



**Economic and Social
Council**

Distr.
GENERAL

ECE/CES/GE.41/2009/9
20 August 2009

Original: ENGLISH ONLY

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

Confidentiality issues in the EU Population and Housing Censuses of 2011

Note by the Statistical Office of the European Communities and by Statistics Netherlands

I. INTRODUCTION

1. The Council of the European Union and the European Parliament have adopted in 2008 the Regulation (EC) No 763/2008¹ on population and housing censuses, acknowledging the Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. This Regulation states that "Member States shall take all measures necessary to meet the requirements of data protection. The Member States' own data protection provisions shall not be affected by this Regulation." The transmission of data subject to statistical confidentiality is governed by specific EU regulations, ensuring the physical and logical protection of confidential data and that no unlawful disclosure or non-statistical use occurs when Community statistics are produced and disseminated. In particular, the new "EU Statistical

¹ Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses (Text with EEA relevance). OJ L 218, 13.8.2008, p.14.

Law² devotes an entire chapter to statistical confidentiality. In other words, what is considered confidential at national level, it remains confidential once transmitted to Eurostat and, if a country wants to transmit confidential data, this has to be done in accordance with the EU regulations in force.

2. The programme for data dissemination that Eurostat is implementing for the census round 2011 is based on an innovative approach. The basic data input is in the form of hypercubes, which are multidimensional tables with several dimensions. The size of these hypercubes has an impact on confidentiality issues: while for a set of predefined common bi- or tri-dimensional tables the disclosure control for census data could be considered (relatively) feasible to implement, such control becomes a real challenge as more dimensions are added.

3. In order to minimise the risk of disclosure, the size of the hypercubes has been kept as small as possible; nevertheless, it remains very likely that methods for confidentiality will have to be applied. If these methods differ from one country to another, the comparability of the data may be affected. Moreover, for the users it would be easier to understand the constraints deriving from the application of one single method for all the countries rather than the consequences of several national methods for confidentiality. It has thus been considered worthy to explore the possibility of a common approach at EU level for disclosure control of census data.

4. In order to assess the state of the art on disclosure control methods for census data in Europe, a small survey was launched in May 2008. The questionnaire was sent to 45 countries and explored the following dimensions:

- (a) whether a specific regulation on the confidentiality of census data existed in the country;
- (b) the implementation of the national confidentiality requirements;
- (c) which institution was responsible for the definition/implementation of the disclosure control;
- (d) the availability of a methodology for confidentiality of the census 2011 data;
- (e) if there was support for an action at EU level aiming to identify a common methodology.

5. The full set of answers from 32 countries is available in the Interest Group “Population” in CIRCA³.

6. Normally, there are not specific provisions for the confidentiality of census data, as this is

² Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics (Text with relevance for the EEA and for Switzerland). OJ L 87, 31.3.2009, p. 164.

³ <http://circa.europa.eu/Public/irc/dsis/population> (access restricted to registered members).

covered by the restrictions applied to all statistical data. If mention is made in the census law, this usually makes reference to the more general national regulations on the subject. The practical implementation of these general provisions is in general a task of the national statistical offices. However, not many of them have defined a methodology for disclosure control for the next census. There was therefore significant support for an action by Eurostat to analyse the feasibility of a common methodology for disclosure control of the census data.

II. THE CHALLENGES OF AN EU HARMONISED APPROACH FOR DISCLOSURE CONTROL

7. On the basis of the survey results, Eurostat has set up a Task Force on "EU Methodology for Census Data Disclosure Control" (CENSDC). The task force consists of experts in the field of disclosure control from Germany, Estonia, Italy, the Netherlands, Portugal and the United Kingdom. Its specific objective is to identify and resolve areas of difficulty relating to the confidentiality data treatment of population and housing census data and adopting or developing a harmonised methodology which respects the national regulations. The Task Force CENSDC has met two times, and it is expected to present the results of its work at the Eurostat Working Group on Demography and Census to be held in the first half of 2010.

8. The Task Force CENSDC has to deal with several conceptual challenges. First of all, the relatively high number of dimensions of the hypercubes makes complicated the application of the standard methods of disclosure control. Secondly, given the approach of the Eurostat Census Hub, it needs to be decided where these controls should take place, in the national databases or "on the fly". Third, consistency of the tables' results should be assured between hypercubes and between extractions. Fourth, as the expertise and tools for disclosure control available in each Member State can be different, a common approach should be as simple as possible to implement. Fifth, the method should be easy to understand for the common user. Sixth, as one of the added values of a census is the availability of detailed information, the loss of data should be minimised. Last, but certainly not the least, each country has its own regulation on confidentiality that has to be respected.

