

Topic (i): new applications of disclosure control methods

**PERFORMANCE OF μ -ARGUS IN DISCLOSURE CONTROL OF UNIQUENESS IN
POPULATIONS**

Submitted by Statistics Catalonia, Spain¹

Invited paper

I. INTRODUCTION

Objective

1. μ -ARGUS is a software oriented towards safe data production; safe data is one of the main goals when a Statistical Office plans a dissemination of microdata. The ARGUS system can be shown like a set of implements, useful for the controller who has responsibility on disclosure risk analysis. The controller must obtain a file that fulfils two requirements: the delivered data must have the highest information significance and the lowest possible disclosure risk; thereby, he/she needs to know the values of two parameters to evaluate if the chosen strategy is correct. These parameters are: a) any measure of the loss of information caused by the application of the control methods, and b) any quantification of risk of disclosure.

2. The aim of this paper is to make a detailed statement of the results obtained through the application of the μ -ARGUS module to a sample of individual records from the Population of Catalonia in 1991; this paper also shows a comparative analysis based on the outputs obtained in the past by a similar operation carried out by the Statistical Institute of Catalonia (IDESCAT).

Control Methods

3. When approaching SDC, the use of the re-identification model, based on a scenario in which an intruder is able to identify a specific person from the previous knowledge of some characteristics corresponding exactly with some values present in the microdata file, is widely spread. In this model, key variables have a major role; key variables are those which have more identification power. A combination of key variables, used to identify an individual, will be considered an "identification key". Then, the crucial issue - working with this model - is to select the identification key to be analysed.

¹ Prepared by Alfons Garín (Polytechnical University of Catalonia) and Enric Ripoll (Statistical Institute of Catalonia (IDESCAT)). The authors wish to thank Mr. Julià Urrútia of IDESCAT for his help with the requirements analysis, gathering sample data and suggestions for improvements.

4. The object of the control methods is to avoid the presence of identification keys with a high disclosure risk in the microdata file. The main control methods are: a) global recoding of variables – applied to all records using usual mechanisms: collapsing and aggregating categories, etc.; b) local suppression of variables which render a record to be dangerous – this procedure is applied to the dangerous record only-; c) modification of data by different procedures: random perturbations, micro-aggregating, etc.

5. On the similar previous operation, IDESCAT did apply the re-identification model using the global recoding procedure and the suppression of singular records that showed high disclosure risk; local suppression wasn't used. The final output was a microdata file that was considered to be the best option between “high information and low disclosure risk”. In this second version - using the Argus system - we start from the same initial conditions (sample fraction and file structure). In the same way, we take advantage of the intermediate results - like the recoding system for some variables, especially for “Place of Residence” and for the variable “Age” grouped in strata of 5 and 10 years - and we take advantage of our knowledge of the most revealing variables too, according with the most common public information about individuals in our country.

II. INITIAL TASKS (preparation of the problem)

II.1 Structure of the microdata file

6. On population microdata files, IDESCAT follows the rule of delivering only samples and never complete census files. The first task is to obtain a simple random sample with an adequate size from the census file; in order to get an adequate significance level regarding the main variables. The structure of the sample is plain, with each row belonging to one individual and each column representing a different variable:

- a) The household characteristics
- b) The personal data: age, sex, marital status, academic level, profession, etc.
- c) The individual location data: place of residence, place of birth, mobility, etc.
- d) Other kinds of information like language, date of arrival to the country, etc.

7. The final size of the sample is of 245,288 records corresponding to a sample fraction of approximately 0.04. The resulting sample underwent a significance test regarding the main population parameters in order to validate the extraction process.

II.2 Definition of the key variables

8. Now, we must choose the most revealing variables from the data matrix. For instance, we can suppose a scenario where the knowledge of the age of an individual, combined with the place of residence and sex, allows the chance to match unmistakably an individual with a unique data record present in the sample; in this case, disclosure risk must be controlled through the analysis of these possible combinations.

9. For this task, the controller did check the following facts:

- a) Experience in previous similar operations to this case;
- b) The type and contents of local files with individual data of public access;
- c) The quality of the information regarding the current value of some variables, the reliability of the matching of codified data, etc.

