Census 2001 SDC experiences and ideas for 2011

Workshop on Census SDC

Luxembourg 19-20 April 2012

Johan Heldal

Statistics Norway

The Norwegian Census 2011

- Completely register based. No forms.
- Values of some statistical variables constructed/-imputed based on register variables and LFS.
  - > The Census is a statistical register
  - > Correct data on individual level not important as long as it provides good statistical population description.
- This will affect disclosure risk, but
- The methods presented here do not consider this.

The Norwegian Statistics Law

- Information, collected by law or given voluntarily, shall not be published in such a way that it that can be traced to the data provider or other identifiable person to his or her harm.
- Interpretation:
  - Main rule: Backtracing not possible.
  - Exeption: Possible if required for an adequate structure of statistics if there is no harm for physical or legal person.

Census 2001 hypercubes

- Based on the census data a predefined set of linked frequency count hypercubes were to be published
  - At national level
  - For each of 19 counties
  - For 434 municipalities (kommuner)
  - For about 13988 base statistical areas

Goals in 2001

- Give adequate protection against
  - Group disclosures
  - Within group disclosures
  - Identification of population uniques
  - Differentiation ?
- Strategy: Rounding small counts in hypercubes (1 and 2) to 0 or 3
- Desired properties:
  - Additive hypercubes
  - Consistency across hypercubes

Results in 2001

- We managed to produce additive hypercubes with small deviations, but
- The solutions were not consistent across cubes.
- The method gave sufficient protection for each hypercube in isolation but
- It could be unpicked by comparing similar tables made from different hypercubes.
- It gave users a perception of protection.

The 2001 method

STEP 1: Reduce the problem:

1. For each hypercube **A** identify a subset **B** consisting of
   - All interior cells in **A** with counts 1 or 2 or
   - all interior cells contributing to 1 or 2 in specified marginals in **A**.
2. Calculate **C = A – B**

STEP 2: Rounding

- Round interior counts in **B** randomly to 0 and 3.
- Recalculate marginals and check their deviations from **B**.
- Repeat until deviations are small enough.
- The rounded cube **B*** is then additive.

STEP 3: Calculate **A* = C + B***, the rounded cube.

- **A*** is then additive with no count 1 or 2.
- Protection offered since true 0, 1 and 2 cells cannot be identified with certainty.

Rounding method used

1. Let $t(\mathbf{B})$ = total count of **B**, e.g. $t(\mathbf{B}) = 76$
2. Determine $t(\mathbf{B}^*) = 75 = \text{round}_3(76)$
3. From the non-zero cells in **B**, select at 75/3 = 25 cells to be rounded to 3.
   - Probabilities: P(2 → 3) = 2·P(1 → 3)
4. Calculate distance $max_c|b_c^* - b_c|$ across marginal cells of **B**.
5. If $max_c|b_c^* - b_c| \leq criterium$, then stop. Else go to 3.

Criteria for 2011

- All census hypercubes are additive.
- The census hypercubes are consistent.
- Specially requested tables should be consistent with the census hypercubes.
- Then only micro data adjustments can meet all requirements.

Method for 2011

STEP 1: Reduce each cube as in 2001.

STEP 2: Merge the reduced cubes (**B**) to microdata D and select the set Q all individual units contributing to at least one reduced cube.

STEP 4: Top down: Sort Q according to

Region × Municipality × Base Statistical Units

× Var1
× Var2
× Var3
× :

× (Var1× Var2)
× Var3

STEP 5: Determine *sample sizes* at every level and for every variable by taking a systematic 1/3 sample from Q.

Method 2011 continued

STEP 6: Multiply sample sizes calculated at step 5 by 3 and we get controlled roundings of population sizes at every level and for every variable in one dimension.

STEP 7: Using the rounded counts from step 6 as balancing variables, select a 1/3 *balanced sample* $\boldsymbol{s}$ from Q for each cube.(Deville & Tillé (2004), Tillé (2006))

- Using the same balancing for the same variables in all cubes provides additive and consistently rounded cubes.

STEP 8: Add the tables from step 7 to the tables from $D - Q$.

Micro data for the tables (?)

STEP 9: Triplicate the balanced sample $\boldsymbol{s}$ and establish $Q^* = 3\boldsymbol{s}$. Replace Q by Q^* in the original dataset D to get D^*.

STEP 10: Run tables from D^*.

Challenges

- The method may need some modifications in detail during implementation.
- Applying the method on individual person units may cause problems in relation to household composition.
- Stats Norway plans to start publishing before all census variables are established (in June).

Thank you for your attention