State Statistical Institute Berlin-Brandenburg

SAFE – a method for anonymising the German Census

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality

(Tarragona, Spain)

26-28 October 2011

Jörg Höhne

SAFE - a pretabular method (1)

SAFE creates an anonymous micro data file.

What are anonymous micro data?

- Micro data are anonymous (confidential), “if they cannot be matched to the concerned person.” (German law of Statistics §16).
- Identification can be avoided by ambiguous records in the micro data file.
- SAFE - micro data files are confidential because of ambiguity in the record set.

SAFE - a pretabular method (2)

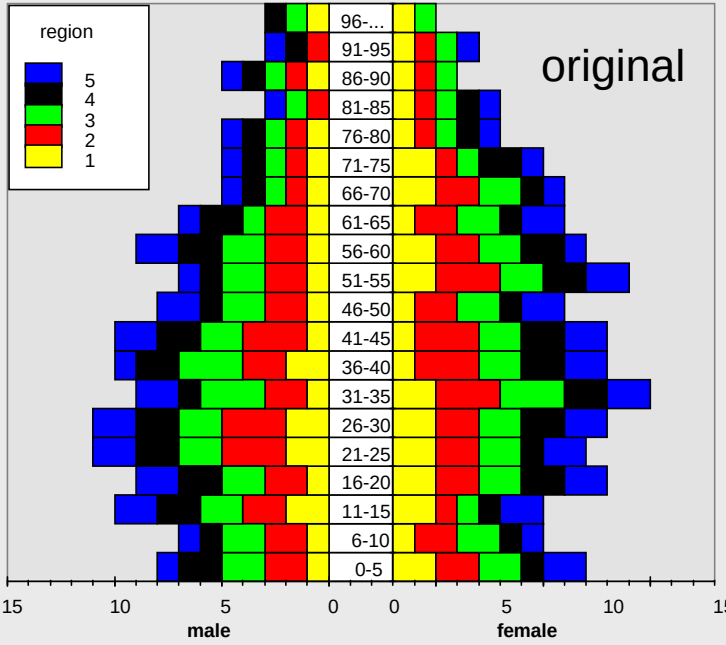
Advantages:

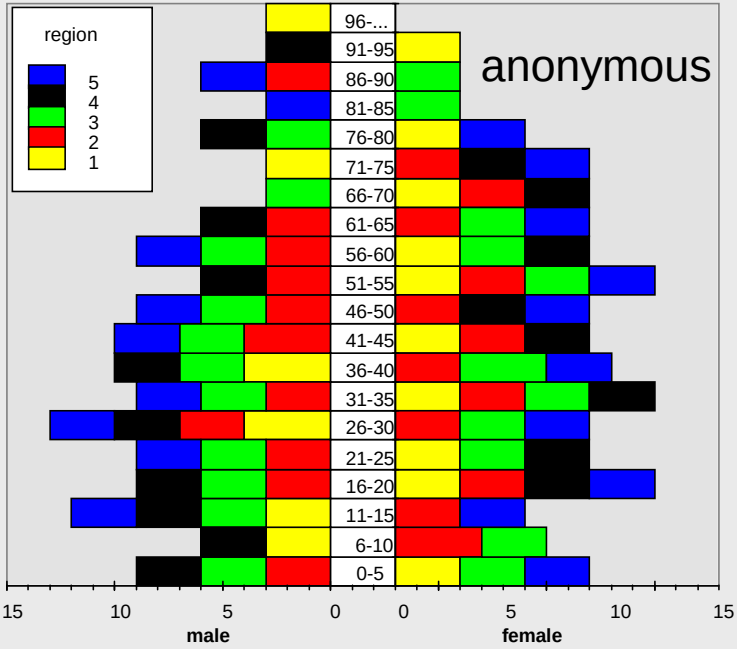
- Solution in one step
- Analysis can reidentify only the anonymous micro data file
- All analysis based on the same source and are consistent
- No cell suppression (primary or secondary) necessary

Disadvantages:

- Change from cell suppression to uncertainty in cell values
- New interpretation of tables
- No easy extension of analysis to not controlled tables
- Calculation effort may be relatively great

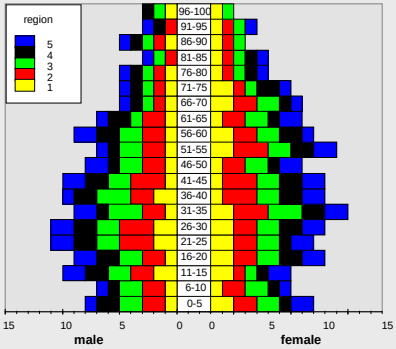
Idea of SAFE - solution (1)

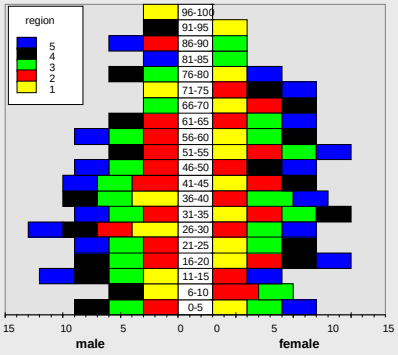


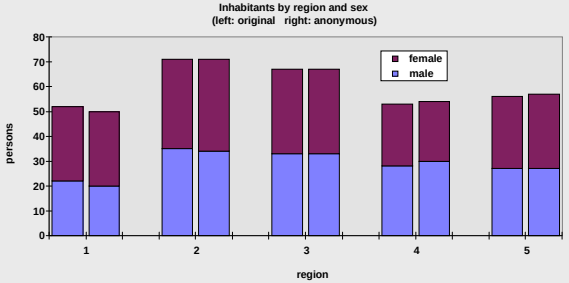


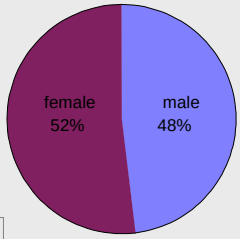
- only triplets
- data attacks (reidentification) lead to more than two objects

