	5.74
Employment PPS (calibrated)	0.26	0.46	0
Employment STRPPS	4.99	0.45	4.87
Employment STRPPS (calibrated)	0.21	0.46	0
Population Count PPS	9.59	NA	10.06
Population Count STRPPS	7.32	NA	7.53

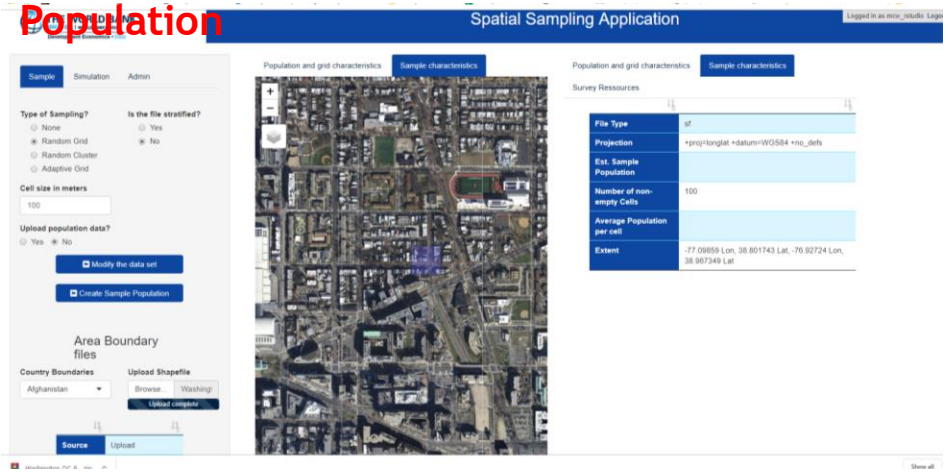
Results: Gridded Population Only

Target Value & Design	MSE	Est. Pop. Mean	CV%
Age PPS	3.96	21.11	3.22
Age PPS (calibrated)	3.29	21.08	0.52
Consumption PPS	2.21	563961.95	1.8
Employment PPS	11.39	0.46	8.96
Employment PPS (calibrated)	8.68	0.46	0

n_psu	Est. Pop. Total	Av. Samp. Size
100	938156	6552
100	922565	6593
100	215671	6598
100	933408	6626
100	965882	6630

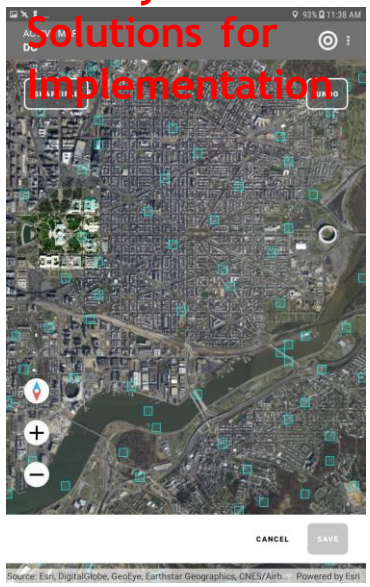
Challenge in Implementation, but with the "right" tool box, even in low skill environment possible.

Shiny Application to Sample from Gridded Population



Rest API

Survey Solutions for Implementation

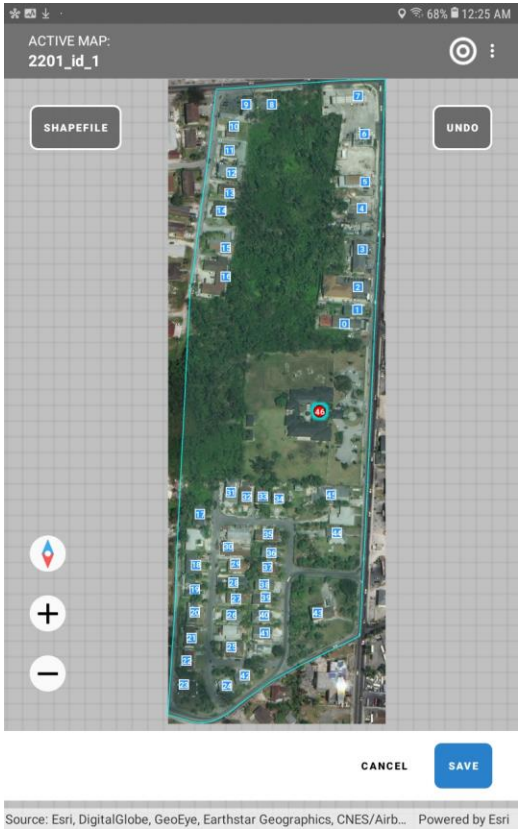


Conclusion (I)

- Using remote sensing data to enhance informativeness for stratification as well as for updating has proven useful in the conducted simulations resulting in more efficient estimators through stratification for efficiency gains, and with similar efficient estimators in the PPS design.
- Since census frame quality deteriorates quickly thereafter and in particular in countries with strong population dynamics its usefulness becomes questionable only a few years after.
- Correcting these shortcoming by using remotely sensed data may therefore be the first line of defense (in the absence of any other data sources)

Conclusion (II)

- Gridded Population data may be used, but degree of precision currently differs strongly between countries
- Connecting ground & sky is therefore mandatory during the upcoming census round and in general for listing operations. The precision of regular tablet GPS may not be sufficient for that, however some Survey Systems allow for using satellite imagery directly inside the standard questionnaire, resulting in highly precise verification data.



Example on how to use building locations in Survey Solutions integrated into the standard census questionnaire.

Challenges

- Capacities in this area can not be build up any longer in a reasonable time, by doing introductory trainings and distributing manuals alone. The two former measures require support in data preparation and pre-processing and through specialize tools using state of the art technologies.
- Auxiliary information, like Remotely sensed data should be preprocesses made available by International Organization to be useable by statistical agencies, as it involves strong economies of scale.
- A caveat of this approach is that through the development of tools, relevant statistical standards can be maintained, and the black-box of data quality finally turns transparent.
- Also the hardware requirements for this kind of data – even preprocessed - may still be prohibitively high, and support through projects like i.e. The World Banks C4D (Cluoud for Development) project may bridge this gap.

Outlook & Points for Discussion

A. What remains to be tested and was not covered in the current simulation are pre-survey household listings for the second stage sampling units, with ample evidence of high errors (i.e. Eckman, 2013). This further deteriorates the quality of our sampling frame and its impact is unknow so far.

B. One additional advantage of well designed low skill requiring Computer Assisted Survey Systems, is the availability of a large number for quality control mechanism, including the use of geo-spatial data for geo-fencing, mapping operations etc. fully integrated into the standard survey workflow (i.e. Survey Solutions).

C. DEC-DG has currently also initiated the process of building up a data base for sampling (and other purposes) and the integration of automated verification system through listing data with the members of the Grid 3 group and in particular i.e. WorldPop. The data is the freely available for statistical agencies, and in the pre-processe format.

Literature

Cochran, W.G., 1977. Sampling Techniques: 3d Ed. Wiley.

Eckman, S., 2013. Do different listers make the same housing unit frame? Variability in housing unit listing.

ESA Climate Change Initiative, 2017, Land Cover project, viewed 22 January 2018, http://2016africalandcover20m.esrin.esa.int .

Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R., 2011. *Survey methodology*(Vol. 561). John Wiley & Sons.

Särndal, C.E. and Lundström, S., 2005. Estimation in surveys with nonresponse. John Wiley & Sons.