CASC PROJECT

Computational Aspects of Statistical Confidentiality

October 2003

Test report on multivariate microaggregation in µ-Argus 3.2

Josep A. Sánchez
Julià Urrutia and
Enric Ripoll

Statistical Institute of Catalonia (Idescat)

Deliverable No: 6-D6

CASC Project
Deliverable 6-D6

Test report on multivariate microaggregation in μ-Argus 3.2

Josep A. Sánchez, Julià Urrutia and Enric Ripoll

Statistical Institute of Catalonia (Idescat)

1. Introduction

As the title of this paper shows, the objective of the deliverable is to publish the results of a test on microaggregation techniques implemented in μ-Argus 3.2 using real data and, more specifically, business data, a new aspect in the analysis. Particulary, it has been tested the microaggregation method that, according to the present bibliography, has provided the best results. Moreover, it's very important to remark that this work analyses the potentiallity of μ-Argus using a very high number of records.

The analysis has been focused on measuring the information loss and the disclosure risk. In this sense, the methods used in order to calculate these measures are based on A. Torres[1] (2003). This empirical work, jointly with the works developed by J. Domingo have allowed to implement the present algorithm that μ-Argus provides about microaggregation.

Moreover, execution times and other remarks on μ-Argus 3.2 are also shown in this deliverable.

2. Data description

The analysis has been done using the Agriculture Census of Catalonia (1999). The original file contains 77839 records, that is to say, a large file that has allowed us to test the potentiallity of μ-Argus and, overcoat, to test the microaggegation techniques.

On the other hand, Argus works under the pressumption that we are using a sample and not a census, so, we have obtained two independent samples with the 5% and 10% of the records. This way, we have performed three analysis under the point of view of the number of records used.

[1] Ángel Torres (2003), *Contribucions a la Microagregació per a la Protecció de Dades Estadístiques (Contributions to the Microaggregation for the Statistical Data Protection).* Doctoral Thesis.

The names of the files with the original data are:

CA_TOTAL.asc
CA_SAMPLE10.asc
CA_SAMPLE5.asc

Further on, we are going to see that the microaggregation methods have been tested under other parameters as the number of variables used in the microaggregation and the number of records per group (k).

The file contains 2 qualitative variables and 7 quantitative variables that have been choosed for the analysis:

- PROV: province (4 values)
- OTE: main product (66 values)
- SUP: total agricultural area
- SAU: utilised agricultural area
- SREG: irrigated agricultural area
- UTA: anual work unit
- UTAA: non-family anual work unit
- UR: livestock units
- MBT: total gross margin

The name of the metadata file used is ARGUS_CA99.rda and the record description is the following:

```
PROV 1  2 99
  <RECODABLE>
  <IDLEVEL>  1
  <SUPPRESSWEIGHT>  50

OTE 3 4 9999
  <RECODABLE>
  <IDLEVEL>  2
  <SUPPRESSWEIGHT>  50

SUP 7  10
  <NUMERIC>
  <DECIMALS>  2

SAU 17  10
  <NUMERIC>
  <DECIMALS>  2
```

```
SREG 27  10
  <NUMERIC>
  <DECIMALS>  2

UTA 37  8
  <NUMERIC>
  <DECIMALS>  3

UTAA 45  8
  <NUMERIC>
  <DECIMALS>  3

UR 53  9
  <NUMERIC>
  <DECIMALS>  3

MBT 62  12
  <NUMERIC>
  <DECIMALS>  1
```

3. Preparation of files

We have talked about the three data files, one census file and two sample files. Obviously, the census file contains all the records but we have had to create the two sample files.

Number of records:

CA_TOTAL.asc	77839
CA_SAMPLE10.asc	7784
CA_SAMPLE5.asc	3892

Two criterions have been applied in the construction of the sample files:

1. All the records with three or more zeros in their numerical variables have been deleted from the original file because they could bias the results from a file with a number of records not very high as sample files are.

2. After that, a random number has been added to each record and the file has been sorted by this new variable. In order to generate the sample files, the first records (5% of the original total), and the last records (10% of the original total) have been selected. This way, none of the records is in the two sample files in order that the comparison of the results would be not affected by a group of common records.

This table shows the main descriptive statistics of the three files:

Table 1.

CA_TOTAL

Variable	N	Mean	Std Dev	Minimum	Maximum
sup	77839	29.60	181.36	0	13653.00
sau	77839	14.86	84.73	0	10894.00
sreg	77839	3.05	16.53	0	1326.57
uta	77839	0.972	1.779	0.002	182.165
utaa	77839	0.307	1.711	0	182.165
ur	77839	35.824	281.035	0	36000.000
mbt	77839	2859942.0	7240791.6	0	682628736.0

CA_SAMPLE10

Variable	N	Mean	Std Dev	Minimum	Maximum
sup	7784	25.10	85.11	0.10	2135.00
sau	7784	15.20	42.06	0	1340.00
sreg	7784	3.98	20.91	0	1250.00
uta	7784	1.146	1.669	0.002	45.000
utaa	7784	0.388	1.627	0	45.000
ur	7784	41.219	194.481	0	8100.000
mbt	7784	3390164.9	6905906.3	0	199217632.0

CA_SAMPLE5

Variable	N	Mean	Std Dev	Minimum	Maximum
sup	3892	25.04	95.41	0.10	3798.59
sau	3892	15.47	44.90	0	1180.64
sreg	3892	4.35	23.30	0	1108.84
uta	3892	1.138	1.558	0.002	33.844
utaa	3892	0.378	1.479	0	33.844
ur	3892	48.729	270.975	0	9000.000
mbt	3892	3508768.7	7292541.5	0	174192640.0

4. Parameters of analysis

At this moment, we have three files to be analysed, but we have mentioned before that other parameters have been introduced in order to do a more exhaustive testing.

1. *Number of variables applied in the microaggregation:* first of all we have had to decide how many and which variables had to be microaggregated. In this sense, three diferent combinations have been applied:
    - all of the 7 numerical variables
    - a combination of 3 + 4 variables (microaggregation in two steps)
    - a combination of 3 + 2 + 2 variables (microaggregation in three steps)

   The decision about which variables are choosen has been product of the previous knowledge about the characteristics of each variable from experts at Idescat. So, the two combinations of variables groupped are:

    - 3 + 4 variables combination:
        - first microaggregation step: SUP, SAU, SREG
        - second microaggregation step: UTA, UTAA, UR, MBT

    - 3 + 2 + 2 variables combination:
        - first microaggregation step: SUP, SAU, SREG
        - second microaggregation step: UTA, UTAA,
        - third microaggregation step: UR, MBT

The microaggregation in more than one step needs to apply μ-Argus as many times as steps there are. This aspect has provided us with some information about possible of μ-Argus 3.2 that will be reported forward.

2. *Number of records per group (k):* three different options have been applied:
    - k=3
    - k=5
    - k=10

Summing up, the total of performed ananalysis is 27 (3 files * 3 combinations of variables * three different k)

We want to remark that we have based our analysis on the previous works developed by the Universitat Rovira i Virgili and so we have selected the parameter values once we have seen the conclusions of the above-mentioned work.