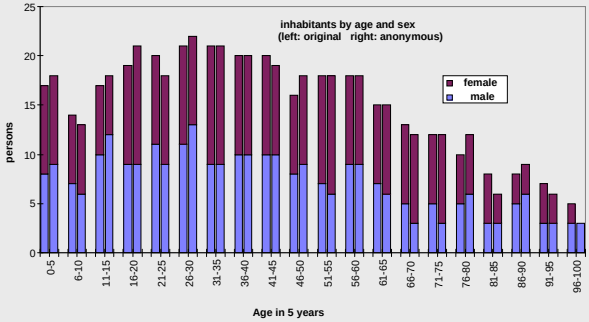
Idea of SAFE - solution (2)











- no real data, only triplets
- but the analysis should be as similar as possible to the original micro data file

Mathematical model of SAFE (1)

$$\min_y \left(\max_{i \in I} \left(\left| d_i \right| - w_i \right) \right)$$

subject to

$$Ay = a + d$$

$$y_j \in \{0,3,4,...\}$$

with:

- y - vector of frequencies of category combinations in the micro data
- a - vector of original frequencies in controlled table cells
- A - linear relation matrix
- d - vector of deviation trough tabulation of anonymised micro data instead of original data d_i as deviation in table cell j
- w - vector of weights for table cells

Adaption for the census

different units:

persons, housings, buildings are different units of the census with hierarchical dependencies.

1. Splitting the micro data set in different variable blocks
   (persons: region, age, sex, profession, ...
   housing: region, heating (oil, gas,...), bath, ...
   building: construction year, size, ...)

2. Counting variable for person, housing, building. Create control-table with counting variables.

Persons were splitted from housing and building information.

Housing and building in one data set with different counting variables.

SAFE – tests with census 1987 West Germany (1)

micro data file for persons register

Records (persons):	63 202 834
Variables:	21 (28)
Controlled tables:	430
Table cells:	10 219 270
Maximal deviation:	9

table cell by size		number of table cells	maximal deviation	Mean in deviation
from	to			
1	9	3 787 507	8	1.64
10	99	3 246 990	8	2.22
100	999	2 022 537	8	2.29
1 000	9 999	887 972	9	2.33
10 000	99 999	233 991	9	2.38
100 000	999 999	37 507	8	2.45
1 000 000	9 999 999	2 727	8	2.61
10 000 000	or more	39	4	1.31

SAFE – tests with census 1987 West Germany (2)

micro data file for persons register

Deviation in table cell	number of table cells by table dimension …				ratio of cells with maximal … deviation in table cell by table dimension …			
	1	2	3	4 or 5	1	2	3	4 or 5
0	7 730	41 472	429 058	764 471	56,7	12,0	11,9	12,2
1	4 600	86 008	1 053 292	2 034 970	90,4	36,8	41,1	44,8
2	1 210	73 607	840 969	1 523 732	99,2	58,0	64,4	69,2
3	104	60 243	594 461	954 158	100,0	75,4	80,9	84,4
4	-	47 035	411 153	608 160	100,0	89,0	92,3	94,1
5	-	26 310	202 569	274 847	100,0	96,5	97,9	98,5
6	-	10 077	66 662	81 715	100,0	99,5	99,7	99,8
7	-	1 823	8 954	9 489	100,0	100,0	100,0	100,0
8	-	80	155	154	100,0	100,0	100,0	100,0
9	-	2	-	-	100,0	100,0	100,0	100,0
10	-	-	-	-	100,0	100,0	100,0	100,0
Other all	13 644	346 657	3 607 273	6 251 696				

SAFE – tests with census 1987 West Germany (1)

micro data file for housing and building

Records (housing):	26 624 252
Variables:	7 (15)
Controlled tables:	119
Table cells:	2 906 234
Maximal deviation:	6

table cell by size		number of table cells	maximal deviation	Mean in deviation
from	to			
1	9	1 358 071	5	1.38
10	99	915 183	6	1.43
100	999	471 606	6	1.38
1 000	9 999	130 859	6	1.39
10 000	99 999	26 334	5	1.44
100 000	999 999	3 838	5	1.46
1 000 000	9 999 999	343	5	1.44

SAFE – tests with census 1987 West Germany (2)

micro data file for housing and building

Deviation in table cell	number of table cells by table dimension …				ratio of cells with maximal … deviation in table cell by table dimension …			
	1	2	3	4 or 5	1	2	3	4 or 5
0	4 391	37 980	155 866	278 551	44.5	17.0	16.4	16.2
1	5 250	84 913	420 560	752 125	97.7	55.0	60.5	59.9
2	223	64 154	260 389	471 625	100.0	83.8	87.9	87.3
3		27 290	86 197	165 421	100.0	96.0	96.9	96.9
4	-	8 555	27 912	51 211	100.0	99.8	99.9	99.9
5	-	425	1 195	1 996	100.0	100.0	100.0	100.0
6	-	-	-	5	100.0	100.0	100.0	100.0
7	-	-	-	-	100.0	100.0	100.0	100.0
Other all	13 644	346 657	3 607 273	6 251 696				

Interpretation of results

SAFE solutions are:

1. Like “noise” added to table cells.
2. Maximal deviation is known and documented.
3. Relative deviation in cells decreases in greater table cell values.
4. Missing combinations are unlikely but not sure not existing.
5. Unique combinations in table row (line or column) do not allow an information gain (group disclosure problem) through not sure uniqueness in the original data.
6. Good preservation of structure in the data. No missing information through complementary cell suppression.

Thank you for your attention!