The European Census Hub hypercubes 2011. Norwegian SDC experiences.

Johan Heldal

Statistics Norway, johan.heldal@ssb.no

Abstract: Statistics Norway used a version of small count rounding as the SDC-method for the hypercubes provided for the Eurostat Census Hub 2011. The method jointly rounds a group of hypercubes with similar breakdowns to additive and consistently rounded hypercubes. There were two main aspects of our procedure

1. A reduction of the problem.
2. The search for solution to the reduced problem.

Although our search for solutions was primitive, we found the results encouraging. We do mean that the way we reduced the problem can be useful in other contexts and an implementation in τ-Argus could extend the feasibility of controlled rounding to larger cubes. But better search methods could have produced better results enabling us to provide more detail or consistency within even larger groups of hypercubes. Abandoning the requirement that rounding should be controlled, such search methods could extend the scope for rounding methods in τ-Argus to cubes of more dimensions than three.

This paper presents the main results of the procedure that Statistics Norway used for the Census Hub hypercubes and suggests some improvements of the method.

1 Background

In 2014 all Member States of the European Union and the EEA countries had to deliver 60 hypercubes from their 2011 censuses to the Eurostat Census Hub. The legal basis for this obligation was the regulations (EC) 763/2008 and CR (EU) 519/2010.

Disclosure control was a central part of the delivery. But it was up to each MS how to protect their own data. The result was that highly different procedures were applied from different Member States, reducing the comparability of the statistics across countries. The ESSnet now works on a harmonized methodology for Statistical Disclosure Control and some countries have tested some proposals. Statistics Norway will test some of these proposals as well as our own version.

The version of small count rounding that statistics Norway used for the 2011 census gave from our point of view satisfactory results but the search for good solutions should be improved before a decision is taken on which method we will go for in the 2021 census.

The method was roughly presented in a paper at the joint UNECE/Eurostat work session in 2013 and implemented during the following months, but the results from the application to the Census Hub hypercubes have not been presented. It handles groups of hypercubes jointly in a way that produced additive and consistent results for the hypercubes within the same group. For completeness the method will be described in detail in section 3.

2 Organizing the hypercubes

The 60 hypercubes (except no. 40, homeless persons) were combined into 17 Super-HyperCubes based on common units and "breakdowns". A complete description of these SHCs are given in appendix. Individual persons were units in 12 of the 17 SHCs while the other 5 SHCs had household, dwellings, families and living quarters.

CR (EU) 519/2010 allowed the Member States to deliver only defined *Principal Marginal Distributions* which result from the cross-tabulation of some but not all breakdowns of the hypercube. The regulation called the cells in the PMDs *primary cells* while hypercube cells that were not in a PMD were called *secondary cells*. When only delivering the PMDs of a hypercube the secondary cells were all set to confidential. In order to make the rounding procedure manageable in the larger SHCs secondary cells, we delivered only PMDs in seven of the 12 SHCs where persons were units.

Within each SHC the hypercubes or PMDs were rounded jointly with small count rounding (base 3) in such a way that all rounded hypercubes and PMDs were fully consistent and additive. For instance, SHC 10_18 was a cross classification of all 13 breakdowns in hypercubes 10 to 18. The 9 hypercubes had defined 33 different PMDs which were all rounded consistently in the sense that aggregations to common marginal distributions for any two PMDs were identical.

3 The rounding procedure

Given a SHC the first stage in the procedure is to reduce the problem. This is done with the following steps.

1. First cross-classify all the breakdowns in an SHC. Call the cross-classification **A**.
2. Then produce all hypercubes or PMDs in the SHC, depending on which level the SHC should be delivered.
3. Find all positive counts less than a base *b* in the hypercubes or PMDs produced in step 2. In the Norwegian Census and for the Census Hub we used *b* = 3.
4. Merge these counts to **A** and identify all small positive counts in **A** that generate small counts in the hypercubes or PMDs produced in step 2. Denote this subset of cells in **A** by **B**. All cell counts in **B** will be less than *b*. Drop zero-cells.
5. Calculate **C** = **A** – **B**.
6. Round the cell counts of **B** to 0 or *b* to produce **B***.
7. The rounded SHC is then **A*** = **C** + **B***.

The challenge is to find the best possible solution in step 6 so that counts at higher aggregate levels are perturbed as little as possible. If the SHC had no more than three dimensions, **B** could have been rounded by τ-Argus. Controlled rounding of the reduced SHC **B** would be much faster than rounding the complete SHC **A**. This would probably have produced the best result possible. To make this more feasible I propose that the reduction of the problem, as described in steps 1-4 above should be implemented as an option in τ-Argus.

