

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, Italy¹

Invited paper

I. Introduction

1. National Statistical Institutes are able to release microdata sets only on the condition that the privacy of respondents is safe 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: achieve the release of what is considered a *safe* microdata set. But, the set of predefined conditions that have to be satisfied to define a safe file changes by changing the methodological framework. So, although a common aim exists the possible different choices adopted 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 set, different approaches have to be considered for different type of data. Large populations, frequently presenting an inherent dependent structure and characterised 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 public domain 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 set of public domain variables in the sample/population), in business data continuous public domain 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 often quantified by a measure called disclosure risk and disclosure control is performed by means of re-categorisation of key variables or local suppression of particularly rare categories of such variables, in the latter case, the concept of risk is not used and 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

¹ Prepared by Luisa Franconi.

addressed and discussed. As well known, most of the methodologies concerning statistical disclosure limitation rely on the definition of threshold values that vary according to the country; in this work we would like to compare the general framework of the different methodologies regardless of the more or less strict parameters adopted.

3. As far as social data are concerned, the common base of all risk estimation models considered for comparison will be the adopted definition of disclosure: in fact, most of the proposed methodologies stem from *re-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), Fienberg and Makov (1996), Skinner (1996), Bethlehem *et al.* (1990), Biggeri and Zannella (1991). The definition of level of safety implied by the different models for risk estimation is analysed and compared in Section 2. In particular *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 completely sample based approach, Willenborg and Hundepool (1998) and the *individual* risk model proposed by Benedetti and Franconi (1998a). The case of panel or longitudinal data will be also addressed (Benedetti *et al.*, 1998).

4. The problem to be tackled in business data is somewhat different. Very rarely, national statistical institutes release business microdata because it is perceived that the possibility of identification would be too high. In this work we consider various families of disclosure limitation methods: in Section III the paper will briefly report on the different philosophy behind different methods and on the type of level of safety reached. Conclusions and suggestions for further work are presented in Section IV.

II. Social Microdata

5. The process of disclosing individual information from a microdata file under the hypothesis of independence among units and using the definition of re-identification disclosure implies a two-step process: in the first step the user identifies the individual by mean of the combination of scores on a set of key variables - i.e. the variables present both in the microdata file to be released and in the public register he has access on - and, in the second step, the user, having established a link between the target record in the released file and the individual in the register, is able to disclose information. If, as in the case of microdata stemming from social surveys, the key variables are categorical then the first step of the identification process is driven by the search for individuals with rare characteristics with respect to the key variables that can be obviously singled out and identified. The extreme case being the unique case in the file to be released i.e. the unit that present a unique combination of scores on the key variables. The categorical nature of the key variables in social microdata has allowed the construction of a framework for the disclosure problem that is completely based on the idea of proxy between the frequency of the combination of values of the scores of the key variables and the concept of identification. In most of the techniques examined a measurement of the possibility of disclosing information, the so-called risk of disclosure, relies completely on the presence of records presenting unique combinations of scores on the key variables. As a consequence most of the proposed methodologies consider the released microdata file *safe* if a certain function of the number of unique cases found in the sample is lower than a given threshold.

6. The approach adopted at CBS (see de Waal and Willenborg (1994) and de Waal and Willenborg (1995)), is completely sample based and is implemented in the software μ -Argus (Argus, 1998). It does not use explicitly a disclosure risk and defines directly the concept of safe record on the observed frequency of selected scores on the key variables; such an evaluation is based on the sole

information contained in the sample. A threshold is given and a unit is considered safe if the number of units in the sample that presents the same combination is greater than such a threshold. Its value is the “most likely” set of records which have to be modified but such value is matter of personal judgement. The disclosure limitation process requires the use of global recoding and local suppression in all unsafe combinations and does not allow the presence in the released microdata file of any frequency greater than the pre-defined threshold.

