UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/AC.6/2008/4
29 February 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population and Housing Censuses

Eleventh Meeting
Geneva, 13-15 May 2008
Item 2 of the provisional agenda

CENSUS QUALITY ASSURANCE AND EVALUATION

Sample results expected accuracy in the Italian population and housing census

Note by the National Institute of Statistics (Istat), Italy

Summary

Istat is considering the use of sampling techniques (based on long form and short form) among many others innovations under evaluation for the next census. In large municipalities (population over 10,000 inhabitants) detailed information would be collected only on a sample of the population (using the long form) whole for the rest of the population only limited information, consisting mainly in demographic variables, would be collected (using the short form). In smaller municipalities, a traditional approach is planned where the long form is submitted to the whole population.

Tests and studies were conducted in order to evaluate the efficiency of sampling estimates. Some preliminary results of these tests are presented in this paper. These results seem to encourage the use of sampling techniques in order to adopt a short/long form data collection strategy for the 2011 Italian population census.

I. INTRODUCTION

1. The Italian National Institute of Statistics (Istat) has recently started evaluating census innovations: to improve the efficiency of the surveys operations and reduce the statistical burden. This work regards the planning of sampling strategies for achieving information on a subset of census variables for the 2011 Population Census in Italy.

2. The main strategy considers a partition of the overall set of census variables into two subsets, the first one containing the demographic census variables and the second one including the remaining variables (educational level, occupational status, commuting). Short and long form versions of the questionnaire would therefore be considered whether to account merely for the first set of variables or for the whole set, respectively. For municipalities with more than 10,000 inhabitants a sample of households will be surveyed by means of a long form collecting the information by a short form on the remaining households. In smaller municipalities every population unit will be instead surveyed through the long forms.

3. As far as the sampling strategy is concerned[1], Istat has to consider carefully many difficulties such as: drawing samples from population lists, defining samples for some suburban or extra-urban areas which considerably increased or decreased their population between two consecutive censuses and defining census domains by aggregating census areas, for which the estimates have to be produced.

4. The main constraint for the definition of the sampling strategy[2] is the precision of the sampling estimates for different territorial levels: the wider the reference territorial is, the greater should be the precision of the estimates for long form variables and their cross-classifications with other variables either belonging to the same class or to the demographic (not sampled) one.

II. SAMPLING STRATEGIES

5. Two different studies were conducted. First of all, the efficiency of Simple Random Sampling of HOUseholds from Administrative Register (SRSHOU) managed by municipalities was studied. Different sampling ratios (10 per cent, 15 per cent, 20 per cent, 33 per cent) were considered in order to evaluate the improvement of the estimates precision for increasing sampling fractions. In the second instance, an area frame sampling based on a Simple Random Sampling of ENumeration Areas (SRSENA) was considered by sampling approximately one third of the population. According to this approach a complete data collection of the households

---

[1] Carbonetti G., De Vitiis C. (2007) "*Efficienza di stime campionarie relative ad un sottoinsieme di variabili di censimento*", Italian Statistical Conference: *"Censimenti generali 2010-2011. Criticità e innovazioni"*. CNR, Rome (Italy) November 2007.

[2] Särndal C.E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag,. New-York.

dwelling in the selected enumeration areas (from Digital Georeferenced Database) was considered.

6. For both cases, the sampling strategies were compared each other through Monte Carlo sampling replications carried out on 2001 census data, in order to assess their properties in terms of the size of the coefficient of variation.

7. The main features of the sampling design are sketched in the following:

(a) domains: "areas" referred to sub-municipal areas[3];
(b) target variables: "variables" related to cross-classification of educational level, employment status, commuting with demographic variables;
(c) sampling units: "households" or "enumeration area" depending on the adopted strategy;
(d) estimator: "calibrated estimators"[4] use final weights that were properly modified[5] in order to make the sample more representative.

III. SIMULATION STUDY

8. The simulation study was carried out on the 2001 population census data. A set of 40 municipalities with different population size and from different regions of Italy were considered in order to allow for the strong differences among the Italian municipalities (table A).

Table A: Distribution of municipalities by geographical area and demographic size.

Geographical area	Classes of population size (a)			Total
	10,000-20,000	20,000-100,000	more than 100,000	
North	4	6	6	16
Center	2	3	3	8
South	4	6	6	16
Total	10	15	15	40

(a) It has been considered the legal (official) population date referred to the 2001 Census of Population.

