Augmentation of the Samples size for the Household Budget Survey

Liljana BOÇI[1], Alma KONDI[2]

[1,2] Instat, Albania

E-mail: [1]lboci@instat.gov.al, [2]akondi@instat.gov.al

ABSTRACT

The sample designs for the Albanian Household Budget Survey (HBS) are based on the around 7900 households of the Census 2011 stratified for prefecture and urban/rural areas with Probability Proportional to size sampling according to the number of households inside the strata. The sample design is not optimized for the new 61 municipalities of Albania. There is a need for publication on small domains like municipalities for important indicators of this survey, at least for the consumptions of the 12 main groups of consumption. Important indicators should be published for the municipalities if the precision is acceptable. Either spatial or temporal aggregation (pooling) are possible solutions for certain (larger) municipalities. This study aims to give solutions for augmentation on the number of PSU to reach the publication on municipalities at least on some of them and save the design effect of HBS survey in Albania.

Keywords

Small Area Estimation, Consumption, Augmentation, Variance, Sample size

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

The Household Budget Survey[1] in Albania is a statistical survey carried out at by Institute of Statistics of Albania (INSTAT). The target groups of this survey are Albanian resident households with the main focus to present the socio-economic panorama of the Albanian households. The first Household Budget survey took place in 1993-1994 which were representative only in Tirana city, then was followed in 1999-2000 which was an annually survey and was representative only in urban areas for whole country. The first HBS representative for whole Albania and also in prefecture level was conducted in 2006-2007, and the second one in 2008-2009. From 2014 and on HBS is a continuous survey and is conducted following EUROSTAT recommendation.

The main purpose of this suvey is to estimate the level and structure of consumption expenditure in the country as a whole as well aggregated by prefecture level. This suvey provides information about the households expenditures by different type of disaggregation like as: prefecture, level of education of the head of households, household size etc.

Regarding this paper the three years data 2016, 2017 and 2018 are taken into consideration to analyses the stability of them and focus on finding the possibilities of publication of the data on the municipality's level. The focus on the analysis will be the calculation of the variance, second on the putting conditions for the precision required and then finding and suggested different methods for publication on municipality's level.

Variance calculations

The variance estimates are based on the inclusion probabilities of the psu $\pi 1i$. The starting points of the sampling weights are taken into account tin the variation of non-response (Laaksonen 2007, Laaksonen and Heiskanen 2014). They are proportional to the number of households in the psu. The allocation to the 24 strata of psu (prefecture × urban/rural areas) is also proportional to the number of households.

The household weight is based on the design weights and an additional calibration. Therefore, to arrive at the household weights the inclusion probability of the households are derived as $\pi_{2i} = 1/(w_{ij} \cdot \pi_{1ij})$. In other words the calibration is not taken into account explicitely. Since a few $\pi 2ij$ had to be censored to 1 the implied weights $1/(\pi_{1i}\pi_{2ij})$ do not sum exactly to $\sum_{i,j\in S} w_{ij}$.

All household members are included in the HBS and, therefore, the person weight is equal to the household weight. The stratification information and the inclusion probabilities and weights are matched to the sample with the help of the svydesign function of the R package survey. For

[1] http://www.instat.gov.al/

the variance calculation the approximation formula for inclusion probabilities to size designs from Brewer is used. The quality of Brewer's it is described by Berger (2004). The calculation of the variance in recursive algorithm it is well described by Bellhouse (1985) . The other authors Cochran (1977) and Sarndal et al (1991) describe the decomposition of the variance into a single-stage between-cluster estimator and a within-cluster estimator, and this is applied recursively.

Estimates for HBS are means of positive quantitative variables. The coefficient of variation is a reasonable measure to judge precision in that case. We did not find a regulation of the European Union on the required level of precision of estimates for the Household budget surveys.

Augmentation of the sample size Household Budget Survey Results

The general idea of the augmentation of the sample size is to work at the level of the first stage, i.e. to increase the number of psu. When increasing the sample size of the households within a psu rapidly, the problem of exhausting the psu could arise, and this would most probably increase the already rather high design effect due to the clustering into psu. Increasing the number of psu should not increase the design effect.

Table 1 : CV-Classification

range of cv	stars	description
$cv < 0.01$	***	Excellent
$0.01 \leq cv \leq 0.025$	**	Good
$0.025 < cv \leq 0.05$	*	sufficient
$cv > 0.05$		insufficient

