SAFE – a method for anonymising the German Census

Dr. Jörg Höhne*

* State Statistical Institute Berlin-Brandenburg, 10315 Berlin Alt-Friedrichsfelde 60, Germany, Joerg.Hoehne@Statistik-bbb.de

Abstract: In Germany the Census 2011 will be provided as a register based census with an additional sample survey to verify the register and include information that is not in the registers used.
To secure the confidentiality, additivity and consistency of the published tables, a pre tabular anonymisation method should be used. The preferred method among the tested ones is a variant of microaggregation and called SAFE. In addition to a short description of the method, the paper focuses on test results with real data of the census of West-Germany in 1987 and an outlook to the implementation for the census 2011.

1 The SAFE – algorithm as a method for anonymising the census

The German census 2011 will partly be register based, and partly be the outcome of a sample survey. This leads of course to limitations in the amount of detail of tables that can sensibly be released, as compared to a full census. Nevertheless, a huge amount of tabular output is going to be published. Publication of tables will to a major part be pre-planned, but there is also a demand for flexible, user driven release of tabular data. In preparation for the German census 2011, a comparative study of several anonymisation methods for census frequency counts was conducted. For the size of planned publications and the dependencies and complexities in the planned table sets non-perturbative methods like cell suppression did not seem to be a good choice. With cell suppression methods a considerable disclosure risk due to incomplete coordination of cell suppression patterns across tables remained or a reduction in the amount of published information was necessary. In addition cell suppression methods revealed to limit flexible tabulation due to the increasing complexity of suppression patterns. Perturbation methods have the advantage that they introduce ambiguity into cells. Therefore the linear dependencies between table cells or specific structure of zero cells, which could be used for attribute disclosure, can no longer be employed for disclosure attempts because the table cells contain noise from the perturbation methods.

Post tabular random perturbation methods suggested in the literature, e.g. Shlomo and Young (2008) or Fraser and Wooton (2006) are simple algorithms for anonymisation but do not lead to consistent results for tables which can be generated by aggregation of two or more independently perturbed tables. The problem of consistency can be solved by the CTA-algorithm. But by solving the consistency in

tables, additivity of tables will be lost. Consistent table cells do not have to be additive between different tables or linear dependencies.

Another possibility is to use perturbation methods in a pre tabular approach. If the micro data file is changed, the results of tabulations are additive and consistent because they are calculated from an identical anonymous source. One pre tabular method is SAFE. It was developed and is in use in the State Statistical Institute Berlin-Brandenburg. The program was tested and will be used for the anonymisation of the German census tables.

In this paper the test results for the SAFE method tests are presented. The methodology of SAFE is briefly described, as far as it is relevant for an application to protect tabulations of population census counts data. The method was tested on the census dataset of the German Census 1987. The micro data file of the census 1987 was the most realistic dataset applicable for the anonymisation tests. The dataset has a high comparability regarding the variable set and is a realistic 75% subset of the actual German dataset as reunification increased the population total. So the dataset allows a good estimation on the calculation time and the quality of the results that can be expected when anonymising the CENSUS 2011 for the whole of Germany.

2 The SAFE method

SAFE is a variant of microaggregation. It can be applied to a data set that consists of categorical variables only. Starting point for the method is a micro data file where all variables are recoded to have the highest degree of detail that is intended to be used in publication. Imagine a variable like age, where perhaps data are collected so that for each person age could be deduced down to the level of age in months, but publications should offer data at most by age in years. Then the variable would be recoded to the level of age in years.

The basic idea of the method is to turn this data set (with N categorical variables n_i ($i=1,\ldots,N$)) into a data set, in which on each of the $n_1 * n_2 * \ldots * n_N$ theoretical combinations of categories at least three records or none exists. In conclusion no tabulation of the dataset can result in confidential table cell counts of 1 or 2, because at least 3 identical objects exist in the micro data set.

