UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/GE.41/2009/10
18 August 2009

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Group of Experts on Population and Housing Censuses

Twelfth Meeting
Geneva, 28-30 October 2009
Item 4 of the provisional agenda

CENSUS QUALITY AND DISCLOSURE CONTROL

Accuracy evaluation of Nuts level 2 hypercubes with the adoption of a sampling strategy in the 2011 Italian Population Census

Note by the National Institute of Statistics, Italy

I. INTRODUCTION

1. Italian National Institute of Statistics (Istat) is considering using sampling techniques in order to adopt a short/long form strategy for the 2011 Italian Population Census. The choice is based on a simple random design for the selection of private household samples from population registers and the calibrated estimators.

2. Because the adoption of a sampling strategy causes the introduction of sampling errors, tests and studies have been conducted in order to evaluate the efficiency of sampling estimates and the accuracy of dissemination hypercubes.

3. The main constraint for the definition of the sampling strategy is the precision of the estimates for different territorial levels: the wider the territorial reference 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.

4. In particular, for a given territorial domain and a specific hypercube it is possible to determine both the percentage of cells where the absolute frequency could be estimated with a

low level of accuracy and the percentage of persons classified in those critical cells. This last indicator expresses the information estimated with a low level of accuracy; lower values of percentage of persons classified in critical cells indicate good quality of the data included in that hypercube.

5. The evaluations of the impact of sampling errors on the quality of the dissemination data have concerned some Eurostat hypercubes with different topics and referred to NUTS level 2.

6. The planning of the 2011 Italian Population Census has taken in account both the critical points of the last census and the possibility to introduce methodological innovations according to the international recommendations[1]. In order to improve the efficiency of the survey operations, to reduce the workload of the municipalities and to minimize the statistical burden for the people involved in the enumeration, many solutions have been taken into consideration. The most important are related to the use of population registers, to a mailing for the census forms and to a mixed mode of data collection mainly based on mail and web.

7. Since success will be assured only if high spontaneous response rates are achieved, to this purpose in the next census there will be adopted a sampling strategy that adopts the use of short and long forms. The strategy will consider simultaneously the use of the two different versions of the questionnaire. The short form helps to reduce the workload while the sample of long forms preserves the richness of the census information. In this way, sampling can be seen as a consequence of the innovations introduced in the planning.

8. Introducing sampling strategy in the next Census round implies not only savings and a smaller amount of data to be managed also provides an opportunity to improve the overall quality of data. It will be possible to set up and execute more checks on census forms and follows up in the field to reduce the non-sampling errors. Another advantage of the sampling strategy is improvements to the Census timeliness, which is a constraint since 2011 Census data have to be provided to Eurostat on 1 April 2014.

9. There is a relevant disadvantage given by the introduction of sampling error and for this it is important to evaluate the efficiency of sampling estimates and the accuracy of dissemination hypercubes (statistical tables given by cross-classification of census variables).

II. THE SAMPLING STRATEGY

10. The approach will consist in the simultaneous use of short and long forms: the short form collecting only demographic and housing variables; the long form collecting the overall set of census variables including educational level, occupational status and commuting. In this way, the demographic data will be collected for the whole population while the information related to the other variables will be surveyed on a sample of households (referring only to *private* households).

---

[1] UNECE (2006). *"Recommendations for the 2010 Censuses of Population and Housing. United Nations Economic Commission for Europe and Statistical Office of the European Communities"*. Conference of European Statisticians. ECE/CES/STAT/NONE/2006/4.

11. The sampling strategy will regard municipalities with population over 20,000 inhabitants; in municipalities smaller than 5,000 persons a traditional approach is planned by submitting the long form to the whole population. In municipalities with populations between 5,000 and 20,000 the adoption of the strategy will depend on the efficiency of the related estimates.

12. Following the need to adopt a very simple strategy, tests and studies were conducted in order to evaluate the efficiency of sampling estimates relating to different sampling designs and estimation methods (direct and indirect).

13. The results of the empirical studies have suggested adopting the following strategy: simple random sampling of households (SRSHOU) from population registers and calibrated estimators[2] which use final weights properly modified in order to make the sample more representative. Another suggestion is to plan the sample to have a good accuracy of estimates referred to Census Areas, sub-municipal domains with about 15,000 inhabitants, whenever it is possible.

