

Working Paper No. 37
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

**SAFE –
A METHOD FOR STATISTICAL DISCLOSURE LIMITATION OF MICRODATA**

Contributed paper

Submitted by the Berlin State Statistical Office, Germany¹

¹ Prepared by Jörg Höhne (j.hoehne@statistik-berlin.de).

SAFE - a method for statistical disclosure limitation of microdata

Jörg Höhne

Berlin State Statistical Office

supporting paper

During the last few years, a method for statistical disclosure limitation of microdata was developed and implemented at the Berlin State Statistical Office. The method is known under the name SAFE (Appel, G.C. 1994). The projected paper will explain the principles of the method and present some test results. The primary aim of developing this method was to have a tool for disclosure protection of multiple linked multidimensional tables. However, as it works on the microdata level, the method is also suitable for the protection of microdata.

The paper will cover the following aspects:

- Which problems of the statistical disclosure control should / can be solved by the method?
- What is the main idea of the method and which steps are taken?
- Short description of the method
- Empirical Results
- Outlook

1. Which problems of the statistical disclosure control should / can be solved by the method?

To ensure disclosure protection, the method creates a safe microdatafile in such a way that the following requirements of statistical confidentiality are fulfilled:

- Any tabulations of the safe microdata file must not contain any sensitive cells: cells with less than 3 contributions, or cells that would allow for inferential disclosure of individual respondent contributions because of dominance problems.
- Exact or inferential disclosure of individual respondent data on basis of differences between cell values should not be possible (secondary disclosure control).
- Consistency of varying tabulations obtained from the same safe file.
- Low information loss

2. What is the basic idea of the method and which steps are taken?

The basic concept of the method is, to create a microdatafile, which consists of records only where three or more records are identical. SAFE reduces the risk that an intruder trying to link records of the safe file to an external file is able to obtain unique matches. There will always be at least three identical records for each combination of the categories of the discrete variables. In addition to that, records are distorted because the values of the continuous variables are constructed as a mean of the original values. Because all tabulations build on the microdatafile, the results are consistent in

multiple tabulations.

The basic idea (creating groups of identical records) is resembles the concept of microaggregation. What is different from microaggregation is the principle used for grouping the data. Microaggregation methods group 'nearest' records, according to some measure of distance. SAFE, on the other hand, constructs groups considering the impact on multidimensional distributions.

3. Short description of the method

The general SAFE-model:

We denote the original microdatafile as matrix B^o . From this microdatafile we construct the set of all possible tables (resulting from the discrete variables, their categories and additive structure).

$T^o = F(B^o)$ vector of the results of the tabulations (for all tables) from the original microdatafile

$t_{p,q,m}^o = f_{p,q,m}(B^o)$ table cell presenting the value of continuous variable m for the combination q of the discrete variables in table p

After set up of all possible tabulations for the given set of discrete variables, we identify sensitive cells and define a protection interval around their original value. Tabulation results of a safe microdatafile must not lie within these interval. The intervals are given by the lower bound ($z_{p,q,m}^l$) and the upper bound ($z_{p,q,m}^u$).

A safe microdatafile B^a must satisfy the following requirements:

$$T^a = F(B^a)$$

where

$$t_{p,q,m}^a = f_{p,q,m}(B^a)$$

and

$$z_{p,q,m}^l \geq f_{p,q,m}(B^a) \vee z_{p,q,m}^u \leq f_{p,q,m}(B^a)$$

$$z_{p,q,m}^l = z_{p,q,m}^u = t_{p,q,m}^o$$

if the table cell relating to the value of continuous variable m for combination q in table p is not sensitive

T^a - vector of the results of all tabulations calculated from the anonymous microdatafile

$t_{p,q,m}^a = f_{p,q,m}(B^a)$ -table cell relating to the value of continuous variable m for combination q in protected table p

In addition to these conditions, the microdatafile B^a should consist of records only, where three or more records are identical. Microdatafiles fulfilling these conditions are considered feasible. Within the set of feasible microdatafiles B^a , we try to identify the one with a minimum distance between T^0 and T^a .

