WP. 22
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 UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (iii) Confidentiality aspects of statistical information taking into account register-based data

CONFIDENTIALITY ASPECTS OF HOUSEHOLD PANEL SURVEYS: CASE STUDY OF THE ITALIAN SAMPLE FROM EU-SILC

Supporting Paper

Submitted by the Italian National Statistical Institute (ISTAT), Italy [1]

[1] Prepared by Lucia Coppola and Giovanni Seri, {lcoppola, seri}@istat.it

Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC

Lucia Coppola* and Giovanni Seri*
* Istat – Italian National Statistical Institute, 00184 Rome, Italy, {lcoppola, seri}@istat.it

Abstract: In this paper we discuss some of the disclosive features to deal with when releasing data collected through a household panel survey. The discussion and the empirical analyses are based on provisional data from the EU-SILC Italian survey. In particular, two structural characteristics are considered: (i) the hierarchical data structure, providing information simultaneously about household and individual characteristics; (ii) the longitudinal data structure, providing information about household and individual specific patterns of change during the period of observation. The disclosive power of these information depends on the nature of the information available to the intruder. We firstly point out a few intruder's attack scenario we consider as reasonable in the Italian specific context. Secondly, we propose an anonymisation strategy to protect micro data against intruders' attack under these scenarios. Such a strategy is based on the estimates of re-identification risk at individual and household level, and on a reduction of household and individual information. This is achieved through global recoding and/or local suppression.

1. Introduction

The Italian National Statistical Institute (Istat) is releasing so called Microdata Files for Research (MFR) since more then ten years. Usually, MFRs consist of individual records representing a sample of the population (MFRs are released only for social surveys). Statistical confidentiality is preserved reducing the information contents of the files minimising the risk of identification of statistical units. Users requiring MFR are asked to sign an agreement with Istat.

Statistical Disclosure Control (SDC) is a relatively recent field of research involving mainly official statisticians. Methodology evolution in these last years has changed the way of producing MFRs. Furthermore, modification of related legislation influenced procedures to access micro data of many National Statistical Institutes.

In order to release a MFR, a reasonable evaluation of the risk of disclosure is needed. Istat recently adopted an approach consisting in estimating for each record the 'risk of re-identification' (Franconi and Polettini, 2004; Benedetti *et al.* 2003) at individual level. A threshold for the re-identification risk is then fixed as a reasonable low level of risk. On the base of this threshold, records are classified as "at risk" or "safe". Consequently, protection methods are applied in order to reduce the risk associated to each record under the given threshold. We only consider protection methods based on data reduction (Domingo-Ferrer and Torra, 2001; Willenborg and de Waal, 2001), particularly "global recoding" and "local suppression" (Willenborg and de Waal, 2001). Data reduction implies loss of information with respect to the original contents of the file. Therefore, main purpose is combining protection methods in order to minimise loss of information, given that the fixed level of re-identification risk is respected.

As a case study, we consider provisional data from the Italian EU-SILC survey. EU-SILC is a panel survey, carried out in different EU member states, and providing every year cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions. Information is collected about both households and household members at the same time. Moreover, households and household members are followed and surveyed yearly, during four years.

In order to measure the re-identification risk for EU-SILC data we need to keep into account that statistical units in the file are not independent. Dependence between units is due to the presence of an household identification number. It is then possible to associate all the individuals belonging to the same household or, equivalently, define a hierarchical structure between households and individuals. Moreover, dependencies between identification variables exist because longitudinal data structure provides information about households and individuals in different periods of observation. Specific patterns of change characterise household and individual during the whole period of observation of the survey. Dependences between units and between variables increase difficulties in computing analytical measure of the re-identification risk (Benedetti and Franconi, 1998; Abowd and Woodcock, 2000). Anyway, the re-identification risk has to be measured under an appropriate disclosure scenario, namely the quality and quantity of the information assumed to be available to the intruder and his/her strategy to re-identify a statistical unit.

In this paper we discuss the disclosure figures belonging to the EU-SILC survey and propose a strategy to produce a MFR from the Italian sample. In particular, in Section 2, we briefly describe the EU-SILC survey, and address the main disclosive figures it implies: household data structure (Section 2.1) and longitudinal data structure (Section 2.2). In Section 3, we describe the intruder's attack scenarios we consider in the Italian specific context. In Section 4, we introduce the SDC method we apply, in order to estimate household and individual re-identification risk. In Section 5, we show the empirical results, and propose a strategy for protecting the EU-SILC micro data. Finally, in Section 6, conclusions are discussed.

2. The EU Survey on Income and Living Conditions (EU-SILC)

The EU-SILC data are organized into four datasets: (i) the *Household Register* file, containing information about every sampled household (including not interviewed households); (ii) the *Household Data* file, containing information about each interviewed household; (iii) *Personal Register* file, containing information about every household member (including temporarily absent members); (iv) *Personal Data* file, containing information about each interviewed household member. These four files can be linked together, through country, household and individual identification codes. It is worth noting that household and individual files can be linked also longitudinally, so that the amount of information at household and individual level is increased yearly.

The survey represents an important source of information about several household and individual living conditions. Moreover, the household and longitudinal data structure allows for tracing specific patterns of household changes, and individual life trajectories, that necessarily become extremely rare or even unique in the population.

2.1 Household data structure

When data are organized in a hierarchical structure, that is households and household members are explicitly linked through identification codes, the following issues have to be considered: (i) household characteristics might be used for identifying an individual; (ii) household members' characteristics might be used for identifying a household; (iii) some household members characteristics might be used for identifying other household members.

