Working Paper No. 17 (Summary)

ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

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

Topic (v): Risk assessment

ASSESSING INDIVIDUAL RISK OF DISCLOSURE: AN EXPERIMENT

Invited paper

Submitted by the National Institute of Statistics (ISTAT), Italy [1]

[1] Prepared by Loredana Di Consiglio, Luisa Franconi (franconi@istat.it) and Giovanni Seri).

Assessing individual risk of disclosure: an experiment

Loredana Di Consiglio, Luisa Franconi and Giovanni Seri
Istat, Servizio della metodologia di base per la produzione statistica,
Via A. Depretis 74/B, 00184 Roma, Italy

National Statistical Institutes routinely release Microdata File for Research (MFR) from their most important social surveys. Now, Commission Regulation no. 831/2002 on "access to confidential data for scientific purposes" introduces the concept of European files of *anonymised microdata.*
An assessment of the risk of disclosure is a crucial step that should always be taken before releasing any microdata. In this paper we adopt the re-identification disclosure definition (Duncan and Lambert, 1986). This occurs when an individual unit is identified (i.e. a one-to-one relationship between a microdata unit in the released file and a target individual in the population is established with some degree of confidence) and then, as a consequence, the user is able to deduce the value of sensitive variables for such individual.

The assessment of the risk of disclosure always implies two steps: first the definition (and evaluation) of a measure to quantify the probability that a user has to disclose individual information and, second, whenever the level of the risk is not acceptable the development of protection techniques that, lowering the information content of the data, decrease the level of the risk.

The methodology we use was initially proposed in a paper by Benedetti, Franconi and Piersimoni (1998) and currently is under implementation into the software μ-Argus. This software package for producing safe microdata files is one of the product of the 'Computational Aspect of Statistical Confidentiality' project founded by the European Union and currently under testing for its final release at the end of 2003. μ-Argus allows to estimate the individual risk for each unit in the file to be released. Whenever is needed μ-Argus performs global recoding or local suppression of all those units presenting a risk higher than a predefined threshold in order to produce a safe microdata file.

The methodology estimates the risk of each single record in the file to be released quantifying the probability of identifying a unit i.e. correctly linking a record to an individual in the population. The identification of an individual is carried out through the *key variables* i.e. variables containing publicly available information. For individuals such key variables are mostly categorical. The question of interest then is mainly the estimation of the frequencies of the combinations of such key variables in the population given the information held in the sample. In the method we used the estimation of these frequencies is made by mean of sampling weights indicating the number of individuals in the population that are represented by each record (Deville and Särndall, 1992).

In this paper we describe the setting and present the results of a large scale assessing exercise carried out to test the accuracy of the estimates of the individual risk of disclosure and the validity of its definition. The experiment reproduces the real instance of the Labour Force Survey in Italy. Samples from different administrative regions have been taken using as population the 1991 Italian Population census. To assess the accuracy of the estimation procedure estimates of the disclosure risk using sampling weights have been compared to the risk using known population parameters. To assess the validity of the individual risk definition, i.e. the extent to which it is measuring what was intended, we compare the estimates of the risk with the probability of an individual being identified given the information contained in the population.

References

Benedetti, R., Franconi, L. and Piersimoni, F. (1998), Per-record risk of disclosure in dependent data, *Proceedings of the Conference on Statistical Data Protection*, Lisbon.

Duncan G.T. and Lambert D., 1986, Disclosure-Limited Data Dissemination (with discussion), *Journal of the American Statistical Association,* 81**,** 393, 10-28.

Deville, J. C. and Särndal, C. E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, vol. 87, 376-382.