The procedure SAFE consists of the following steps:

1. create a safe solution for the discrete variables
2. match the original continuous variables to this solution
3. find a solution without violations of the primary sensitivity criteria
4. optimise the solution
5. build groups of three or more records while setting values to the mean.

3.1 Create a safe solution for the discrete variables

Starting point is the multidimensional frequency table.

Let:

X - vector of frequency x_i of combinations of categories of discrete variables $i ; i=1,2,\dots,n$
(where n = number of different combinations in the database)

R - vector of all table-frequencies (number of objects in each table cell) for all tables of a certain set of so called 'controlled tabulation tables'

$r_j ; j=1,2,\dots,k$ (k =number of table cells in all controlled tabulation tables)

A - matrix of structure of objects - $a_{ij} = 1$ - if cell j relates to combination i .
 $a_{ij} = 0$ - if cell j does not relate to combination j

The matrix A consists of blocks of unit-vectors in each block $A_t ; t=1,2,\dots,T$ (T - number of controlled tabulation tables)

$$A = \begin{Bmatrix} A_1 \\ A_2 \\ \vdots \\ A_T \end{Bmatrix} \quad \left(\begin{array}{l} a_{ij} = 1 \text{ - if cell } j \text{ relates to} \\ \text{combination } i \text{ in table } t \end{array} \right.$$

mit

$a_{ij} = 0$ - else

$$A_t = \begin{Bmatrix} \begin{array}{c|c|c} a_{111}=1 & a_{11j}=0 & a_{11n}=0 \\ a_{121}=0 & a_{12j}=0 & a_{12n}=0 \\ \hline a_{1i1}=0 & a_{1ij}=1 & a_{1in}=0 \\ \hline a_{1(m-1)1}=0 & a_{1(m-1)j}=0 & a_{1(m-1)n}=0 \\ a_{1m1}=0 & a_{1mj}=0 & a_{1mn}=1 \end{array} \end{Bmatrix}$$

Thus the relations between statistical objects and table cells are given by:

$$AX=R$$

For safe datafile $x_i \in \{0,3,4,5,\dots\}$ holds., e.g. all existing objects have a frequency of at least 3.

According to the definition above, the solution contain only existing combinations, because all x_i are real existing objects. To avoid that each record in the solution relates to a real existing object, the vector X

and the matrices A are extended by synthetic (plausible but not existing) combinations. The starting frequency x_i for these combinations is 0.

There is usually no unique solution to this linear system. Therefore it makes sense to introduce a vector of distortion D (d_j ; $j=1,2,\dots,k$ - distortion in the table cell j).

The set of all possible safe solutions can be stated as

$$\begin{aligned}
 AX + D &= R \\
 \sum_{i=1}^n x_i &= g \\
 x_i &\in \{0,3,4,5,\dots\} \\
 i &= 1,2,\dots, n \\
 j &= 1,2,\dots, k
 \end{aligned}$$

where g is the number of all statistical objects.

To find a unique solution we need an objective function. Empirical experience justifies use of the following objective function:

Minimise the maximal distortion of the table cells.

The anonymous microdatafile is the solution with the smallest maximal distortion.

$$\begin{aligned}
 Z &= \min \left(\max_j (abs(d_j)) \right) \\
 AX + D &= R \\
 \sum_{i=1}^n x_i &= g \\
 x_i &\in \{0,3,4,5,\dots\} \\
 i &= 1,2,\dots, n \\
 j &= 1,2,\dots, k
 \end{aligned}$$

The problem can be solved by a hill-climbing method (Höhne, J. 2003).

3.2 Match the original continuous variables to this solution

As result of step 3.1 we have a solution for the discrete variables. Now we link the original records to this solution to get a starting solution for the continuous variables. We link the records in such a way that the changes in the microdatafile are minimised.