7. The adoption of the solution of this sample based approach, which has been very strict, allows us to be sure, given a low value of the threshold, on the level of safety reached in the released microdata. However we know very well that not all the sample uniques are also population uniques and therefore it is possible that such method overprotect the microdata file. The measure used can be improved if we are able to distinguish sample from population uniques. This is what has been tried when a step of inference from the sample to the population has been proposed and a series of probabilistic models has been applied. For such an approach, the aggregated “model based” approach, the set of conditions to be met in order to reach a safe microdata file depends on the expected number of uniques in the population as evaluated through a probabilistic model - Poisson log-normal as suggested by Skinner and Holmes (1992), modified Negative binomial-Gamma as proposed by Crescenzi (1993) and so on. In fact, under the hypothesis of independence among individuals the approach let the number of uniques in the file be a random variable with a Poisson distribution with parameters depending on various factors affecting the possibility of a “link” (quality of the variables, time lag etc.) and operates an inference from the sample to the population via a probabilistic model. Under such framework if the expected number of uniques in the population is below a certain threshold then the file is safe and can be released. However also in this case the solution is not ideal because we express the level of safety by mean of the *expected number* of uniques and not by mean of an individual evaluation based on the real situation. Moreover, in such approach, the attention is restricted only to the unique cases. A more general situation could be considered where a risk is evaluated also for records that appears twice in the released file, or if the situation requires it also three or more times, to test the risk of such units. A solution to all these problems is the adoption of an individual risk of disclosure that, on the base of the actual values observed on the key variables of each individual in the released file, is able to attach a disclosure risk. In the last year few proposals have been made. Fienberg and Makov (1996) and Skinner and Holmes (1998) propose, with different motivations, a log linear model for the estimation of the individual risk. Benedetti and Franconi (1998a) propose a methodology for individual risk estimation based on the usual instrument that National Statistical Institutes evaluate to allow for inference from the sample to the population: the sampling weight. Such quantity, attached to each unit in the microdata file, and indicating the number of units that such record represents in the population is a natural tool to evaluate the rareness of an individual in the population whatever the observed frequency we obtain in the sample (unique, two cases and so on). In particular, the individual risk in such a framework is defined as the probability of identification of a unit in the sample where for identification we indicate the fact that the unit in the released file and the unit in the register of the user belongs to the same individual in the population. The hypotheses made are that the intruder will always try to match a record and a unit in the register (i.e. if the record is not unique he will try a probabilistic link). Obviously, other factors influencing the identification are taken into account in the framework; for further details see Benedetti and Franconi (1998a) or Benedetti and Franconi (1998c).

8. The second important characteristic of social microdata is the inherent hierarchical structure of the data that allows us to recognise *groups* inside the population (the most typical case being the household) where for group we mean a set of units linked by some kind of relationship. This strong dependence among units can not be forgotten when defining a risk for an individual because of the much higher degree of information content that enables to link units together. Obviously, to make use of such dependence structure to disclose individual information the user needs to have access to a type of register with the same structure as the released file (for example hierarchical or longitudinal). In fact most of the microdata sets released by national statistical institutes retain their hierarchical structure

even though till now only approximate solutions of the above-mentioned global measure of risk have been implemented. The methodology proposed by Benedetti and Franconi (1998b) to deal with dependencies in the data - not only among units but also among variables - is based on the idea of several components of the risk that can be added to take into account more and more complex situations (independent data, hierarchical data, independent panel data and hierarchical panel data). The structure of the risk in case of dependence among units stems from a rationale where the intruder starts comparing the scores on the key variables on the single target record of interest and then, if this tentative is not successful, he makes use of the information relative to the units belonging to the same group (in general a household) and tries to establish a link for each single unit j in the group. If even this strategy is not successful he considers all the combinations of couples of units in the group of interest that can be linked to couples of individuals in the register, and so on with triplets etc. Under the hypothesis of independence among temptatives, and if we consider only single individuals, the risk in the case of dependence among units, r_i^{hier} , can be written in an additive way as:

$$r_i^{\text{hier}} = r_i^{\text{ind}} + r_i^{o(1)} + r_{\delta_i}^{\text{group}}.$$

where r_i^{ind} is the risk for individual i in the case of independence as evaluated through any individual risk model, $r_i^{o(1)}$ is the probability of being identified for each record j in the group of the target individual i . This, being an application of the general probability formula, is equal to:

$$r_i^{o(1)} = \sum_{j=i+1}^{(i+s_i-1) \bmod s_i} \left(\prod_{h=i}^{j-1} (1 - r_h^{\text{ind}}) \right) r_j^{\text{ind}}.$$

The term $r_{\delta_i}^{\text{group}}$ refers to the group, the household, as unit of interest (notice that households are independent from each other). The key variable for the unit “group” is the size of the household, therefore the subpopulation k refers to the subpopulation of all the groups sharing the same size. In order to make use of the dependence structure, the intruder will have to rely on the quality of the variable that defines the group in his register, $d(\delta_i)$, and the quality of the same variable in the released file, $d^*(\delta_i)$. In the example of the hierarchical file this variable will be the relationship with the head of the household; the risk (Franconi and Benedetti 1998b) is then given by:

$$r_i^{\text{hier}} = r_i^{\text{ind}} \rho_i + \sum_{j=i+1}^{(i+s_i-1) \bmod s_i} \left(\prod_{h=i}^{j-1} (1 - r_h^{\text{ind}}) d_{k(h)}^* d_{k(h)} \right) r_j^{\text{ind}} \rho_j + r_{\delta_i}^{\text{group}}$$

Such an approach to deal with a dependence structure among units has been further developed to take into account also the dependence among variables (Benedetti and Franconi 1998c), the idea being the addition to the above structure for the risk of a term that takes into account the risk due to a panel or longitudinal structure.

