Statistical Disclosure Control of frequency tables and the Census Hub

Eric Schulte Nordholt
Senior researcher and project leader of the Census
Statistics Netherlands
Division Social and Spatial Statistics
Department Support and Development
Section Research and Development

e.schultenordholt@cbs.nl

Lecture at the Census SDC Workshop in Luxembourg (April 2012)

Contents

Census hypercubes

Different kinds of tables:

- Magnitude tables
- Frequency tables
  - Stating the problem(s)
    - Sensitive categories
    - Group disclosure
  - Possible Criteria
- Linked tables
- Hierarchical tables

Microdata availability

The 2011 Census Round

Census hypercubes

The 2001 Census: forty different tables (in total 2000 tables for the Netherlands)

- Twenty-eight at country level
- Nine at regional level (NUTS 3)
- Three at municipality level (NUTS 5)

The 2011 Census: sixty different detailed tables (so-called hypercubes) (in total 3000 tables*) for the Netherlands)

- Five at country level
- Thirty-six at provincial level (NUTS 2) (two on commuting)
- Four at provincial level (NUTS 2) for those who work
- Ten at regional level (NUTS 3)
- Five at municipality level (LAU 2)

*) If no subhypercubes are produced, otherwise 10,000 tables result for the Netherlands

Magnitude tables

Magnitude table:

each cell value represents the sum of the scores of the respondents that fall into that cell

Frequency tables (I)

Frequency table:

each cell value represents the number of respondents that fall into that cell

Example: Dutch population, 1/1/2006

	Male	Female	Total
North	845 667	855 671	1 701 338
East	1 716 457	1 743 635	3 460 092
West	3 750 224	3 875 594	7 625 818
South	1 766 540	1 781 721	3 548 261
Total	8 078 888	8 256 621	16 335 509

Frequency tables (II)

Cell value not sensitive

Spanning variables:

- identifying
  - (Region, gender, type of business,…)
- sensitive
  - (Sexual behaviour, criminal offence, …)

Frequency tables (III)

(Spanning) variables:
one sensitive
remaining identifying

Example: number of shipowners

	Environmental offence		
Region	Yes	No	Total
…			
A	9	0	9
...			

Frequency tables (IV)

Group disclosure:

All shipowners in region A committed an environmental offence

Conclusion:

Not all respondents should score on a sensitive category

Frequency tables (V)

Example, continued

number of shipowners

	Environmental offence		
Region	Yes	No	Total
…			
B	14	2	16
…			

Frequency tables (VI)

Still:

non-offensive shipowners know quite surely that all other shipowners in region B committed an environmental offence

Conclusion:

There should not be too many respondents that score on a sensitive category

Frequency tables (VII)

Possible criterion:

Fraction of respondents that score on a sensitive category should be less than *p*%

to increase the uncertainty

E.g., $p = 40$

Frequency tables (VIII)

Example, continued

number of shipowners

	Environmental offence		
Region	Yes	No	Total
…			
C	1	1	2
…			

Frequency tables (IX)

Still:

Non-offensive shipowner knows that the other one committed an environmental offence

Possible criterion:

If respondents score on a sensitive category, at least *n* respondents should score on *non*-sensitive categories

Frequency tables (X)

Example, continued

number of shipowners

	Environmental offence		
Region	Yes	No	Total
…			
D	1	9	10
...			

Non-offenders now do not know *which* other shipowner committed the offence

Linked tables (I)

Tables sharing cells

- Example: Consider the tables Gender $\times$ Municipality and Gender $\times$ Provinces, the marginals of first table are the interior of the second table

Tables that can be considered to be parts of a higher dimensional table

Linked tables (II)

Number of booksellers: Gender × Region × Criminal record

Table 1

	Amsterdam	Rotterdam	Total
Male	21	12	33
Female	16	19	35
Total	37	31	68

Table 2

(Criminal record)	Yes	No	Total
Male	23	10	33
Female	8	27	35
Total	31	37	68

Table 3

(Criminal record)	Yes	No	Total
Amsterdam	11	26	37
Rotterdam	20	11	31
Total	31	37	68

Linked tables (III)

Number of booksellers: Gender × Region × Criminal record

Denote cell values of three dimensional table by x_{GRC} where

G : *M* (= Male)