5. Analysis phases

The analysis done in order to get the aims of this deliverable is divided in two stages:

1. Creation of the microaggregated file (μ-Argus)
2. Measure of the information loss and disclosure risk (SAS)

5.1 Microaggregation files:

In order to improve the analysis of the results, the names of the 27 files created by Argus are shown in the next table:

Table 2.

	number of	*number of variables*		
File	*records per group*	7	3 + 4	3 + 2 + 2
CA_TOTAL	k=3	MAT_7_3	MAT_34_3	MAT_322_3
	k=5	MAT_7_5	MAT_34_5	MAT_322_5
	k=10	MAT_7_10	MAT_34_10	MAT_322_10
CA_SAMPLE10	k=3	MA10_7_3	MA10_34_3	MA10_322_3
	k=5	MA10_7_5	MA10_34_5	MA10_322_5
	k=10	MA10_7_10	MA10_34_10	MA10_322_10
CA_SAMPLE5	k=3	MA5_7_3	MA5_34_3	MA5_322_3
	k=5	MA5_7_5	MA5_34_5	MA5_322_5
	k=10	MA5_7_10	MA5_34_10	MA5_322_10

The number of times that μ-Argus has been run is bigger than the number of files created because of the steps needed in the two and three steps microaggegation; so μ-Argus has been executed 54 times

Each file has been processed twice with m-Argus, in order to check the existence of variations in execution times. These variations do exist but they are not significant. A table with mean execution time is shown.

Table 3.

File	Mean Execution time
MAT_7_3	30'55"
MAT_7_5	29´30"
MAT_7_10	28´40"
MA10_7_3	15"
MA10_7_5	15"
MA10_7_10	15"
MA5_7_3	3"
MA5_7_5	3"
MA5_7_10	3"
MAT_34_3	16'40" + 19'45"
MAT_34_5	16'10" + 19'10"
MAT_34_10	15'35" + 18'35"
MA10_34_3	5" + 7"
MA10_34_5	5" + 7"
MA10_34_10	5" + 7"
MA5_34_3	1,5" + 1,5"
MA5_34_5	1,5" + 1,5"
MA5_34_10	1,5" + 1,5"
MAT_322_3	16'40" + 15'15" + 15'15"
MAT_322_5	16'00" + 14'55" + 14'55"
MAT_322_10	15'35" + 14'35" + 14'35"
MA10_322_3	5" + 3" + 3"
MA10_322_5	5" + 3" + 3"
MA10_322_10	5" + 3" + 3"
MA5_322_3	1,5" + 1,5" + 1,5"
MA5_322_5	1,5" + 1,5" + 1,5"
MA5_322_10	1,5" + 1,5" + 1,5"

Hardware used has been a Pentium III (Intel), 127 MB RAM.

5.2 Measure of the information loss and disclosure risk

Quality of a microaggregation method can be obtained from information loss due to the publication of non-original data and disclosure risk.

Let's suppose a set of microdata corresponding to n individuals $I_1, I_2, ..., I_n$ and p continous variables $Z_1, Z_2, ..., Z_P$.

Let X be a matrix of n rows and p columns representing the original microdata set and X' a matrix of n rows and p columns representing the modified microdata set.

$\overline{X}, \overline{X}'$ are p-dimensional vectors corresponding to X and X' means.

V, V' are the $p \times p$ covariates matrices corresponding to X and X'.

R, R' are the $p \times p$ correlation matrices corresponding to X and X'

5.2.1 Information loss measures (PI):

Information loss can be measured as a function of structural differences between X and X'. Five different measures are defined:

PI1: mean variation of data $X - X'$

$$PI1 = 100 \cdot \frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\left| x_{ij} - x'_{ij} \right|}{\left| x_{ij} \right|}}{np}$$

Note: if $x_{ij} = 0$ and $x'_{ij} \neq 0$, then divide by $\left| x'_{ij} \right|$. If $x_{ij} = x'_{ij} = 0$, the term is not added to the sum. This rule is applied to the rest of formulas.

PI2: mean variation of data means $\overline{X} - \overline{X}'$

$$PI2 = 100 \cdot \frac{\sum_{j=1}^{p} \frac{\left| \overline{x}_{j} - \overline{x}'_{j} \right|}{\left| \overline{x}_{j} \right|}}{p}$$

PI3: mean variation of data covariates $V - V'$

$$PI3 = 100 \cdot \frac{\sum_{j=1}^{p} \sum_{1 \le i \le j} \frac{\left| v_{ij} - v'_{ij} \right|}{\left| v_{ij} \right|}}{\frac{p(p+1)}{2}}$$

PI4: mean variation of data variances $S - S'$

$$PI4 = 100 \cdot \frac{\sum_{j=1}^{p} \frac{\left| v_{jj} - v'_{jj} \right|}{\left| v_{jj} \right|}}{p}$$

PI5: Mean absolut error of data correlations $R - R'$

$$PI5 = 100\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\left|r_{ij} - r'_{ij}\right|}{\frac{p(p-1)}{2}}$$

Global information loss PI is defined and computed as a weighted mean of PI1, PI2, PI3, PI4 and PI5:

$$PI = \frac{1}{3}PI1 + (\frac{1}{6}PI2 + \frac{1}{6}PI4) + (\frac{1}{6}PI3 + \frac{1}{6}PI5)$$

5.2.2 Disclosure risk measures:

The analysis of a disclosure control method cannot be reduced to information loss: the loss of confidentiality due to the disemination of modified microdata must be also analysed. Three different measures are used in order to obtain a global disclosure risk measure: ERD, ICN and ICD. The global disclosure risk measure is defined as

$$PC = \frac{1}{2}ERD + (\frac{1}{4}ICN + \frac{1}{4}ICD)$$

<u>ERD: Record matching based on distances:</u>

This measure is based on the idea of matching records. This method supposes that an intruder has a set of external microdata Y containing a subset of key variables that are common to the modified microdata X'. The intruder tries to match the modified microdata X' and the external microdata Y using the subset of common variables in order to discover original data X. For every record x'_i in the modified file, distances to all the records $\{x_k\}_{k=1\ldots n}$ in the original file are calculated using only the subset of common variables (variables are normalised prior to the calculation of distances). If the modified record x'_i and its closer original record x_j are the same record, $(i = j)$, a match is produced. ERD is defined as the percentage of matched records.

<u>ICN: Confidentiality interval on number of records:</u>

This measure is based on the idea of building intervals on the modified microdata.

Each variable X'_j is sorted independently of the others. For each value x'_{ij} and variable X'_j a centered interval I'_{ij} is built containing at most q% of the number of records (q is fixed). A record x_i is matched if for all its variables j=1,...,p, $x_{ij} \in I'_{ij}$. ICN is defined as the percentage of matched records.

<u>ICD: Confidentiality interval on standard deviation:</u>

This measure is similar to ICN, building centered intervals I'_{ij} of range at most q% of the standard deviation of each variable X'_j (q is fixed). A record x_i is matched if for all its variables j=1,...,p, $x_{ij} \in I'_{ij}$. ICD is defined as the percentage of matched records.

5.2.3 Quality global measure:

A global measure must give the same importance to information loss and disclosure risk. So, we define the global measure as

$$MG = 0.5 \cdot PI + 0.5 \cdot PC = 0.5 \cdot PI + 0.25 \cdot ERD + 0.125 \cdot ICN + 0.125 \cdot ICD$$

Every term of the sum MG belongs to the interval $[0,100]$, except PI that could be higher than 100. The next rule could be useful for the understanding of MG values: publishing the original microdata using no disclosure control method, would produce no information loss (PI=0) but a total revelation risk (PC=100). In that case, MG=50. As a consequence, any method with a value of MG greater than 50 would be useless; and the lower the value of MG, the better.