[3] Astorri P., Bianchi G., Di Pede F., Esposito N., Patruno E., Reale A., Ronchi I., Talice S. (2007) "*Metodi di determinazione delle aree di censimento a livello sub comunale*", XXVIII Italian Conference on Regional Sciences, Bolzano (Italy) September 2007.

[4] Deville J.C., Särndal, C.E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, vol. 87, pp. 367-382.

[5] Software Genesees v3.0 developed in Istat has been used in order to compute final weights in the calibration process.

9. The attention focused on cross-classification of educational level, employment status, commuting and gender for a total amount of 90 simple estimation cells.

10. Calibration constraints were defined by cross-classifying gender by age, and gender by civil status for a total amount of 40 reference frequency.

11. A computational algorithm consisting in the following steps was implemented by SAS code for each municipality and for each alternative sampling design:

1) selection of a sample (of households or enumeration areas);
2) computation of final weights;
3) estimation of the relative frequencies *p* for each dichotomous target cell;
4) iteration of steps 1), 2) and 3) for a fixed number of replications (1,000 sampling replications were chosen for practical reasons);
5) computation for each dichotomous estimation cell X, of the mean p_x and standard error $\sigma(\hat{p}_x)$ of the estimates from the simulated sampling distribution.

12. In order to compare the sampling strategies the coefficient of variation evaluation criterion $cv(\hat{p}_x) = \frac{\sigma(\hat{p}_x)}{p_x} \cdot 100$ was considered. In fact it represents an accuracy measurement of the sampling strategies in order to estimate the unknown value of *p* for each X.

13. The distribution of the empirical CV's for all the 90 target cells was determined. After having classified the target cells depending on their value *p* the distribution of the CV's related to the variable in the same group was studied.

14. In table B it is described the amount of units involved in the simulation study.

Table B: Size of the simulation in terms of areas, enumeration areas, households and individuals involved.

	Sampled Unites	Universe	%
Areas	497	3,347(*)	14.85%
Enumeration areas	30'890	382'534	8.08%
Households	2'243'511	21'810'676	10.29%
Individuals	5'537'582	56'594'021	9.78%

(*) Estimated numbers

IV SOME RESULTS

15. In table 1 the median CV per cent among the areas is showed for different classes of estimated *p* and for each of four tested sampling ratio. In order to take the simulation study manageable, each sampling ratio was tested on 10 municipalities, given that 5 of them were common for all the tests and the others were chosen in the same class of population. When a sampling ratio of 33 per cent is adopted, percentages *p* between 0,5 per cent and 1 per cent can be estimated with a median CV per cent of about 16 per cent. The median CV per cent increases twofold when the sampling ratio is 10 per cent. This behaviour remains similar for all the classes of *p*. In general, the tested sampling ratios show a fairly constant gain in term of CV decrease for almost all the classes of *p*.

Table 1: Distribution of median *cv* for classes of *p*. Comparison of 4 different sampling ratios in the SRSHOU design.

Classes of *p*	sampling ratio= 10%	sampling ratio= 15%	sampling ratio= 20%	sampling ratio= 33%
	170 areas	140 areas	111 areas	204 areas
< 0.05%	220.51	157.20	142.00	98.21
0.05% ├0.1%	111.48	87.22	74.20	51.14
0.1% ├0.25%	75.57	59.83	49.97	34.76
0.25% ├0.5%	50.70	39.92	33.97	23.44
0.5% ├1%	35.54	28.10	23.74	16.56
1% ├2.5%	23.62	18.56	15.33	10.68
2.5% ├5%	15.50	12.29	10.09	7.04
5% ├10%	10.46	8.26	6.93	4.82
10% ├15%	7.06	5.40	4.40	3.13
15% ├20%	5.57	4.27	3.54	2.42
20% ├30%	4.50	3.48	2.84	1.93
≥ 30%	3.20	2.42	1.94	1.34

16. Table 2 gives the same results already showed in table 1, with a further classification of the areas in three classes of population (less than 10,000; between 10,001 and 12,000; more than 12,000 inhabitants). It appears quite clearly that for areas larger than 12,000 inhabitants a remarkable decrease of the CV size is obtained.

Table 2: Distribution of median *cv* for classes of *p* and three classes of area (according to population size). Comparison of 4 different sampling ratios in the SRSHOU design.

