

**STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE  
EUROPEAN COMMUNITIES**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROSTAT**

**Joint ECE/Eurostat Work Session on  
Statistical Data Confidentiality**  
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 3 (Summary)  
English only

Topic (i): new applications of disclosure control methods

**PERFORMANCE OF  $\mu$ -ARGUS IN DISCLOSURE CONTROL OF UNIQUENESS IN  
POPULATIONS**

Submitted by Statistics Catalonia, Spain<sup>1</sup>

**Invited paper**

**Summary**

1. The immediate antecedent of this analysis is the work done by the Statistical Institute of Catalonia (IDESCAT) between 1994-95 with the aim of obtaining a microdata file with an acceptable degree of security related to the potential risk of statistical disclosure. The resulting file was a random sample with individual anonymised registers coming from the Population Census of Catalonia 1991.

2. The coincidences of the model used in the operation carried on by IDESCAT with the methodological approach built-in  $\mu$ -Argus module constitutes an attractive field to compare the results and plan improvements in further operations of statistical disclosure control (SDC). Concretely, all this could allow to verify the performance of  $\mu$ -ARGUS in that field providing a double perspective:

(a) To check the effectiveness of  $\mu$ -ARGUS in a real situation of unique populations, with large files and reidentification keys (from six to eight variables).

(b) Comparative analysis of the obtained results with  $\mu$ -ARGUS related to the IDESCAT sample, by means of the protection criteria applied, the level of the information lost and the measuring of the resulting risk of statistical disclosure.

**A. Models of statistical disclosure: a short description**

3. The  $\mu$ -Argus program uses an identification model by applying the threshold rule: a safe microdata file doesn't contain any combination of keywords below a pre-established level. IDESCAT also applies an identification model based on the key-variables scenario, but, in this case, it is used as

---

<sup>1</sup> Prepared by Alfons Garín and Enric Ripoll.

an estimator of the probability that a unique combination in the sample is also unique in the population, subject to a safety rule consisting that its probability is very low.

4. The coincidence of both models allows to apply  $\mu$ -Argus to a population sample obtained with the same sampling fraction (4% approximately) by using the experience resulting in the operation done by IDESCAT related to the identification ability of the available variables. However, there are some remarkable differences in the procedures, as well as the safety rule mentioned above.

## **B. Statistical disclosure control procedures**

5. Global recoding consists of adding categories or strata when the level of key-variable desaggregation produces unsafe combinations. It is the unique control procedure used by IDESCAT;  $\mu$ -ARGUS implements it as well.

6. In local suppression, the variable value of a register is suppressed when it is the cause of a high disclosure risk (in the case of  $\mu$ -ARGUS when it causes that the frequency of variable combination is below the level pre-established and according to the priorities of suppression marked by the controller).

7. The variable suppression process provides a larger flexibility to the controller activity given that the two procedures are complementary but knowing the loss of information in each case is a strategic matter. On the other hand, the statistical model applied by IDESCAT accepts unique combinations in the samples (below level = 2), taking into account that the probability of establishing that it will be also unique in the population is very low. If the threshold rule is applied, it would be convenient to limit the number of keywords in order not to obtain an excessive loss of information; in this sense,  $\mu$ -ARGUS envisages a high flexibility in the manipulation of the rda file.

## **C. The process**

8. The general stages of the process carried out are the following:

- (a) Sample obtaining; generation of a simple random sample, using the Bernoulli method, with  $f = 0.04$  from Population Census 1991 file, where  $n = 248.000$  individual aprox. for  $N = 6.053.568$
- (b) Sample preparation; depuration and recoding of some variables according to the results of the previous operation.
- (c) Determination of identifying variables considered as possible “key-variables”: place of residence, place of birth, age, sex, marital state, occupation, academic titulation, activity.
- (d) Rda file (metadata file) preparation.
- (e) To fix “place of residence” and “age” as variables that will take part in all the levels of the analysis of keyword combinations.
- (f) Run  $\mu$ -Argus.