Notes on the implementation of measures:

For implementing ERD (Record matching based on distances), 7 different scenarios have been defined, depending on the variables known by the intruder:

ERD-1: One common variable: SUP
ERD-2: Two common variables: SUP, SAU
ERD-3. Three common variables: SUP, SAU, UTA
ERD-4: Four common variables: SUP, SAU, UTA, UR
ERD-5: Five common variables: SUP, SAU, UTA, UR, MBT
ERD-6: Six common variables: SUP, SAU, UTA, UR, MBT, SREG
ERD-7: Seven common variables: SUP, SAU, UTA, UR, MBT, SREG, UTAA
ERD is defined as the weighted average of these 7 scenarios:
ERD = (ERD-1 + ERD-2 + ERD-3 + ERD-4 + ERD-5 + ERD-6 + ERD-7) / 7

For implementing ICN and ICD a value of q=5% has been used. In the work by A.Torres, values of q=1%, 2%, ..., 10% have been used, producing ICN-1, ICN-2, ..., ICN-10 and defining ICN = (ICN-1 + ICN-2 + ... + ICN-10) / 10 (analogous for ICD). According to the results by A. Torres we have observed that ICN-5 $\cong$ ICN and ICD-5 $\cong$ ICD. So, we have defined ICN = ICN-5 and ICD = ICD-5.

The calculation of the set of measures has been performed in SAS in a mainframe Alpha server ES40 with a 500 Mhz CPU. Most of calculation time is due to the calculation of ERD as it searches the closest record for each record according to 7 different distances. A table for mean execution times is included.

Table 4.

	mean execution time
CA_TOTAL	11 h 40'
CA_SAMPLE10	10' 40''
CA_SAMPLE5	5' 30''

The following table shows all the measures that have been calculated:

Table 5.

File	sample%	variables	k	mg	pi	pc	erd	icn	icd
MAT_7_3	5	7	3	20,18	11,27	29,09	12,35	27,94	63,72
MAT_7_5	5	7	5	18,88	16,00	21,75	6,91	15,62	57,57
MAT_7_10	5	7	10	20,37	24,49	16,25	3,44	7,71	50,39
MAT_34_3	5	3 + 4	3	35,19	3,54	66,84	53,81	76,56	83,18
MAT_34_5	5	3 + 4	5	31,92	5,30	58,54	47,56	60,70	78,33
MAT_34_10	5	3 + 4	10	28,45	8,87	48,02	39,51	40,67	72,40
MAT_322_3	5	3 + 2 + 2	3	40,17	1,68	78,66	63,60	92,80	94,64
MAT_322_5	5	3 + 2 + 2	5	38,79	3,60	73,99	60,22	84,65	90,87
MAT_322_10	5	3 + 2 + 2	10	37,26	4,74	69,78	56,39	77,87	88,49
MA10_7_3	10	7	3	25,29	32,19	18,39	13,85	8,31	37,55
MA10_7_5	10	7	5	32,02	52,06	11,99	7,82	2,90	29,39
MA10_7_10	10	7	10	48,71	90,07	7,35	3,98	0,80	20,64
MA10_34_3	10	3 + 4	3	31,41	8,86	53,96	56,92	39,95	62,06
MA10_34_5	10	3 + 4	5	29,23	15,49	42,96	48,63	20,94	53,66
MA10_34_10	10	3 + 4	10	29,11	26,51	31,72	37,57	6,89	44,86
MA10_322_3	10	3 + 2 + 2	3	38,94	5,73	72,14	67,39	74,38	79,41
MA10_322_5	10	3 + 2 + 2	5	37,11	9,70	64,52	63,32	59,17	72,25
MA10_322_10	10	3 + 2 + 2	10	35,91	18,32	53,50	57,14	37,63	62,09
MA5_7_3	5	7	3	25,7	34,46	16,93	14,38	5,42	33,56
MA5_7_5	5	7	5	30,55	50,48	10,62	8,32	1,49	24,33
MA5_7_10	5	7	10	41,12	76,03	6,21	4,19	0,15	16,29
MA5_34_3	5	3 + 4	3	32,44	15,15	49,73	55,81	30,68	56,63
MA5_34_5	5	3 + 4	5	34,42	30,13	38,70	46,39	14,54	47,48
MA5_34_10	5	3 + 4	10	38,66	49,42	27,90	35,03	3,24	38,28
MA5_322_3	5	3 + 2 + 2	3	39,62	11,29	67,94	66,90	64,16	73,79
MA5_322_5	5	3 + 2 + 2	5	36,42	13,37	59,46	62,29	48,07	65,21
MA5_322_10	5	3 + 2 + 2	10	37,49	27,09	47,90	55,08	25,77	55,65

A table with the rank for each measure is also included:

Table 6.

File	sample%	variables	k	mg rank	pi rank	pc rank	erd rank	icn rank	icd rank
MAT_7_3	5	7	3	2	10	10	7	14	17
MAT_7_5	5	7	5	1	15	8	4	11	14
MAT_7_10	5	7	10	3	17	5	1	8	10
MAT_34_3	5	3 + 4	3	15	2	22	16	24	24
MAT_34_5	5	3 + 4	5	11	5	19	14	21	22
MAT_34_10	5	3 + 4	10	6	8	15	12	18	20
MAT_322_3	5	3 + 2 + 2	3	25	1	27	25	27	27
MAT_322_5	5	3 + 2 + 2	5	22	3	26	22	26	26
MAT_322_10	5	3 + 2 + 2	10	19	4	24	19	25	25
MA10_7_3	10	7	3	4	21	7	8	9	6
MA10_7_5	10	7	5	12	25	4	5	4	4
MA10_7_10	10	7	10	27	27	2	2	2	2
MA10_34_3	10	3 + 4	3	10	7	18	20	17	15
MA10_34_5	10	3 + 4	5	8	14	13	15	12	11
MA10_34_10	10	3 + 4	10	7	18	11	11	7	8
MA10_322_3	10	3 + 2 + 2	3	23	6	25	27	23	23
MA10_322_5	10	3 + 2 + 2	5	18	9	21	24	20	19
MA10_322_10	10	3 + 2 + 2	10	16	16	17	21	16	16
MA5_7_3	5	7	3	5	22	6	9	6	5
MA5_7_5	5	7	5	9	24	3	6	3	3
MA5_7_10	5	7	10	26	26	1	3	1	1
MA5_34_3	5	3 + 4	3	13	13	16	18	15	13
MA5_34_5	5	3 + 4	5	14	20	12	13	10	9
MA5_34_10	5	3 + 4	10	21	23	9	10	5	7
MA5_322_3	5	3 + 2 + 2	3	24	11	23	26	22	21
MA5_322_5	5	3 + 2 + 2	5	17	12	20	23	19	18
MA5_322_10	5	3 + 2 + 2	10	20	19	14	17	13	12

Some graphics based on these results are shown. They are focussed on two kinds of comparisons:

Figures A: comparison between number of microaggegated variables 7, 3+4 and 3+2+2