Classes of *p*	Population by area (thousands)	sampling ratio= 10%	sampling ratio= 15%	sampling ratio= 20%	sampling ratio= 33%
< 0.05%	< 10	233.15	177.67	149.83	102.20
	10 ├12	225.61	145.74	143.95	88.14
	≥ 12	184.05	141.61	117.45	82.77
0.05% ├0.1%	<10	136.29	109.01	88.57	61.14
	10 ├12	107.97	85.49	72.90	50.61
	≥ 12	100.88	79.37	64.90	46.30
0.1% ├0.25%	<10	90.00	71.27	59.48	40.20
	10 ├12	76.23	60.03	50.45	34.41
	≥ 12	66.65	53.04	43.51	30.58
0.25% ├0.5%	< 10	60.88	47.23	40.24	27.29
	10 ├12	50.87	39.73	34.10	23.40
	≥ 12	45.22	35.68	29.93	21.01
0.5% ├1%	< 10	43.11	33.50	28.97	19.53
	10 ├12	35.08	27.46	22.99	16.48
	≥ 12	31.25	24.95	20.97	14.85
1% ├2.5%	< 10	28.85	21.82	18.11	12.42
	10 ├12	23.37	18.00	14.92	10.62
	≥ 12	21.10	16.46	13.45	9.48
2.5% ├5%	< 10	19.12	14.68	12.22	8.25
	10 ├12	15.58	12.36	9.89	7.08
	≥ 12	14.00	10.98	9.06	6.35
5% ├10%	< 10	12.70	9.77	8.14	5.65
	10 ├12	10.36	8.01	6.90	4.82
	≥ 12	9.37	7.43	6.21	4.36
10% ├15%	< 10	8.78	6.44	5.22	3.67
	10 ├12	7.00	5.46	4,41	3.13
	≥ 12	6.29	4.79	3.89	2.83
15% ├20%	< 10	6.79	5.03	4.31	2.85
	10 ├12	5.47	4.18	3.41	2.31
	≥ 12	4.99	3.80	3.09	2.17
20% ├30%	< 10	5.46	4.16	3.44	2.27
	10 ├12	4.57	3.42	2.94	2.01
	≥ 12	4.05	3.20	2.59	1.77
≥ 30%	<10	3.85	2.84	2.32	1.56
	10 ├12	3.14	2.30	1.88	1.32
	≥ 12	2.83	2.17	1.74	1.18

17. The median CVs are then described in table 3 with respect to three specific municipalities located in different Italian geographical areas (Bologna in the North, Perugia in the Centre and Trapani in the South). Even considered that further analyses should be conducted in order to test

for eventual geographical differences of the estimated CV, does not appear any evidence of large differences in terms of CVs with respect of these municipalities.

Table 3: Distribution of median *cv* for classes of *p*. Comparison of 4 different sampling ratios (s.r.) in the SRSHOU design. Municipalities of Bologna, Perugia and Trapani.

Classes of p	Bologna (32 areas)				Perugia (10 areas)				Trapani (5 areas)			
	s. r.= 10%	s. r.= 15%	s. r.= 20%	s. r.= 33%	s. r.= 10%	s. r.= 15%	s. r.= 20%	s. r.= 33%	s. r.= 10%	s. r.= 15%	s. r.= 20%	s. r.= 33%
< 0.05%	206.09	164.99	137.07	96.52	199.94	164.00	137.12	94.68	223.25	173.66	144.99	102.00
0.05% ├0.1%	107.70	85.78	72.14	50.67	105.78	83.96	70.74	49.14	120.35	94.89	78.63	56.37
0.1% ├0.25%	76.59	60.19	50.74	35.00	76.12	61.17	52.04	36.44	68.70	51.88	44.30	31.03
0.25% ├0.5%	52.27	41.09	34.45	24.17	49.04	38.20	32.20	21.87	49.69	38.13	32.47	22.49
0.5% ├1%	36.50	28.93	24.35	16.85	34.32	27.24	22.77	15.97	32.52	25.37	20.83	14.86
1% ├2.5%	23.28	18.27	15.28	10.74	22.44	17.41	14.53	10.24	24.20	18.58	15.95	10.96
2.5% ├5%	15.47	12.03	9.98	7.07	14.86	11.48	9.66	6.76	15.57	12.13	10.15	7.05
5% ├10%	10.43	8.16	6.76	4.75	10.50	8.21	6.76	4.77	10.62	8.40	6.98	4.83
10% ├15%	6.83	5.28	4.38	3.09	6.54	5.11	4.22	2.97	6.68	5.17	4.44	2.96
15% ├20%	5.19	4.02	3.41	2.38	5.45	4.26	3.53	2.41	5.01	3.94	3.29	2.31
20% ├30%	4.23	3.26	2.78	1.92	4.20	3.26	2.74	1.92	4.63	3.67	3.03	2.15
≥ 30%	2.84	2.25	1.86	1.32	2.62	2.07	1.77	1.21	3.30	2.51	2.09	1.41

18. Table 4 replicates table 2, where the median CVs are compared with respect to different classes of *p*, classes of area populations and sampling ratios, considering in more details the municipality of Bologna. Given that Bologna was chosen because its 32 sampled areas allow to properly compute the whole cells of table 4, it does not appear any remarkable difference between the CVs reported in this table when compared against the table 2.