With far more than three dimensions controlled rounding was not feasible. The search method that we used was far from optimal but produced solutions that we saw as satisfactory. But it was what we could come up within a few months before delivery. The algorithm we used for step 6, which is a kind of systematic sampling, went as follows.

a. Calculate a control set C of marginal distributions of **B** on which deviations between $\mathbf{B}^*$ and **B** can be tested. In the Norwegian Census Hub delivery all one and two dimensional distributions of the PMDs or hypercubes were specified for the test.
b. Start with setting the rounded counts $y_c = 0$ for all cells *c* of $\mathbf{B}^*$.
c. Calculate the total count $N_{\mathbf{B}}$ of **B**.

d. Sort the cells in **B** by the *k* "most important" breakdowns *Var*_1 *Var*_2 … *Var_k* … in a priority order. It is not necessarily optimal to include all breakdowns in a SHC in the sort. Within the last levels specified by the sort, the cells were sorted in a random order. For the Norwegian SHCs the breakdowns common to all the hypercubes in the SHC, typically GEO and SEX were taken as *Var_1* and *Var_2*.
e. Consider each cell to have a *length* in the sorted sequence which is equal to its cell count. Aggregate the cell counts along the sorted cell sequence. Denote the aggregated count at cell *c* by t_c.
f. Calculate $m_{\mathbf{B}} = N_{\mathbf{B}}/b$ and round either down to $m_{\mathbf{B}}^* = \lfloor m_{\mathbf{B}} \rfloor$ or up to $m_{\mathbf{B}}^* = \lfloor m_{\mathbf{B}} \rfloor + 1$. The rounding can be done either deterministically or randomly with the probability $p_{up} = (N_{\mathbf{B}} - \lfloor m_{\mathbf{B}} \rfloor b)/b$ for rounding up. $N_{\mathbf{B}}^* = m_{\mathbf{B}}^* b$ is then the rounded total for **B** which is obtained by rounding exactly $m_{\mathbf{B}}^*$ of the positive counts in **B** to *b* and the rest to 0.
g. Calculate a *step length* as $s = N_{\mathbf{B}}/m_{\mathbf{B}}^*$ and generate a random start value u uniformly distributed in the interval $(0, s]$. Let $d = u$.
h. Move along the sorted list of cells. Whenever $t_c \geq d$, do
    i. Set $y_c = b$.
    ii. Update $d \leftarrow d + s$.
i. Step h produces a version of $\mathbf{B}^*$. Calculate the marginal distributions of $\mathbf{B}^*$ corresponding to the control set C defined in step a. and calculate a measure of distance $d(\mathbf{B}, \mathbf{B}^*)$ between **B** and $\mathbf{B}^*$ based on the cells in C.
j. If the $d(\mathbf{B}, \mathbf{B}^*)$ is smaller than in any previous iteration of steps d. to i., keep $\mathbf{B}^*$ as the best solution found.
k. Repeat steps d to j until $d(\mathbf{B}, \mathbf{B}^*)$ is small enough or a maximum number of iterations has been carried out. Keep the best solution as the final solution.

This algorithm will always round exactly $m_{\mathbf{B}}^*$ cells of **B** to *b* so that $|N_{\mathbf{B}^*} - N_{\mathbf{B}}| < b$. The probability that at cell *c* in **B** is rounded up will always be proportional to the cell count x_c. The small cell rounding will therefore be unbiased in each iteration, $E(y_c) = x_c$ for all cells *c*. The algorithm does not round the small counts independently. For *Var*_1 the solution will be controlled at *Var*_1 level in the sense that $|N_c - N_c^*| < b$ for all cells *c* in the one-way marginal distribution for *Var*_1. It will also be controlled for *Var*_2 within each level of *Var*_1 and then for *Var*_3 within each level of *Var*_1 by *Var*_2 and so on. But it will not be controlled for the marginal distributions of *Var*_2, *Var*_1 by *Var*_2, *Var*_3 etc. If *Var*_1 has few categories, the deviations for the marginal distribution of *Var*_2 will never the less be small. To find solutions for the Census Hub we did 10000 iterations for each SHC and the solution with the smallest absolute deviation (utility measure), $d(\mathbf{B}, \mathbf{B}^*) = max_{c \epsilon C} |N_c - N_c^*|$ was chosen. The Hellinger distance, as suggested by Shlomo et al (2007), could have

been a relevant alternative, but its calculation would be slower since it cannot be calculated based on **B** and $\mathbf{B}^*$ alone. After completing $\mathbf{A}^*$ the hypercubes and PMDs to be delivered were calculated for each SHC.

