ELTE | FACULTY OF SOCIAL SCIENCES

Modified adaptive cluster sampling for improved field data collection

Dr David Simon
assistant professor
Department of Statistics,
Faculty of Social Science, ELTE

Jamie Burnett
Head of Methods for
Multinational Surveys
Verian

CSDI Workshop, Berlin 18.03.2024

Partially supported by Erasmus+

Content

- Sampling problems of rare and elusive population
- Adaptive cluster sampling
- Sampling design of Roma Survey 2020-2021
- Modified adaptive cluster sampling
- Experiences with modified adaptive cluster sampling
- Conclusions

Sampling problems of rare and elusive population

- Definition of Spreen (1992)
  - low proportion in general population
  - respondents hiding identity
- Increasing importance: healthcare, migration, minority groups, miscellaneous social groups (with decreasing visibility)
- From sampling perspective:
  - no sampling frame
  - fishy definition of target group
  - lack of reliable population data
  - target population difficult to access
  - large gross samples, large costs

Adaptive cluster sampling

- Different sampling solutions based on probability sampling increasing inclusion probability: stratified sampling with uneven probability, time and space sampling, network based sampling and adaptive cluster sampling (etc.)
- Adaptive cluster sampling
  - proposed by Thompson (1990) for different situations (for example ecology, geology, epidemiology), used for human research for example by Verma and Gagliardi (2010)
  - usefull when target population assumed to be clustered
  - in essence:
    - random sample is drawn from the general population
    - when member of the target population identified
    - neighbourhood included in sample as well until non-eligible member was found
  - advantages: larger sample with smaller cost (even comparing effective sample sizes), no bias

Adaptive cluster sampling

- No recent review about the method (last review was published in 2005 by Turk and Borkowski)
- Practical usage can be difficult combined with random walk
  - in these cases neighbourhood considered to be those households that were not asked on the route before and after the given eligible households in row until a non-elligible HH was found or until the next or previous core household
  - during EU MIDIS II (2011-13) issues with applying rules as it seemed to be to complicated (some interviews had to be deleted)

Sampling design of Roma Survey 2020-2021

- Goal: sampling 16+ Roma population in 10 EU countries (Portugal, Spain, Italy, Czechia, Romania, North Macedonia, Serbia, Greece, Croatia and Hungary) with highest coverage, random sample, no bias, low design effect, CAPI
- Background research: no population register, Census (~2010) or other Roma research
- Random PSU selection – excluding PSU withe <5-10% Roma population
- Mapping Roma population by experts based on 250m2, 500m2 or 1km2 grid -> SSU
- Random selection of SSUs – similar exclusion criteria (coverage should be above 90%)

Sampling design of Roma Survey 2020-2021

- In selected SSUs
  - HH level screening
  - Gross sample size was set based on estimated proportion of Roma population (eligibility rate) and estimated resopnse rate
- Random walk HH selection
  - Starting point: Random GPS coordinates, reverse geocoding, moving to nearest residential area if necessary
  - Selection interval: NT/Ng where NT is the total number of households in the SSU and Ng the gross sample (capped at 10)
  - Random direction matrix with three ordered options (Bauer, 2017)

Sampling design of Roma Survey 2020-2021

- Adaptive cluster sampling (ACS)
  - in SSUs where the estimated Roma concentration was less than 25%
- Stopping and dropping rule
  - Dropping rule: drop SSUs where the eligibility rate during fieldwork was estimated to be significantly lower than the Roma concentration estimated based on the SSU frame (probability of the concentration rate correct <10% - based on binomial distribution)
  - Stopping rule: if three times of the expected number of interviews were achieved no further new addresses were visited in the given SSU
- Within household selection
  - Random algorithm

Modified Adaptive Cluster

- Modified ACS, including only allocating all addresses between the eligible core address and the next core address (no backward selection)
- Questions (pre-calculations)
  - Expected cluster size (gain of ACS) assumed to be smaller
  - Expected design effect
  - Expected loss of sampling size