Figures B: comparison between files 5%, 10% and 100%

Fig A.1: MG in the microagggregation over a sample of 5%

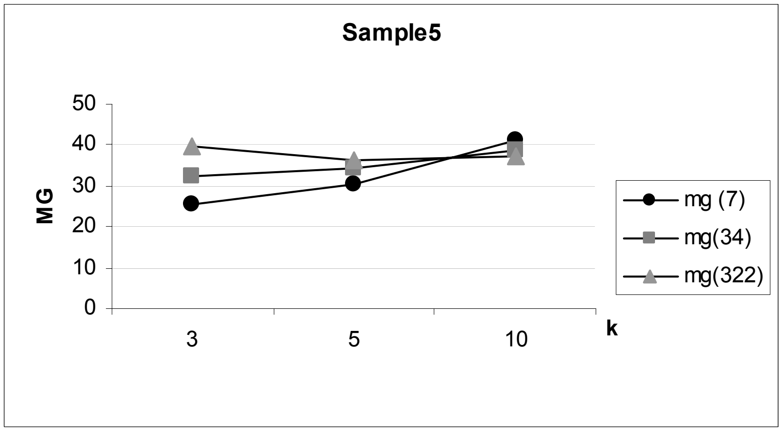


Fig A.2: MG in the microagggregation over a sample of 10%

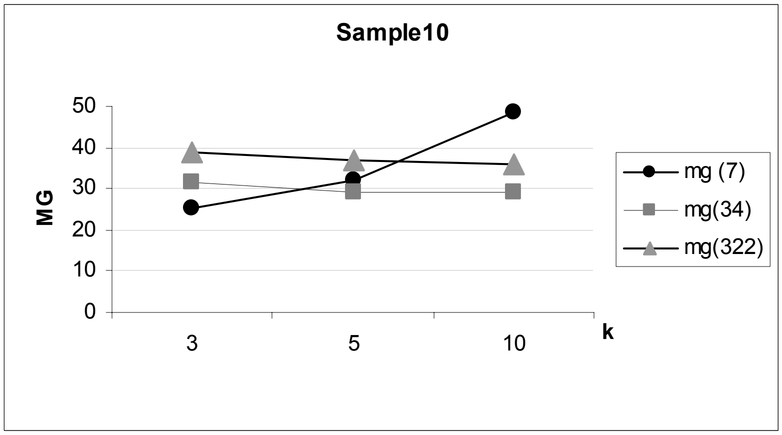


Fig A.3: MG in the microagggregation over the census file

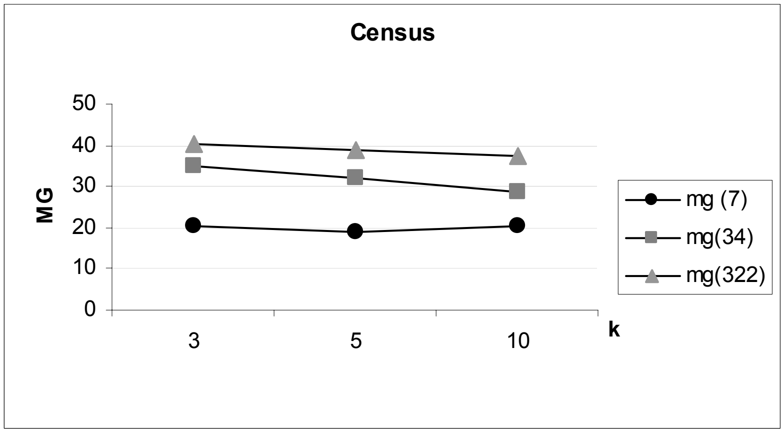


Fig A.4: PI in the microagggregation over a sample of 5%

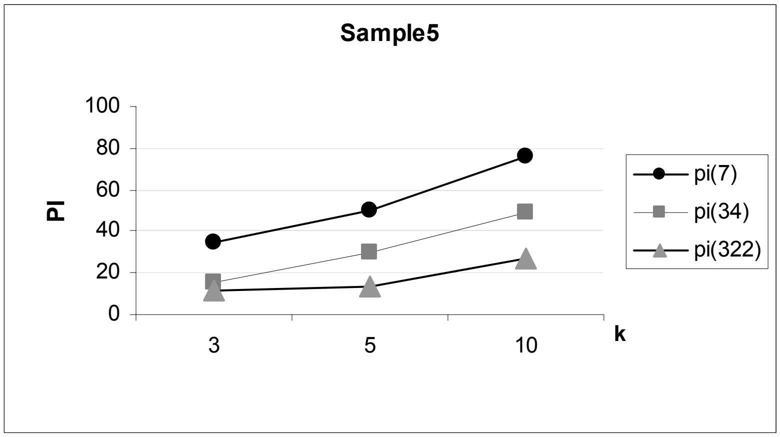


Fig A.5: PI in the microagggregation over a sample of 10%

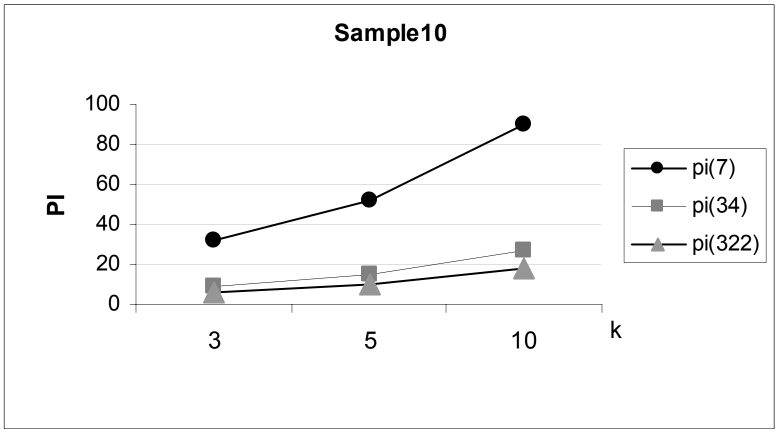


Fig A.6: PI in the microagggregation over the census file

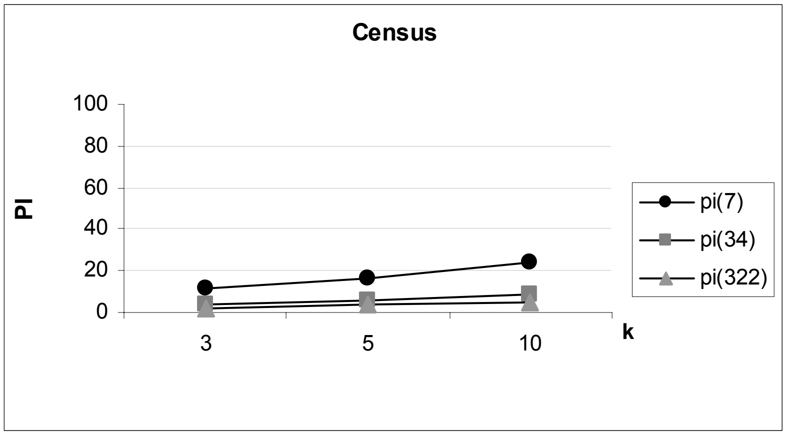


Fig A.7: PC in the microagggregation over a sample of 5%

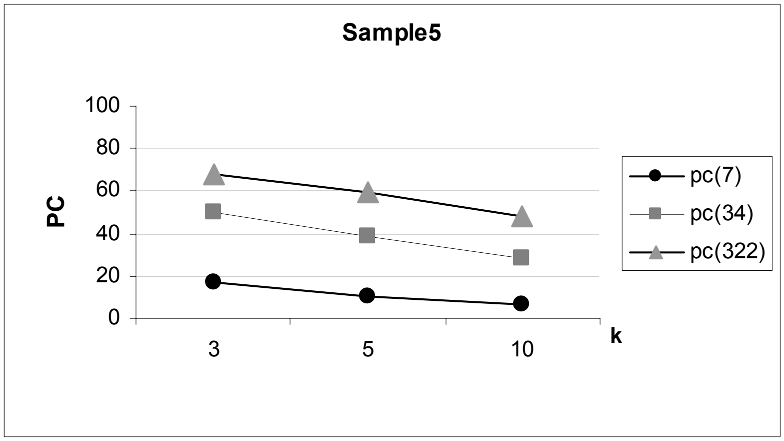


Fig A.8: PC in the microagggregation over a sample of 10%

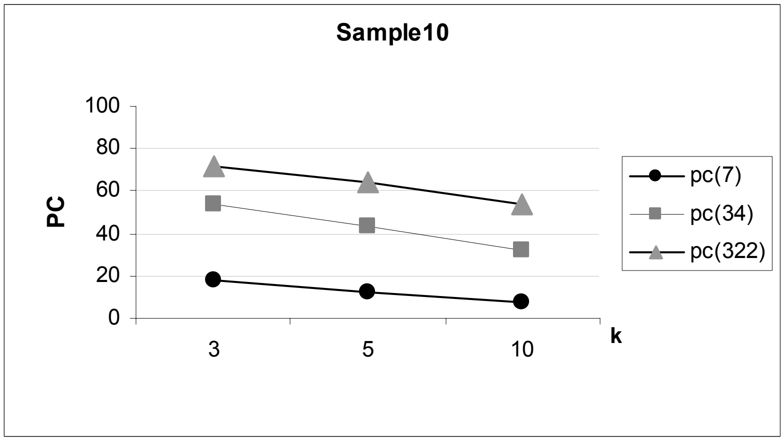