With respect to data quality, the method aims to preserve as many as possible cell counts in a pre-defined set of tables. For those tables (called ‘controlled tables’), the method yields results that are in some sense ‘optimal’. If any other table is derived from the perturbed data set, differences between original counts and those computed on the perturbed data set can be much larger than differences that arise for the controlled tables. The experience is that the method is usually able to achieve a maximum deviation in cell counts between 4 and 10 for a sensibly defined set of controlled tables.

2.1 The SAFE mathematical model

The algorithm computes a heuristic solution for the problem of minimizing the maximum absolute deviation between true and perturbed cell values in the controlled tables.

Instances are defined by the following parameters:

- A set of linear relations $Ay=a$ defining the table cells of the controlled tables as sums of cells of an elementary table consisting of all combinations of categories of all variables in the micro data set.
- Vector a, $a=(a_i,\ i \in I)$ denotes the original frequencies presented in the controlled tables, and vector y, $y=(y_j, j=1,\ldots,N)$ the entries (e.g. frequencies of category combinations) of the cells of the elementary table resulting from the micro data. In a valid solution, vector y does not contain any entries of 1 or 2.
- Vector w of weights associated to perturbations of the table cells of the controlled tables. For example, we may want to allow larger perturbations for larger cells, or avoid them for cells that are rated "highly important".

The objective of the model is to minimize the maximum entry of vector $d=(d_i,\ i \in I)$, $d_i \in ¢$, denoting the deviations of original and perturbed cell counts in the controlled tables. With these definitions, broadly, the model is as follows:

Solve the problem

$$\begin{aligned} &\min_y \quad \max_{i \in I} \left(|d_i| - w_i \right) \\ &\textit{subject to} \quad Ay = a + d \\ &\qquad\qquad y_j \in \{0,3,4,5,\ldots\} \quad j = 1,\ldots,N \end{aligned} \tag{1}$$

This statement of the problem resembles a huge non-linear integer optimization problem which is computationally intractable. Therefore, an efficient heuristic algorithm has been developed that gives near optimal solutions at reasonable expense of computer resources.

Beginning with the (infeasible) initial solution given by $d=0$, i.e. where cell values are kept at their original value, a first feasible solution is obtained. This solution is optimized later on.

2.1.1 A first feasible solution

In addition to the above parameters, we define now

- Vector $b=(b_i,\ i \in I)$, $b_i \in B$ of bounds for maximum allowed deviations. In practice, B consists of two values only, one stating the maximum deviation to be allowed for cells in one-dimensional tables, the other one stating the maximum allowed deviation for the other cells, e.g. cells defined as cross-combination of categories of two or more variables,

- Vector $x=(x_j, j=1,\ldots,N)$, $x_j \in \{0,1\}$ is 1, if elementary table cell j is "unsafe", e.g. if $y_j \in \{1,2\}$ and 0 otherwise.

The problem to be solved is

$$\begin{aligned} &\min_y \quad \sum_{j=1,\ldots,N} x_j \\ &subject\ to \quad |Ay-a| < w+b \\ &\qquad y_j \in \{0,1,2,3\ldots\} \\ &x_j = 1; \quad if \quad y_j \in \{1,2\} \\ &x_j = 0; \quad if \quad y_j \notin \{1,2\} \quad j=1,\ldots,N \end{aligned} \tag{2}$$

A feasible solution is obtained when the objective function is zero.
Minimizing the number of "unsafe" frequencies, using a heuristic, the algorithm step by step changes critical frequencies of 1 and 2 into uncritical frequencies 0,3,4,… If the process stagnates, the statement of the problem is modified automatically by increasing the vector of bounds b, e.g. $b=b+1$.

2.1.2 Optimizing the solution

Once a feasible solution has been obtained, the method will seek to improve the solution by reducing the maximum allowed perturbation, e.g. b and eventually w. Usually the number of cells where the deviation is identical or near-identical to the respective bound is relatively small. In the optimization step, after changing (reducing) b or w, some of the constraints in model (2) will be violated. Accordingly, we define now

- Vector $z=(z_i, i \in I)$, $z_i \in \{0,1\}$ is 1, if for controlled tables cell i the bound constraint of model (2) is violated and 0 otherwise.

