Empirical evidences on protecting populations uniqueness at Idescat*

Julià Urrutia[1] and Enric Ripoll[1]

[1]Statistical Institute of Catalonia (Idescat)

Abstract. This paper describes the process of statistical disclosure analysis and control applied by the Statistical Institute of Catalonia (Idescat) to microdata samples from census/surveys with some population uniques. Since 1995, by means of models which allows calculation of the risk and data protection procedures, some empirical evidences have been achieved in order to check the performance of μ-ARGUS in a real situation of unique populations, with large files and re-identification keys. The analyzing way used preferably is the measuring of dimensions (to a maximum of six) and the recodification on the changes of information loss versus the disclosure risk variations in the dissemination of anonymised registers. These results should be systematically extented to both social and business (micro)data so as they could allow us to test the effectiveness of new μ-ARGUS features in the undergoing CASC project.

Keywords: disclosure risk, information loss, unique populations, anonymised records, re-identification and subsample methods

1 Introduction

The advances of the methodology guided to the preservation of the statistical secret, specially intense in the decade of the nineties, have not always gone accompanied by the appropriate empirical contrasts of its eventual power on their main beneficiaries side: the official statistical offices. This necessity has been accentuating as these techniques are diversified in a growing way and, mainly, in view of the absence of a conceptual framework (including the own definition of statistical disclosure, what they are sensitive data or the measure of the loss of information) and an empirical one (oneself group of data qualitative and /or quantitative) stable enough.

The difficulties are also enlarged when articulating an integral and coherent policy in this field on the statistical institutes' side. This way, the analytic power

* This work was partly subsidised by the 5th Framework program of European Commission under grant no. IST-2000-25069

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Statistical Institute of Catalonia (Idescat)

of the users of data, a software on statistical disclosure control not very capable still for agile and complete processing of big volumes of data (interrelated) and the problem of the protection of the accessible databases (necessarily) from Internet represent substantial challenges to continue maintaining the simultaneous trust of offerers and users of information.

In this context, the objective of this paper is double: to point out the trajectory of the solutions rehearsed and/or adopted by the Statistical Institute of Catalonia (from now on, Idescat) in the environment of the microdata diffusion and also to present some of the results in the validation of those approaches that have been revealed more promising. For all of it, we count on the experience accumulated from 1995 around effective levels of microdata protection and the corresponding loss of information (to the statistical effects) whose position has counted on a systematic analysis of the main variables (for the most part demographic and social) from the 1991 Population Census of Catalonia.

This way, the most complete whole in the tasks on statistical disclosure control done by the Idescat, that are commented later on, corresponds to a stable but concrete scenario and, therefore, limited: the treatment of physical individuals' sociodemographic variables generally associated to an exhaustive operation of basic characteristics of the population -census- and of mainly qualitative or categorical nature.

In this sense, the document is structured in four parts. In section 1 the nature of the main investigations carried out by the Idescat are exposed, from its creation in 1990, about the microdata cession previous to the adoption of the SDC pattern using μ-ARGUS. It also includes a reference to the works on the macrodata controlled dissemination based on the random perturbations with compensation and the experimental use of the privacy homomorphisms for microdata statistical confidentiality developed outside the Idescat.

Section 2: the empirical mark of the re-identification pattern adopted by the Idescat is described to evaluate the probability of risk of statistical revelation by means of the use of the available options in μ-ARGUS (local suppressions), with which, in 1999, superior results were already obtained to the strict use of the re-identification models.

Section 3: the treatment of the loss of information associated to the procedures that offer smaller risks of statistical revelation in the dissemination of microdata files is approached, contrasting the trade-off between both concepts as the level of dimensions and thresholds of the included variables increase, and analysing its evolution and its profile in particular.

Section 4: contains some final remarks and comments on the last two sections of this paper, which benefit of the "stability" of the conceptual and empirical framework that the Idescat has maintained in its rehearsals about microdata protection, with independence of its intrinsic value and the limitations of its eventual extrapolation and validity for other scenarios.