14. An important issue related to the choice of the sampling ratio is that the larger is the sample more accurate are the estimates. The intention is to choose a sampling ratio of 33 per cent with the aim to preserve the richness of the census information as much as possible. In the case of severe budget constraints, the sampling ratio could be reduced considering a trade-off between needed financial savings and accuracy required at different territorial domains.

15. Regarding the choice of the estimation methods, calibrated estimators assure the coherence between the estimates and the demographic data collected on the whole population. Indirect methods based on small area estimation techniques could be adopted to produce more accurate estimates referred both to the smallest territorial levels and to rare populations. Early results from ongoing experiments[3] seem to be encouraging since, for absolute frequencies fewer than 150 units, a reduction of Coefficient of variation (CV) between 40 and 80 per cent is obtained.

III. ACCURACY OF ESTIMATES

16. A simulation study was carried out on the 2001 population census data[4]. A set of 40 municipalities with different population size and from different NUTS2 areas of Italy were considered to allow for the strong differences among Italian municipalities. In particular, the study considered a little more than 10 percent of households and a little less than 10 percent of persons enumerated in Italy in the last census.

---

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

[3] Borrelli F., Carbonetti G., De Felici L., Solari F. (2008) *"Metodologie di stima per piccole aree applicabili a variabili di censimento rilevabili tramite questionario long form"*. XXIX Italian Conference on Regional Sciences, Bari (Italy), September 2008.

[4] Carbonetti G., Fortini M. (2008) *"Sample results expected accuracy in the Italian population and housing census"*. Joint UNECE/Eurostat Meeting on Population and Housing Censuses. UN, Geneva (Switzerland), May 2008. ECE/CES/AC.6/2008/4.

17. About 90 cell counts resulting from census multiway tables were estimated through calibration methods. Cell counts were computed for each of the 497 tested sub-municipal areas drawn between 5,000 and 15,000 persons. The properties of sampling estimates were assessed by means of CV statistic computed on 1,000 sampling replications for each tested sampling rate. 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.

18. Table 1 shows some of the simulation results[5] referred to the SRSHOU. It reports, for each class of cell counts and for each tested sampling ratio, the average and the maximum CV estimated over the sub-municipal areas.

Table 1
Distribution of average and maximum CV per cent for classes of cell counts for the three tested sampling ratios (SRSHOU design)

Classes of cell count	sampling ratio = 10%		sampling ratio = 20%		sampling ratio = 33%	
	Average CV%	Max CV%	Average CV%	Max CV%	Average CV%	Max CV%
<10	**143.3**	191.8	**101.4**	123.7	**66.5**	95.8
10 ├30	**75.9**	85.1	**48.4**	54.6	**33.8**	38.5
30 ├50	**51.8**	57.1	**31.8**	37.1	**23.4**	25.6
50 ├100	**38.6**	41.3	**22.3**	28.4	**17.4**	19.1
100 ├250	**25.4**	28.5	**15.7**	19.6	**11.4**	12.8
250 ├500	**16.1**	18.3	**10.4**	12.5	**7.5**	8.1
500 ├1,000	**11.8**	12.8	**7.5**	8.2	**5.3**	5.9
1,000 ├2,500	**7.5**	8.9	**4.7**	5.9	**3.3**	3.9
2,500 ├5,000	**4.9**	5.4	**3.0**	3.6	**2.0**	2.5
5,000 ├10,000	**3.2**	3.8	**2.0**	2.5	**1.3**	1.9

19. For instance, in case of 10 per cent sampling ratio, the average CVs fall under 10 per cent when counts go over 1,000; when sampling ratio increases up to 33 per cent the threshold decreases to 250. Higher values of CV for small frequencies are expected. However, higher values of CV for small counts correspond to smaller differences in absolute terms. As expected, the most accurate estimates are obtained for largest sampling ratios. In general, the gain of efficiency measured as relative difference of CVs is about 33 – 38 per cent when the sampling ratio goes from 10 to 20 per cent and about 53-58 per cent when the sampling ratio increases from 10 to 33 per cent.

[5] Carbonetti G., Fortini M., Solari F. (2008) *"Innovations on methods and survey process for the 2011 Italian population census"*. European Conference on Quality in Official Statistics, Roma (Italy), 2008.