9. For the next census round, there is an ambitious programme of census data dissemination. More information than in past census rounds from all the EU Member States (and other countries willing to be part of the census dissemination programme) will be at the disposal of the users by means of a single interface, the Eurostat Census Hub. This system transmits any user's query to the national databases, retrieves the information, and displays it all together. From the confidentiality point of view, the other side of the coin is that, given the high number of dimensions and the freedom of the user to build the tables of interest (including repeated queries), the risk associated with standard methods for disclosure control need to be carefully assessed.

10. One basic decision is whether the national data on which the extraction is made have to be "pre-cleaned" for confidentiality, or if the disclosure control can be made "on the fly", just before the data are displayed to the user. The latter option would be justified by the fact that the number of dimensions displayed to the user would be much less than the total number of dimensions available in the related hypercube, thus reducing the risk of disclosure. On the other

side, such an approach would mean that confidential data are somehow being transferred from the national database (with all the implications from the IT security point of view), and that the risks connected to multiple queries by potential intruders, are increased. The levels at which the data can be treated for confidentiality are three: microdata, hypercube or extraction.

11. No matter the level at which the disclosure control is implemented, it is considered important that the results disseminated to the users are consistent between selected tables and between extractions. The large use made of census data and the fact they are queried for a long time periods require that the users will not be confronted with different results depending on the time of the extraction or on the hypercube of reference. Although some confidentiality methods could be very effective, it will need to be assessed whether they generate undesired consequences in terms of data comparability.

12. Methods and tools for disclosure control have reached a certain level of complexity. Several European projects have been devoted to this domain, among which are the CASC project (2000-2003, see <http://neon.vb.cbs.nl/casc/>), the CENEX project (2006) and the ESS net project (2008-2009), and UNECE and Eurostat organise regular Work Sessions on Statistical Data Confidentiality. However, it can not be assumed that the same level of expertise is available in all statistical offices. If the statistical disclosure control has to be applied at national level based on a common approach, then the harmonisation of the confidentiality methods has to take into account additional requisites such as the ease of implementation and the availability of tools (possibly without excessive costs). In periods of scarcity of resources, countries can not be asked to sustain significant additional expenses.

13. Besides the above challenges relevant to a harmonised approach, there are also other elements to be considered as part of the exercise, irrespective of whether the method is applied to all countries or is country-specific. Whatever the disclosure control applied, the user should be informed of its characteristics and consequences on the data. The easier a method, the easier for the common user to understand the implications (and likely the easier to communicate this information). For the sake of transparency and overall data quality, this aspect should not be neglected.

14. One of the major features of the census is that it provides information at small geographical level and/or for small groups of persons, and is sometimes the only available source. Such richness should be preserved as much as possible while respecting the need to ensure data confidentiality. Although the extensive application of rigorous methods of statistical disclosure control may prevent confidentiality breaches, at the same time it can significantly reduce the availability of information. It is in the interest of the users to minimise the impact of disclosure control methods on data availability. The filters for confidentiality should be applied to a reasonable extent, bearing in mind the related unavoidable loss of (detailed) information.

15. Finding a satisfactory solution for each of the above-listed challenges risks becoming a hopeless task, and adding the national constraints makes things even more complicated. However, despite of the large number of national requirements on data confidentiality, it still makes sense to look for a common solution because the national regulations are often only setting general principles, leaving in many cases the practical implementation (and related methodological choices) to the national statistical offices. If these statistical offices agree on a harmonised approach, there is no infringement of the national provisions, as these bodies are the

technical responsible of the appropriate disclosure control to the national data. The wide support expressed by the statistical offices of a joint action at EU level on confidentiality of census data can be seen as an expression of the need of exchange of experiences and/or assistance on technical issues. Thus the Task Force CENSDC can certainly play an important role.

III. TECHNICAL ASPECTS OF AN EU METHODOLOGY

16. The Regulation (EC) No 763/2008 is output oriented, i.e. it is open to the use of different data sources, but requires the respect of the essential features of population and housing censuses, the use of harmonized definitions, technical specifications, topics and breakdowns. The Census regulation foresees unified reporting years (the first being 2011), a common EU dissemination programme, technical standards for the data transmission and the establishment of quality reports for European purposes. Concerning the statistical confidentiality, the following aspects are of particular importance:

- (a) Article 4 (2) foresees that the "Member States shall take all measures necessary to meet the requirements of data protection. The Member States' own data protection provisions shall not be affected by this regulation." That means that the protection of census data comes under the responsibility of the Member States, and has to be done at their level rather than by the Commission. Article 4 (2) provides further that the European Commission is not entitled to issue legislation on the disclosure protection of census data on the basis of the Census regulation. However, Article 6 (4) stipulates "The Commission (Eurostat), in cooperation with the competent authorities of the Member States, shall provide methodological recommendations designed to ensure the quality of the data and metadata produced, acknowledging, in particular, the Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing". Consideration 3 to the regulation explains that "in view of methodological and technological developments, best practices should be identified and the enhancement of the data sources and methodologies used for censuses in the Member States should be fostered";
- (b) Article 5 (2) of the Census regulation foresees that the "Member States shall provide the Commission (Eurostat) with final, validated and aggregated data (...)" . This excludes the transmission of microdata. Although aggregated data are not necessarily protected against disclosure of sensitive data, the spirit of Article 5 (2) implies that no confidential data shall be transmitted to Eurostat;
- (c) Considerations 5 and 7 stipulate that the Statistical Law, respectively the European Statistics Code of Practice, constitute the framework for the Census regulation, both containing provisions on statistical confidentiality;
- (d) Consideration 6 recalls the regulations on the transmission of data subject to statistical confidentiality. This means that, if Member States transmit data they feel is subject to statistical confidentiality, Eurostat has to ensure the physical and logical protection and that no unlawful disclosure or non-statistical use occurs when Community statistics are produced and disseminated. However, the census

regulation does not foresee the transmission of confidential census data from the Member States to Eurostat. In a broad sense, Consideration 6 reminds indirectly that everything must be done to avoid inadvertent disclosure of any confidential data.

17. In principle, the Task Force CENSDC follows up two major branches of thinking:

- (a) A recommendation on the pre-tabulation noise protection at the microdata level. This seems to have advantages in the context of both a national and a European dissemination of 2011 Census results. However, this protection can only be done at the NSI (National Statistical Institute) level and Eurostat would have no means of even verifying that such a protection has been executed;
- (b) A recommendation on post-tabulation protection (hypercube level). For the time being, the work is split into "cell suppression" and "post-tabulation noise protection". A simple solution would be to check which cells cannot be published (the so-called primary suppressions) and protect in addition a number of cells to prevent recalculations from the margins (the so-called secondary suppressions). However, the Task Force might also consider whether synergies between these two methodologies are achievable — given that the objective is limited to preventing the identification of individuals, i.e. to prevent certainty about cell values in frequency tables. This prevention action should ideally take place with minimum information loss.

18. As the real data of the 2011 Censuses are of course not yet ready, test hypercubes of a few countries are being used by the Task Force CENSDC. In the two examples below two dimensional subtables of higher dimensional Italian test hypercubes are shown⁴. For obvious reasons variable names are replaced by names like var2, var3 and so on. In the first example (Figure 1) some cells have to be protected, but the confidentiality problems seem to be solvable. If even more cells contain no observations a proper protection strategy will probably lead to many cells without a real frequency score.

19. In the second example (Figure 2) even all interior cells are zeroes or have to be suppressed. It is clear that this problem will occur more often for countries that make use of survey information (instead of complete enumeration or register information) for some of the 2011 Census variables.

20. What is the lesson that can be learnt from these pictures? For all Census hypercubes it will be important to verify whether most cells can be published. It does not make sense to produce and publish hypercubes with only or mainly zeroes and suppressions. Although confidentiality rules may differ between countries, this problem plays a role in all European countries. Larger countries tend to have more regions and thus face the same problem at a more detailed level as the smaller

⁴ These pictures were produced by Federal Statistical Office, Germany.