2 Approaches to the Macrodata and Microdata Statistical Control

Although the central purpose of this paper is limited to the controlled dissemination of microdata files through the adoption of a concrete strategy, it is worthwhile to highlight two experiences relatively innovative examined or promoted by the Idescat between 1993 and 1996: the adaptation of new masking procedures for contingency tables and the potentiality of the cryptographic techniques, fully emergent in the electronic exchanges of information.

2.1 Random Perturbations by Compensation in Macrodata on Population Census

The first of the previous experiences rehearsed by the Idescat refers to the introduction of random perturbations by compensating developed by Appel and Hoffman [1], intending an implementation of this focus that simplified the original proposal. Starting from its presentation in the first "International Seminar on Statistical Confidentiality 1992" held by ISI-Eurostat in Dublin, the work of Turmo [2] evidenced that, in the case of the variables "Place of birth" and "Academic level" for all the municipalities of Catalonia, two refinements could be incorporated: to previously treat the unique populations in the environment row/column and to adjust the marginal totals of the tables through compensating methods.

This first approach to the statistical disclosure control on tabular data was clearly placed in the segment of well-known procedures as of "modification" of confidential data and the disclosure risk of this modified method was irrelevant and it kept to high level of information (also users allowed to work with methods they are used to).

Despite the kindness and the effectiveness of the procedures applied to a set of statistical tables from 1991 Population Census of Catalonia, their applicability was committed by the concurrence of diverse elements. On one hand, the reticences on an undesirable distortion of the data original value from the users, aware of the existence of alternative and equally valid procedures guided to the "reduction" of information instead of its "modification". On the other hand, the non-existence of an appropriate computing support by the middle nineties, for the agile and powerful processing of the information that should be tabulated (including the case of hierarchical and/or linked tables) frustrate good part of the automated approaches.

Finally, the progressive orientation of the Idescat towards the articulation of the control politics of the statistical revelation in the microdata cession is justified in two elements:

- The individual records are potentially the more requested type of data by the users and/or researchers due to the growing power of their treatment teams and analysis of the primary information. This tendency is accentuated particularly in the case of data with reduced geographical levels (i.e., statistical information on European regions).

- A significant proportion of the applicable "solutions" for the preservation of the microdata statistical confidentiality is expandable to the protection of tabulated data. On the other hand, some research lines are centred in the possibility of disseminating tables of results based on a previous treatment of the individual records that tabulate in two or more dimensions (either by means of their reduction or modification).

This way, in the Idescat, procedures of statistical data protection on macrodata were being settled down based on the focus principles in favour of the reduction of the confidential information, instead of subjecting it to the opportune "modifications"[1].

2.2 Use of Privacy Homomorphisms for Macro/microdata Statistical Confidentiality

In the field of statistical data protection for macro/microdata it is necessary to point out a second relevant experience, consisting of the development of prototypes for delegation and computing data based on cryptographic techniques. The research carried out by the team of J. Domingo-Ferrer [3] allowed to evidence the possibility that the classified level (i.e. statistical institute) can recover exact statistics from statistics obtained at an unclassified level (i.e. subcontracted) on disclosure-protected dates by means of privacy homomorphisms (PHs) for multilevel processing of classified statistical data.

Although the application field of the cryptographic techniques can extend to the multiple steps of the statistical data protection -data collection, data processing or data dissemination-, like is showed in Domingo, Mateo and Snchez [4], the emphasis and the contrasts of its use by the Idescat were centred in their benefits in data processing. The experiences carried out in this last field starting from 1996 were of great utility for the treatment of specially sensitive data (e.g. individualised data of health).

At the same time, some practical limitations of the developed prototype were revealed and they derived from two facts:

- Strong restrictions on the analytic or computing capacity on the part of an unclassified level as far as doesn't allow to make bivariant statistical analysis (or of superior dimension) of the corresponding variables in a direct way.

[1] In general terms, starting from the pointed out empiric evidences, the procedures on statistical data protection applied to tabular data at the Idescat concentrate on the following rules: a) rule of the minimum frequency or minimum value of a cell (all the cells below this limit are considered confidentials). b) the dominance rule (or concentration rule) that establishes that a cell is dangerous if a minimum number of individuals contribute in more than a certain percentage of the total of the cell (it is considered as a case of predominance the one that allows to one of these individuals to deduce the value corresponding to the rest of the contributers of the cell). On the other hand, the techniques applied for the elimination of confidential cells are "primary and secondary cell supression" and changes in the outline of the table (reduction of the table dimension or recoding of its features).