10. In accordance with these considerations, we defined the most revealing variables: *Place of residence, Place of Birth, Age, Sex, Marital Status, Academic Level, Profession, and Activity Situation*.

II.3 The *rda* files in the Argus System

11. The **.rda* files contain a part of the metadata of the process: the structure of the microdata file and the characteristics of the variables subject to the disclosure control. In these files, the columns 1, 2 and 3 define the structure; columns 4 and 5 inform about the missing codes; columns 6 to 9 have a strong influence on the output of the process:

- a) The 6th column specifies the revealing power of each key variable; ARGUS can generate combinations in according with the composition of this column. This is the instrument used to minimise the disclosure risk.
- b) The 7th column indicates the suppression strategy when a record represents one or more dangerous combinations. This is the instrument for minimising the importance of the loss of information.
- c) The 8th column is useful to mark the variables containing location data; the information of this column allows us to connect the microdata file with the land coding system when the controller runs the global recoding procedure.
- d) The 9th column identifies which variable can be truncated; this is an easy way of global recoding.

12. In order to construct our **.rda* file we included the key variables only, plus a **Number ID** that will be used to merge the rest of the data matrix in the final file composition.

We used the following **.rda* file through all the iterations:

	1	2	3	4	5	6	7	8	9
NumberId *	1	11	0	0	0	0	0	0	0
PlaceResid	12	2	0	0	1	8	2	0	0
PlaceBirth	14	1	0	0	2	4	2	0	0
AgeStrata	15	2	99	99	1	7	0	0	0
Sex	17	1	00	0	3	6	0	0	0
MaritalSta	18	1	00	0	3	5	0	0	0
Profession	19	2	00	0	2	1	0	0	0
AcademicL	21	2	00	0	2	2	0	0	0
ActivitySit	23	2	00	0	3	3	0	0	0

RDA File

II.4 Global Recoding: **.rdc* files in the Argus System

13. **.rdc* files contain the rest of the metadata system that the process uses. The **.rdc* files are the correspondence tables between different information levels for a set of variables. We used this procedure on two variables: Place of Residence and Age.

a) **Place of Residence.** The original information level (Municipality) is too low given the very small size of a lot of municipalities. The inconvenient distribution forced us to apply two alternative recoding sets: i) aggregation in 4 categories corresponding to broad administrative divisions; ii) aggregation in 16 categories corresponding to groups of local counties with low distribution variance.

b) **Age Strata.** The original data contain the date of birth of the individual. We know that one-year strata are dangerous. Therefore we establish two aggregation levels: i) categories of 5 years; ii) categories of 10 years.

In short: we analyse the four combinations resulting of the use of four alternative recoding systems.

Place of Residence	5 years strata	10 years strata
4 Provinces	SAM_4_5 (file)	SAM_4_10 (file)
16 Regions	SAM_16_5 (file)	SAM_16_10 (file)

II.5 Analysis of the variable combinations

14. μ -ARGUS offers two basic options to analyse key variable combinations with potential disclosure risk:

a) Automatic table generation using information from the sixth column of the **.rda* file. The maximum dimension of the generated tables is 3, and the controller defines the threshold, common for all tables.

b) Manual table generation, where dimension of tables and threshold are defined by the controller.

15. The final output depends on this option, taking into account that the larger dimension of the table to be analysed, the higher the number of local suppressions that will be necessary to protect data. It will be interesting to know the expected inverse relation between both parameters: the loss of information and the disclosure risk, through the observation of the joint variation of these values.

II.6 Phases of the process

16. Given the high number of analysis options available, we decided to study a limited number of variants using, in all cases, the four combinations of the **.rdc* files formed above. Therefore we followed the following sequence of tasks:

- To generate automatic tables according to the sixth column of the **.rda* file with a threshold of 2.
- To define combinations with dimension = 4 and threshold = 2.
- To calculate the disclosure risk for the eight resulting safe data files, considering an identification key composed by the eight key variables defined in the sixth column.
- To show the parameter's values (number of suppressions and disclosure risk) in a table for every case, with percent variations as an indicator of recoding procedure efficiency.
- To compare ARGUS' outputs with previous IDESCAT outputs.

III. THE ARGUS SYSTEM'S OUTPUTS

17. The ARGUS system produces two final outputs:

- a) A safe data file, with local suppressions imputed through missing code values (eighth column of *.rda file).
- b) A report on lost information (*rep* files), specifying the local suppression distribution between variables.

