The European Census Hub hypercubes 2011 Norwegian SDC Experiences

Joint UN/UNECE Work session on Data Confidentiality, 20-22 September 2017

Johan Heldal
Statistics Norway

Statistisk sentralbyrå
Statistics Norway

Eurostat Census Hub

- In 2014 all EU MS and EEA countries had to deliver 60 HyperCubes from their 2011censuses to the ECH.
- The HCs have up to eight dimensions or «breakdowns».
- SDC was a central part of it, but it was up to each MS how to to it.
- The ESSnet on SDC has proposed a harmonized metodology for the 2021 censuses.
- Statistics Norway will test the proposals on our 2011 data but also try to improve our 2011 method.

This talk

1. will tell how Norway did it for ECH 2011and
2. what the results look like.

- The method we planned for has been presented earlier but
- The results have not been presented
- For completeness I will sketch the method before describing the results.
- Although simple, Statistics Norway found the results satisfactory
- and we do mean the method, with some improvements, can have a potential.

Organization of the HyperCubes

- The 60 HCs (except one) were combined into 17 SuperHyperCubes (SHCs) based on common units and breakdowns.
- 12 had individual persons as units.
- For seven of the 12 we delivered only Principal Marginal Distributions (PMDs).
- A version of small count rounding was used within each SHC.
- PMDs (or HCs) in a SHC were rounded jointly with complete consistency and additivity for all PMDs in the SHC.

Example

- SHC 10_18 was a cross classification of all 13 breakdowns in HC 10 to 18
- The nine HCs had defined 33 PMDs which were all rounded in a consistent and additive way.

Step 1: Reduce the problem

1. For each SHC **A** identify a subset **B** consisting of all interior ("secondary") cells in **A** with counts $0< x < b$ (=3) *contributing to small cells in the PMDs* or *HCs* of **A**.
2. Calculate **C = A – B**

Step 2: Search for rounding

a) Calculate the total count in **B**, $N_{\mathbf{B}}$.
b) Sort the cells of **B** by the breakdowns in a priority order: $Var_1 * Var_2 * Var_3 * \cdots$
c) With systematic pps-sampling, select $N_{\mathbf{B}}\,/b$ cells in **B** to be rounded to b, resulting in **B***.
d) Calculate a distance d(**B**, **B***) on control set of marginals *C*.
e) Repeat c) and d) until d(**B**, **B***) is small enough or until a maximum number of iterations.
f) Calculate **A* = C + B***, the rounded cube.

- More details in the paper.

Properties of the solution

- PMDs or HCs of SHC **A*** are additive and consistent.
- With respect to priority order sorting,

$$Var_1 * Var_2 * Var_3 * \cdots,$$

  the solution is *controlled* for Var_1, for Var_2 within each level of Var_1, for Var_3 within each level of $Var_1 * Var_2$ etc.
- *Not controlled* for $Var_2, Var_3, \cdots$ marginally.
- A breakdown that occurs in two SHCs $\mathbf{A}_1$ and $\mathbf{A}_2$ will be rounded differently.

Example results SHC 1_9x5

- In **A**: #(1) = 203 699, #(2) = 53 383
- In the 30 PMDs: #(1) = 32 369, #(2) = 17 427
- In **B:** #(1) = 19 763, #(2) = 469
- ⇒ #(persons in **B**) = 19 769 + 2·469 = 20 701
- ⇒ 20 701/3 ≈ 6 900 cells are rounded to 3 with pps-sampling and the rest to 0.
- After 10 000 iterations: $d(\mathbf{A}, \mathbf{A}^*) = d(\mathbf{B}, \mathbf{B}^*) = \max_{c \epsilon C}(b_c - b_c^*) = +85$ = 0.009 % of "true".
- Occurred for breakdown LOC = 500 000 – 999 999.

Example results SHC 10_18

- In **A**: #(1) = 679 599, #(2) = 139 801
- In the 33 PMDs: #(1) = 17 347, #(2) = 9 310
- In **B**: #(1) = 19 103, #(2) = 185
- ⇒ #(persons in **B**) = 19 103 + 2·185 = 19 473
- ⇒ 19 473/3 ≈ 6 490 cells are rounded to 3 with pps-sampling and the rest to 0.
- After 10 000 iterations: $d(\mathbf{A}, \mathbf{A}^*) = d(\mathbf{B}, \mathbf{B}^*) = \max_{c \epsilon C}(b_c - b_c^*) = +91$ = 0.051 % of "true".
- Occurred for breakdown POB_M='ASI', .

Worst result

- Largest deviations for any SHC occurred for SHC 38_39 with 140.
- Occurred for GEO_L = ‘NO01’, POB_L =‘NEU’.
- This was 0.099 % of “true” value.

Limitations and possibilities

- The reduction of the problem (step 1) is a condition for finding satisfactory solutions.
- The search for solutions is far from optimal
- If dimensionality were 3 or less, the best possible solution would be achieved by using τ-Argus on **B**.
- A «reduce the problem» option in τ-Argus would extend its capacity to larger problems.

Other search methods

- Simulated annealing?
- Branch and cut?
- Balanced sampling instead of systematic?
- Other suggestions?
- We are now rewriting the program from SAS to R to take advantage of available methods in CRAN.

Comparison to the ESSnet/ABS method

- The ESSnet/ABS method produces
  - small deviations and consistency across HCs,
  - non-additive tables without a fix that removes consistency.
  - Requires a record key for each unit.
- Our method
  - will produce some larger deviations than ABS.
  - ensures consistency and additivity for HCs/PMDs within SHCs, but not between them.
  - does not require record keys.

Thank you!