- The scarce acceptance by the potential beneficiaries due, among other reasons, to the relative technical complexity that supposes the management and maintenance of this kind of platforms.

2.3 Re-identification Model for the Release of Microdata (Sub)samples

The growing microdata demand from the users on statistics and the commitment that the offices of official statistic have of giving the available information, made that the Idescat began a cession policy of these data mainly to the research centres of the Catalan universities. Evidently the cession of data should be carried out maintaining the commitment of preservation of statistical secret and therefore, controlling the disclosure risk through some of the well-known techniques. The Idescat opted to use reduction techniques due to the previous experiences in the field of the masking techniques mentioned and to the possibility of appearance of distortions not wanted in the ceded data.

This way, in 1995 the Idescat made out a microdata file with an acceptable disclosure risk level by means of the following process:

- The file contained a sample with a sample fraction of 0.4%, corresponding to more than 245.000 records, of the 1991 Population Census of Catalonia.
- Modification of the level of aggregation of the geographical and conceptual variables as instruments of control of the disclosure risk instead of using data distortion or alteration methods.
- Calculation of the disclosure risk by means of some appropriate pattern.

Once elaborated the file and before beginning their dissemination, the Idescat should make sure that the disclosure probability was small enough as to guarantee the population's privacy.

A model based on the frequency analysis was chosen, starting from the contained data in the sample, where the selection of the useful variables for the identification would be crucial in the final analysis.

Measuring the disclosure risk required the calculation of an identification probability serial based on the frequencies of unique populations in the population and in the sample. Starting from the existent methodological literature on the topic, a method based in a subsampling technique[2] presented by Zaiatz [5] was chosen.

Finally, with the encouraging results of the SDC project from the 4th Framework of European Union and, especially, the readiness of the first versions of the software packages μ-Argus and τ-Argus, the Idescat was almost forced to reconsider its methodological strategy of the statistical disclosure control for the case of micro and macrodata.

[2] This method is based on the use of a subsampling obtained from the sample of census individual records. The idea is to obtain an estimator of the proportion of unique cases in the population, starting from the empiric data given by the analysis of the observations behaviour of the subsampling regarding the sample.

This way, in 1998, the Idescat began to work with Argus, comparing the model used by Idescat for the microdata protection process and the methodological approach built in μ-Argus in order to make it possible to evaluate the results and plan for further operations in the field of statistical disclosure control. This comparison enables us to test the performance of μ-Argus from two perspectives, to check the effectiveness of μ-Argus in a real situation of unique populations, with large files and re-identification keys, and to carry out a comparative analysis of the results obtained using μ-Argus on the Idescat sample, by considering the protection criteria applied, the level of the information lost and the resulting risk of disclosure. These results will be shown in the next section of this paper.

3 Advances and Evidences in the Sure Processing of Microdata

3.1 Empirical Framework

The aim of this part is to present the results obtained when μ-Argus has been applied to a sample of individual records from Population of Catalonia in 1991 and to extend the analysis reported by Garín and Ripoll [7] in 1999. The sample size is 245.288 records corresponding to a sample fraction of approximately 0.04, and it has been disseminating between different universities and research centres applying previously a statistical disclosure control based on re-identification techniques.

But not all the variables of the file have been introduced in the analysis, only eight of them, those that have been considered as with more identification power are part of the analysed file. These variables are: Place of Residence, Place of Birth, Age Strata, Sex, Marital Status, Profession, Academic Level and Activity Situation.

It's necessary to consider that the file contains only qualitative data and that, therefore, the analysis level is considerably limited since any quantitative variable is not included. But to keep on working with the same microdata has been decided, on one hand to be able to carry out comparative valuations of the results with the μ-Argus application and on the other hand to be able to enlarge the results obtained previously.

The results obtained in 1999 showed that, indeed, the risk of statistical disclosure decreases with μ-Argus after the application of suppression techniques, in exchange for a loss of information. Now, it is sought to analyse, among other questions, how this loss of information is distributed.