The outcome of 3.1 is a table where the frequency (X^a). for every combinations of values of discrete values is either 0, or exceeds 2. We extend the original microdatafile by adding a column for the information that the record is linked (starting value not matched)

Firstly, we select a continuous variable as pivot variable. We need this variable as help-key for describing the size of the objects. The variable will be used for a priority, if more than one respondent has the same combination of discrete values. In this case, the smallest respondent (according to the value of the pivot variable) should get the largest changes in discrete values.

The following algorithm will do the linkage:

1. Create a priority list of the discrete variables and their hierarchies. Using this priority list we can create a combined key by concatenation of the discrete variables according to the priority list. The starting length s is the length of the combined key. The key length s will be reduced during the iteration of the algorithm.
2. The microdatafile will be sorted by the combined key in the actual key-length s , identical records in descending order of the pivot variable.
3. For every non-linked record in the microdatafile we look in the discrete solution table for a matching row with free capacity. If there is a matching row, the values of the discrete variables for this record of the microdatafile will be turned into the corresponding values of the matching row in the table. The capacity of the row in the table will reduced by 1 and the record of the microdatafile is marked as linked.
4. After a loop trough the microdatafile the length s of the combined key is reduced and the algorithm continues with step 2 until all records are linked.

In that way, the records of the microdatafile will be linked by their similarity and records with the same similarity by their size. Large changes in discrete values should be on the smallest objects.

From this new microdatafile we derive all aggregation tables. With the tables we store the original and the new frequency and the original and the new magnitudes for continuous values. The last iteration of 3.1 and 3.2 leads to a protected microdatafile with no non-zero frequency less than 3 in any possible tabulation. Thus, apart from dominance problems all tabulations will be safe now. For any sensitive cells in the tabulations, we derive an upper and a lower bound for the protection interval.

3.3 Protection against inferential disclosure

In the next step the solution of 3.2 should be modified as to prevent inferential disclosure of predominating units. In step 3.2 we have obtained all new tabulations and the protection intervals for cells if they exists.

Starting point is the multidimensional tabulation with all discrete variables.

For each row of this table we can check whether it relates to a tabulation cell with an unsafe cell value, i.e. a cell value located inside the protection interval. If so, we check if the actual row relates to the dominating unit. If so, the value of the continuous variable will be changed in such a way that the value of

the tabulation cell is set to the upper or lower bound of the protection interval. After some loops through the table the values of all tabulation cells will be located outside the protection intervals.

3.4 Optimise the solution

If the safe microdatafile is generated as safe data base for multidimensional tabulations, the next step to follow step 3.3 is an optimisation of the feasible solution resulting from 3.3. If, on the other hand, the purpose of the safe file is to be supplied as scientific use file, this step is not required. The differences between original and safe tables resulting from the previous steps are reduced by changing the magnitudes in the multidimensional table in 3.3. For the optimisation we use a distance function for all table cells (i.e. the sum of squares of all distances). Using this objective function, and considering the constraints resulting from the requirement that cell values of sensitive cells must be located outside the protection intervals, we optimise the solution using a hill climbing method.

3.5 Construct groups of three or more records while setting values to the mean

As final step we create sets of identical records. As result of the previous steps we have a microdatafile with perturbed discrete values and original continuous variables, and a table with "optimal" values for the continuous variables. Now we can group the records in the microdatafile by the discrete values and their size (according to the pivot variable) to groups of at least 3 elements. Using the results of 3.4 perturbed values for the continuous variables are computed. Continuous variables are replaced by their corrected mean.

As result we have an protected microdatafile, which consists of records only where three or more records are identical. All possible tabulations have a 'minimal' perturbation in the frequency of objects and the perturbed magnitudes should be close to the original magnitudes. Sometimes perturbations in particular cells of the tables are felt to be too strong. The data user can improve the situation if he computes higher level table cells.

4. Tests

Recently, we have tested the method using data of the "monthly statistic of manufacturing and mining" ("Monatsbericht im Bergbau und Verarbeitenden Gewerbe"). The method of step 3.1 (finding a feasible solution for discrete variables) has also been tested with data of the Berlin register of inhabitants.