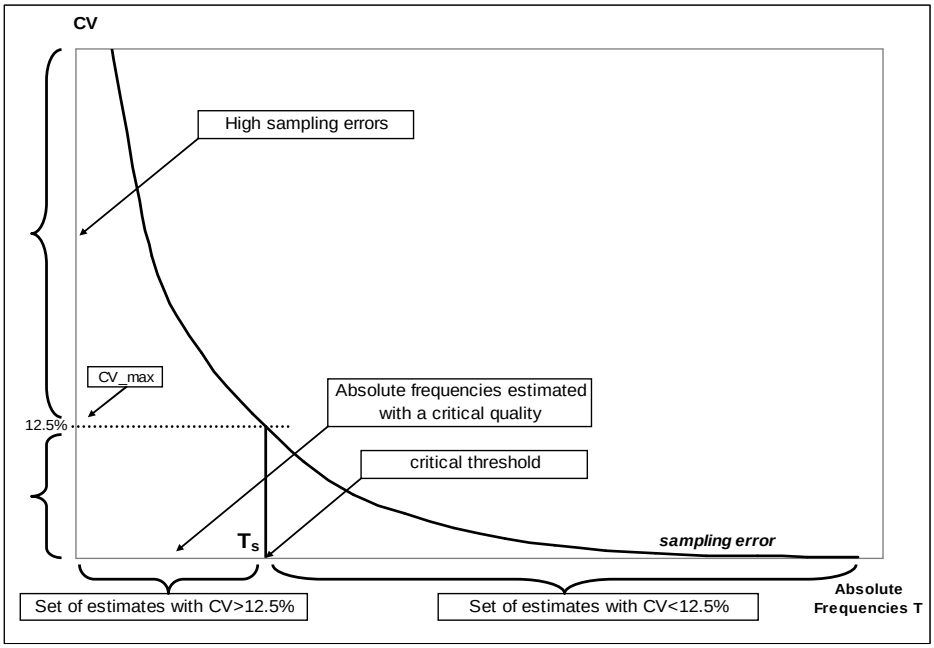
IV. IMPACT OF SAMPLING ERRORS ON DISSEMINATION HYPERCUBES

20. The main constraint for the definition of the sampling strategy is the precision of the estimates for different territorial levels: the wider is the territorial reference, 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.
In this section the potential impact of sampling strategy on the quality of the dissemination hypercubes is studied.

21. In correspondence of an acceptable value of CV the curve of sampling errors drawn by the simulation results[6] allows the setting of the absolute frequency *threshold* so every frequency greater than this is estimated with a CV smaller than the fixed value.

22. Graph 1 shows how it is possible to determine the sets of estimates with critical accuracy. For a fixed CV equal to 12.5 per cent, by means of the sampling errors curve, it can be determined the critical threshold under which all the absolute frequencies are estimated with a CV more than the fixed value. On the one hand these can be considered as critical cases resulting in high sampling error. On the other hand, frequencies larger than the threshold will have a smaller sampling error.

Graph
Identification of the sets of estimates with "critical accuracy" by means of the sampling errors curve and the critical threshold T_S related to a fixed CV



[6] Carbonetti G., Dardanelli S., Fiorello E., Mastroluca S., Verrascina M., (2008) "*Ipotesi di innovazione per il censimento della popolazione del 2011: una valutazione degli effetti su un possibile piano di diffusione*". XXIX Italian Conference on Regional Sciences, Bari (Italy), September 2008.

23. If estimates are referred to domains partially surveyed where not all the population is eligible for sampling, the errors curve reduces thus encouraging efficiency gains and reductions of the critical threshold.

24. In this way, for a fixed territorial domain and a dissemination hypercube it can be computed both the percentage of cells where the estimated absolute frequency is less than the observed critical threshold and the percentage of persons classified in those critical cells. In particular, the last indicator quantifies the amount of data estimated with a low level of accuracy; low percentages of persons classified in critical cells indicate good quality of the information referred to that hypercube. For instance, 10 per cent can be fixed as an acceptable value of the indicator.

