Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

Confidentiality protection of large frequency data cubes

Prepared by Johan Heldal and Svetlana Badina, Statistics Norway

Confidentiality protection of large frequency data cubes

Johan Heldal and Svetlana Badina

Statistics Norway, Kongens gate 6, N-0152 Oslo, Norway
E-mail: johan.heldal@ssb.no, svetlana.badina@ssb.no

Abstract: In 2014 all EU + EEA countries will have to submit 60 hypecubes from their 2011 censuses to Eurostat. All the cubes are frequency counts and in most of them individuals are counting units. Each country is responsible for disclosure control of their own cubes according to their own legal definitions and with their own choice of disclosure control method. As the only of 32 countries Norway has decided to use rounding of small counts as the preferred disclosure control method. This paper describes why we want to do it that way and how we want to carry it out. The paper presents the results so far in the development of the method and describes our plans for improvements.

1 Background[1]

The Norwegian Census 2001 used a small count rounding procedure to protect the Census cubes against potential disclosure. The method used produced additive rounded cubes but the results were not consistent across cubes (Statistics Norway 2003). The procedure was relatively simple but was considered successful and the same method will be used for our national dissemination of the 2011 Census tables. But because of extensive demands for detailed tables based on our population registers we need to make improvements in the capacity of the method.

According to Regulations (EC) 763/2008 and Commission Regulation (EU) No 519/2010, all EU and EEA countries have to submit 60 frequency count hypercubes from their 2011 censuses to Eurostat in 2014. These cubes pose similar challenges to confidentiality as our national register cubes will do and it was decided to work with them for the development and testing. 50 hypercubes have individuals as counting units. They are spanned by four to nine variables of various detail and many of them have millions of cells, most of them empty. It is left to each Member State to carry out disclosure control according to its own laws and preferred methods. Eurostat has chosen hypercube 06, which is spanned by eight variables, as a preferred test cube for which confidentiality challenges should be analyzed and methods tested. Hypercube 06 will therefore be used as illustration in this paper.

Section 2 will describe how Norway defines the confidentiality challenge and why we have chosen to use rounding as our disclosure limitation tool. Section 3 will

[1] The opinions expressed in this paper are those of the authors and do not reflect the official policy of Statistics Norway.

describe Hypercube 06 and some concepts defined in the Regulations. Section 4 presents some outcomes of simulations and section 5 provides a discussion. Section 6 suggests work for further improvement.

2 The problem

Statistics Norway applies the following definition of disclosure for frequency count cubes:

Disclosure is assumed to have taken place if for any subset A of variables spanning the cube there is a combination of values a that can be traced to an identifiable statistical unit and for which there exists only one combination of values b with positive count for a disjoint subset B of the spanning variables where $A \cap B = \emptyset$.

Table 1 illustrates the definition. **0** is a vector of zeroes. We can call the combination of variables *A identifying* and the combination *B sensitive* although the legal framework in Norway does not distinguish identifying from sensitive variables. *A* and *B* may therefore be any disjoint pair of subsets of the spanning variables. The count 7 in cell (*a*,*b*) is completely arbitrary. It does not have to be a small count. *A* and *B* may change roles and the number of combinations will often be too large to consider every one of them separately.

A	***B***				
	1	…	*b*	…	*L*
1					
:					
a	0	**0**	7	**0**	0
:					
K					

Table 1 Illustration of disclosure definition

The definition is somewhat ambiguous since the situation described in table 1 can arise from the combination of two or more values of *B* where more than one may have positive counts in combination with *a*. This may be seen as a paradox since combining values in rows or columns always reduces the information content in the table and should therefore make it less disclosive. Never the less we use the definition. The paradox can be resolved by defining aggregation of the values of the *B*-variables that are too crude to any longer be considered sensitive. But the

Regulations that define the hypercubes to be submitted for each cube do not allow combining values of the spanning variables except for what is being called "Principle Marginal Distributions" of the cube. Principle Marginal Distributions will be described in section 3.

A count of 1 or 2 in a cell is not considered disclosive if it does not occur alone in a row or column of table 1. Rows with two positive counts are sometimes considered disclosive if one of the counts is 1 (or 2) since the singleton (or doubleton) theoretically can be able to disclose values of other units with combination a on A. Real disclosure of identifiable units this way is rather unlikely to occur in the census hypercubes. Anyway, the proposed method will provide some protection for this case as well.