The algorithm derives a heuristic solution to

$$\begin{aligned} &\min_y \quad \sum_{i \in I} z_i \\ &subject\ to \quad |Ay-a| - (w+b) < z \\ &\qquad y_j \in \{0,3,4\ldots\} \end{aligned} \tag{3}$$

If a solution is obtained where $\sum_{i \in I} z_i = 0$, the constraints will be tightened further (i.e. reduce b or w), and model (3) will be solved again. This step is repeated until either an expected level of optimality (in the bounds) is reached, or further attempts seem to be rather unpromising.

3 Application to census 1987

A special quality of census datasets is that it is not a “normal” statistic in a way that it is an analysis of one type of statistical object. Census datasets are analyses of mixed statistical objects. Most tables generate from a census are tabulations of persons. Additionally there exist also summations of households, flats and buildings (as statistical units). For the micro data set in the above model, there are two ways to store the hierarchical dependencies.

The first way is to block the dataset in logical blocks. Variables which are only characteristics of persons can be stored independent from characteristics of buildings or flats. There are going to be two independent micro data files, which can be independently anonymised with the above model. Tabulations on persons would be calculated from the micro data file on persons. Tabulations on buildings from the micro data file on buildings.

The second way is to describe the greater hierarchies as a subset of the finer one. Household are summations of one ore more persons which live together. Analogically one ore more households live in a flat and one ore more flats are situated in a building. Defining subgroups means that the finest dataset will also store the variables of the higher hierarchy and additionally will be extended by a variable which can be used as a “counting variable” for the higher hierarchy. If flats and buildings should be stored in the same data file, every first flat in the building will get the value 1 and the other flats the value 0 as “counting variable building”. Tabulations on flats can be created by counting all rows of the datasets in combination with the interesting variables (like e.g. region). Tabulations on buildings can be created by summarizing the “counting variable building” in combination with other variables of interest.

For the German tests of anonymisation methods, both ways of data storage were used. Information about persons is in the German Census 2011 that is collected as a register based census. Registers of inhabitants, social security register and other sources are collected, matched together and integrated into a great file with information on inhabitants. The information of flats and buildings is collected through a survey. The two different ways of getting the information is used to create two independent micro data files. The information on flats and buildings (investigated through a survey) will be stored together and anonymized in one step.

In the test scenario, an anonymisation was calculated for the personal data. A micro data file for 63.2 million people with 21 variables was created. For the 21 variables, hierarchies for aggregated tabulation were defined (e.g. regions in the hierarchies of city, district, county, states, and the variable age in the hierarchies of 1-, 5- and 10-year groups, additionally under 18, 18-65 and other 65). Including hierarchies there were 28 variables. For this set of variables and variable-aggregations, a set of 430 controlled tables was defined. The controlled tables describe a combination of

variables which are crossed to generate one dimensional and up to five dimensional tables over the whole dataset. For the publication a table of age by sex is to be calculated for each city (over 8 000 cities). For SAFE, only one controlled table is defined as a table of region (with 8 000 categories at city-level) by age by sex. All 430 tables together contain 10.2 million cells, which are controlled to diverge at most by the maximal deviation allowed.

The second example was an anonymisation of the German housing census (flats and buildings) of 1987. The micro data file consists of 26.6 million flats with 11 variables (20 variables when including hierarchies). The micro data file was extended by a counting variable for the building as explained above. As controlled tables where defined a set of 119 tables with 2.9 million controlled cells.

The weighting function in both models was defined as

$$w_i = \text{int}(\log_{10}(a_i)) \qquad ; i \in I \tag{4}$$

with int(..) as integer- function. Therefore bigger cell values will be allowed an additional deviation w_i depending on the cell size with

table cell range values	w_i
1 – 9	0
10 – 99	1
100 – 999	2
1 000 – 9 999	3
… – …	…

Table 1 Additionally allowed deviation w_i in anonymous cell values

To find the first feasible solution for model (2), the starting parameter vector B was defined as $b_i=3$ for table cells in one-dimensional tables and $b_i=5$ for table cells in more-dimensional tables.