9. The strategy behind this approach is the mimic of the reasoning of a user who attempts to breach confidentiality making use of the dependence structure in the data and the power of technological tools to match records with equal characteristics. The dependence structure built in the individual risk evaluation procedure allows us to reach a higher level of safety for the released microdata file than the one that consider the sole case of independence.

III. Business Microdata

10. Disclosure limitation in business microdata is somehow a different problem with respect to the one concerning social microdata. Sparse population, extremely skewed distributions, particular sampling design added to the different and far more dangerous motivations for trying to breach

confidentiality and far more accurate information held in public registers makes it a delicate and, at the same time, a difficult problem as already reported by Cox (1995).

11. The common feature of all the approaches presented to limit disclosure is the “invasive” nature i.e. the modification of the original values to create artificial enterprises that try to maintain, as far as possible, the characteristics of the original ones. These can be generated by masking procedures, Cox (1994) and Duncan and Pearson (1991), where the microdata file to be released is represented as a matrix X where each row is a respondent and each column is a variable and where the masking procedures are transformations of X of the form $AXB+C$, $A, B \neq 0$. Such a powerful matrix representation of the problem allows to view several different techniques for disclosure limitation under a unique framework. The list of methods that can be obtained by matrix transformation comprehends local suppression, global recoding, top and bottom coding, aggregation in various forms including microaggregation procedures, Defays and Nanopoulos (1992), data swapping, see also Fienberg *et al.* (1996) for relationship with log linear models, rounding and perturbation. Another way to create artificial enterprises is by simulation from relevant distributions, McGuckin and Nguyen (1988) and Fienberg (1994) to build “imaginary” units that should retain characteristics of the original data but whose actual attributes are generated via a stochastic or a non stochastic modelling process.

12. The impact of such different techniques on the level of safety can be distinguished according to the obtained kind of output, i.e. the released microdata file. For all the techniques that involve aggregation of units in various form such, for example, single axis microaggregation where the idea is to release pseudo-units obtained by averaging values of k original enterprises with similar characteristics, the disclosure limitation problem for microdata can be assimilated to the disclosure problem in case of aggregated data i.e. tabular data. In fact microaggregated data by single axis methods can be viewed as the release of a very detailed table where each cell has frequency k . But national statistical institutes publish masses of tabular data according to some standard criteria such as threshold values (in this case the size k of the microaggregates) and, often, a dominance rule that can be applied to each microaggregates without much effort. If these criteria are valid for tabular data then they will be valid as well for microaggregated data and an initial form of pseudo microdata (enterprises) release can be implemented without further problems.

13. For other matrix masking procedures that change attributes of the original microdata maintaining an individual profile of the units special care should be taken. In fact, there are controversies on the level of perturbation needed to reach a satisfactory level of security. Methods that are seen as safe such as microaggregation on multiple axis because of their aggregation nature are in reality dangerous in several cases. The problem stems from the impossibility of knowing *a priori* the level of security i.e. of perturbation reached in each file. An extensive research carried out during the SDC project, see Pagliuca and Seri (1998), has revealed that in the file of the system of large size enterprises the percentage of units that after individual ranking did not change their values or changed it only for a small fraction (only 1% with respect to the original values) reached in some cases 90%. Therefore, a check on the level of perturbation reached in some cases is necessary.

14. As far as synthetic data are concerned, the methodological development in this field is still at an early stage and further experiences and studies are needed to evaluate the possibility of implementation of such methods in a National Statistical Institute.