F (= Female)

R : *Am* (= Amsterdam)

Ro (= Rotterdam)

C : *Y* (= Criminal record Yes)

N (= Criminal record No)

Linked tables (IV)

Number of booksellers: Gender × Region × Criminal record

Equalities can be derived:

E.g.,

	Amsterdam	Rotterdam	Total
Male	21	12	33
Female	16	19	35
Total	37	31	68

Male Booksellers in Amsterdam =
Male Booksellers in Amsterdam with Criminal Record Yes +
Male Booksellers in Amsterdam with Criminal Record No

i.e., $21 = x_{MAmY} + x_{MAmN}$

Linked tables (V)

Number of booksellers: Gender × Region × Criminal record

Equations following from Table 1:

$$x_{MAmY} + x_{MAmN} = 21$$
$$x_{MRoY} + x_{MRoN} = 12$$
$$x_{FAmY} + x_{FAmN} = 16$$
$$x_{FRoY} + x_{FRoN} = 19$$

Equations following from Table 2:

$$x_{MAmY} + x_{MRoY} = 23$$
$$x_{FAmY} + x_{FRoY} = 8$$
$$x_{MAmN} + x_{MRoN} = 10$$
$$x_{FAmN} + x_{FRoN} = 27$$

Equations following from Table 3:

$$x_{MAmY} + x_{FAmY} = 11$$
$$x_{MAmN} + x_{FAmN} = 26$$
$$x_{MRoY} + x_{FRoY} = 20$$
$$x_{MRoN} + x_{FRoN} = 11$$

Linked tables (VI)

Number of booksellers: Gender × Region × Criminal record

Solving these equations with assumptions

$$x_{GRC} \geq 0$$

$$x_{GRC} \text{ integer}$$

we get

$$x_{MAmY} = 11 \qquad x_{MAmN} = 10$$

$$x_{MRoY} = 12 \qquad x_{MRoN} = 0$$

$$x_{FAmY} = 0 \qquad x_{FAmN} = 16$$

$$x_{FRoY} = 8 \qquad x_{FRoN} = 11$$

Hierarchical tables (I)

Hierarchical tables: special case of linked tables

One or more of spanning variable are hierarchic, i.e. its categories contain several (sub)totals

E.g.: region (province/county/district/municipality)
branch of economic activity (NACE)

Hierarchical tables (II)

Groningen	21
Friesland	x
Drenthe	23
Overijssel	27
Gelderland	41
Flevoland	x
Utrecht	32
Noord-Holland	54
Zuid-Holland	67
Zeeland	38
Noord-Brabant	44
Limburg	39
The Netherlands	417

Hierarchical tables (III)

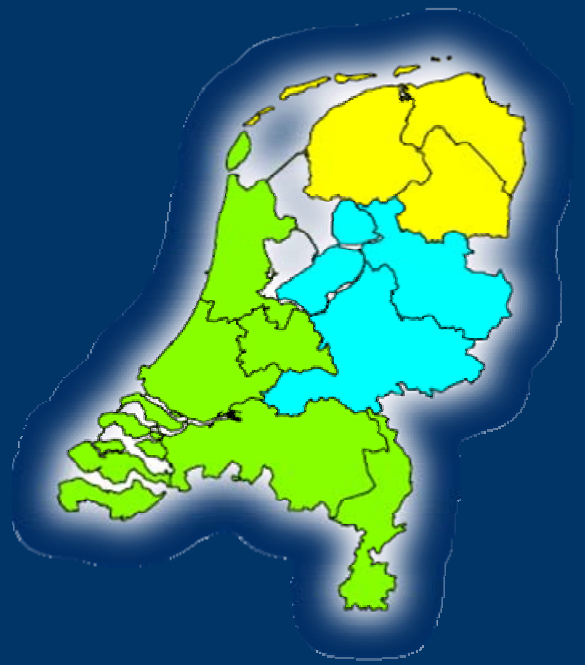
Groningen	21
Friesland	x
Drenthe	23
Overijssel	27
Gelderland	41
Flevoland	x
Utrecht	32
Noord-Holland	54
Zuid-Holland	67
Zeeland	38
Noord-Brabant	44
Limburg	39
The Netherlands	417