19. Table 5 describes the percentage distributions of the estimates *p* by classes of CV for each tested sampling ratio. When the sampling ratio increases from 10 per cent to 33 per cent it can be observed an uniform increase of the percentage distribution below a CV-threshold of 10 per cent compared with a corresponding decrease over that value.

Table 4: Distribution of median cv for classes of *p* and three classes of area (according to population size). Comparison of 4 different sampling ratios in the SRSHOU design. Municipality of Bologna.

Classes of p	Population by area (thousands)	sampling ratio= 10%	sampling ratio= 15%	sampling ratio= 20%	sampling ratio= 33%
< 0.05%	< 10	219.16	174.43	147.14	98.84
	10 ├12	220.94	172.48	144.22	99.28
	≥ 12	180.79	140.52	118.01	83.60
0.05% ├0.1%	<10	138.44	101.63	88.69	60.32
	10 ├12	106.08	83.85	70.45	49.64
	≥ 12	100.56	79.05	64.11	46.23
0.1% ├0.25%	<10	89.74	69.56	58.36	40.89
	10 ├12	76.74	60.76	51.22	35.53
	≥ 12	66.63	52.92	44.46	30.78
0.25% ├0.5%	< 10	58.78	46.06	39.06	27.05
	10 ├12	52.00	41.09	34.58	23.91
	≥ 12	45.87	36.52	30.67	21.71
0.5% ├1%	< 10	45.08	34.37	29.01	20.22
	10 ├12	35.41	28.45	23.52	16.34
	≥ 12	31.89	24.85	20.86	14.88
1% ├2.5%	< 10	27.27	21.54	18.01	12.53
	10 ├12	23.00	18.07	15.13	10.61
	≥ 12	20.74	16.39	13.90	9.63
2.5% ├5%	< 10	18.54	14.49	11.88	8.29
	10 ├12	15.42	12.04	9.88	7.05
	≥ 12	13.70	10.74	8.99	6.23
5% ├10%	< 10	12.12	9.35	7.76	5.44
	10 ├12	10.38	8.09	6.65	4.72
	≥ 12	9.43	7.42	6.10	4.29
10% ├15%	< 10	8.27	6.29	5.21	3.64
	10 ├12	6.93	5.33	4.46	3.19
	≥ 12	6.10	4.71	3.90	2.77
15% ├20%	< 10	6.38	4.98	4.18	2.88
	10 ├12	5.06	3.91	3.16	2.23
	≥ 12	4.54	3.59	3.02	2.13
20% ├30%	< 10	5.12	3.95	3.34	2.33
	10 ├12	4.33	3.34	2.83	1.98
	≥ 12	3.72	2.93	2.45	1.72
≥ 30%	<10	3.46	2.69	2.23	1.50
	10 ├12	2.75	2.14	1.80	1.27
	≥ 12	2.73	2.12	1.78	1.23

Table 5: Distribution of the estimates referred to areas larger than 12,000 inhabitants for classes of *cv*. Comparison of percentage frequencies for 4 different sampling ratios in the SRSHOU design.

Classes of *cv*	sampling ratio= 10%	sampling ratio= 15%	sampling ratio= 20%	sampling ratio= 33%
< 2%	0.57	2.69	6.39	13.14
2% ├5%	13.04	17.53	18.40	23.64
5% ├10%	16.18	18.02	26.28	28.64
10% ├20%	29.14	30.16	23.54	16.20
20% ├50%	25.09	19.71	16.75	13.32
50% ├100%	9.32	7.21	5.69	3.44
100% ├200%	4.40	3.65	2.00	1.61
≥ 200%	2.25	1.03	0.95	-

20. The table 6 makes a comparison between the SRSHOU and the SRSENA sampling schemes with respect to the sampling ratio of 33 per cent. The median CVs for the two sampling frame by various classes of *p*, for each of 4 different municipalities, were computed. Due to the cluster effect, the sampling designs on households achieve sampling errors uniformly smaller than those obtained for the area sampling. However, being the SRSENA design not so worse than the SRSHOU in terms of efficiency estimates, it could represent a solution to obtain estimates with an acceptable quality level when reliable household registers are not available.