Figure 1

Table: Region x var1 x var2 x var3 x var4 | frequenza

var2	var3	var6	tot	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	▲
tot		tot	tot	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
tot	46.191	2.292	2.064	1.655	1.821	4.785	7.895	7.234	5.237	3.160	2.089	1.622	1.297	1.142	880	723	7		
8101	114	6	6	3	3	9	14	26	20	5	3	4	4	3	4	2			
8102	213	9	19	12	11	24	43	23	31	9	4	2	3	3	7	1			
8103	86	1	4	2	1	11	20	20	11	4	3	1	2	2	1	2			
8104	228	5	9	10	21	25	30	30	25	9	11	3	6	7	6	4			
8105	131	9	7	5	7	18	24	17	13	5	7	4	3	5	1	4			
8106	402	12	17	9	9	49	65	55	37	38	22	13	18	20	4	4			
8107	86	2	3	4	6	15	14	10	9	6	2	5	1	1	3				
8108	88	-	1	3	3	17	14	12	11	9	4	5	2	4	1	1			
8109	72	6	2	2	3	8	17	12	8	3	2	1	-	-	4	1			
8110	47	3	1	3	4	7	5	7	10	4	-	1	-	-	1	-			
8111	238	13	12	10	12	15	38	23	33	20	14	12	12	9	3	2			
8112	601	33	25	29	25	71	107	97	56	39	29	18	19	7	9	6			
8113	51	3	2	1	1	8	8	10	3	1	2	4	4	-	1	-			
8114	160	1	5	5	8	15	24	22	12	13	8	7	5	4	6	4			
8115	1.100	75	45	50	38	102	202	194	146	77	49	30	20	12	20	6			
8116	402	19	23	24	23	34	87	56	37	24	20	18	7	6	7	1			
8117	394	24	19	20	14	43	72	61	43	26	11	6	12	11	10	8			
8118	401	22	16	18	16	41	58	63	47	26	14	15	11	11	11	7			
8119	237	11	8	11	12	32	46	36	21	14	15	5	8	6	2	2			
8120	89	2	6	1	3	10	12	12	10	1	4	2	3	2	4	-			
8121	739	40	41	27	39	76	127	109	76	47	39	31	25	20	8	7			
8122	555	38	32	20	18	51	91	92	73	38	13	22	14	19	12	6			
8123	243	14	8	13	16	24	38	43	24	24	11	10	5	4	2	3			
8124	32	-	-	-	1	3	8	4	2	2	-	1	1	-	1	-			
8125	731	28	27	22	34	85	131	108	65	40	29	31	20	14	12	18			
8126	151	7	4	8	2	18	30	21	16	9	6	13	2	6	3	1			
8127	405	22	23	13	22	46	69	54	58	29	16	6	9	5	12	6			
8128	431	26	15	25	25	47	81	60	37	38	27	9	16	5	8	6			
8129	28	1	2	-	1	7	6	4	4	-	-	-	-	-	2	-			
8130	186	8	9	5	6	20	34	34	21	13	12	4	4	5	4	2			
8131	164	13	10	6	3	10	33	28	17	10	5	4	9	3	2	3			
8132	54	3	2	4	3	4	9	5	10	3	3	-	-	3	1	-			
8133	31	1	1	1	1	8	3	4	-	4	2	4	-	1	-	-			
8134	18	-	-	-	-	2	-	2	2	-	3	-	1	2	3	3			
8135	230	15	6	11	10	19	35	37	30	10	13	9	7	11	4	3	▼		

Figure 2

21. The Task Force will compare different confidentiality rules that countries apply and analyse the effects on the primary suppressions. To prevent recalculating primary suppressions from marginal totals some extra cells have to be suppressed. These extra suppressed cells are called secondary suppressions. A common method to decide on the secondary suppressions in an optimal way is the so-called hypercube method which is implemented in Tau-ARGUS and used by many countries. The software package Tau-ARGUS can be downloaded free of costs from the website <http://neon.vb.cbs.nl/casc/>. The website also contains test data and the manual of the software can be found and downloaded.

IV. SOME PROVISIONAL CONCLUSIONS

22. On the basis of the outcomes of the first two meetings of the Task Force CENDSC, and considering the timetable of the censuses, the current orientation is towards a simplified approach. Unless a clear and full agreement is quickly reached by all Member States on a proposed methodology, the EU implementing regulation regarding the statistical data to be transmitted to Eurostat will not contain any provision on the disclosure control method to be applied to census data. This means that the countries can send cells of the hypercubes blanked for confidentiality reasons.

23. However, it should be taken into account that not all topics may be considered by the Member States as confidential. For instance, thoughts should be given to assess whether characteristics like sex or age have to be subject to disclosure control according to the national regulations. It may well be indeed that some topics are more "sensitive" than others. Considering that the topics listed in the EU regulation – thus mandatory for the Member States - are in fact the CES core topics, which do not include any characteristic on – e.g., health or income, it may be worthy to consider – already at national level and respecting the national provisions – whether confidentiality applies to all topics and all enumeration units⁵.

24. The Task Force CENDSC will continue to work on a harmonised approach to be recommended for adoption to the countries, taking into account the as far as possible the constraints expressed above. Whether this will lead to an increased comparability of the census data will depend also on the degree of flexibility the national statistical offices will apply in considering the "fit-for-all" proposal of the Task Force. In any case, its outcome will contribute to the discussion in this domain.

⁵ If confidentiality is clearly an issue for persons, this is not straightforward for data on other kinds of enumeration units, like households, dwellings, etc.