The carried out empirical experiences[3] are guided to compare, in different ways, the changes produced in the loss of information done by μ-Argus. In the first place, the variation in the number of suppressions after applying the global recoding is analysed, that is to say, in function of the codification used in the two

[3] It's been used μ-Argus version 2.5 due to problems with version 3.0 installation.

main variables. On the other hand, it will be interesting to observe what happens when the number of dimensions or of combinations of tables is increased following the information introduced in the identificative level (3, 4, 5 and 6 dimensions). On the other hand, with the increment of the treshold (treshold=1, treshold=2). Finally, the loss of information is analysed from the analysis of the change in the distributions of the variables through a brief analysis of homogeneity for the cases of 3 and 4 dimensions.

3.2 Recoding of Variables

We would like to comment that, like in the previous works, the techniques of statistical disclosure control used have been, basically, the global recoding of variables and the local suppression. Also, two of the eight key variables contained in the file, (Place of Residence, Age Strata) are subjected to recode:

Place of Residence: two alternative recoding sets:
1. aggregation in four categories corresponding to broad administrative divisions
2. aggregation in sixteen categories corresponding to groups of local counties with low distribution variance

Age Strata: we know that one-year stratas are dangerous, therefore we established two aggregation levels:
1. categories of 5 years
2. categories of 10 years

The combination of these variables produces the following four files:

	Age Strata	
Place of Residence	5 years strata	10 years strata
4 categories	File 4_5	File 4_10
16 categories	File 16_5	File 16_10

3.3 The Identification Level and Treshold

The most revealing variables from the data matrix have been chosen because the disclosure risk must be controlled through the analysis of these possible combinations and their identification levels. To make a decision, the following aspects have been taken into consideration:

- Experiences in previous similar operations
- The type and contents of local files that contain individual data and are accessible to the public
- The quality of the information regarding the current value of some variables, the reliability of the matching of codified data, etc.

On the other hand, it has been already mentioned that this analysis contains a 3 to 6 dimensions comparison. So, in the sixth column of the metafile, we have defined an identification level distribution for each dimension. We also should establish that, if we want to compare the results between dimensions, the distribution of the identification levels mustn't change so much. So we only change the identification level of four variables in order to obtain the fourth, fifth and sixth dimension.

Remember that, to the highest identification level corresponds the smallest value and that μ-Argus always requires that the identificative levels are correlative:

3 dimensions	**IL**	**4 dimensions**	**IL**
Place of Residence	1	Place of Residence	1
Place of Birth	2	Place of Birth	2
Age Strata	1	Age Strata	1
Sex	3	**Sex**	**4**
Marital Status	3	Marital Status	3
Profession	2	Profession	2
Academic Level	2	Academic Level	2
Activity Situation	3	Activity Situation	3

5 dimensions	**IL**	**6 dimensions**	**IL**
Place of Residence	1	Place of Residence	1
Place of Birth	2	Place of Birth	2
Age Strata	1	Age Strata	1
Sex	**5**	**Sex**	**6**
Marital Status	**4**	**Marital Status**	**5**
Profession	2	Profession	2
Academic Level	2	**Academic Level**	**3**
Activity Situation	3	**Activity Situation**	**4**

Where in bold appear the variables that have changed their identification level in relation to the previous combination of dimensions.

It has been necessary to force a change in the variable Academic Level to maintain an identification level similar to 3 since the original metafile (3 dimensions) presents only three variables with an identification level equal to 3. For this reason, the increment in the last combination of dimensions presents the change in four variables. It could be noted that the two variables subjected to recode and which produce the different four files to analyse are those that have a superior identification level (IL=1).

The treshold is the other big decision that the controller should take in order to obtain a safety file. In this paper we will compare the increment of suppressions when we increase the treshold from 1 to 2.

4 Empirical evidences: the distribution of the information loss

4.1 Number of Suppressions

The following tables show the number of suppressions that the program μ-Argus carries out to obtain a safe file and how this number varies in function of the used file (according to codification), the number of dimensions and the treshold value used.