Fig A.9: PC in the microagggregation over the census file

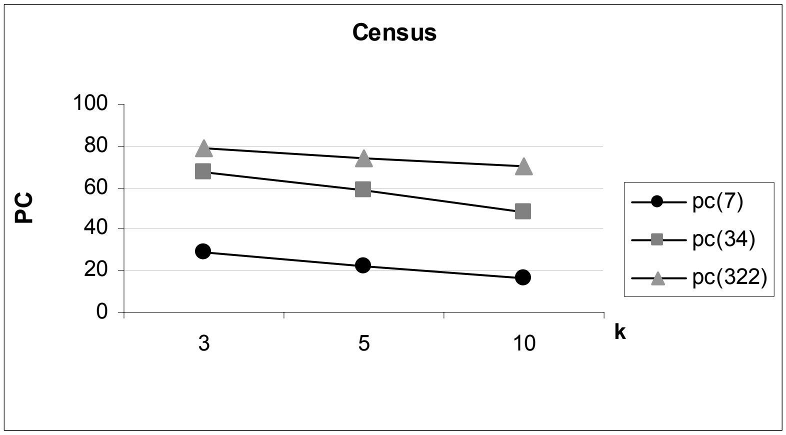


Fig B.1: MG in a microagggregation in one step with 7 variables

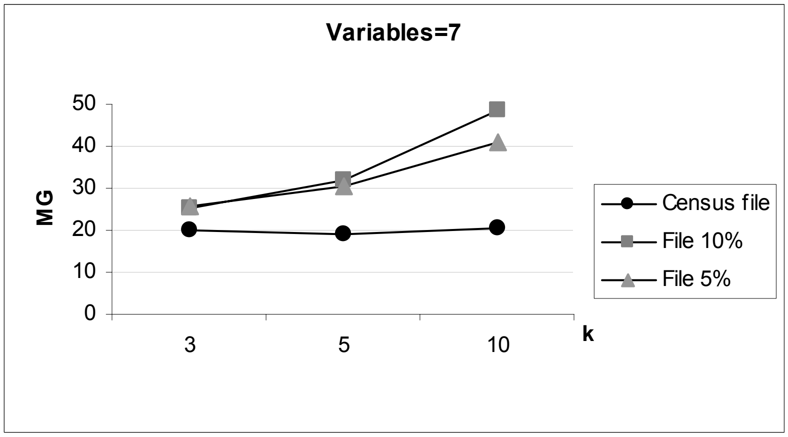


Fig B.2: MG in a microagggregation in two steps with 3+4 variables

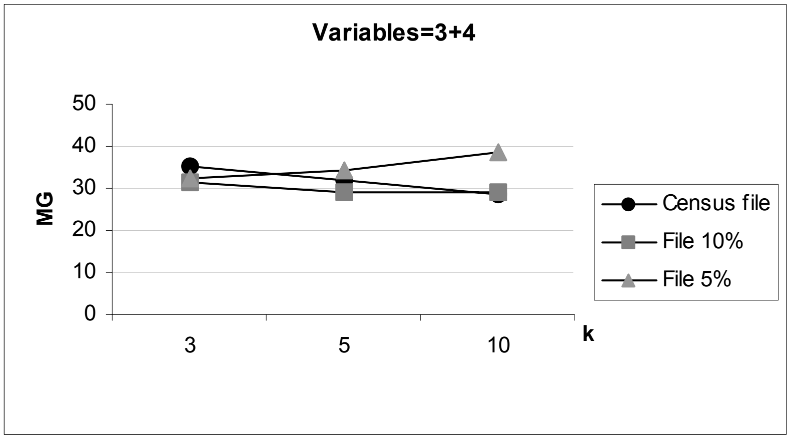


Fig B.3: MG in a microagggregation in three steps with 3+2+2 variables

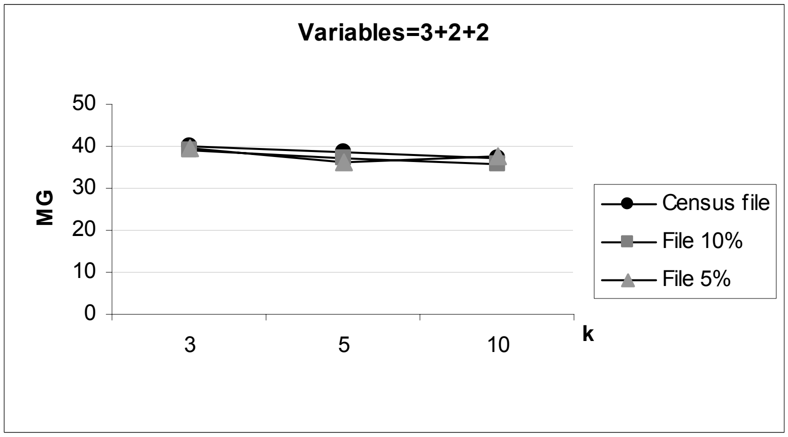


Fig B.4: PI in a microagggregation in one step with 7 variables

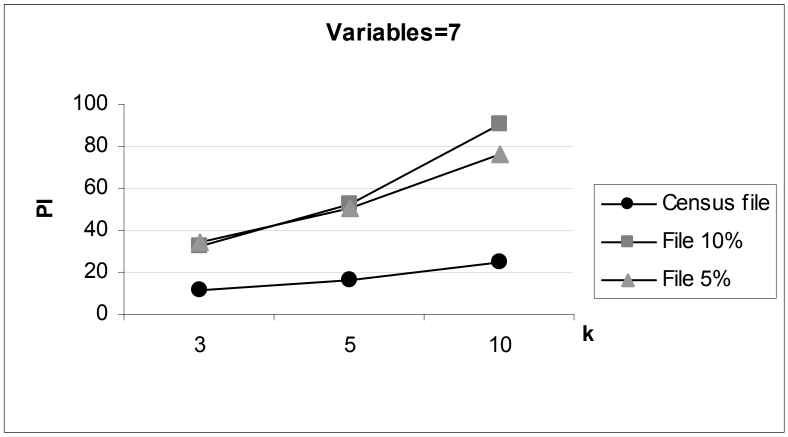


Fig B.5: PI in a microagggregation in two steps with 3+4 variables

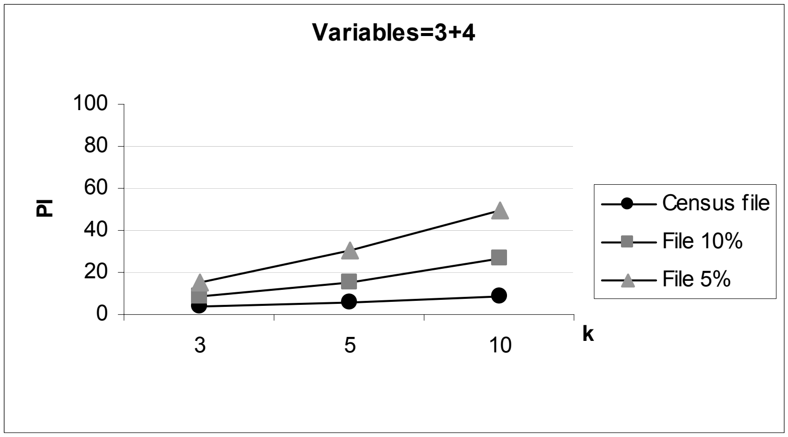


Fig B.6: PI in a microagggregation in three steps with 3+2+2 variables

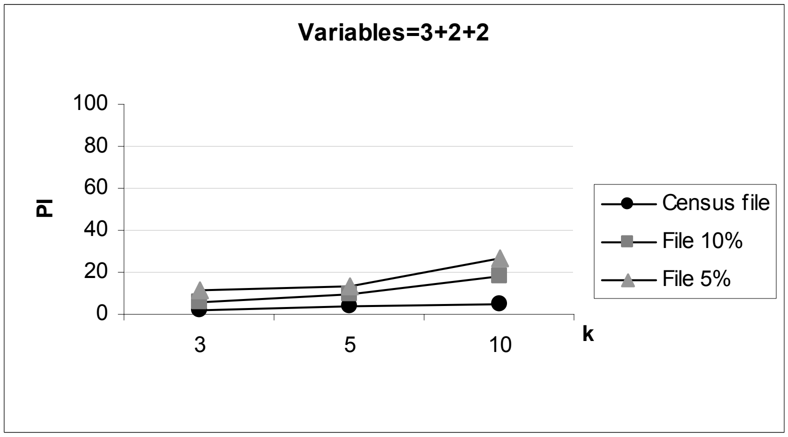


Fig B.7: PC in a microagggregation in one step with 7 variables

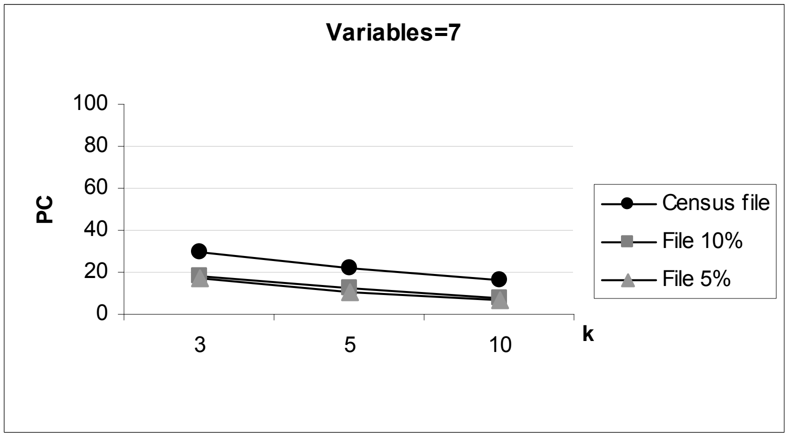