V. EXPECTED QUALITY OF NUTS 2 HYPERCUBES

25. Evaluations[7] about the impact of sampling strategy on the quality of the dissemination data have concerned some hypercubes with different information details and referred to NUTS level 2. In this section, after some discussion about the choice of hypercubes included in the Eurostat dissemination programme, some results on the expected quality are shown.

A. Hypercubes considered in the study

26. From the draft version of dissemination programme to be approved for 2011, 8 hypercubes (Table 2) were selected containing topics from the 2001 Italian census but also, and especially, containing topics with breakdowns for our past dissemination close, in terms of number and information content, to breakdowns provided for the next census round and with topics definitions near to those foreseen at international level. For example, we excluded hypercubes containing the topic "ever resided abroad" that Italy will collect for the first time in 2011, but also topics and breakdowns related to households and families, in some cases broken down in Italy differently from that provided at international level for the next census.

27. The use of hypercubes with topics related to households, families, population by household status and family status would have implied a redefinition and new counting of related breakdowns, since the new definitions and breakdowns are, in some cases, different from those provided for the Italian dissemination and it would have lengthened the time of our studies. We selected hypercubes which considered only the characteristics of the population.

28. LAU2 hypercubes haven’t been considered since they include (demographic) topics we are planning to include in the short form, so information will be available for the entire population. We therefore concentrated our attention on topics required only at NUTS 2 level, topics we are planning to include in the long form and then on hypercubes for the geographical national, NUTS 1, NUTS 2 level. We have attempted to select hypercubes that were different from each other and covering the various topics proposed for the survey. A set of topics is

[7] Carbonetti G. (2009) *“Use of sampling strategy in the Italian population census and accuracy of estimates for different territorial domains”*. ITACOSM09 - First Italian Conference on Survey Methodology, Siena (Italy), June 2009.

common to all, concerning sex and age, even though the latter is in some hypercubes broken down by single age, in others by a more aggregated classification (M or S).

Table 2
Eurostat hypercubes considered in the study of the impact of sampling strategy on the dissemination tables

Hypercube number	Code	Name	Enumeration base	Topics (Breakdowns)
1st hypercube	H.B1.E0.R1	Single ages - "Current activity status"	Total population	• Sex (M) • Age (L) • Current activity status (M)
2nd hypercube	H.B1.E0.R2	Single ages - "Occupation"	Total population	• Sex (M) • Age (L) • Occupation (M)
3rd hypercube	H.B1.E0.R3	Single ages - "Industry"	Total population	• Sex (M) • Age (L) • Industry (branch of economic activity) (M)
4th hypercube	H.B1.E0.R4	Single ages - "Status in employment"	Employed persons	• Sex (M) • Age (L) • Status in employment (M)
5th hypercube	H.B1.E0.R5	Single ages - "Educational attainment"	Total population	• Sex (M) • Age (L) • Educational attainment (highest completed level) (M)
14th hypercube	H.B1.E1.R2	Employment (people where they live) - "Occupation"	Total population at their usual residence	• Sex (M) • Age (M) • Current activity status (L) • Occupation (M) • Educational attainment (highest completed level) (M)
15th hypercube	H.B1.E1.R3	Employment (people where they live) - "Industry"	Total population at their usual residence	• Sex (M) • Age (M) • Current activity status (M) • Industry (branch of economic activity) (M) • Educational attainment (highest completed level) (M)
16th hypercube	H.B1.E1.R4	Employment (people where they live) - "Occupation" by "Industry"	Total population at their usual residence	• Sex (M) • Age (S) • Occupation (M) • Industry (branch of economic activity) (M) • Educational attainment (highest completed level) (M)

29. A number of cells for each hypercube was originally calculated as product of the number of categories for each topic included in the hypercube A new size of the hypercubes was then introduced, the “relevant size”, calculated by excluding from the first the number of categories corresponding to the totals, the sub-totals and the categories "Not stated".

30. On the selected hypercubes, we calculated (Table 3) the “relevant size” multiplying the number of categories utilized in the Italian dissemination for the 2001 census and we obtained a number of potential cells. Then we excluded the "structural zeros", i.e. the number of cells that are certainly zero because they correspond to impossible outcomes. This led to the number of acceptable cells by excluding cases of inability to cross (e.g. Age=15 years and Educational Attainment=ISCED 5a).