Small count rounding means that counts less than a certain threshold in either elementary cells or at some aggregate levels (e.g. "Principal Marginal Distributions") will be rounded according to a rounding base. In the rounding of the Norwegian Census hypercubes both the threshold and the rounding base will be chosen as 3. This means that counts of 1 and 2 will be rounded to 0 or 3. In the released cubes the rounded counts will be indistinguishable from the real counts. This creates an uncertainty about whether a single positive count in a row of table 1 is really alone. An intruder will try to evaluate the likelihood of this being the case. The intruders' analysis will depend on his or her prior knowledge and may best be described in Bayesian terms. However, such an analysis will be beyond the scope of this paper. We consider the small count rounding as giving a sufficient protection against the means that we reasonably can expect an intruder to be willing to use.

We know that many EU Member States plan to use cell suppression or record swapping on the hypercubes. In Statistics Norway we consider cell suppression as a method for magnitude tables and unsuitable for frequency counts. Cell suppression of frequency count tables requires secondary suppression and cells available for secondary suppression may have large counts or they may be zero. It may be transparent which cells are the primary and which are the secondary suppressions, in particular if marginal counts are retained. A sufficient suppression may induce large information losses, rendering the released cube almost worthless. The small count rounding procedure described in section 3 has a much smaller information loss. It will not be necessary to flag any released cell as 'confidential' or 'not available'.

3 The census hypercubes

Regulation CR (EU) 519/2010 provides the following definitions which will be used in this paper

'hypercube' means a multidimensional cross tabulation of breakdowns which contains a cell value for the measurement of each category of each breakdown cross-tabulated by each category of any other breakdown used in that hypercube;

'principal marginal distribution' means a subset of a given hypercube which results from the cross tabulation of some but not all of the breakdowns of the hypercube;

'primary cell' means any cell which is part of at least one principal marginal distribution in a given hypercube. In hypercubes for which no principal marginal distribution is defined all cells are primary cells;

'secondary cell' means a hypercube cell that is not a primary cell in a given hypercube;

'cell value' means the information transmitted in a hypercube cell. A cell value can be either a 'numerical cell value' or a 'special cell value';

Hypercube 06 is spanned by eight variables each having the number of categories shown in table 2. So far we have worked with a preliminary version of the cube.

Variable	Explanation	No. of groups
GEO.L	Region of residence according to NUTS2	7 regions
SEX	Sex	2
FST.H	Family status. High detail	6
LMS.	Marital status	4
CAS.L	Activity status. Low detail	3
POB.M	Country of birth. Medium detail	9
COC.M	Citizenship. Medium detail	9
AGE.M	Age. Medium detail (5-year groups)	21

Table 2 Spanning variables and groups in hypercube 06

The total population count in Norway at the census date is 4 979 955 persons.

The eight variables span 1 714 608 'secondary' cells.

Regulation 519/2010 gives each country the option either to submit the entire cube or a given set of 'Principle Marginal Distributions' from them. For hypercube 06 the given set consists of six PMDs that must be submitted. They are presented in table 3 and denoted 6.1-6.6.
As will be demonstrated in section 4, submitting only the six PMDs will significantly reduce the size of both the disclosure and the rounding problem.
The hypercubes 07, 08 and 09 differ from hypercube 06 only in that the variable LMS is repaced by other variables. The PMDs 6.4, 6.5 and 6.6 are common PMDs for the four hypercubes.

	Breakdowns							
6.	**GEO.L.**	**SEX.**	**FST.H.**	**LMS.**	**CAS.L**	**POB.M.**	**COC.M.**	**AGE.M.**
6.1	GEO.L.	SEX.	FST.H.	LMS.				AGE.M.
6.2	GEO.L.	SEX.	FST.H.	LMS.	CAS.L	POB.M		
6.3	GEO.L.	SEX.	FST.H.	LMS.	CAS.L		COC.M	
6.4	GEO.L.	SEX.	FST.H.		CAS.L			AGE.M.
6.5	GEO.L.	SEX.	FST.H.			POB.L		AGE.M.
6.6	GEO.L.	SEX.	FST.H.				COC.L	AGE.M.