Table 2: Total Consumption 2017 per Municipality

Code of Municipality	name	Mean	SE	deff	cv	stars
1	BERAT	774914	46964	2.1	0.061	
2	URA VAJGURORE	826267	41329	0.8	0.050	
3	KUÇOVË	603211	25837	0.8	0.043	*
4	SKRAPAR	465858	9211	0.2	0.020	**
5	POLIÇAN	471290	108993	4.4	0.231	
6	DIBËR	579722	39278	3.0	0.068	
7	BULQIZË	431733	61281	6.3	0.142	
8	MAT	347018	14576	2.3	0.042	*
9	KLOS	268683	18117	1.0	0.067	
10	DURRËS	816191	46183	4.0	0.057	
11	SHIJAK	761505	63482	1.7	0.083	
12	KRUJË	507735	38627	2.1	0.076	
13	ELBASAN	556784	19725	1.8	0.035	*
14	CËRRIK	553348	35177	0.9	0.064	
15	BELSH	500695	40959	1.3	0.082	
16	PEQIN	825501	28025	0.8	0.034	*
17	GRAMSH	552647	49132	1.3	0.089	
18	LIBRAZHD	676083	51440	0.9	0.076	
19	PRRENJAS	447357	56121	3.0	0.125	
20	FIER	650524	43580	3.7	0.067	
21	LUSHNJE	704590	34525	1.4	0.049	*
22	PATOS	967851	102246	1.3	0.106	
23	ROSKOVEC	529222	83097	3.5	0.157	
24	DIVJAKË	872111	77760	1.3	0.089	
25	MALLAKASTËR	798085	39881	1.0	0.050	*
26	GJIROKASTËR	576785	29558	1.1	0.051	
27	LIBOHOVË	342017	16850	0.2	0.049	*
28	PËRMET	585970	81128	2.1	0.138	

29	KËLCYRË	524203	12233	0.1	0.023	**
30	TEPELENË	870490	61840	1.0	0.071	
31	MEMALIAJ	835956	52519	1.4	0.063	
32	DROPULL	254739	1297	0.0	0.005	***
33	KORÇË	762635	59563	2.7	0.078	
34	POGRADEC	825724	27289	0.8	0.033	*
35	MALIQ	635820	63657	3.6	0.100	
36	PUSTEC	461368	29674	0.1	0.064	
37	KOLONJË	809771	121899	4.1	0.151	
38	DEVOLL	464674	36257	1.4	0.078	
39	KUKËS	727853	60188	2.5	0.083	
40	TROPOJË	516621	28403	0.8	0.055	
41	HAS	687939	60668	3.9	0.088	
42	LEZHË	776694	54763	1.6	0.071	
43	MIRDITË	864022	131673	2.2	0.152	
44	KURBIN	637330	20804	1.0	0.033	*
45	SHKODËR	703826	43616	2.8	0.062	
46	VAU I DEJËS	816127	133084	4.9	0.163	
47	MALËSI E MADHE	853115	46713	0.6	0.055	
48	PUKË	624266	26782	0.4	0.043	*
49	FUSHË ARRËS	467533	11095	0.1	0.024	**
50	TIRANË	913298	25350	2.4	0.028	*
51	KAMËZ	737464	41923	2.5	0.057	
52	VORË	351042	25420	2.6	0.072	
53	KAVAJË	762191	81975	2.7	0.108	
54	RROGOZHINË	973569	77023	0.4	0.079	
55	VLORË	532874	34895	1.9	0.065	
56	HIMARË	515572	7863	0.0	0.015	**
57	SARANDË	845295	75365	1.3	0.089	
58	KONISPOL	666910	54829	0.9	0.082	
59	DELVINË	457221	35536	0.5	0.078	
60	FINIQ	655254	54967	0.1	0.084	
61	SELENICË	495844	28764	0.5	0.058	

Actually the augmentation factor, the square of this ratio: $a = \frac{cv^2}{cv^2_{derised}}$

is used to increase the number of psu per municipality and save the design effect(deff).

It turns out that an increase of psu to yield a particular level of cv for all municipalities would mean a very large number of new psu to include in the sample. Therefore, two alternatives are considered: Increase of the largest 9 municipalities only, and increase of the largest 6 municipalities only. The largest 9 municipalities in terms of number of sampled psu in 2017 are DURRËS, ELBASAN, FIER, LUSHNJE, KORÇË, SHKODËR, TIRANË, KAMËZ, VLORË. For the largest 6 municipalities LUSHNJE, KORÇË and KAMËZ are dropped from this list. Also in terms of number of households in 2011 these are the 6 resp. 9 largest municipalities. In addition to restricting to the largest municipalities, the variability of the coefficients of variation must be smoothed in order to arrive at a stable prediction of the precision. Figure 1 gives the coefficient of variation of total consumption for 2017 for the municipalities with more than 7 sampled psu. The explanatory variable is $1/\sqrt{nspsu}$, the inverse of the square root of the number of sampled psu. There is a good correlation visible, though obviously the variability increases with lower number of psu (to the right of the scatterplot). The red observations are the 6 largest municipalities and the three black observations are the ones that are added for the 9 largest municipalities. Instead

of the observed (estimated) cv we use the prediction from a linear fit (the blue line in the scatterplot), i.e. the predicted cv instead of the observed cv.