Table 1. Number of suppressions

File 4_10				
	3 dimensions	4 dimensions	5 dimensions	6 dimensions
Treshold=1	930	6016	12480	16179
Treshold=2	1894	10230	19666	25313

File 4_5				
	3 dimensions	4 dimensions	5 dimensions	6 dimensions
Treshold=1	1631	9548	18489	23010
Treshold=2	3343	15953	28530	35324

File 16_10				
	3 dimensions	4 dimensions	5 dimensions	6 dimensions
Treshold=1	2330	13563	26480	33293
Treshold=2	4709	21963	39642	49736

File 16_5				
	3 dimensions	4 dimensions	5 dimensions	6 dimensions
Treshold=1	3415	19909	37115	45157
Treshold=2	6901	32123	55124	66893

Obviously, the number of suppressions increases with the number of dimensions as well as with the increment in the treshold value. On the other hand, the following tables are arranged of smaller to more number of suppressions according to the analysed file.

This way, it's noted that the file previously surer is the File 4_10 and the one that needs more suppressions, that is to say, the less sure file, is the File 16_5. This aspect works together with the number of categories corresponding to the two recoded variables. The number of combined categories of the variable Place of Residence and Age Strata in the four files is:

File 4_10 15
File 4_5 25
File 16_10 27
File 16_5 37

In the analysed cases, the number of suppressions increases when increases the level of detail of the microdata.

It could also be seen that the variable Place of Residence has bigger importance in the increment of the number of suppressions because the two files with a more added stratification regarding this variable are located in the first two places of the ordination.

But the number of suppressions doesn't agree with the number of records affected by some suppression since there are records that have suffered more than one local suppression. The fact of suppressing more than one variable for each record favours the fact that a bigger percentage of records remain intact but, on the other hand, is neither advisable to have an excess of records with lack of information in more than one variable. In this sense, Table 2 shows the percentage of records with more than one local suppression.

Table 2. Percentage of records with more than one suppression

		dim 3	dim 4	dim 5	dim 6
File 4_10	treshold 1	0.11	1.28	2.49	2.68
	treshold 2	0.37	2.07	4.02	4.66
File 4_5	treshold 1	0.18	1.51	2.87	3.17
	treshold 2	0.48	2.12	4.95	5.80
File 16_10	treshold 1	0.39	2.12	3.54	3.78
	treshold 2	0.71	3.52	5.78	6.53
File 16_5	treshold 1	0.71	2.71	4.48	4.72
	treshold 2	1.29	4.74	7.40	8.33

It could be seen how the increment in the detail level of the information in the number of dimensions and in the number of the treshold value, increases the percentage of records with more than one local suppression reaching relatively high values in some cases.

It could be also pointed out that in the analysis to 3 dimensions, the maximum number of record local suppressions is 2, and starting from the dimension 4, the maximum number of record local suppressions is 3. On the other hand, it has not been noted in any case that more than 3 variables in a record have been suppressed.

4.2 Distribution of the Suppressions by Variables

Another way to analyse the number of local suppressions is from the point of view of how this suppressions change in each one of the eight variables. In this sense, Table 3 contains the ratio of variation suppressions when increasing one dimension. The table is shown only for one of the four files analysed since almost the same behaviour pattern in all of them repeats.

It could be noted that in the step from 3 to 4 dimensions the variables Place of Birth, Sex Marital and Status possess a very superior ratio than the other

Table 3. Variation ratio in the variables by dimension changing

File 4_10	3 to 4	4 to 5	5 to 6
Place of Residence	4.27	3.78	1.26
Place of Birth	25.81	2.76	1.21
Age Strata	3.03	2.07	1.10
Sex	46.75	4.82	1.96
Marital Status	13.30	1.60	1.11
Profession	2.73	1.54	1.10
Academic Level	5.67	2.41	1.32
Activity Situation	4.21	1.27	1.06

ones. It is also observed how the ratios are equalled in the step from 5 to 6 dimensions. The same behaviour repeats when we work with a treshold=2.

Another question to think about is how the suppressions are distributed among the variables, that is to say, which variables suffer more suppressions. The following table shows this aspect through the percentage of suppressions regarding the total.