Table 3 Principle Marginal Distributions of hypercube 06

4 The rounding procedure

The rounding procedure is carried out on selected 'secondary cells' with counts 1 or 2 at the most elementary level of the 8-way hypercube. The extent of rounding depends on whether the entire hypercube or only the six PMDs are to be delivered. The procedure can be described as follows:

1. Reduce the problem:
   a. For a hypercube **A** identify a subset **B** consisting of
      - All 'secondary' cells with counts 1 or 2 or
      - All 'secondary' cells contributing to cells with counts 1 or 2 in the PMDs.
   b. Calculate **C** = **A** - **B**
2. Let n_1 and n_2 be the number of cells with value 1 and 2 in B respectively. The total value of **B** is $n_B = n_1 + 2n_2$.
3. Select $[n_B/3]$ cells from **B** to be rounded to 3 and round the rest to 0 where $[\cdot]$ means rounding to the nearest integer. Call the rounded cube $\mathbf{B}^*$.
4. Calculate $\mathbf{A}^* = \mathbf{C} + \mathbf{B}^*$, the rounded cube.

The marginal distributions of $\mathbf{A}^*$ are calculated directly by aggregating from its 'secondary' cells. $\mathbf{A}^*$ will therefore be additive along with its marginals and the

marginals will be consistent with each other. The deviations $\mathbf{A}^* - \mathbf{A} = \mathbf{B}^* - \mathbf{B}$ for all cells. The total value of $\mathbf{A}^*$ will deviate from that of $\mathbf{A}$ with $|3[n_B/3] - n_B| \leq 1$.

In the Norwegian hypercube 06 there are 53 550 secondary cells with positive cell value of which n_1 =18728 have the value 1 and n_2 = 7 095 have the value 2. They represent n_B = 32 918 individuals. Although the small cells represent 48.2 percent of all positive cells they represent only 0.66 per cent of the population. When applying the rounding method, [32 918/3] = 10 973 secondary of the $n_1 + n_2$ = 25 823 small secondary cells will have to be rounded to 3 and the rest to 0.

The six PMDs contain are m_1 = 1 891 cells with value 1 and m_2 = 1 157 cells with value 2. These $m_1 + m_2$ = 3048 small 'primary' cells are generated by $n_1 + n_2$ = 2941 'secondary' cells of which n_1 = 2 683 have the value 1 and n_2 = 258 have the value 2. The n_B = 3 199 persons in these cells represent 0.064 percent of the total population. When rounding the six PMDs only, [3199/3] = 1 066 cells will have to be selected from the 2941 to be rounded to 3 and the rest to 0. Some of the small count secondary cells contribute to more than one small count primary cell. The rounding will perturb primary cells that are not small as well as those who are small.

The next challenge is how to select a sample of small secondary cells to be rounded to 3 in the best possible way to keep the perturbances at aggregated level as small as possible. So far we have only tried repeated random search where the samples were taken as probability samples with probabilities for selection proportional to the cell value (1 or 2). This makes the selection unbiased. The deviations at aggregate level were checked for a specified set of marginal distributions. The details in rounding procedure under step 3 above were carried out as follows:

a) $Iter = 1$.

b) From the non-zero cells in **B**, select a probability sample of size $[n_B]/3$ witout replacement whose values are replaced by 3 so that Prob[2→3] = 2Prob[1→3]. Replace the values of the rest of the cells with 0. The new cube is called $\mathbf{B}^*$.

c) Aggregate **B** and $\mathbf{B}^*$ to a given set of marginals distributions M. Calculate $m_{Iter} = \max_{c \in M} | b_c^* - b_c |$ across all cells $c \in M$.

d) If $m_{Iter} < m_i$ for $i < Iter$ or if $m_{Iter} = \min_{i<Iter} m_i$ and $\#(c \in M : \max | b_c^* - b_c | = m_{Iter})$ is smaller than for any previous iteration, then keep $\mathbf{B}^*$ as the best solution so far. Otherwise reject $\mathbf{B}^*$.

e) If $\max_{c \in M} | b_c^* - b_c | \leq criterium$ or Iter ≥ maxit, then stop. Else $Iter = Iter + 1$ and go to b).

There are several ways of selecting a probability sample under item b). For the experiments presented in table 4 we have used systematic PPS-sampling with randomized ordering within strata. Balanced sampling (Deville and Tillé (2004),

Tillé (2006)) is an interesting option that will be tried in further work. Balanced sampling has been implemented in the R-package 'sampling' freely available from CRAN. Using software for Mixed Integer Linear Programming (e.g. Rsymphony in R) is another possible route that we want to investigate. Computing time is an issue, in particular when several large cubes are to be rounded.