4 Results

The search method was not able to handle all 49 hypercubes with individual persons as units jointly with acceptable results. This was why we divided them into 12 SHCs and decided only to deliver PMDs for the hypercubes in seven of them. A complete overview is given in appendix A. This means that cells that occur in more than one SHCs may be rounded differently. However, such cells only occur at an aggregation level defined by common breakdowns for the SHCs and should not be harmful from a disclosure control point of view. The priority sorting described at step d. of the algorithm in section 3 has great impact on the aggregate rounding errors for the higher level marginal distributions. The priority sorting for SHC 1_9X5 was specified as

```
%LET SORT = geo_l sex age_l age_m hst_l hst_m fst_l fst_h
hst_h cas_l
```

For SHC 10_18 it was specified as

```
%LET SORT = geo_l sex age_l age_m cas_l occ ind_l
```

Both SHCs had *Var*_1 = GEO.L, *Var*_2 = SEX, *Var*_3 = AGE_L and *Var*_4 = AGE_M. One may have to experiment with the number of breakdowns to include in the sort. Including all breakdown of the SHC in the sort leaves too little freedom of variation between iterations to find good solutions. The rounded marginal distributions for these breakdowns are shown in tables 1 to 3.

Table 1. Unrounded and rounded counts for GEO_L in SHC 1_9x5 and SHC

		SHC 1_9x5		SHC 10_18	
GEO_L	Original	Rounded	Difference	Rounded	Difference
NO01	1 167 195	1 167 194	-1	1 167 195	0
NO02	379 741	379 740	-1	389743	2
NO03	948 806	948 806	0	948 806	0
NO04	727 819	727 821	2	727 819	0
NO05	854 291	854 290	-1	854 2991	0
NO06	430 688	430 688	0	430 688	0
NO07	469 697	469 697	0	468696	-1
NOZZ	1 718	1 719	1	1 717	-1
Norway	4 999 955	4 999 955	0	4 999 955	0

Table 2. Unrounded and Rounded counts for SEX in SHC 1_9x5 and SHC

		SHC 1_9x5		SHC 10_18	
SEX	Original	Rounded	Difference	Rounded	Difference
F	2 484 178	2 484 179	1	2 484 177	-1
M	2 495 777	2 495 776	-1	2 495 778	1
Norway	4 999 955	4 999 955	0	4 999 955	0

Table 3. Unrounded and Rounded counts for AGE_L and AGE_M in SHC

			SHC 1_9x5		SHC 10_18	
AGE_L	AGE_M	Original	Rounded	Diffe-rence	Rounded	Diffe-rence
Y00_14		923766	923762	-4	923759	-7
	Y00-04	310523	310524	1	310521	-2
	Y05-09	300625	300630	5	300619	-6
	Y10-14	312618	312608	-10	312619	1
Y15-29		975390	975395	5	975388	-2
	Y15-19	324682	324686	4	324686	4
	Y20-24	329537	329541	4	329535	-2
	Y25-29	321171	321168	-3	321167	-4
Y30-49		1400977	1400970	-7	1400990	13
	Y30-34	325241	325234	-7	325246	5
	Y35-39	352008	352017	9	352004	-4
	Y40-44	373191	373188	-3	373194	3
	Y45-49	350537	350531	-6	350546	9
Y50-64		913383	913382	-1	913372	-11
	Y50-54	322729	322737	8	322718	-11
	Y55-59	304335	304336	1	304339	4
	Y60-64	286319	286309	-10	286315	-4
Y65-84		652583	652592	9	652584	1
	Y65-69	247451	247459	8	247453	2
	Y70-74	166680	166680	0	166681	1
	Y75-79	130209	130211	2	130205	-4
	Y80-84	108243	108242	-1	108245	2
Y_GE85		113856	113854	-2	113862	6
	Y85-89	74057	74053	-4	74056	-1
	Y90-94	32243	32240	-3	32239	-4
	Y95-99	6825	6829	4	6830	5
	Y_GE100	731	732	1	737	6
Norway		4 999 955	4 999 955	0	4 999 955	0

Obviously, the differences $|N_c^* - N_c|$ for higher aggregation level cells *c* increase with lower priority of the breakdowns defining the cell *c*.