In our process, we made the assumption that the loss of information caused by the global recoding procedure is constant whether we check options a) or b) from 2.6

Disclosure risk analysis in a safe data file

18. We need to complete the outputs available from the ARGUS system, through an external process by applying an algorithm to calculate the risk of disclosure of the safe data file if it were delivered. It's necessary to have at least a risk indicator, suitable to compare results. Therefore, we define the probability of identification:

$$p(\text{identification}) = f * p(\text{unique individuals in the population}) * p(\text{unique sample record is unique within population})$$

19. We define the following variables:

PUP : Proportion Unique individuals in Population

PUS : Proportion Unique records in Sample

PUSUP: Proportion Unique records in Sample that are also Unique in Population.

20. Our problem is that we don't know PUP; so we must estimate this factor of $p(\text{identification})$. Using the subsample method [1] we obtain the proportion of unique records in subsample that are also unique in the sample, like an estimator of PUSUP; then, $PUP(\text{estimate}) = PUS * PUSUP$. Finally, $p(\text{identification}) = f * PUP * PUSUP$.

IV. FINAL RESULTS OF THE PROCESS

21. Distribution of loss of information:

- a) analysis of key variable combinations of dimension: 1..3;
- b) analysis of key variable combinations of dimension: 1..4;

a) **local suppression; combinations: dimension=1..3, threshold=2 (all tables)**

	RegioResid	RegioBirth	AgeStrata	Sex	MaritalSt	Professio	Academ	Activity	Totals
Sam_4_10	515	40	84	9	157	558	41	490	1894
Sam_4_5	584	100	66	33	392	1084	198	886	3343
Sam_16_10	262	148	457	27	628	1679	96	1412	4709
Sam_16_5	1282	187	107	45	833	2272	256	1919	6901

table 1

b) *local suppression* ; combinations: dimension=1..4, threshold=2 (all tables)

	RegioResid	RegioBirth	AgeStrata	Sex	MaritalSt	Professio	Academ	Activity	Totals
Sam_4_10	3030	1371	199	698	1921	1336	279	1793	10627
Sam_4_5	4345	1959	257	1005	2903	2492	740	2535	16236
Sam_16_10	1390	3237	1059	1104	4142	6402	575	4358	22267
Sam_16_5	4157	4339	490	1373	5514	9589	1245	5610	32317

table 2

Table of distribution of disclosure risk (subsample method), and the total of loss of information:

- Residence Regio: **4** categories; **16** categories.
- Age: **10-year** strata; **5-year** strata.
- Combinations analysed: table's dimension = 3; table's dimension = 4 (threshold = 2 for all tables)

FileName	Total Suppres	Risk %	Subsample Size	Unique sample	Unique subsamp	True Unique	PUP	PUS
Safe_4_10_3dim	1894	0,0134	9559	16821	2998	660	0,0150	0,0685
Safe_4_10_4dim	10627	0,0091	9918	15295	2995	571	0,0118	0,0623
Safe_4_5_3dim	3343	0,0254	10025	23263	3801	979	0,0244	0,0948
Safe_4_5_4dim	16236	0,0175	9788	21030	3709	835	0,0193	0,0857
Safe_16_10_3di	4709	0,0489	9862	34075	4563	1347	0,0410	0,1389
Safe_16_10_4di	22267	0,0345	9912	30463	4675	1226	0,0325	0,1241
Safe_16_5_3dim	6901	0,0887	9825	44835	5354	1856	0,0633	0,1827
Safe_16_5_4dim	32317	0,0537	9976	38988	5332	1543	0,0459	0,1589

table 3

* **f** (sample fraction) for sample and subsample = 0.0404. Sample size = 245288.

PUP : Proportion of unique individuals in the population (estimation):

$$\text{PUP (estimate)} = (\text{Usample} / \text{SampleSize}) * (\text{TrueUnique} / \text{Usubsam})$$

PUS: Proportion of records which are unique in the sample (Usample / SampleSize).