Table 4. Variation ratio in the variables by dimension levels

File 4_10	3 dim	4 dim	5 dim	6 dim
Place of Residence	27.6	21.1	33.2	32.3
Place of Birth	2.9	13.4	15.4	14.4
Age Strata	3.2	1.7	1.5	1.3
Sex	0.9	7.2	14.4	21.8
Marital Status	8.8	21.0	14.0	12.0
Profession	28.0	13.6	8.8	7.4
Academic Level	1.9	2.0	2.0	2.0
Activity Situation	26.7	20.1	10.6	8.7

When working with 3 dimensions, we see how the 80% of the suppressions concentrate on the variables Place of Residence, Status Profession and Activity Situation.

A singularity must be highlighted: the variables Profession and Activity Situation reduce their weight in the number of suppressions as the number of dimensions is increased, while the variable Sex suffers the opposing phenomenon. It increases the percentage of suppressions regarding the total as advances in the number of analysed dimensions are carried out. This behaviour pattern repeats in the four files and when we work with a treshold=1 or treshold=2.

Only another one more aspect to highlight. In the File 16_10 the percentage of suppressions regarding the total in the variable Place of Residence decreases until values inferior to 5%, although, intuitively, we could think that with an increment in the variable detail an increment in the number of suppressions would take place.

4.3 Increase of suppressions in dimension and treshold values changing

It is obvious that with the increment in the number of dimension combinations an increment of the number of local suppressions takes place because the number of combinations analysed is bigger as well as the identification risk. The same effect takes place with an increment in the treshold value because the records in danger of being not identified are not only those which are unique in the population. It's enough that only two records were equal (treshold=2) to mark this records as conflicting.

In this section it is analysed, by means of increment ratios and for each one of the four files, how the loss of information or number of local suppressions vary *when we increase the dimensions* from 3 to 6 on one hand and from 1 to 2 in the treshold value. The ratio built is simply the quotient between two numbers of local suppressions to be able to see by which value has been multiplied the suppression number as the Table 5 shows.

Table 5. Ratio variation in dimension changing with treshold=1 and treshold=2

Treshold=1	**3 to 4**	**4 to 5**	**5 to 6**
File 4_10	6.5	2.07	1.3
File 4_5	5.9	1.94	1.2
File 16_10	5.8	1.95	1.3
File 16_5	5.8	1.86	1.2

Treshold=2	**3 to 4**	**4 to 5**	**5 to 6**
File 4_10	5.4	1.9	1.3
File 4_5	4.8	1.8	1.2
File 16_10	4.7	1.8	1.3
File 16_5	4.7	1.7	1.2

By observing the previous tables it seems interesting to highlight:

- A very high ratio is produced in the step from 3 to 4 dimensions, but in the following step, from 4 to 5 dimensions, this ratio falls to practically the third part and it continues falling, although in smaller measure, in the step from 5 to 6 dimensions where the ratio for each one of the four files is equalled. That is to say, as we advance in the number of analysed combinations, the increment of suppressions in ratio terms spreads to be constant.
- When we increase the treshold value from 1 to 2, it is observed that from 3 to 4 dimensions the ratio has decreased in a constant similar to 1.1, of 4 to 5 dimensions the ratio decreases in a constant of 0.15 and in the step from 5 to 6 dimensions the ratio stays unalterable.

These results allow figuring the number of suppressions that will take place when applying μ-Argus with an increment of dimensions or of the treshold value.

That is to say, if the ratios were known, the program could be applied only for a combination of dimensions equal to 3 and a treshold value equal to 1 and to estimate the changes that would take place with more dimensions or higher treshold values.

Now we will measure in Table 6 the changes in the number of suppressions in each one of the *files when passing from a treshold=1 to a treshold=2*. That is to say, for each file, it will be analysed for each fixed level of dimensions, how the fact of increasing a unit the treshold value affects the result.

Table 6. Ratio variation in treshold changing with dimension levels

	3 dim	4dim	5 dim	6 dim
File 4_10	2.04	1.70	1.58	1.56
File 4_5	2.05	1.67	1.54	1.54
File 16_10	2.02	1.62	1.50	1.49
File 16_5	2.02	1.61	1.49	1.48

It is quickly observed how, inside each dimension level, when passing from a treshold value 1 to 2, the ratio doesn't almost vary with the used code level. This way, when working with a dimension level 3, the number of suppressions duplicates in the four files. With a dimension level equal to 4, it multiplies by approximately 1.65 and in higher levels also takes place a new interesting fact: the ratio stays with what seems to be a limit located in approximately 1.5.