Table A1 shows that SHC 1_9x5 had 203 699 cells with count = 1 and 53 583 cells with count = 2. In the 30 PMDs belonging to SHC 1_9x5 there were 32 369 primary cells with count = 1 and 17 427 with count = 2. The procedure to reduce the problem size, steps 1 to 4 in section 3, detected that these cells were aggregations of 19 763 secondary cells with count = 1 and 469 cells with count = 2 with 19 763 + 2·469 = 20 701 individuals. This means that $\lfloor 20\,701/3 \rfloor = 6900$ secondary cells had to be rounded to *b* = 3 and the rest to 0. This is never the less a large rounding task and the maximum difference between original and rounded counts that occurred in any one or two-dimensional distribution of breakdowns (*C*) was +85. This occurred for the LOC = '500 000 – 999 999', a breakdown that occurred in only six of the 30 PMDs of SHC 1_9x5 and was not even included in the sort priority list. 85 was however only 0.009 per cent of the "true" value.

The procedure can produce new small counts because a count larger than *b* (= 3) in a PMD can be a sum of smaller counts of which not all are selected for rounding in the reduction process. For instance, a primary count equal to 4 in a PMD can be the sum of two secondary cells, 4 = 2 + 2, where only one of the '2's is selected for rounding by the reduction procedure. So 4 = 2 + 2 can be replaced by 2 + 0 = 2. In the PMDs in SHC 1_9x5, 1229 new 1-cells and 469 new 2-cells were generated this way. However, this phenomenon can *only* occur for cells that originally had counts larger than *b*.

In SHC 10_18, the 33 PMDs contained 17 347 primary 1-cells and 9310 primary 2-cells which aggregated from 19106 secondary 1-cells and 185 secondary 2-cells with 19 106 + 2·185 = 19 476 persons. Of these cells $\lfloor 19\,476/3 \rfloor = 6\,491$ cells were rounded to 3 and the rest to 0. The maximum difference that occurred was +91 at POB_M='ASI' which was 0.051 per cent of the true count. The breakdown POB_M was not included in the sort priority list for SHC 10_18.

The largest difference that occurred in any SHC was 140 for the two-way breakdown GEO_L = 'NO01', POB_L = 'NEU' in SHC 38_39. This was 0.099 percent of the true value. For SHC 38_39 the complete hypercubes were delivered, not only the PMDs.

Results for all SuperHyperCubes are presented in the appendix.

5 Does the method protect adequately?

We do not consider many of the breakdowns as highly sensitive. Some categories of the breakdown POB may be the most sensitive. The main disclosure risk comes in the form of group attribute disclosure (where a single individual is also considered as a "group"). I this context the problem is not the positive count, but the zeroes that occur if the category with positive count is the only non-zero value given the values of all other breakdowns.

The fact that the small count rounding generates more zeroes protects in two ways

1. Some small groups that could be subject of group attribute disclosure risks may disappear.
2. New rounded zeroes cast doubt on whether zeroes really are zeroes.

The first situation takes place if all small counts contributing to the group are rounded to zero, which is most likely for the smallest counts, but may happen even for groups of three and four.

The second way introduces an uncertainty on whether an apparent group attribute disclosure really is a group attribute disclosure. The degree of uncertainty will depend on the subjective judgement on the hand of the person who reads the statistics and will be influenced by the number of original zeroes in the hypercube or PMD. If the number of original non-structural zeroes in the hypercube is large compared to the possible or likely number of rounded zeroes, which is often the case for the Census Hub hypercubes, it will be considered less likely that a zero is a rounded zero and not a real one. This topic should be analyzed from a Bayesian perspective of the intruder and be a part of the risk analysis. This is another reason why Statistics Norway has chosen to deliver only PMDs for many of the hypercubes. The much larger number of real zeroes in many of the hypercubes decreases the (posterior) probability that a published zero is a rounded zero and so reduces the protective effect of rounding against group attribute disclosure. This is also relevant for the risk analysis that should be done for the method proposed by the ESSnet for the 2021 hypercubes. But that is beyond the scope of this paper.

Anyway, it is our perception that a disclosure attack on the rounded Norwegian Census Hub hypercubes will require a considerable effort to give interesting output and that the cost of such an enterprise will be far more than it pays.

6 Discussion

In Statistics Norway we consider the results for the Census Hub disclosure control as satisfactory. However, the details and the sizes of the hypercubes in combination with the desire to make the results as consistent as possible across hypercubes stretched the method to its limits. Whether we will use this method for the Census Hub 2021 depends on the results of testing of other methods as well, among them the variant of the ABS method (Thompson et al 2013) that the ESSnet has come up with more recently. We will also consider another variant of the ABS that we have developed for another project. The method proposed by ESSnet will produce consistent figures across hypercubes but the hypercubes will need an extra adjustment to become be additive. This extra adjustment will be at the price of consistency.

