Dissemination Strategy and Statistical Disclosure Control of the 2011 Register-based Census in Slovenia

Danilo Dolenc[1]

1. Introduction

The confidentiality issues of the population and housing census data were for the first time raised in the 1991 Census. In previous censuses no statistical disclosure control (hereinafter SDC) was applied at all. In the 1991 Census only data on nationality (ethnic affiliation) for the level of settlement were suppressed, but only for settlements with less than 30 inhabitants. But the aggregation of data (breakdown) was the same, irrespective of the geographical level of dissemination.

Publishing of 2002 Census data represents an important step in the development of SDC in the Statistical Office of the Republic of Slovenia (hereinafter SORS). Slovenia was the only EU Member State which sent protected census data to Eurostat. The selected method of SDC was absoluteminimum threshold valid in every case (irrespective of geographical level or aggregation level of individual variable) and secondary protection was applied with the aimof not allowing the reconstruction of original data. But there was no cross-table protection, so it was easy to compare some tables and recalculate protected cells. Besides, for some variables (e.g. ethnic affiliation, religion) additional rules were applied, such as no data at the level of settlement are disseminated or only data at the highest level of aggregation are available. Due to very strict SDC, the feedback of our users was very negative and SDC represented a huge additional workload for IT and methodology staff. But at the same time similar data (e.g. age and sex structure) for the same reference date (31 March) from the regular statistical survey on population prepared on the basis of administrative sources (Central Population Register, Register of Foreigners) were available without any SDC.

2. Dissemination strategy for the 2011 Register-based Census data

The following points were taken into account preparing the new dissemination strategy:

- Users' expectations and needs, including their dissatisfactionwith the dissemination of 2002 Census data;
- Information privacy of individuals should be respected according to the national legislation and good practices from SORS and from abroad;
- To assure maximum information value of disseminated data for all users and for different purposes;
- Data that are available to the public in an individual form in any of the administrative source used for the 2011 Register-based Censusshould be treated in the same way for dissemination (valid only for variables that are the same in the source and the census);
- Determination of sensitive variables;
- No cross-tabulation protection (SDC applied to single tables only);
- More and more demand for small area statistics and access to micro-data.

[1]Danilo Dolenc, Statistical Office of the Republic of Slovenia, Litostrojska 54, 1000 Ljubljana, Slovenia (danilo.dolenc@gov.si)

The facts that effect additional consideration about SDC are:

- Smallness of population (2,050,189);
- Administrative division of Slovenia with 211 municipalities (ranging from 300 to 280,000 inhabitants) and almost 6,000 settlements (one third of them with fewer than 60 inhabitants);
- New developments in web-mapping using square grid cells with the smallest size of 100m x 100m;
- The availability of tools for more efficient SDCincluding the human resources.

3. Dissemination policy and statistical disclosure control

General principles of the dissemination policy were adopted by the responsible bodies of SORS (Project Council of the Register-based Census, Data Protection Committee, General Methodology Department).

3.1 Geographical level and level of aggregation of variables

Interdependence of territory and breakdowns of variables is the basic principle in the dissemination policy. At higher territorial levels more detailed breakdown can be shown and at the same time more variables can be included in aggregation. Breakdown atlower territorial levelscan be the same as or lower than at higher territorial levels. At the same timethe number of variables in a single aggregated table is limited to maximum five (including geographical area) at the highest geographical level and to two at the lowest (settlement, square grid cells). At the level of settlements and smallest size of square grid cells the threshold is set up to preserveinformation privacy and for units below a minimum threshold (30 inhabitants) all values are suppressed and also secondary protection is applied due to the hierarchical structure of the territorial classification. The only exception is the basic demographic data by age (5-year age groups, 85+) and sex, which are available without any SDC. The same is true for most of the other published tables as the aggregation of categories could be presumed as a guarantee which enables information privacy of individuals. As an example the dissemination of educational data in our SI-STAT Data Portal is shown in Table 1. The code-book for the Education variable consists of22 categories in origin.

Table 1 Dissemination of educational data by age, sex and NUTS level, SI-STAT Data Portal

Variable	Geographical area				
	NUTS 0	NUTS 2	NUTS 3	NUTS 5 (LAU 2)	NUTS 7
	Slovenia	Cohesion region	Statistical region	Municipality	Settlement
AGE	1 - year age groups, 15-85+	5 - year age groups, 15-65+	5 - year age groups, 15-65+	no	no
SEX	yes	yes	yes	yes	no
EDUCATION	11 categories	9 categories	7 categories	7 categories	3 categories

3.2 Sensitive data in the Register-based Census 2011

In general, the sensitive characteristics have a subjective dimension (such as ethnicity or religion), can be methodologically questionable (disability for example) or correspond to small population groups. In case of the register-based method used in the 2011 census round the latest could be the only point for discussion. We found the variable citizenship as problematic from the confidentiality point of view. In total 137 different citizenships are recorded (half of them with fewer than 10 people). The next example is persons not living in private or institutional households, including homeless persons, but the distinction to other persons in this category is not possible. The number of persons in this group is 1,395.