Fig B.8: PC in a microagggregation in two steps with 3+4 variables

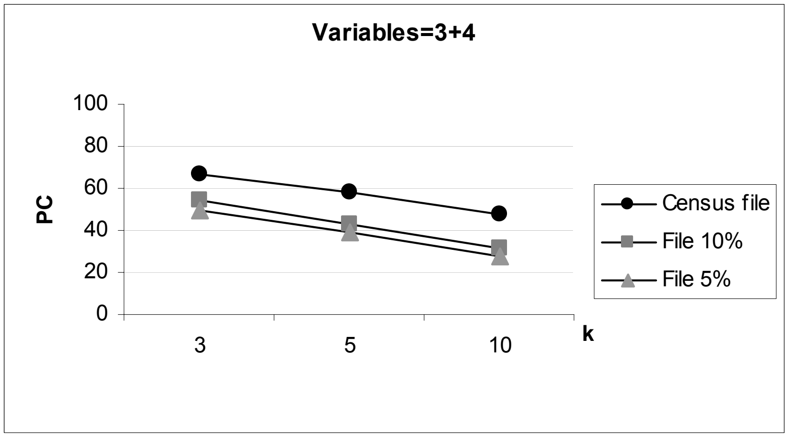
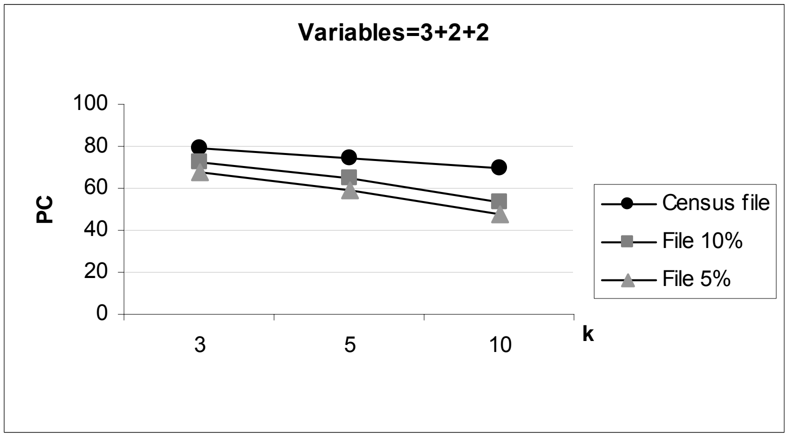


Fig B.9: PC in a microagggregation in three steps with 3+2+2 variables



Some comments can be made about confidentiality loss measures. ERD can be a subjective measure as it depends on the information the intruder has. We have supposed 7 different scenarios for the intruder (producing ERD-1,..., ERD-7) but there are many possible scenarios. ERD-i increases with the number of variables known by the intruder, with a critical point in multi-step aggregations, as can be seen in figures C. On the other hand, figures B show how multi-step aggregation increases PC for any file. As confidentiality is an important subject for statistical offices, multi-step aggregation in μ-argus should be used carefully.