Table 3
Number of potential cells and acceptable cells for each hypercube considered

code	Number of potential cells	Number of acceptable cells
H.B1.E0.R1	**1,212** *(2x101x6)*	**1,062**
H.B1.E0.R2	**2,020** *(2x101x10)*	**1,922**
H.B1.E0.R3	**3,434** *(2x101x17)*	**3,126**
H.B1.E0.R4	**1,212** *(2x101x6)*	**1,032**
H.B1.E0.R5	**1,414** *(2x101x7)*	**1,342**
H.B1.E1.R2	**23,520** *(2x21x8x10x7)*	**3,810**
H.B1.E1.R3	**29,988** *(2x21x6x17x7)*	**5,574**
H.B1.E1.R4	**30,940** *(2x13x10x17x7)*	**26,350**

31. It can be seen from the first 5 hypercubes that cross demographic topics with only one socio-economic topic are very manageable with a limited number of potential cells. The number is reduced further when we consider only the acceptable cells, those for which we could expect a frequency. When it comes to the second block of hypercubes, however, even if the age (no longer single years of age but five or ten years classes) is more aggregated, the number of crossed topics increases and the number of involved socio-economic topics increases, consequently the potential size of hypercubes increases greatly, even if the size diminishes substantially in the case of acceptable cells. The hypercube with the largest size is the last one that crosses, other than Sex and Age (ten years classes), Occupation, Industry and Educational Attainment.

B. Some results

32. Some results of the quality analysis related to the most complex hypercubes among those considered in the study and filled with data available from the 2001 Italian population census are presented below.

33. The first example is referred to the hypercube H.B1.E1.R3 which crosses Sex, Age, Current Activity Status, Industry and Educational Attainment. The number of *acceptable cells* is 5,574 (no structural zeroes are considered). Relating to this hypercube, Table 4 shows for each tested sampling ratio (adopting the simple random sampling of households from population registers) and for three Italian NUTS2 areas with different size (*Molise* is one of the smaller,

Marche has a medium size and *Sicilia* is one of the larger) the critical thresholds, the percentages of critical cells and the percentages of individuals in critical cells. It's easy to see gains of quality increasing the sampling ratio and higher quality for larger areas.

34. For instance, considering *Sicilia* area, in case of 33 per cent sampling ratio, the related critical threshold is 100, 59.4 per cent of cells have absolute frequencies lower than 100 but in these cells only 1.0 per cent of eligible people is classified.

Table 4

Hypercube H.B1.E1.R3. Quality indicators (CV=12.5 per cent) related to three NUTS2 areas of Italy: Molise, Marche and Sicilia

	Molise			**Marche**			**Sicilia**		
Sampling ratio	Critical threshold Ts	% of critical cells	% of individuals in critical cells	Critical threshold Ts	% of critical cells	% of individuals in critical cells	Critical threshold Ts	% of critical cells	% of individuals in critical cells
10%	**100**	**79.2**	**10.7**	**250**	**78.8**	**6.9**	**500**	**75.9**	**4.2**
20%	**50**	**71.0**	**5.8**	**100**	**68.4**	**3.0**	**250**	**68.7**	**2.1**
33%	**30**	**63.6**	**3.4**	**50**	**59.8**	**1.5**	**100**	**59.4**	**1.0**

35. Table 5 shows the results related to a more complex hypercube (H.B1.E1.R4) which crosses Sex, Age, Occupation, Industry and Educational Attainment. The number of *acceptable cells* is 26,350.

Table 5

Hypercube H.B1.E1.R4. Quality indicators (CV=12.5 per cent) related to three NUTS2 areas of Italy: Molise, Marche and Sicilia

	Molise			**Marche**			**Sicilia**		
Sampling ratio	Critical threshold Ts	% of critical cells	% of individuals in critical cells	Critical threshold Ts	% of critical cells	% of individuals in critical cells	Critical threshold Ts	% of critical cells	% of individuals in critical cells
10%	**100**	**91.9**	**14.9**	**250**	**91.1**	**11.2**	**500**	**91.8**	**7.3**
20%	**50**	**86.5**	**9.3**	**100**	**84.4**	**6.0**	**250**	**87.6**	**4.5**
33%	**30**	**81.3**	**6.5**	**50**	**77.1**	**3.4**	**100**	**79.4**	**2.2**

36. The considerations are similar to the previous case. The difference is a lower quality since this hypercube has a bigger number of cells. However, there is a slight reduction of accuracy and the overall quality of the statistical table remains acceptable.