The Berlin register of inhabitants provides data on around 3.5 million objects (inhabitants). We have split the database by the top level categories of the regional variable ("Bezirke"; there are 23 "Bezirke" in Berlin) and carried out computations for each subset separately. As set of controlled tabulation tables we select all one-, two- and three-dimensional table cells as resulting from tabulating according to the following set of variables:

variable	number of key values	steps of controlled aggregations
region	340 000	5
age	116	3
sex	2	0
nationality	217	4
kind of residence	5	0

The protected microdatafile may be used for any tabulations. The most important uses of the inhabitant-register are ad-hoc-tabulations by different regions. Thus we have used the register with the bottom level regional information (street and number).

The quality of the result is shown in the following table:

region Bezirk	inhabitants	different combinations of values	controlled one-dimensional table cells	controlled more-dimensional table cells	maximum perturbation in frequency one-dimensional	maximum perturbation in frequency multi-dimensional	probability that multi-dimensional table cell has error on frequency <=3 in %	probability that multi-dimensional table cell has error on frequency <=5 in %
1	78 791	68 904	3 873	1 957 608	2	9	96.7	99.8
2	95 322	88 257	3 767	2 541 910	2	8	96.4	99.8
3	160 409	150 043	6 558	4 169 830	2	8	96.8	99.8
4	137 688	124 056	6 297	3 188 990	2	9	97.0	99.9
5	101 762	90 297	4 272	2 288 020	2	8	98.0	99.9
6	154 924	143 357	5 468	3 853 928	2	8	96.5	99.8
7	186 863	174 566	10 046	5 264 223	2	9	96.5	99.8
8	221 749	209 030	25 232	6 857 729	2	9	97.5	99.9
9	148 757	141 001	9 168	4 421 452	2	10	96.6	99.8
10	104 240	100 421	18 824	4 126 193	2	9	97.7	99.9
11	155 347	145 525	6 558	4 242 445	2	10	96.5	99.8
12	197 864	188 242	20 866	6 408 948	2	11	96.8	99.8
13	194 517	184 114	22 806	6 089 881	2	10	97.6	99.8
14	315 823	296 680	27 343	9 059 132	2	9	97.1	99.9
15	114 600	108 545	15 704	3 575 280	2	9	98.6	99.9
16	118 077	110 532	16 974	3 744 034	2	9	98.6	99.9
17	158 293	157 551	8 253	3 459 676	2	10	97.4	99.9
18	74 773	70 601	10 595	2 413 301	2	9	98.6	100.0
19	124 281	116 657	13 582	3 680 527	2	8	98.3	100.0
20	255 671	241 141	34 884	8 231 061	2	8	97.7	99.9
21	141 814	124 541	9 500	2 902 714	2	8	98.3	100.0
22	111 258	99 086	6 610	2 291 383	2	7	99.0	100.0
23	128 795	120 901	15 506	3 638 217	2	8	98.7	100.0

The "monthly statistic of manufacturing and mining" ("Monatsbericht im Bergbau und Verarbeitenden Gewerbe") provided the data for another empirical test. For Berlin, it is a small survey with ca. 1 000 respondents. Variables in the data set were: classification, Workers, others employees, working hours, wages, salaries, sales (domestic sales and non domestic sales).

An example for the quality of the result is presented by the table below. The quality of the result depends on the original distribution. If protection intervals have to be considered, or the number of companies in the original data-set is below three, larger perturbations are required. The data user can improve the situation if he computes higher level table cells.

classification WZ93	Firms		Workers			
	original value	anonymous value	original value	anonymous value	lower limit*)	upper limit*)
1	165	166	13 471	13 471	0	0
12	149	150	10 318	10 318	0	0
1208	1	0	7	0	5	9
1209	28	29	1 273	1 278	0	0
1210	2	3	62	39	38	78
1211	1	0	15	0	11	19

classification WZ93	Firms		Workers			
	original value	anonymous value	original value	anonymous value	lower limit*)	upper limit*)
1212	1	0	30	0	23	38
1213	2	3	179	228	122	224
1214	2	3	51	36	38	64
1215	1	0	9	0	7	11
1216	67	68	2 592	2 613	0	0
1217	3	3	410	411	0	0
1218	11	11	2 172	2 170	0	0
1219	1	0	20	0	15	25
1220	6	6	510	510	0	0
1221	2	3	156	195	109	195

*) The lower and the upper bound of the protection interval around of the original value if there is a dominance problem. In one case that the perturbed value is still within the interval (WZ93 = 1210) due to integer rounding while replacing values by their average.