To solve the model on small computers the micro data file of persons was splitted by region into 4 subsets. On bigger machines is it possible to solve it in one run. The model (2) was solved and stopped with a maximal deviation of $b_i=4$ for table cells in one-dimensional tables and $b_i=7$ for table cells in higher dimensional tables.

table cell range values	w_i *on correction step …*					
	0	*1*	*2*	*3*	*…*	*7*
1 – 9	0	0	0	0		0
10 – 99	1	0	0	0		0
100 – 999	2	1	0	0		0
1 000 – 9 999	3	2	1	0		0
… – …	…	…	…	…		0

Table 2 Deviation w_i for the correction steps

The model (3) for optimizing the solution was then solved with the changed weighting parameters (see table 2). On correction step 7 is $w_i=0; \forall i \in I$. Through the correction step (model 3) the allowed higher deviation for bigger table cells could be dropped. In result the quality can be independently interpreted of the cell size.

3.1 Results on the personal micro data

For the micro data file of personal data, the maximum deviation over all controlled table cells after the anonymisation was 9. The deviation over the number of combined variables is displayed in table 3.

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				

Table 3 Distribution of deviations in controlled table cells personal data

table cell by size		number of	maximal
from	to	table cells	deviation
1	9	3 787 507	8
10	99	3 246 990	8
100	999	2 022 537	8
1 000	9 999	887 972	9
10 000	99 999	233 991	9
100 000	999 999	37 507	8
1 000 000	9 999 999	2 727	8
10 000 000	or more	39	4

Table 4a Distribution of deviations in controlled table cells by size for personal data

One-dimensional tabulations have a distance of maximal 3 between original and anonymised table cells. Only 104 of the 13 644 table cells calculated in one-dimensional analyses of the dataset have a distance of 3 between original and

anonymised table frequency. For any more-dimensional tabulation, the table cell value can be securely interpreted if a distance of 9 is acceptable. If a distance of at maximum 5 is acceptable in a table cell the risk of 'useless information' is lower then 1.8%.

table cell by size		ratio of table cells with maximal deviation of … in percent									
from	to	0	1	2	3	4	5	6	7	8	9
1	9	10,8	53,4	81,3	92,9	98,2	99,8	100,0	100,0	100,0	100,0
10	99	13,0	37,8	59,8	78,3	91,3	97,8	99,8	100,0	100,0	100,0
100	999	12,9	36,7	57,9	76,4	90,5	97,4	99,7	100,0	100,0	100,0
1 000	9 999	13,0	36,9	57,8	74,9	88,6	96,5	99,4	100,0	100,0	100,0
10 000	99 999	13,3	36,9	57,6	73,9	86,3	94,6	98,9	100,0	100,0	100,0
100 000	999 999	13,2	36,8	56,6	72,4	84,6	93,0	98,4	99,9	100,0	100,0
1 000 000	9 999 999	12,7	37,7	55,5	69,1	80,2	88,8	96,0	99,2	100,0	100,0
10 000 000	or more	25,6	64,1	82,1	97,4	100,0	100,0	100,0	100,0	100,0	100,0

Table 4b Distribution of deviations in controlled table cells by size for personal data

In the range of the 430 controlled tables are also many cells with low cell frequencies and so potential confidentiality problems. In the case of 1 or 2 in the original table cell counts after anonymisation the cells became 0 or 3. So for this cells exists a high probability to get changed by 1 or 2 (see table 4b). While the data attacker don't know for a zero-cell that it is a real not existing combination or it was deleted by the method this pre tabular method has a higher confidentiality level like a pure cell suppression. In the case of cell suppression a complementary suppression is necessary.

3.2 Results for a hierarchical micro data set of housing and buildings

For the micro data file of housing and building the model after anonymisation the maximal deviation in controlled tables was 6.