37. In Table 6 a summary of the results about the percentage of individuals in critical cells has illustrated, related to all the 20 Italian NUTS2 areas, for all the hypercubes considered in the study and for the three tested sampling ratios.

Table 6
Distribution of Italian NUTS2 areas by the percentage of individuals classified in critical cells (CV=12.5 per cent) for some Eurostat hypercubes

Number of NUTS2 Areas	s.r. = 33%		s.r. = 20%			s.r. = 10%				
Eurostat Hypercubes (acceptable cells)	**Percentage of individuals in critical cells**									
	<5%	5-10%	**<5%**	5-10%	10-15%	**<5%**	5-10%	10-15%	15-20%	>20%
H.B1.E0.R1 (1,062)	**20**	0	**20**	0	0	**19**	1	0	0	0
H.B1.E0.R2 (1,922)	**20**	0	**20**	0	0	**15**	4	1	0	0
H.B1.E0.R3 (3,126)	**20**	0	**17**	3	0	**11**	4	4	1	0
H.B1.E0.R4 (1,032)	**20**	0	**16**	4	0	**7**	8	5	0	0
H.B1.E0.R5 (1,342)	**20**	0	**20**	0	0	**15**	5	0	0	0
H.B1.E1.R2 (3,810)	**20**	0	**18**	2	0	**12**	6	2	0	0
H.B1.E1.R3 (5,574)	**20**	0	**17**	3	0	**10**	5	5	0	0
H.B1.E1.R4 (26,350)	**15**	5	**8**	10	2	**1**	9	6	3	1

38. It is possible to observe that for 33 per cent sampling ratio in seven of the eight hypercubes considered, the percentage of individuals classified in critical cells is less than 5 per cent for all the areas, while for the largest hypercube H.B1.E1.R4 the indicator is less than 5 per cent in 15 areas and the value is between 5 and 10 per cent in the remaining 5 areas.

39. Sampling with a ratio of 20 per cent implies a slight loss of quality, in fact the percentages of individuals in critical cells related to the different areas remain less than 10 per cent except for the last hypercube where in two areas (the smallest) the indicators is between 10 and 15 per cent.

40. The quality could be acceptable if a sampling strategy of 10 per cent sampling ratio is adopted. In this case, for a large number of areas, quality indicators less than 10 per cent for the different hypercubes can be observed; some problems could occur for more complex hypercubes and for smallest areas.

VI. CONCLUSIONS

41. The results strengthen the introduction of sampling techniques for the 2011 Italian Population Census. The strategy to adopt a simple random sampling of households from

population registers and calibrated estimators produces accurate estimates and allow the reduction of non-sampling errors. Small area estimators can improve the accuracy of estimates referred to small size domains or to rare populations.

42. Italian Census definitions and classification are fully compliant with Eurostat requirements so allowing the production of all the requested hypercubes. Nevertheless, the adoption of a sampling strategy will cause a certain degree of variability into some of the delivered Census pictures. More precisely, all the hypercubes required at LAU2 level will be assured by short forms and consequently free of sampling variability, whereas the hypercubes for the national, NUTS 1, NUTS 2 level, that include topics which will be subjected to sampling through long forms, will be affected by sampling variability and consequently will be provided with sampling variance estimates.

43. Experiments to evaluate the impact of sampling strategy on the quality of some hypercubes referred to NUTS level 2 highlighted the effectiveness of the new approach for the Italian Population Census. The main result is that hypercubes with more than 20,000 cells can be estimated with a low percentage of units in critical cells even for lower sampling ratio. Some problems are observed just for smaller areas (for example: Val d'Aosta).

44. Since a trade-off between the sample size and the costs of the data collection exists, the final choice will depend on the balance between the financial budget and the minimum accuracy level required at the various territorial domains.

45. Another item related is the opportunity to produce data with a wider level of information at lower territorial reference. Different approaches suggest to use *breakdowns* with a reduced number of categories or to adopt alternative indirect estimators. In fact, small area methods seem to improve the accuracy of the estimates for smallest territorial levels in which the sample could not be representative and for very small cell counts in the greater domains.

*****