North	63
East	80
South	83
West	191
The Netherlands	417

Hierarchical tables (IV)

Groningen	21
Friesland	19
Drenthe	23
Overijssel	27
Gelderland	41
Flevoland	x
Utrecht	32
Noord-Holland	54
Zuid-Holland	67
Zeeland	38
Noord-Brabant	44
Limburg	39
The Netherlands	417

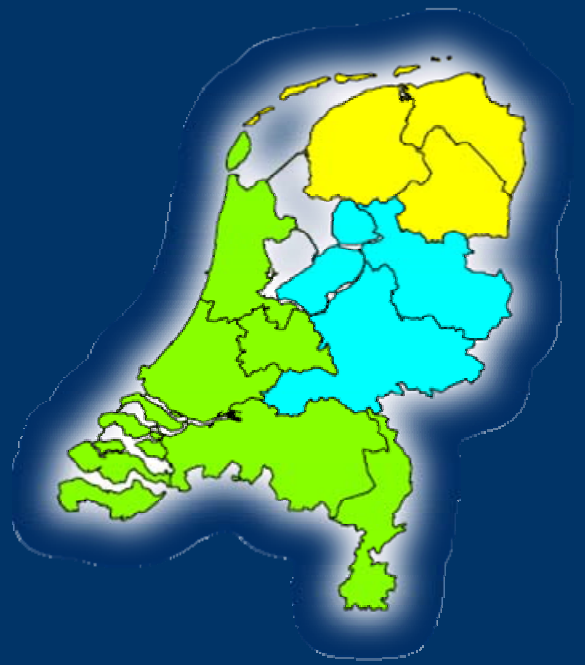
North	63
East	80
South	83
West	191
The Netherlands	417



Hierarchical tables (V)

Groningen	21
Friesland	19
Drenthe	23
Overijssel	27
Gelderland	41
Flevoland	12
Utrecht	32
Noord-Holland	54
Zuid-Holland	67
Zeeland	38
Noord-Brabant	44
Limburg	39
The Netherlands	417

North	63
East	80
South	83
West	191
The Netherlands	417



Microdata availability (I)

One percent samples for three years (1960, 1971 and 2001)
IPUMS (Integrated Public Use Microdata Series):
http://www.ipums.org/international/index.html

Weighting to population totals

Protecting according to rules for public use microdata files with μ-ARGUS

Microdata sets for all three years available!
DANS (Data Archiving and Networked Services):
http://www.dans.knaw.nl/en/

Microdata availability (II)

Public use microdata files:

1. Microdata must be at least one year old
2. No direct identifiers or direct regional variables
3. Only 1 kind of indirect regional variables. Values of indirect regional variables sufficiently scattered. Each area should contain at least 200,000 persons in the target population and should consist of municipalities from at least six of the twelve provinces. No dominating municipality in any area.
4. At most 15 indirect identifiers
5. No sensitive variables

Microdata availability (III)

Public use microdata files (continued):

6. Sampling weights should not provide additional identifying information
7. Rule against spontaneous recognition: at least 200,000 individuals in the population for each category of an identifying variable
8. Another rule against spontaneous recognition: at least 1000 individuals in the population for each category in the crossing of two identifying variables
9. At least 5 households per combination of categories of household variables
10. Records should be in random order

Microdata availability (IV)

Microdata for remote analyses

- Remote execution:

  Scripts are sent (on line) to Statistics Netherlands and applied to the microdata; SDC is applied before returning the results

  (Expensive, only used for some important customers)

- Remote access:

  On-line access to anonymised microdata sets

  (As much detail as at the on-site facility, no travel costs)

The 2011 Census Round (I)

Four Census 2011 regulations:

– Regulation No 763/2008 on Population and Housing Censuses ('European Census Act')