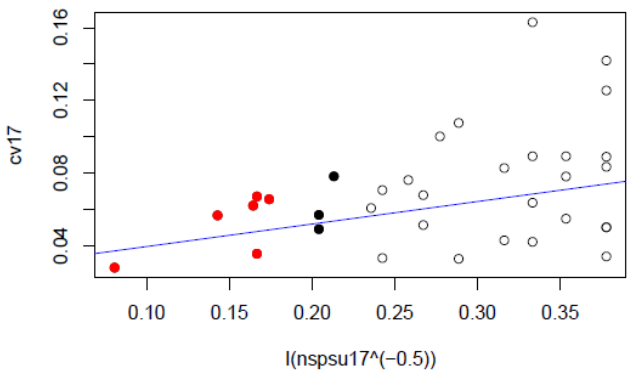

Figure 1: CV for total consumption 2017 vs. $1/\sqrt{n_i}$ reduce the cv.

To evaluate a few scenarios with different levels of requirement on the desired predicted cv we use the limits 2:5%;4% and 5% as desired cv. For example the second scenario would require a predicted cv less than 4% either for all, for the largest 9 or only for the largest 6 municipalities. The number of additional psu is given in Table 2.

Number of additional psu needed to achieve a desired precision

desired cv	all	large9	largest6
0.25	3643	910	669
0.4	971	124	73
0.5	423	7	0

A minimal scenario would draw just 7 psu more in the largest 9 municipalities to achieve a predicted cv below 5% in these municipalities. A somewhat more ambitious scenario would draw 73 new psu in the largest 6 municipalities. This would achieve a cv of maximum 4% for these 6 municipalities. A still bit more ambitious scenario would draw 124 new psu in the 9 largest municipalities to bring the predicted cv down to 4%.

Sample design

An obvious sample design for the augmentation would be to give the municipalities which are to be augmented an own stratum, to allocate the sample proportional to the number of households and to draw the total number of psu with inclusion probability proportional to the number of households. Thus the change would be dealt with by an additional stratification.

A second version for the sample design would be to draw a first sample S_1 according to the existing sample design. This would leave the existing design untouched. Then for each municipality with augmentation, a second sample S_2 is drawn among the remaining psu with inclusion probability proportional to the number of remaining households. The total inclusion probability of a psu in an augmented municipality then is

$$\pi_{1i} = P[i \in S_1] + P[i \notin S_1] * P[i \in S_2 | i \notin S_1] = \frac{n_1 M_i}{\sum_U M_i} + \left(1 - \frac{n_1 M_i}{\sum_U M_i}\right) \frac{n_2 M_i}{\sum_U M_i - \sum_{S_1} M_i},$$

where M_i is the number of households of psu i , n_1 and n_2 are the sample sizes of the first and second sample in the municipality. With this second version of the design, the inclusion probability of a psu depends on the first sample. However, if the sampling fraction of psu is low, the deviation from $(n_1 + n_2)M_i / \sum_U M_i$ would be small. A variant would be to not withdraw the units of the first sample when drawing the second sample. Then it may happen that a unit is drawn twice, of course. But the inclusion probability would be $(n_1 + n_2)M_i / \sum_U M_i$.

Conclusion and Recommendations

The sample design based on the municipalities is reasonable. It will obtain similar precisions as for the present sample design. Small strata (say below 40 PSU in the population) should be avoided. Too small strata do not leave enough freedom to rotate the sample or the renew a sample with low probability to select the same PSU. If strata within municipalities are too small they may be collapsed. This results in a sub-stratification which is limited to the larger municipalities.. A minimal sample size of 4 may be considered. The samples of the largest 9 municipalities may be augmented to reach a chosen precision in terms of cv. The augmentation of 124 psu to reach 4% coefficient of variation of the largest 9 municipalities is recommended. Alternatively only the largest 6 municipalities could be augmented with 73 new psu. Pooling the HBS over three years would bring down the coefficient of variation from 4% to approximately 2.5%. Results with less than 5% coefficient of variation can be published when standard errors are accompanying the estimates.

References

[1] Bellhouse DR (1985) Computing Methods for Variance Estimation in Complex Surveys. Journal of Official Statistics. Vol.1, No.3, 1985

[2] Berger, Y.G. (2004), A Simple Variance Estimator for Unequal Probability Sampling Without Replacement. Journal of Applied Statistics, 31, 305-315.

[3] Cochran, W. (1977) Sampling Techniques. 3rd edition. Wiley.

[4] Laaksonen, S. (2007) Weighting for Two-Phase Surveyed Data. Survey Methodology, December Vol.33, No. 2, pp. 121-130, Statistics Canada.

[5] Laaksonen, S. and Heiskanen, M. (2014). Comparison of Three Modes for a Crime Victimization Survey, Journal of Survey Statistics and Methodology 2 (4): 459-483 doi:10.1093/jssam/smu018

[6] Sarndal C-E, Swensson B, Wretman J (1991) Model Assisted Survey Sampling. Springer