Unique sample (Usample): Unique records in the sample

Unique subsam (Usubsam): Unique records in the subsample

True Unique: Unique records in the subsample also unique in the sample

The analysis of variation of disclosure risk vs. variation of loss of information

A) *impact of the dimension of the analysed combinations*

the change of loss of information vs. the disclosure risk variation

"Safe(x,y)_3 dimension analysis files vs. "Safe(x,y) 4 dimension analysis files

x: Place of Residence (4 regions or 16 regions); **y**: Age Strata (5 years or 10 years)

File(1) vs. File(2)	Risk Variation	Loss of Information Variation	Factor RV / LIV *100
Safe_16_10_3dim v Safe_16_10_4dim	-30%	372%	8
Safe_16_5_3dim vs. Safe_16_5_4dim	-39%	368%	10,5
Safe_4_10_3dim vs. Safe_4_10_4dim	-31,80%	461%	6,8
Safe_4_5_3dim vs. Safe_4_5_4dim	-45,40%	385%	8

table 4

The factor RV/LIV = 10,5, means that the most favourable change of 3 dimension analysis to 4 dimension analysis is done on Safe_16_5.

B) impact of recodification

Increments of disclosure risk and loss of information with a constant analysis level (3 or 4 dimensions)

Files	Suppress 3dim	Risk 3dim	Suppress 4 dim	Risk 4 dim
Safe_4_10 vs. Safe 4_5	76%	89%	52,70%	91,73%
Safe_16_10 vs. Safe 16_5	46%	81%	45%	55,80%
Safe_4_10 vs. Safe 16_10	148,60%	264,30%	109,50%	276,40%
Safe_4_5 vs. Safe 16_5	106,40%	249,07%	99,00%	206%

table 5

In the change: Safe_16_10 to Safe_16_5 the best relation is obtained (minimum loss of information and minimum increment of the risk).

C) the distribution of the risk (IDESCAT 1995)

Place of Residence	Age Strata		
	Strata 1 year	Strata 5 years	Strata 10 years
Catalonia	risk = 0.0455 %	risk = 0.0068%	risk = 0.0051%
4 Provinces	risk = 0.1093%	risk = 0.0194%	risk = 0.0144%
16 Regions	risk = 0.3264%	risk = 0.0636%	risk = 0.0443%

table 6

(using the same algorithm for risk estimation that was used in table 3)

V. GLOBAL ASSESSMENTS

22. IDESCAT (1995) delivered a microdata file corresponding to the configuration: **SAM_16_5** (Place of Residence stratified in 16 regions and Age Strata of 5 years), with an estimate disclosure risk of **0.0636%**; the suppression method was applied on single records, mainly corresponding to rare combinations of Profession Class or Academic Level with small regions. The outputs of ARGUS are similar but improved: **safe_16_5_4dim** from table 3 with an estimate disclosure risk of **0.0537%** is better option than IDESCAT SAM_16_5.

23. Regarding performance, our first experience with ARGUS has been satisfactory. However, the following issues should be highlighted:

- The high flexibility of the ARGUS system to change initial specifications, is of great help to the controller.
- About the speed of the process: in our experience (more than 245.000 records), we obtained a high speed for the process; though, when analysing tables of dimension 5 or higher, it decreases dramatically.
- It would be desirable to include the option of exporting the file **.rep* (report about loss of information) to any structured file: a Text Fixed Format (columns: Variable name, number of suppressions), MS Excel, Lotus, MS Access, etc.
- It would be useful to have a new statistical module to integrate the post-analysis, taking into account that a safe data file is an intermediate result: we need to test the population parameter estimates from the safe file, the disclosure risk variation, etc.; in this way, ARGUS appears as an isolated phase in the global statistical process when a microdata dissemination is planned. We suggest that it would be a good alternative to develop, for instance, a SAS or a SPSS application though the high cost of the runtime could be dissuasive; however, now we have WEB utilities to share networking software, and we have also developed encrypted techniques for data transmission an restoration.

24. **Documentation.** We do not have a very rich documentation at hand for understanding completely the inner operations of the ARGUS system; for instance, of the treatment of missing values. Any feeling of a black box inside the system should be eliminated by means of good documentation.

[1] **Voshell Zaiatz, L.** (1991). *Estimation of the percent of unique population elements on a microdata file using the sample.* Bureau of Census. Statistical Research Division Report Series. Census/SRD/RR-91/08.