Table 6: Distribution of median *cv* for classes of *p* for SRSHOU design and SRSENA design (both with sampilng ratio=33%). Comparison of 4 municipalitiies.

Classes of p	**Milano (111 areas)**		**Bologna (32 areas)**		**Padova (18 areas)**		**Livorno (13 areas)**	
	SRSHOU	SRSENA	SRSHOU	SRSENA	SRSHOU	SRSENA	SRSHOU	SRSENA
< 0.05%	97.78	94.12	96.52	94.31	99.65	98.34	102.21	101.61
0.05% ├0.1%	51.61	51.59	50.67	49.54	51.70	54.13	50.69	52.06
0.1% ├0.25%	34.67	34.92	35.00	35.20	35.37	36.03	35.08	35.67
0.25% ├0.5%	22.96	24.38	24.17	24.73	25.58	26.45	23.70	24.37
0.5% ├1%	16.86	18.71	16.85	18.32	16.95	17.81	17.16	18.72
1% ├2.5%	10.61	12.21	10.74	11.95	11.07	12.00	11.34	12.90
2.5% ├5%	7.02	8.53	7.07	8.25	7.35	8.48	7.17	9.00
5% ├10%	4.84	5.97	4.75	5.74	5.05	5.85	4.88	6.39
10% ├15%	3.17	4.41	3.09	4.09	3.19	4.37	3.06	4.82
15% ├20%	2.44	3.46	2.38	3.12	2.44	3.14	2.44	3.39
20% ├30%	1.89	2.61	1.92	2.48	2.08	2.73	2.05	2.88
≥ 30%	1.35	1.78	1.32	1.60	1.39	1.72	1.40	2.00

21. The estimates concerning the territorial domains larger than the area (municipality, Nuts3, Nuts2 and Nuts1 level) were evaluated as aggregation of areas.
For example, for a large domain *R* consisting in *k* areas, the *p* estimate is given by:
$\hat{p}_R(x) = \sum_{a \in R} w_a \hat{p}_a$ where $w_a = N_a / N_R$ represents the weight in term of population of the generic area *a* on the large domain *R*. It can be showed that, for a given level of percentage *p* (concerning both a single area and a more large domain) the formula $cv(\hat{p}_R) \cong \frac{1}{\sqrt{K}} cv(\hat{p}_a)$ hold with an

expected percentage reduction of the related CV $\text{red}\% \cong \left(1 - \frac{1}{\sqrt{K}}\right) x100$.

In case in which a part of the large domain *R* is totally surveyed, an expected reduction of the CV will be given by the formula:

$$\text{rid}\% \cong \left(1 - \frac{\gamma}{\sqrt{K}}\right) x100$$

where $\gamma = N_C / N$ represents the quote of population where a sampling survey is arranged.

V. CONCLUSIONS

22. Istat is considering the use of sampling techniques among many others innovations under exam for the next census. More precisely, only few information, mainly regarding demographic variables, would be thoroughly collected by means of a short form, whereas an extensive questionnaire (long form) containing the complete information would be filled in by a sample of respondents. Moreover it is planned to sample only among dwellings located into municipalities larger than 10,000 inhabitants which represent about 40 million of people. A more traditional approach where the long form is submitted to the whole population is instead considered for those living in municipalities smaller than 10,000 people that account for 18 million of people.

23. Tests and experiments were conducted in order to evaluate the efficiency of sampling estimates regarding frequencies of very different size. Monte Carlo simulations of the sample spaces were carried out for simple random sampling of households with different sampling ratios and for area frame sampling.

24. Early results seem to encourage the use of sampling techniques in order to adopt a short/long form data collection strategy for the 2011 Italian population census. The most accurate estimates were of course observed for the simple random sampling of households from administrative registers and for the largest sampling ratio. Since a trade-off exists between the sample dimension and the costs of the data collection, the final choice will depend on the balance between the financial budget and the minimum accuracy level required at the various territorial domains.

25. Since more accurate estimates are observed for the largest areas in terms of their population size, a suggestion for planning the sample design would be to define areas of about 15,000 people.

26. Thought the area frame sampling is less efficient than the SRSHOU the simulation results indicate it could be an alternative solution where reliable administrative registers are not available.

27. Further efforts will be devoted to study alternative approaches based on small area estimation techniques both for the smallest territorial levels and for rare populations.