Summing up, these results seem to indicate that when the treshold value is increased, the quotient among the number of suppressions tends to not depending on the number of increasing changing dimensions.

4.4 Disclosure Risk Decreasing

Two kinds of analysis have been made to measure how the disclosure risk decreases when we obtain a safe file by means of μ-Argus. In fact, it is an approaching attempt to the protective effectiveness of μ-Argus software by calculating the decrease in the disclosure risk. For this reason, firstly, the results have been compared on the four files applying a local suppression in 3 and 4 dimensions and applying the calculation of the disclosure risk through re-identification techniques.

Evidently the identification risk decreases as more combinations of variables are analysed but in exchange for increasing the level of lost information. A considerable decrease on disclosure risk takes place when applying a dimension increment (see Table 7).

On the other hand, an experiment has been carried out consisting on obtaining two independent samples (10,000 records) from our original file with the same sampling fraction (0.404) that the one that represents this file regarding the total population, as we have seen in section 2. Later on, to one of the samples, the program μ-Argus has been applied with a treshold equal to 1 and with

Table 7. Disclosure risk decreasing in dimension changing from 3 to 4

	%
File 4_10	32.09
File 4_5	31.10
File 16_10	29.45
File 16_5	39.46

combinations of 4 dimensions. After calculating the disclosure risk in the two samples, the decrease of this risk has been measured, as it could be seen in Table 8.

Table 8. Disclosure risk decreasing with vs without μ-Argus performance (treshold=1, dim=4)

	%
File 4_10	66.98
File 4_5	72.60
File 16_10	59.71
File 16_5	69.56

4.5 Homogeneity of the Variables Distribution

One must keep in mind that when a file with microdata is spread, what is sought is that the analysts obtain some results the most similar possible to the original information, that is to say, the variables should maintain the distribution function.

To see how the option of local suppression acts in μ-Argus a test of homogeneity has been carried out comparing the variables distributions in the original file with the distributions of the safe file produced by μ-Argus. This test has been carried out for each one of the four analysis files, with combinations of 3 and 4 dimensions, with a treshold value equal to 1 and at a level of analysis univariant.

Table 9. Homogeneity test in 3 dimensions

3 dimensions

	Homogeneity?			
	File 4_10	File 4_5	File 16_10	File 16_5
Place of Residence	Yes	Yes	Yes	Yes
Place of Birth	Yes	Yes	Yes	Yes
Age Strata	Yes	Yes	Yes	Yes
Sex	Yes	Yes	Yes	Yes
Marital Status	Yes	Yes	No	No
Profession	Yes	Yes	No	No
Academic Level	Yes	Yes	Yes	Yes
Activity Situation	Yes	Yes	Yes	No

Subsequently, two tables are presented (Tables 9 and 10) with the result of the test in terms of to reject or not to reject the variable, that is to say, if the function of variable distribution has suffered important changes or stays between some acceptable margins.

It is observed that most of the changes are not significative, which means that the null hypothesis that the distribution function has not varied is accepted. The files with an Age Strata with 10 year-old strata don't reject any variable while the files with strata of 5 year-old age present problems in the variables Marital Status, Profession and Activity situation, this last one only in the File 16_5.

Table 10. Homogeneity test in 4 dimensions

4 dimensions	Homogeneity?			
	File 4_10	File 4_5	File 16_10	File 16_5
Place of Residence	No	No	Yes	No
Place of Birth	No	No	No	No
Age Strata	Yes	Yes	No	Yes
Sex	No	Yes	Yes	Yes
Marital Status	Yes	No	No	No
Profession	Yes	No	No	No
Academic Level	Yes	Yes	Yes	Yes
Activity Situation	Yes	No	No	No

In view of the previous results, we can highlight the following aspects:

1. When passing from 3 to 4 dimensions, the increment of rejected variables is considerable, one of them is rejected systematically (Place of Birth). On the other hand, the Education Level stays unalterable and the other ones depend on the file we are working with.
2. If the distribution of the variables are analysed one to one, we can see that the variables more easily rejected are those distributed internally in a heterogeneous way and, however, the suppressions are not carried out in a similar proportion but are distributed homogeneously among the different categories. This causes that some category maintains its weight practically intact while, in another one, it could have varied considerably.
3. It has also been observed that it is more frequent that a variable with many categories would be rejected easier than a variable with few categories, although this condition is not always completed like could be observed with the variable Education Level (21 categories) or with the variable Place of Birth (3 categories).
4. Bivariate analyses have been carried out but in most of the cases the homogeneity test doesn't support the null hypothesis.
5. Although it has not still been tested empirically, we intuitively suppose that if the number of dimensions or the treshold value increases more, the number of rejections will continue increasing.

6. We also point out the possibility to carry out other kind of homogeneity tests for qualitative data where the variation of marginal totals would be taken into consideration, as McNemar test, for instance.

5 Global Assessments

In this paper some empirical results in the application of the software package μ-Argus have been showed, and supplement other works carried out in the Idescat. Of those results the following valuations can be extracted as the most important:

1. For the construction of the safe files, and the consequent rising decrease of the disclosure risk, the options of global recoding and local suppression have been used. A decrease of the disclosure risk has been achieved indeed and the obtained files can be considered sure enough.
2. On the other hand, with the increment in the number of dimensions and the treshold value, the quantity of lost information tends to be too high as for not keeping in mind that the safe file cannot be satisfactory. In fact, we have seen how the probability distributions of certain variables could be modified, especially the more homogeneous the suppressiones would be among the categories in a heterogeneous variable.
3. Argus works under the hypothesis that the processed file refers to the global population and not to a sample of the same one. When working with samples the disclosure risk decreases in great measure and, therefore, it would maybe not be necessary to carry out a number so high of suppressions since the unique populations in the sample shouldn't be unique in the population. It would be interesting to be able to introduce a parameter that will take into account this fact.
4. It would be convenient, on one hand, more precise and extensive documentation on different options, algorithms and proceedings used by Argus and, on the other hand, some analysis procedures in the same software package to evaluate its effectiveness, like basic descriptive analysis, analysis of the suppressed data distribution, and so on..

As partner of the CASC project in the 5th Framework program of European Commission, the Idescat hope to continue testing the software package Argus in order to extend the results to both social and business data and to improve our policy and guidelines about the statistical disclosure control.

References

1. Appel, G. and Hoffman, D.J.: "Perturbation by compensation". Preproceedings of the International Seminar on Statistical Confidentiality, ISI-Eurostat. Dublin, September 1992. Final version published by Eurostat in 1993.
2. Turmo, J.: "Random perturbations by compensation: a method to protect confidential statistical information" [in Catalan]. Quaderns d'Estadística i Investigació Operativa (Qüestiió), **17**, 3, 1993, 413–435.

3. Domingo-Ferrer, J.: "Privacy homomorphisms for statistical confidentiality". Quaderns d'Estadística i Investigació Operativa (Qüestiió), **20**, 3, 1996, 505–521.
4. Domingo-Ferrer, J., Mateo, J.M. and Snchez, R.X.: "Cryptographic techniques in statistical data protection". Joint ECE/Eurostat Work Session on Statistical Confidentiality. Thessaloniki, Greece, March 1999. Final version published by Eurostat in 1999.
5. Zaiatz, L.V.: Estimation of the percent of unique population elements on a microdata file using the sample. Bureau of Census. Statistical Research Division Report Series (1991), Census/SRD/RR-91/08.
6. Garín, A.: "Control of statistical disclosure risk in dissemination of demographical microdata, applied on a sample of individual records from Catalonia's Population Census 1991" [in Catalan]. Quaderns d'Estadística i Investigació Operativa (Qüestiió), **20**, 3, 1996, 523–546.
7. Garín, A. and Ripoll, E.: "Performance of μ-Argus in disclosure control of uniqueness in populations". Joint ECE/Eurostat Work Session on Statistical Confidentiality. Thessaloniki, Greece, March 1999. Final version published by Eurostat in 1999.