– Regulation No 1201/2009 on variables and categories

– Regulation No 519/2010 on the hypercubes

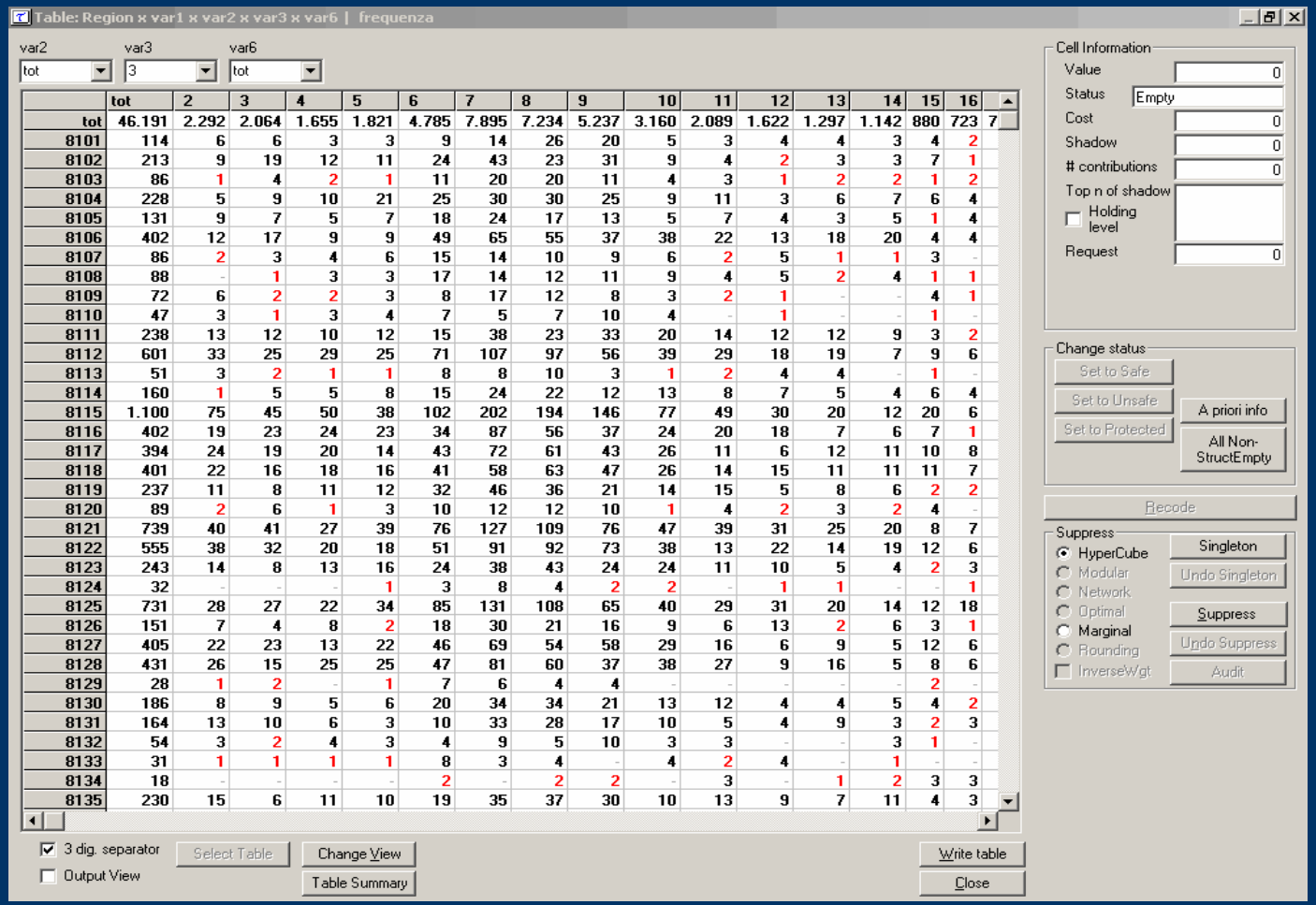
– Regulation No 1151/2010 on SDMX and the quality report

– Detailed output in high-dimensional tables (hypercubes)

– How well are Census tables with these variables filled?