5. Outlook

The SAFE method is one of the methods considered in the German national research project "Factual Anonymisation of Business Microdata". Within this project, we compare the SAFE method to common alternative methods for statistical disclosure limitation of microdata. Because the main aim of the project is to compare methods for generation of scientific use files, on the one hand the distance between the original and the anonymous microdatafile was an important quality criterion. On the other hand the quality of tabulations (information loss through perturbation, and disclosure risk aspects) are not evaluated in the quality-scores for microdata anonymization methods. Therefore, for the purpose of comparing methods for microdata protection, steps 3.3 and 3.4 of the algorithm were not applied in the tests. Within the project five different databases of the official statistics will be considered. At the time of writing this, we have been working on two of those databases: the data of the structure of costs survey ("Kostenstrukturerhebung") and of business tax statistics ("Umsatzsteuer"). The databases consist of 16 918 objects with 4 discrete and 33 continuous variables (structure of costs survey) and 2 909 150 objects with 5 discrete and 21 continuous variables (business tax statistics).

The IAW (institute for applied economic research - "Institut für angewandte Wirtschaftsforschung" Tübingen) will analyse the quality of the protected microdatafiles. They compared SAFE results with results of the methods Microaggregation, Rankswapping and Latin Hypercube Sampling (LHS) (for description see Dandekar, R.A. 2002). The following table exhibits some results for the structure of costs survey (see Rosemann, M. 2003).

Method	Mean variation in averages	Mean variation in variances	Mean variation in covariances	Mean error in correlation (x100)	Mean error in rank correlation (x100)
MA 1g	3,494	21,297	75,759	5,806	8,952
MA 2g	2,449	23,432	61,987	4,826	6,767
MA 33g	0,035	5,909	21,211	2,408	0,005
RSWP (10%)	0	0	131,906	35,404	1,639
RSWP (5%)	0	0	130,752	34,453	0,485
RSWP (1%)	0	0	147,638	31,627	0,087
LHS	1,004	0,592	219,644	36,234	0,754
SAFE	2,814	46,931	88,567	4,401	6,607

- MA ng - Microaggregation that grouped the continuous variables to n groups.
- RSWP ($n\%$) - Rankswapping that swapped the values in the range of $n\%$ of the database.
- LHS First naive application of the method. We admit that there is probably room for some improvement in the results, when parameters of the method are chosen in a more sophisticated way.

(source: Rosemann, M. 2003)

Other evaluation criteria were "the impact of disclosure control methods on the averages and ranks of branches in specific coefficients" and "the impact on the coefficients in regressions models". Results of show that SAFE is one of the best performing methods.

At the time of writing, results for business tax statistics were not yet available.

References:

Appel, G.C. (1994), 'Anonymization of Microdata, 1st Practical Experience with the SAFE-Programme', proceedings of the Second International Seminar on Statistical Confidentiality, Luxemburg, 1994

Dandekar, R.A., M. Cohen und N. Kirkendall (2002): Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique, 2001. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Dandekar, R.A., J. Domingo-Ferrer und F. Sebé (2002): LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection, 2001. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Höhne, J. (2003), 'SAFE - ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben', Berliner Statistik, Statistische Monatsschrift Nr. 3 2003

Mateo-Sanz J.M. und J. Domingo-Ferrer (1998): A Comparative Study of Microaggregation Methods. Studie erhältlich auf der Homepage von Domingo-Ferrer unter <http://www.etse.urv.es/~jdomingo/>

Rosemann, M. (2003), 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten - Strategien, Vorgehen, erste Ergebnisse', paper presented at the FiDASt-workshop, Berlin, 7. march 2003