Deviation in table cell	number of table cells by table dimension …				ratio of cells with maximal deviation of … 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	9 864	223 317	952 119	1 720 934				

Table 5 Distribution of deviations in controlled table cells housing data

One-dimensional tabulations have a distance of maximal 2 between original and anonymised table cells. Only 5 of the 2 906 234 table cells calculated by analyses of the dataset by one variable have a distance of more than 5 between original and anonymised table frequency. For any more-dimensional tabulation a table cell can be sure interpreted if a distance of 6 is acceptable. If a distance of 4 is maximal acceptable in a table cell the risk of getting 'useless' information is lower then 0.2%.

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

Table 6a Distribution of deviations in controlled table cells by size housing data

table cell by size		ratio of table cells with maximal deviation of … in percent						
from	to	0	1	2	3	4	5	6
1	9	12,6	62,1	89,7	97,6	100,0	100,0	100,0
10	99	19,4	57,5	84,6	95,8	99,8	100,0	100,0
100	999	20,1	58,6	86,3	96,9	99,9	100,0	100,0
1 000	9 999	20,7	58,7	85,0	96,4	99,9	100,0	100,0
10 000	99 999	20,4	57,4	82,8	95,7	99,7	100,0	100,0
100 000	999 999	20,8	57,7	81,6	94,6	99,6	100,0	100,0
1 000 000	9 999 999	25,9	61,8	80,2	90,4	97,4	100,0	100,0

Table 6b Distribution of deviations in controlled table cells by size for housing data

The better results for the housing data in comparison to the personal data set seem to be a result of the smaller number of controlled tables. The lower the number of controlled tables the smaller is the maximal distance between original and anonymised table cell counts.

3.3 Concluding remarks

As there is no combination of response with a frequency less than 3 in the anonymous vector *y* (see model 1) all variable combinations can flexibly be analysed and will always result in secure tables with no cell counts of 1 or 2.

While the quality of results for controlled tables are good as inherent to and documented by the program, the tabulation of other variable combinations is not

controlled for. In not controlled table cells higher deviations are possible. If the original cell frequency is very small, there is a high probability that the table cell is changed to nearest cell count dividable by 3. This effect results in a non-symmetric bias for very small cell counts. For larger table cells, the deviation in not–controlled-for tables tends to a normal distribution. As shown in the comparative study by Gießing and Höhne (2010) the post-tabular methods result in smaller perturbations than SAFE. On the other hand, as a pre-tabular method SAFE preserves additivity and consistency, and is easier to implement in a flexible online table generator environment. The tests show that it is able to keep the maximum deviation in a set of pre-specified tables acceptably small. These are important properties and may be worth "less optimal" performance regarding data quality to some degree. While the perturbation caused by SAFE tends to be stronger than those caused by non-additive post-tabular approaches, the tests shows that they tend to be normally distributed, e.g. large deviations are relatively unlikely, even so for cells that are not contained in the set of pre-specified, controlled tables.

References

Castro, J. (2011). Extending controlled tabular adjustment for non-additive tabular data with negative protection levels, *Statistics and Operations Research Transactions*, vol. 35, no. 1.

Fraser, B., Wooton, J. (2006). A proposed method for confidentialising tabular output to protect against differencing, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302

Giessing, S., Höhne, J. (2010). Eliminating Small Cells From Census Counts Tables: Some Considerations on Transition Probabilities, in *J. Domingo-Ferrer and E. Magkos, eds., Privacy in Statistical Databases*, 52-*56*. New York: Springer-Verlag. LNCS 6344

Höhne, J. (2003a). SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung Statistischer Einzelangaben, in *Berliner Statistik - Statistische Monatsschrift 3/2003*, Statistisches Landesamt Berlin.

Höhne, J. (2003b). SAFE - a method for statistical disclosure limitation of microdata, paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Luxembourg, December 2003, available at www.unece.org/stats/documents/2003/04/confidentiality/wp.37.e.pdf

Leaver, V., (2009). Implementing a method for automatically protecting user-defined Census tables, paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Bilbao, December 2009, available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Shlomo, N., Young, C. (2008). Invariant Post-tabular Protection of Census Frequency Counts. In *J. Domingo-Ferrer and Y. Saygin, eds., Privacy in Statistical Databases, 77-89*. New York: Springer-Verlag. LNCS 5262