Fig C1: ERD1-ERD7 for aggregation of variables=7

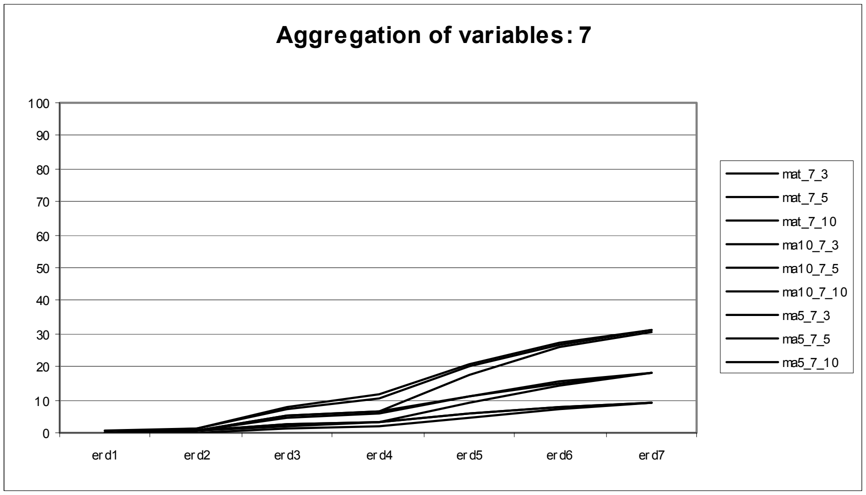


Fig C2: ERD1-ERD7 for aggregation of variables=3+4

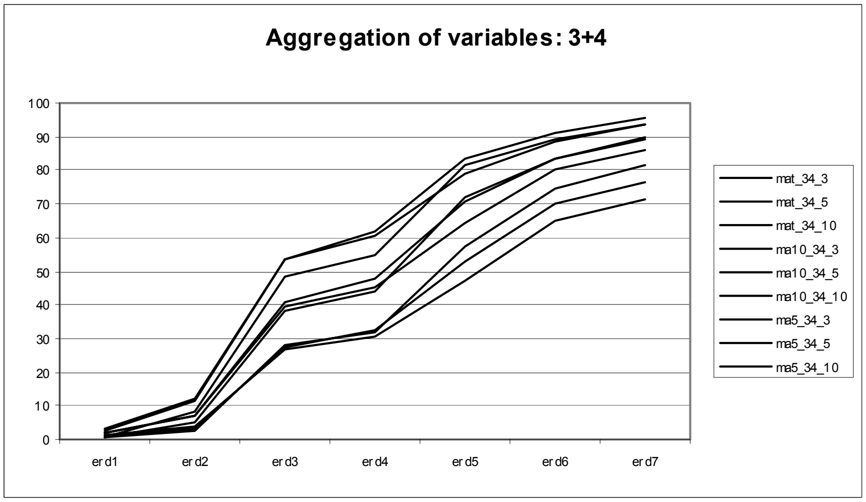
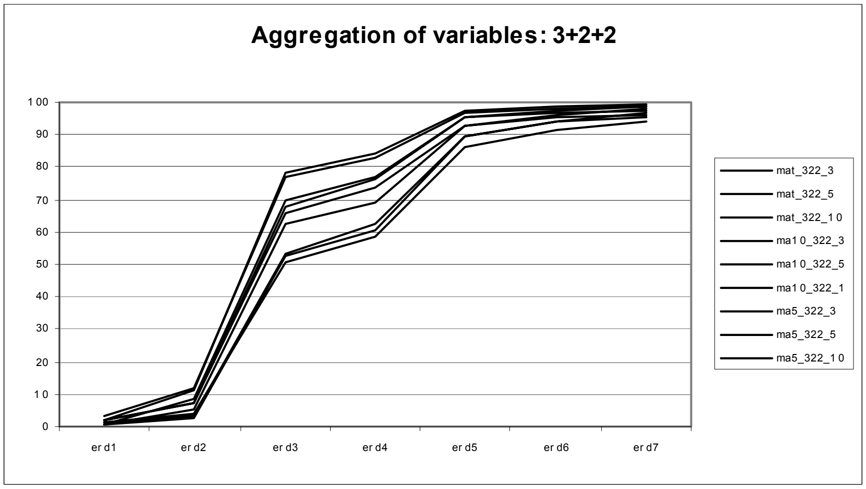


Fig C3: ERD1-ERD7 for aggregation of variables=3+2+2



Finally, figures D show the variation of the MG measure for different k and multi-step aggregation scenarios.

Fig D1: MG crossed by k and variables in the census file

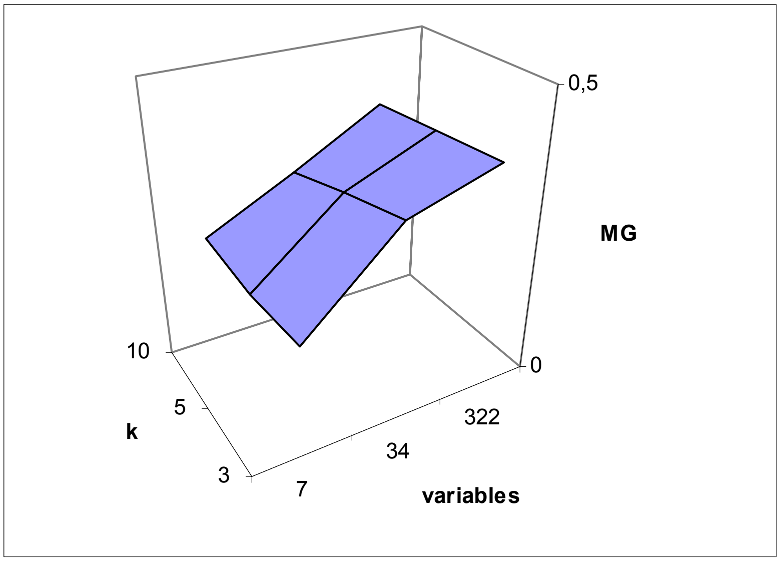


Fig D2: MG crossed by k and vars in the sample 10% file

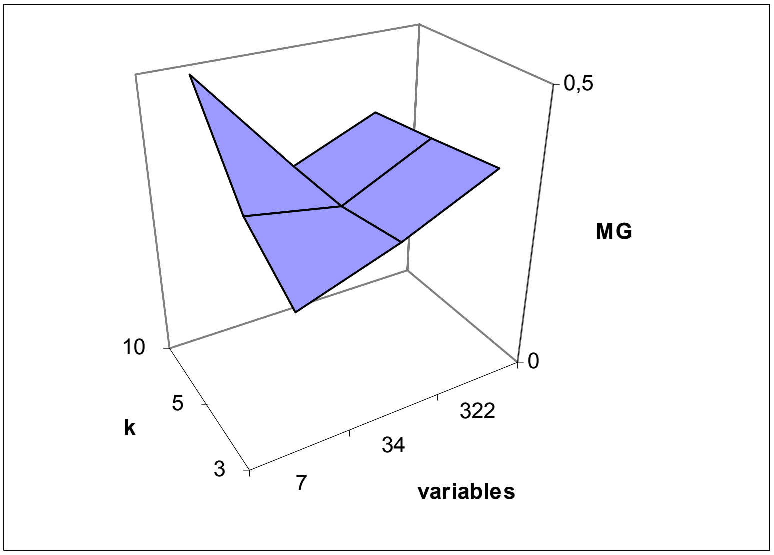
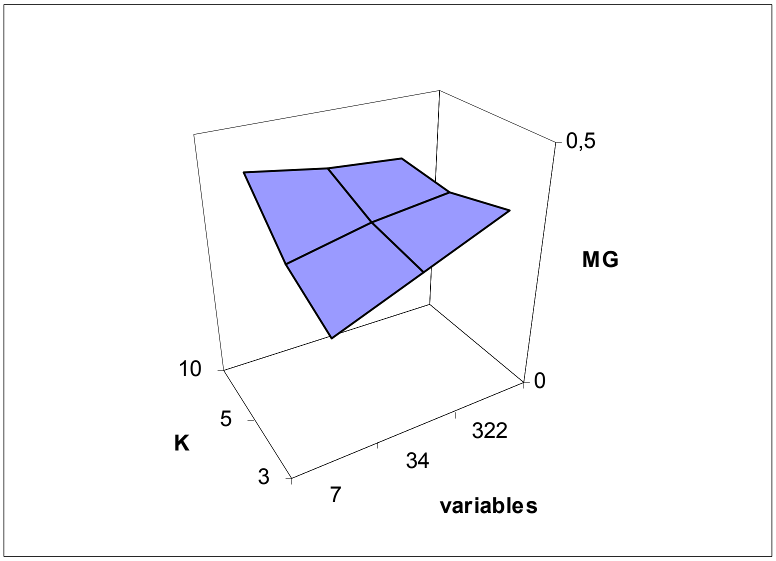