Modified Adaptive Cluster

	P_e=0.25			P_e=0.18			P_e=0.1			P_e=0.01		
	P_c=0.6	P_c=0.8	P_c=0.90	P_c=0.6	P_c=0.8	P_c=0.90	P_c=0.6	P_c=0.8	P_c=0.90	P_c=0.6	P_c=0.8	P_c=0.90
i=2	1,60	1,80	1,90	1,60	1,80	1,90	1,60	1,80	1,90	1,60	1,80	1,90
i=3	2,01	2,45	2,71	2,00	2,45	2,71	1,98	2,44	2,71	1,96	2,44	2,71
i=4	2,34	3,00	3,45	2,29	2,98	3,45	2,23	2,97	3,44	2,18	2,95	3,44
i=5	2,63	3,47	4,12	2,52	3,43	4,11	2,42	3,40	4,11	2,32	3,36	4,10
i=6	2,89	3,88	4,74	2,73	3,81	4,72	2,57	3,75	4,70	2,40	3,70	4,69
i=7	3,15	4,24	5,31	2,93	4,15	5,28	2,69	4,05	5,25	2,45	3,96	5,22
i=8	3,40	4,58	5,83	3,11	4,44	5,79	2,81	4,31	5,74	2,49	4,17	5,70
i=9	3,65	4,89	6,32	3,30	4,71	6,26	2,91	4,53	6,19	2,52	4,35	6,13
i=10	3,91	5,19	6,78	3,48	4,96	6,69	3,02	4,72	6,61	2,54	4,49	6,52

Loss of sample size

Average (based on average cluster size): 9.8%
Min: -42.6%
Max: 37.1%

Pe	proportion of eligible HHs
Pc	conditional probability that a household is eligible if the previous household is eligible
i	selection interval
red	smaller cluster size than non-modified ACS
green	larger cluster size than non-modified ACS

Modified Adaptive Cluster

Design effect estimation

	$E(C) = 1.5$	$E(C) = 2.0$	$E(C) = 2.5$	$E(C) = 3.0$	$E(C) = 3.5$	$E(C) = 4.0$	$E(C) = 4.5$	$E(C) = 5.0$	$E(C) = 5.5$	$E(C) = 6.0$
ρ = 0.02	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.10
ρ = 0.06	1.03	1.06	1.09	1.12	1.15	1.18	1.21	1.24	1.27	1.30
ρ = 0.12	1.06	1.12	1.18	1.24	1.30	1.36	1.42	1.48	1.54	1.60

E(C) cluster size

ρ intraclass correlation for clusters (estimated)

	Non-modified ACS ($E(C) = 3.95$)	Modified ACS ($E(C) = 3.56$)	Loss of effective sample size
ρ = 0.02	3,73	3,39	9.1%
ρ = 0.06	3,36	3,09	8.0%
ρ = 0.12	2,92	2,72	6.6%

ICC values based on ESS technical documentation

Experiences with modified adaptive cluster

- MACS: 2727 of the total sample of 8773
- Average cluster size: 3,5
- ICC, design effect, sample gain, sample loss for (some) key indicators

	ICC	Design effect	Effective sample gain compared to no AC	Effective sample loss compared to ACS
Discrimination	**0.62**	2.55	1071	40
Paid work	**0.43**	2.07	1320	75
Deprivation	**0.74**	2.86	953	24

Conclusions

- Simpler rule for ACS combined with random walk: including all HHs between an eligible HH and the following core HH
- Modified ACS can be a useful alternative to ACS samplig rare and elusive population
- Small decrease in sample gain might occur, partially counterballanced by smaller design effect
- The larger the expected intraclass correlation the smaller the loss of sample size
- The more clustered the target population the lower the loss of sample size (it can even be negative – further gain of sample size)
- **ICC can be significantly larger than expected**

Thank you for your attention