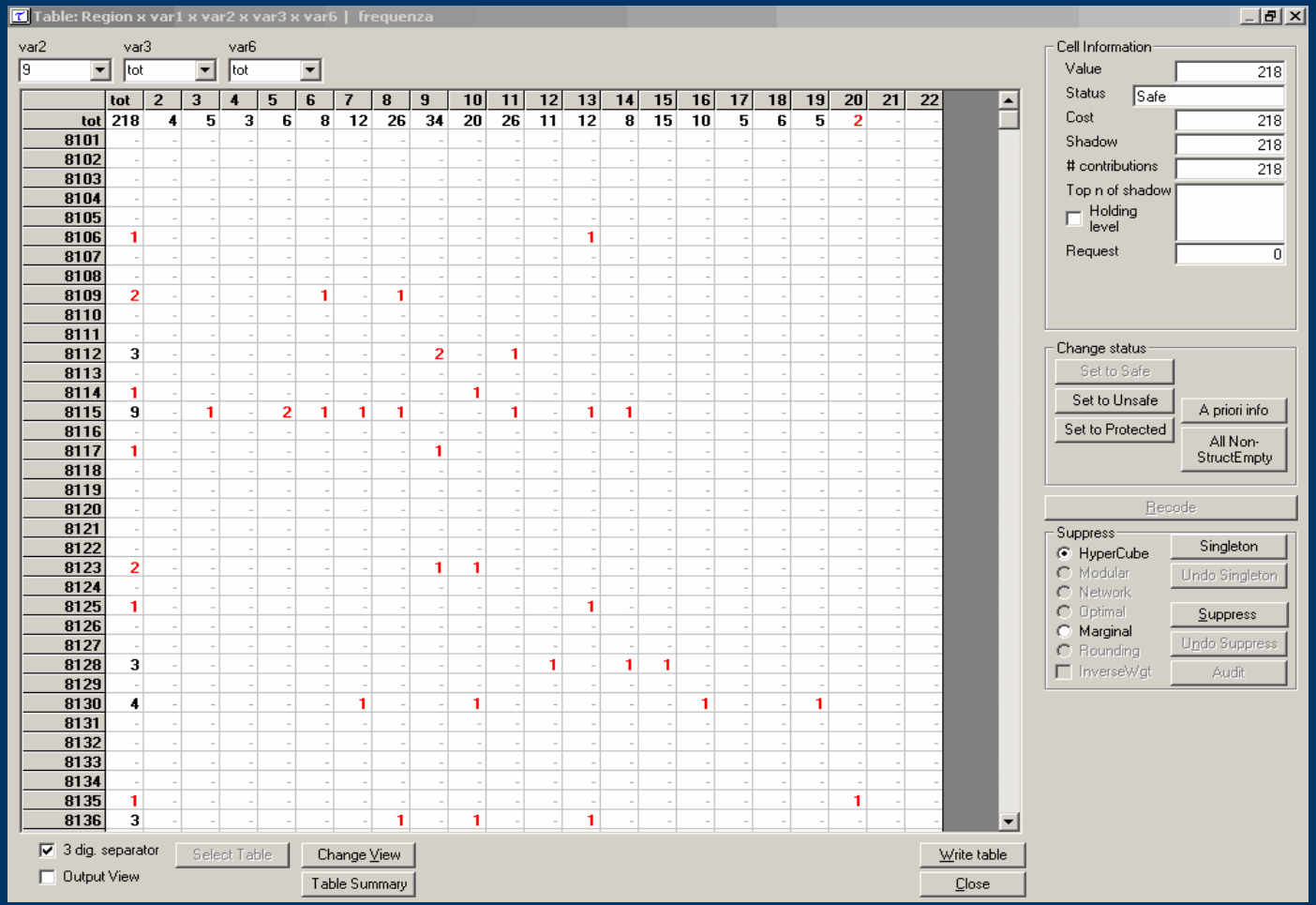
– Census Hub is not part of the four regulations but a means to transmit the data from the member states to Eurostat

The 2011 Census Round (II)