For the testing on hypercube 06 we have chosen the set *M* defined in item c) as all one- and all two-dimensional marginal distributions that can be generated from the six PMDs defined for hypercube 06. Subtotals corresponding to the low-detail classifications FST.L, POB.L, COC.L and AGE.L were included. This totals 1985 marginal cells. 10000 iterations were used to find a best possible solution.

The improvements obtained in maximum deviation m_{Iter} by iteration in a set of random searches are shown in table 4. Table 4 clearly shows that only considering the six PMDs gives much smaller maximum deviations than when considering the entire hypercube and that stratification has an effect. The stratification was done in a hierarchical way that guarantees minimum deviation of at most 2 for the one-dimensional level GEO.L, the two-dimensional GEO.L*SEX, three-dimensional GEO.L*SEX*AGE.M and four-dimensional GEO.L*SEX*AGE.M*FST.H, but not for one.dimensional SEX, AGE.M or FST.H. This would require balanced sampling.

	Full hypercube			PMDs only		
	Stratification			Stratification		
Iter	None	GEO.L	GEO.L SEX AGE.M FST.H	None	GEO.L	GEO.L SEX AGE.M FST.H
		SEX			SEX	
1	263	149	198	68	62	57
10	186	145	154	62	56	52
100	153	133	123	50	50	40
1000	140	133	100	45	45	37
10000	133	121	100	43	41	35

Table 4 Improvements in m_{Iter} by iterations in random search

A deviation of 133 in a marginal cell for the entire hypercube may seem large, but it is only 0.064 per cent of the unrounded cell value of 208 553 (CAS.L = 'not economically active', AGE.L = '30-49 years'). The maximum deviation of -35 in the best solution for a delivery with the six PMDs is from a total of 85 383 (-0.04 per

cent, not economically active living in consensual union). This may be less than the day-to-day variation of such a value. Smaller values tend to have smaller absolute deviations but larger relative deviations and some of smallest counts may disappear completely.

5 Discussion

Statistics Norway has not yet evaluated the results officially. But although the result may be satisfactory for hypercube 06 alone, the method used has its limitations.

One problem with random search is that the result is subject to chance. For a given cube one may have luck or bad luck with the solution. We will need to find better and more stable search engines. There is also a diminishing return from increasing the number of iterations. The better the result that has been found, the longer we must expect to wait for a new improvement. An advantage compared to many other methods is that it provides consistent solutions for linked cubes of some size. However we wish to be able to produce consistent results for even larger problems than hypercube 06, such as joint consistent rounding of several hypercubes.

The method can be generalized to rounding bases of more than three, but this will generate larger deviations at aggregate level.

Tau-Argus offers controlled rounding of three-way tables and linked two-way tables, but cannot take tables with more than three dimensions. The tau-Argus solution requires that all cells, not only the small counts, are rounded. The reduction of the table described in section 4 to consider only small counts is what makes our method feasible. However, in cubes with no more than three dimensions tau-Argus could be applied on a reduced cube.

6 Further work

The results presented in section 4 are as far as our work has come at the end of August 2013, the deadline for submission of the paper. Our next step will be to test other sampling methods and more advanced mixed integer linear programming. We wish to try to round the PMDs of hypercubes 06, 07, 08 and 09 simultaneously trying to produce consistent cubes and hope to be able to present further improvements at the workshop. Another idea is to try to merge the reduced rounded cells back into the micro data they have been generated from, in order to generate a micro data set from which safe consistent cubes can be generated. But this will probably still be some time into the future.

References

Chauvet, G. & Tillé, Y. (2006). *A Fast Algorithm for Balanced Sampling*. Computational Statistics 21: 53-61. Physica-Verlag.
Deville, J.-C. & Tillé, Y. (2004). *Efficient balanced sampling*: the cube method, Biometika 91, 893-912.

Harter, R., Hornik, K. & Theussl, S. (2013). *Rsymphony*. Software package for R available at CRAN (http://www.r-project.org/).

Matei, A. & Tillé, Y. (2012). *sampling*. Software package for R available at CRAN (http://www.r-project.org/).

Statistics Norway (2003). *Rounding as a confidentiality measure in Statbank Norway*. Working Paper 26, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.
http://www.unece.org/fileadmin/DAM/stats/documents/2003/04/confidentiality/wp.26.e.pdf

Tillé, Y. (2006). *Sampling Algorithms*, Springer Series in Statistics, N.Y.