3.3Users feedback

By now 67 aggregated tables have been published in our SI-STAT Data Portal and no SDC have been applied at the level of municipality or higher due to adequate structure of tables. Only settlement tables (6 of them) are statistically protected. We assume that our users appreciate our new dissemination policy as there were no requests for more detailed data at any geographical level. Beside that tables with higher number of variables are difficult to understand and also more complicated by retrieving data from statistical databases.

4.Hypercubes

The average number of variables of the hypercubes for total territory and NUTS 2 level is more than 7, for NUTS 3 level almost 5 and for LAU 2 level 3. From our SDC point of view only the hypercubes in LAU 2 seem to be harmonized with our dissemination policy.

Closer to our dissemination standards are principal marginal distributions (hereinafter PMD) with an average of more than 5 variables for total territory and NUTS 2 level and almost 4 for NUTS 3 level, while for LAU 2 level there are no PMDs.

From the SDC point of view and comparing the national practice with EU Regulation on the programme of the statistical data the PMDs are the only solution, but, on the other hand, the workload is moredemanding as there are 202 PMDs in contrary with 60 hypercubes (in fact 12 of them equal to PMD).

4.1 SDC in the case of hypercube 2

The main problem of hypercubes and PMDs from the confidentiality point of view is in fact the possibility to disclose several individual data of persons belonging to small population groups. The best examples are PMDs with single-year of age breakdown cross-tabulated with several other variables or PMDs concerning country of birth or country of citizenship at the H level.

For the purposes of this paper we select hypercube Nr. 2 with 8 variables (geographical are NUTS 2, sex, the most detailed household status, educational attainment, current activity status at higher level, country of birth and country of citizenship at the most aggregated level, 5-year age groups). τ-ARGUS version: 3.5.0, which is used for the SDC in our Office, was not able to produce the requested protection for the hypercube as 6 variables is a maximum input.

The final results from the 2011 Register-based Census using threshold 3 are presented in Table 2 for each PMD of the hypercube. The hypercube method has been selected in τ-ARGUS; the

frequency range was set to 0%. There was no cross-protection between individual PMDs. The information loss was from still acceptable 18.2% (PMD 4) to almost unbelievable 69.8% (PMD 2). And what is even more concerning – also aggregated data at the top level were suppressed (for example total population in PMDs 2 and 3). Due to the small population and also very rare combination of some topics (for example the age of child by household status is mostly below 50 years), the proportion of empty cells exceeds 30% in PMD 1.

Table 2 Principal marginal distributions (PMD) of hypercube Nr. 2 - indicators[2]

	PMD_1	PMD_2	PMD_3	PMD_4	PMD_5	PMD_6
Nr. of cells	**26208**	**23400**	**23400**	**16380**	**16380**	**16380**
Empty cells	8103	3978	5197	4893	1929	2727
Numerically safe cells	10310	3086	2621	8504	8379	5905
Numerically unsafe cells	7795	16336	15582	2983	6072	7748
Primary confidential cells	945	1237	1476	206	540	886
Secondary suppressions	6850	15099	14106	2777	5532	6862
Information loss (%)	29,7	69,8	66,6	18,2	37,1	47,3
Maximum value suppressed	226148	2050189	2050189	293905	355435	575287

5. Conclusion

We expected that the main output of the ESSnet Workshop on Statistical Disclosure Control of Census Data will be guidelines to countries how to send to Eurostat to the greatest extentvaluable but at the same time non-confidential data according to the national practice. Even more important is to find an appropriate tool that could produce efficient but reasonably protected hypercubes or PMDs for European comparison and EU-27 totals at least at higher levels of aggregation. But the question is if the European Commission really needs so detailed data that are requested in many of the hypercubes and even PMDs.

According to the implementing regulation as regards the modalities and structure of the quality reports and the technical format for the data transmission, the deadline for delivery of data and metadata to Eurostat is 31 March 2014 or more than 3 years after our census reference date. In the meantime we are going to produce another ‘’census’’. Data from 2011 Register-based Census will be already outdated and will have more historic character than could be used for relevant up-to-date economic, social or other political decisions. The modern society is changing faster! Besides, there is only output harmonization (hypercubes) but more important input harmonization is determined only by the obligatory topics and methodology.

[2] Categories that are not relevant for Slovenia (e.g. not stated, stateless, other countries) are not included. Therefore, the total number of cells is below the maximum possible number of product of all categories in selected breakdowns according to Commission Regulation (EC) No 1201/2009. All subtotals and totals for obligatory categories are included as defined in every single breakdown in the Regulation.