The ABS method operates with "record keys" for each unit. This is not feasible in all situations. In such situations rounding methods like the one Statistics Norway applied for the 2011 Census and controlled rounding as in τ-Argus will still be needed. In that context the method for reduction of the high-dimensional cubes that focus of the tables or cubes that will actually be published, is of interest. Better search procedures to find good, but not necessarily controlled, solutions should be tried. Methods like Simulated Annealing or Branch and Cut are candidates to try. We will try to rewrite our program from SAS to R in order to make it easier to take advantage of the free software for these methods available there.

References

Regulation (EC) 763/2008

Commission Regulation (EU) 519/2010

Heldal, J. and Badina, S. (2013). *Confidentiality protection in large frequency data cubes.* Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada.

Shlomo, N. (2007). *Statistical Disclosure Control Methods for Census Frequency Tables.* International Statistical Review, 75, 2,199-217.

Thompson, G., Broadfoot, S., Elazar, D. (2013). *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics.* Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada.

Appendix

Table A1. The composition of the SuperHypercubes, their dissemination level, number of small count and largest avsolute and relative differences.

						No of counts in SHC			No. of counts to be rounded in HC		No. of new small counts in HC/PMD				
SHC	#HC	# PMD	# break-downs[2]	Units	Dissem-ination level	# >0	# 1	# 2	# 1	# 2	# 1	# 2	Max diff	% diff	Occurs at
1_9x5	8	30[1]	12	Persons	PMD	383845	203699	53584	19763	469	1229	2703	+85	.009	LOC='500000-999999'
10_18	9	33	13	Persons	PMD	1041624	679599	139801	19103	185	540	1438	+91	.051	POB_M='ASI'
19_22	4	17	10	Persons	PMD	240217	119705	33357	11506	454	272	763	-65	-.102	SIE='SAL', COC_L='NEU'
													-65	-.005	SEX='M', SIE='SAL'
													+65	.067	SEX=M, SIE='SELF_NS'
23-24	2	4[1]	9	Persons	PMD	241368	118914	33187	10420	593	113	434	-48	-.030	OCC='OC2', AGE_L='Y50-64'
													-48	-.008	LPW_N='NO', IND_L='G-I'
25	1	9	7	Persons	PMD	99093	38862	14725	15052	2286	202	631	+80	.046	POB_L='NEU', CAS_L='EMP'
26_28	3	6	6	Persons	HC	38934	17146	5069	9895	2596	104	343	+64	.030	POB_L=POB_M='EU_OTH', COC_L='EU_FOR'
29_35	7	16[1]	11	Persons	PMD	409816	246579	53607	24974	555	980	2260	-107	-.010	EDU='ED2', COC_L=COC_M='NAT'
36_37	2	3[1]	9	Persons	HC	47424	18494	6440	10396	2653	119	368	+72	.272	YAT='Y_GE2000, OCC='OC3'
38_39	2	7	11	Persons	HC	114244	47162	16144	37298	8016	62	415	140	.099	GEO_L='NO01', POB_L='NEU'

Table A1. The composition of the SuperHypercubes, their dissemination level, number of small count and largest avsolute and relative differences.

						No of counts in SHC			No. of counts to be rounded in HC		No. of new small counts in HC/PMD				
SHC	#HC	# PMD	# break-downs[2]	Units	Dissem-ination level	# >0	# 1	# 2	# 1	# 2	# 1	# 2	Max diff	% diff	Occurs at
41, 54	2	7	10	*Occupied* conventional dwellings	HC	26319	7337	3030	5140	864	13	134	-42	-.016	NOC_M=NOC_L='3', TOI='TOIL'
42_45	4	10	13	Persons	PMD	1646435	1168463	211379	14188	199	152	518	+73	.028	IND_L='M-N'
46, 47, 50	3	17	9	Persons	HC	107024	44518	15061	29698	5200	87	428	-113	-.138	GEO_L='NO03', ROY='CHG_IN3'
48, 51, 55, 56	4	4	5	Persons	HC	268075	74754	38365	1066	72	0	29	-16	-.057	HST_M= HST_H= 'IST'
													+16	.0008	SEX='M', HST_M='FAM'
5, 49, 57	3	3	4	Private households	HC	26488	5419	2761	461	78	2	9	±11		7 occurences
52, 58	2	2	3	Families	HC	11667	1727	916	346	97	0	5	+12	1.62	TFN_L=TFN_H='F1_CH', SFN_M='6-10'
53, 60	2	2	4	Conventional dwellings	HC	26725	2763	2071	43	17	0	0	±4		6 occurences
59	1	1	2	Living quar-ters	HC	828	101	80	101	80	0	0	±2		62 occurences

[1] Number of different PMDs. Different hypercubes in the same SHC can have some identical PMDs.

[2] The same breakdown can occur with different levels of detail (L. M H), but is counted as one in this table.