IV. Conclusions and suggestions for further work

15. As for microdata from social surveys, although methodologies are available to estimate the individual risk of disclosure (also in the more common case of dependence among units and variables) and the evaluation of a sound level of safety for each unit in the microdata file to be released is feasible,

still there are areas that need further development. A primary need is the development of a software able to implement such methodologies and make them available to each national statistical institute. Given the flexibility of the individual risk methodologies such software could be shared among national statistical institutes each one adopting its own threshold according to the final use of the file (public use file or microdata for research). The potential effect of common methodologies and a common software would be to harmonise those that now are different practises, while enhancing each national statistical institute options to apply these methods. The SDC project (Esprit no 20462) has given the basis for the construction of a common software and the definition of a methodological framework for some aspect of statistical disclosure limitation, however software development and methodological research still need to be done in these areas to reach a comprehensive and flexible tool. In particular, work is needed to reach a definition of algorithms to protect single units when dependencies are present. The problem of taking into account the dependent structure of the data has already arisen in several areas of statistics, for example data editing for hierarchical and panel data, and common efforts in this direction will bring to a solution also in the area of statistical disclosure limitation. The solution seems necessarily a mixed strategy of global recoding and local suppression, as already implemented by μ -Argus. However, new and more specialised methods, such as conditional recoding, and algorithms are necessary to deal with dependencies and identify patterns that minimise information loss keeping the level of security high. Moreover, ways to improve precision of the estimation process for the individual risk should be addressed.

16. As far as the field of business statistics is concerned, research, testing of new methodologies and software development seem all essential to reach a solution that satisfies both the maintenance of the characteristics of the original population and the requirement of safety. Also in this field a general software that implements several different and specialised techniques would be welcome. From the point of view of the methodological research to be carried out we have seen how using some techniques such as microaggregation by single axis methods moved the attention from the difficult area of microdata to the usual and well known field of tabular data. However, in such a change the information loss is quite high and alternative and more advanced methods that reduce the price we have to pay to released such type of data should be investigated. Moreover careful considerations on the various methodological and computational aspects inherent to the framework of the “imaginary” units as well as the implications that such techniques have on the level of safety could be further explored.

17. Finally, we have to remember that also in the field of business data the problem of dependencies is present and it should be tackled given the rising importance that longitudinal and panel data have acquired lately in several areas of economics.

References

- μ -ARGUS (1998). User's manual. *Deliverable TT-2/D, Statistical Disclosure Control Project*, Esprit no.20462.
- 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. (1998a). An estimation method for individual risk of disclosure based on sampling design, *submitted for publication to Survey Methodology*.
- Benedetti, R. and Franconi, L. (1998b). Statistical and technological solutions for controlled data dissemination. *Proceedings of the Conference New techniques and Technologies for Statistics*, 1, 225-232.
- Benedetti, R. and Franconi, L. (1998c). Applied issues on disclosure avoidance complex microdata files. *Deliverable MI2-D2, Statistical Disclosure Control Project*, Esprit no.20462.
- 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.
- Cox, L.H. (1994). Matrix masking methods for disclosure limitation in microdata. *Survey Methodology*, 165-169.
- Cox, L.H. (1995). Protecting confidentiality in business surveys. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds), Wiley; New-York.
- 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.C.R.J. (1994), Minimizing the number of local suppression in a microdata set. *Report Statistics Netherlands, Voorburg*.
- de Waal, T. and Willenborg, L.C.R.J. (1995), Optimal global recoding and local suppression. *Report Statistics Netherlands, Voorburg*.
- de Waal, T. and Willenborg, L.C.R.J. (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.
- Defays, D. and Nanopoulos, P. (1992). Panels of enterprises and confidentiality: the small aggregates method. *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*.
- Duncan, G.T. and Pearson, R.W. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science*, 6, 219-239.
- Fienberg, S.E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. *Technical report no.611*, Department of Statistics, Carnegie Mellon University, Pittsburg, Pennsylvania.
- Fienberg, S.E. and Makov, U.E. (1996), Confidentiality, uniqueness and disclosure avoidance in categorical data, *Third International Seminar on Statistical Confidentiality*, Bled, 165-174.
- Fienberg, S.E., Russel, J.S. and Makov, U. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models.
- Franconi, L. and Benedetti, R. (1998). Some aspects of disclosure avoidance in complex microdata files, *Research in Official Statistics*, 1,0, 59-70.
- McGuckin, R.H. and Nguyen, S.V. (1988). Use of "surrogate file" to conduct economic studies with longitudinal microdata. *Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census*, 20, 193 - 211.
- Pagliuca, D. and Seri, G. (1998). Some results of individual ranking method on the system of enterprise accounts annual survey. *Deliverable MI3-D2, Statistical Disclosure Control Project*, Esprit no.20462.
- 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.
- Skinner, C. J and Holmes D.J. (1998). Estimating the re-identification risk per record in microdata, *to appear in Journal of Official Statistics*.
- Willenborg, L.C.R.J. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Springer-Verlag: New-York.
- Willenborg, L.C.R.J. and Hundepool, A. (1998). ARGUS for Statistical Disclosure Control, *Statistical Data Protection '98*, Lisbon.