STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

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

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

Working Paper No. 4 (Summary)
English only

Topic (i): new applications of disclosure control methods

LEVEL OF SAFETY IN MICRODATA: COMPARISONS BETWEEN DIFFERENT DEFINITIONS OF DISCLOSURE RISK AND ESTIMATION MODELS

Submitted by the National Statistical Institute, Italy[1]

Invited paper

Summary

1. The National Statistical Institute is able to release microdata sets only on the condition that the privacy of respondents is secure and that the event of breach of confidentiality is extremely unlikely. Different institutions adopt different definitions of disclosure, of disclosure risk, and use different models to estimate such risk and protect microdata set. In all cases the aim is the same: to achieve the release of what is considered a *safe* microdata set. Although a common aim exists, the possible different choices adopted in different institutes have consequences on the definition of the *level of safety* in the released data set. In this paper we compare different definitions of disclosure risk and different models for the estimation of such risk from the point of view of the level of safety achieved.

2. Although the concept of safety is shared by all released microdata sets, different approaches have to be considered for different types of data. Large populations, frequently presenting an inherent dependent structure and characterized by key variables of categorical nature, such as the case of social data, present completely different problems compared to small populations with skewed distribution and key variables mainly continuous, such as the case of business data. Whereas, in social data, categorical key variables allow us to tackle the problem of disclosure limitation via the concept of unique case (i.e. an individual that presents a unique combination of values of the key variables in the sample/population), in business data continuous key variables make the same concept inappropriate (practically all the units considered would be unique cases). Therefore, whereas for the former type of data disclosure limitation is performed by mean of re-categorization of key variables or local suppression of particularly rare categories of such variables, in the latter case perturbation methods are normally adopted. Such differences obviously are reflected on the definition of safety of the microdata set. In this paper both types of problems will be addressed and discussed.

3. As far as social data are concerned, the common base of all the models considered for comparison will be the adopted definition of disclosure: all the methodologies stem from *re-*

[1] Prepared by Luisa Franconi.

identification disclosure (it is possible to establish, with some degree of confidence, a one-to-one relationship between a microdata record and a target individual and, as a consequence, the value of a sensitive variable for such individual is deduced). The reason for such a choice is the fact that this definition is currently used by most national statistical institutes and researchers in the field: see for example de Waal and Willenborg (1996), Fienber and Makov (1996), Skinner (1996), Bethlehem *et al*. (1990), Biggeri and Zannella (1991), Franconi and Benedetti (1998). The definition of level of safety implied by the different definitions of disclosure risk and different models for its estimation is analysed and compared. In particular the results obtained by *aggregated* models such as the one currently adopted at Istat, Crescenzi (1993), and the one proposed by Skinner and Holmes (1992) will be compared to the results obtained by the software μ-Argus, Willenborg and Hundepool (1998) and the *individual* risk model proposed by Benedetti and Franconi (1998). Such a comparison will be performed on data from Italian survey data and both in the case of independence among individuals and of dependence (hierarchical structure) among them (Benedetti *et al*., 1998). The case of panel or longitudinal data will be also addressed.

4. The problem to be tackled in business data is somehow different. Most National Statistical Institutes apply perturbation methods to release safe business microdata. In this work we consider a particular family of perturbation methods: microaggregation techniques (Defays and Anwar, 1995). The paper will briefly report on the different philosophy behind different microaggregation methods and on the level of safety reached by mean of experiments carried out on Italian business data.

References

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

Benedetti, R. and Franconi, L. (1998). An estimation method for individual risk of disclosure based on sampling design, *submitted for publication to Survey Methodology*.

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990), Disclosure Control of Microdata, *Journal of the American Statistical Association*, 85, 38-45.

Biggeri, L. and Zannella, F. (1991), Release of microdata and statistical disclosure control in the new national system of Italy: main problems, some technical solutions, experiments, *Proceedings of the 48th ISI session*, Cairo.

Crescenzi, F. (1993), On estimating population uniques. Methodological proposals and applications on Italian Census data. *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin, 247-260.

de Waal, T. and Willenborg, L. (1996), A view on statistical disclosure for microdata. *Survey Methodology*, 22, 1, 95-103.

Defays, D. and Anwar, N. (1995), Microaggregation: a generic method. *Proceedings of the Second International Seminar on Statistical Confidentiality*, Luxemburg, 69-78.

Fienberg, S.E., Makov, U.E. (1996), Confidentiality, uniqueness and disclosure avoidance in categorical data, *Third International Seminar on Statistical Confidentiality*, Bled, 165-174.

Franconi, L. and Benedetti, R. (1998). Some aspects of disclosure avoidance in complex microdata files, *Research in Official Statistics*,1,0, 59-70.

Skinner, C. J. (1996). Estimating the re-identification risk per record in microdata, *Third International Seminar on Statistical Confidentiality*, Bled, 123-129.

Skinner, C. J and Holmes D.J. (1992). Modelling population uniqueness, *International Seminar on Statistical Confidentiality*, Dublin.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Springer-Verlag: New-York.

Willenborg, L. and Hundepool, A. (1998). ARGUS for Statistical Disclosure Control, *Statistical Data Protection '98*, Lisbon.