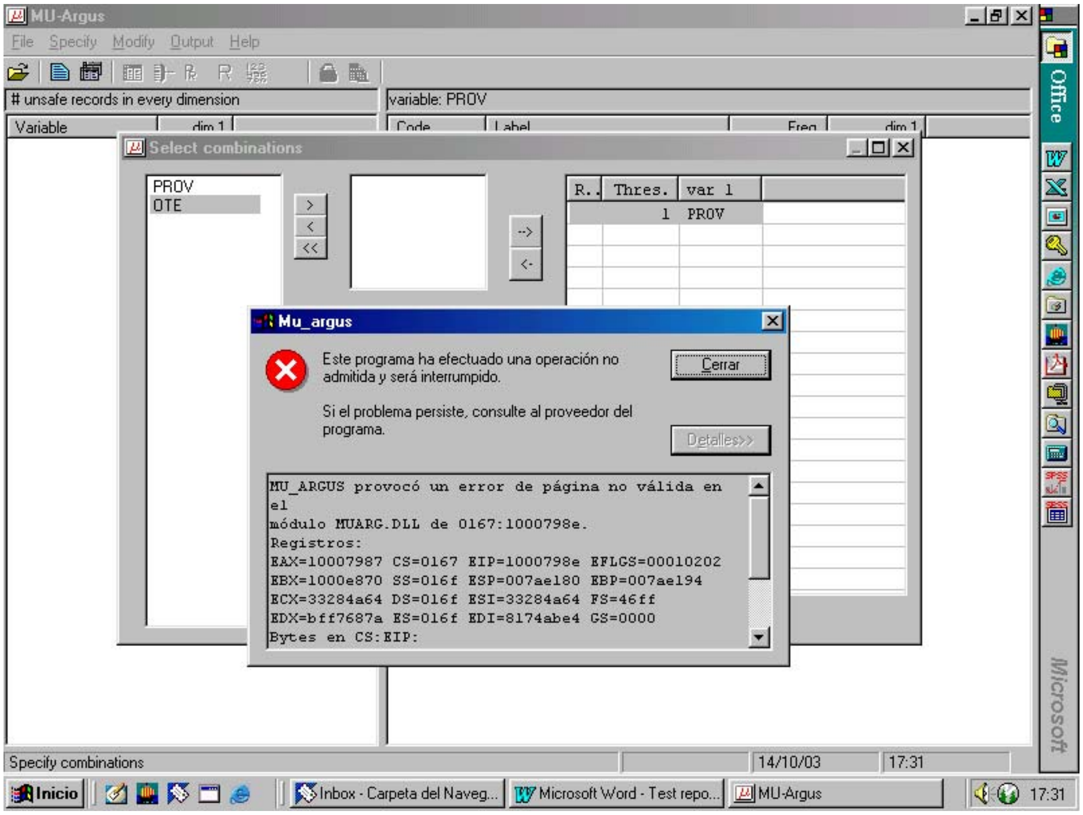
Fig D3: MG crossed by k and vars in the sample 5% file



6. Report on μ-Argus 3.2 problems

In the next lines, some information is reported about problems detected in the execution of μ-Argus or options to improve in the development of new versions.

1. When a safe file has been created, μ-Argus doesn't allow to specify new combinations; an error message appears and the user is expelled. So, it's a little bit tedious to begin again when a new microaggregation should be done.



Este programa ha efectuado una operación no admitida y será interrumpido.

Si el problema persiste, consulte al proveedor del programa.

```
MU_ARGUS provocó un error de página no válida en
el
módulo MUARG.DLL de 0167:1000798e.
Registros:
EAX=10007987 CS=0167 EIP=1000798e EFLGS=00010202
EBX=1000e870 SS=016f ESP=007ae180 EBP=007ae194
ECX=33284a64 DS=016f ESI=33284a64 FS=46ff
EDX=bff7687a ES=016f EDI=8174abe4 GS=0000
Bytes en CS:EIP:
```

2. We had a problem when we have done a microaggregation in more than one step. In the second step, the new file to be microaggregated is the safe file created in the first step. At the beginning, we read this safe file with the ".saf" extension and μ-Argus worked or at least, it seemed to do it but actually the microaggregation was not done.

   In order that the second step would work, the extension of the name created in the first step has to be changed for an ".asc" extension and then this file can be read. It would be useful that a safe file could be read as an original file.

7. Conclusions

Results of the measures obtained with the microaggregation method

1. PI measures information loss. PI increases when k increases. For a fixed k and file, microaggregation 3+2+2 is the best and microaggregation 7 is clearly the worst. For a fixed k and aggregation of variables, the census file has always lower values of PI. The worst results of PI are obtained for files 5% and 10% with k=10 and aggregation of 7 variables.

2. PC measures disclosure risk. PC decreases when k increases. For a fixed k and file, aggregation of 7 variables is clearly the best method and 3+2+2 the worst. For a fixed k and aggregation of variables, file 5% is always the best and census file the worst. The best results of PC are obtained for aggregation of 7 variables and the worst for 3+2+2.

   In fact, these are the expected results: increasing values of k and increasing number of aggregated variables produce worse results of PI and better results of PC.

3. MG is a global measure giving the same importance to information loss and confidentiality loss but other global measures could be defined. Best results are obtained for census file with aggregation of 7 variables and worst results are for sample 10% with aggregation of 7 variables and k=10. Aggregations 3+4 and 3+2+2 produced stable results of MG, always between 30% and 40%. On the other hand, results corresponding to aggregation 7 are much more sensitive to k for 5% and 10% files, because of values of PI. This can be represented on a three-dimensional graphic (Figures D): best results of MG are obtained for the vertex (variables=7;k=3); moving in any direction makes MG increase. Except for the census file, best results are always obtained for a compromise between k and 'variables'. These results could change with a different weighting of PI and PC to obtain MG.

4. It's important to notice that different conclussions on MG can be obtained depending on the size of the file. For a high number of records, aggregating 7 variables is always the best method, independently of values of k. For a 5% and 10% file, aggregating 7 variables can be the best or the worst method, depending on values of k. Best results can be obtained for a compromise between k and 'variables': K=3 or k=5 with microaggregation 7 or 3+4. Tests with other files should be done to verify the hipothesis that using k=10 and aggregating all the variables can be a bad decission for a file under 8000 records.

Execution time

1. The execution time of the microaggregation method is good, the maximum values have been obtained, obviously, in the census file but half an hour is not a big period when working with 77.000 records and 7 variables.

2. The size of the file is the most influential parameter over the execution time. The number of microaggregated variables is also related with the execution time but differences are not so important. For exemple, changing microaggregation from 2 to 3 variables increases the execution time in more or less 1’ 30’’, from 3 to 4 variables in 3’ and from 4 to 7 between 10’-12’.

3. With regard to the execution time in the calculation of the measures, it can be observed again that the size of the file is very important and too many hours are needed in order to analyse the census file, but not the sample files.

μ-Argus

Two aspects should be improved related with the microaggregation method in μ-Argus 3.2. They are not related with the algorithm implemented, which can be justified through the results shown in this deliverable, but they are related from the point of view of the user. These two problems have been documented in the point 6 of this deliverable.

Bibliography:

Ángel Torres (2003), *Contribucions a la Microagregació per a la Protecció de Dades Estadístiques (Contributions to the Microaggregation for the Statistical Data Protection).* Doctoral Thesis. Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya.

Josep Maria Mateo-Sanz and Josep Domingo-Ferrer (1998), *A comparative study of microaggregation methods*, Qüestiió, vol. 22, no. 3, pp. 511-526

Josep M. Mateo-Sanz, Josep Domingo-Ferrer, Francesc Sebé, Antoni Martínez, Àngel Torres and Narcís Macià (2002), *Microaggregation algorithms (software, documentation and related papers),* Deliverables 1-1.D5 and 1-1.D6 of the CASC project (IST-2000-25069)

Anna Oganian (2003), *Security and Information loss in Statistical Database protection*. Doctoral Thesis. Departament de Matemàtica Aplicada 4, Universitat Politècnica de Catalunya.