	tot	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
tot	46.191	2.292	2.064	1.655	1.821	4.785	7.895	7.234	5.237	3.160	2.089	1.622	1.297	1.142	880	723	7
8101	114	6	6	3	3	9	14	26	20	5	3	4	4	3	4	2	
8102	213	9	19	12	11	24	43	23	31	9	4	2	3	3	7	1	
8103	86	1	4	2	1	11	20	20	11	4	3	1	2	2	1	2	
8104	228	5	9	10	21	25	30	30	25	9	11	3	6	7	6	4	
8105	131	9	7	5	7	18	24	17	13	5	7	4	3	5	1	4	
8106	402	12	17	9	9	49	65	55	37	38	22	13	18	20	4	4	
8107	86	2	3	4	6	15	14	10	9	6	2	5	1	1	3	-	
8108	88	-	1	3	3	17	14	12	11	9	4	5	2	4	1	1	
8109	72	6	2	2	3	8	17	12	8	3	2	1	-	-	4	1	
8110	47	3	1	3	4	7	5	7	10	4	-	1	-	-	1	-	
8111	238	13	12	10	12	15	38	23	33	20	14	12	12	9	3	2	
8112	601	33	25	29	25	71	107	97	56	39	29	18	19	7	9	6	
8113	51	3	2	1	1	8	8	10	3	1	2	4	4	-	1	-	
8114	160	1	5	5	8	15	24	22	12	13	8	7	5	4	6	4	
8115	1.100	75	45	50	38	102	202	194	146	77	49	30	20	12	20	6	
8116	402	19	23	24	23	34	87	56	37	24	20	18	7	6	7	1	
8117	394	24	19	20	14	43	72	61	43	26	11	6	12	11	10	8	
8118	401	22	16	18	16	41	58	63	47	26	14	15	11	11	11	7	
8119	237	11	8	11	12	32	46	36	21	14	15	5	8	6	2	2	
8120	89	2	6	1	3	10	12	12	10	1	4	2	3	2	4	-	
8121	739	40	41	27	39	76	127	109	76	47	39	31	25	20	8	7	
8122	555	38	32	20	18	51	91	92	73	38	13	22	14	19	12	6	
8123	243	14	8	13	16	24	38	43	24	24	11	10	5	4	2	3	
8124	32	-	-	-	1	3	8	4	2	2	-	1	1	-	-	1	
8125	731	28	27	22	34	85	131	108	65	40	29	31	20	14	12	18	
8126	151	7	4	8	2	18	30	21	16	9	6	13	2	6	3	1	
8127	405	22	23	13	22	46	69	54	58	29	16	6	9	5	12	6	
8128	431	26	15	25	25	47	81	60	37	38	27	9	16	5	8	6	
8129	28	1	2	-	1	7	6	4	4	-	-	-	-	-	2	-	
8130	186	8	9	5	6	20	34	34	21	13	12	4	4	5	4	2	
8131	164	13	10	6	3	10	33	28	17	10	5	4	9	3	2	3	
8132	54	3	2	4	3	4	9	5	10	3	3	-	-	3	1	-	
8133	31	1	1	1	1	8	3	4	-	4	2	4	-	1	-	-	
8134	18	-	-	-	-	2	-	2	2	-	3	-	1	2	3	3	
8135	230	15	6	11	10	19	35	37	30	10	13	9	7	11	4	3	

The 2011 Census Round (III)



	tot	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
tot	218	4	5	3	6	8	12	26	34	20	26	11	12	8	15	10	5	6	5	2	-	-
8101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8102	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8103	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8104	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8105	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
8106	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
8107	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8108	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8109	2	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8110	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8111	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8112	3	-	-	-	-	-	-	-	2	-	1	-	-	-	-	-	-	-	-	-	-	-
8113	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8114	1	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
8115	9	-	1	-	2	1	1	1	-	-	1	-	1	1	-	-	-	-	-	-	-	-
8116	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8117	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
8118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8119	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8120	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8121	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8122	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8123	2	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-
8124	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8125	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
8126	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8127	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8128	3	-	-	-	-	-	-	-	-	-	-	1	-	1	1	-	-	-	-	-	-	-
8129	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8130	4	-	-	-	-	-	1	-	-	1	-	-	-	-	-	1	-	-	1	-	-	-
8131	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8135	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
8136	3	-	-	-	-	-	-	1	-	1	-	-	1	-	-	-	-	-	-	-	-	-

The 2011 Census Round (IV)

Questions:

– How to prevent disclosure of individual sensitive information?

– Which (categories of) variables are sensitive?

– What protection measures are feasible (preferably European wide co-ordinated)?

– To prevent recalculations from marginal totals: are secondary suppressions (across a hypercube and between hypercubes) with τ-ARGUS allowed?

– What about aggregation problems when cells are suppressed?

– Do we have other protection methods (rounding microdata, rounding tables, record swapping, additive noise)?

– What about the consistency of rounded data?

These issues have to be worked out further!

Thank you for your attention!

Time for questions and discussion