Firstly, household characteristics might have a strongly identifying power. For instance, a household composed by 15 individuals might be very rare (especially when provided with other variables, as place of residence). Thus, depending on the distribution of households by size in the population, such a variable alone might be enough to identify a specific household and all its members. However, protection criteria as local suppression, global recoding, or perturbation cannot be applied to household size, because such an information is implied by the data structure.

Secondly, individual characteristics might be so rare to identify a household. As an example consider three or more twins belonging to the same household (same age and same parental relationship). In this case, joint information of the household members might allow for household identification.

Finally, a household member showing a frequent combination of individual characteristics might be identified, for living in the same household with an individual showing a rare combination of characteristics. As an example we might consider a household composed by two individuals: a man, 20 years old and never married, and a man, 18 years old, and widow. In the case of the second man, the combination of individual characteristics might be rare or unique in the population. In the example, the combination of individual and household characteristics might not be enough to identify the first individual, but they are probably enough to identify the second individual and consequently the first one.

2.2 Longitudinal data structure

As we said, EU-SILC provides cross-sectional and longitudinal data at the same time. In the Italian case, both data sets belong to the same sample of households. These data sets provide almost the same set of variables, and might be easily linked. Thus, the same protection criteria have to be applied simultaneously to both data set.

Otherwise, an intruder might use cross-sectional data to increase the detail of information provided by the longitudinal data, and vice versa.

Some anonymisation criteria often applied to cross-sectional data, might not be used when dealing with longitudinal data. An example is provided by the aggregation of age in classes. If such a variable is provided yearly for the surveyed individuals, as soon as an individual moves form one class to the next one, the exact age of the individual might be easily deduced. Top coding, instead, might be applied to longitudinal data, allowing to protect elder people al least.

It is also worth considering that in the case of longitudinal data, if a variable is treated through local suppression (or other protection methods), in a specific record and at a given year of survey, the same treatment has to be coherently applied in the following years. Thus, if local suppression is chosen, the suppression of a variable for some records during the whole period of observation should be applied. Similarly, when perturbing the value of a variable, we have to be aware of the consequences on the analyses of these variables over the period of observation. Given that most of the analyses carried on longitudinal data deal with “changes”, the quality of longitudinal research might not be guaranteed.

Thus, when dealing with longitudinal data, using local suppression or perturbation methods might be not convenient. The recoding of some variables might be more appropriate and more easily handle, although such a solution is not always feasible. The analysis of the disclosure risk at individual and household levels would be useful for identifying the categories more “at risk”, as for instance individuals older than a certain age, or household bigger than a certain size. These categories could be protected, for instance through top-coding.

3. The intruder’s scenarios: available information and attack strategy

The definition of a scenario is a first step towards the development of a strategy for producing a “safe” MFR. Indeed, a scenario synthetically describes (i) which is the information potentially available to the intruder, and (ii) how the intruder would use such information to identify an individual: i.e. the intruder’s attack means and strategy. We observe that the intruder’s chance of identifying an individual depends on the External Archive main characteristics, such as completeness, accuracy, data classification, etc.

In this paper, we refer to the information available to the intruder as an *External Archive*, where information is provided at individual level, jointly with directly identifying data, such as name, surname, etc. We refer to the information we aim to protect as the *EU-SILC micro data*, containing information belonging to individuals, as well as to households. Finally, we refer to the anonymised data set as the *EU-SILC user data base*.

3.1 The Nosy Neighbour scenario

In the Nosy neighbour scenario (Nosy scenario), we assume that the intruder has many information about a single individual (or a few individuals), and the information is based on personal knowledge of individuals. It might be the case of a *neighbour*, or a colleague, or anybody else the intruder knows. We are not able to know how many and which individual or household characteristics the intruder knows, for depending on his/her personal knowledge. Nevertheless, we assume that the intruder does not know that the individual or household is in the data set we want to protect. In this case, the intruder's attack would be the spontaneous recognition of the individual in the EU-SILC micro data. That is, the intruder might recognize an individual in the EU-SILC micro data, for showing the same characteristics as the person he/she knows.

For protecting the EU-SILC micro data against this kind of attack, we propose to reduce the information, for instance dropping or recoding variables with a high identifying power. Consequently, the intruder can not be confident that a given combination of information is unique or rare in the population. Moreover, the further protections suggested in the following sections are needed against spontaneous recognition.

3.2 The Individual Archive scenario

The individual archive scenario is based on the assumption that the External archive available to the intruder is an individual archive. That is, for each individual directly identifying variables, and some other variables are available. Some of these further variables are assumed to be available also in the data set we want to protect. The intruder's strategy would be matching the information in the individual archive with that in the EU-SILC user data base. A "match" would be considered as correct only if all the matching variables assume the same value in both data sets. We refer to these matching variables as *key* or *identifying variables*. Broadly speaking, given a match between the two archives the probability that the match is "correct" depends on how many individuals show the same set of characteristics in the population. Therefore, the selection of key variables is crucial.

The Individual Archive we consider to be worth of attention is represented by the *Electoral Registers*. These are based on the Population register, and provide information about individuals having electoral right. Electoral Registers are public, but available only at municipality level.

Information provided at the same time by the electoral registers, and the EU-SILC micro data are: place of birth (at municipality level), place of residence, date of birth, sex, marital status, occupation and educational level. We suggest to drop information about the place of birth and provide information only about the current place of residence to be recoded at least at regional level. Date of birth is reliable, and has a strongly identifying power. Thus we recommend to reduce the information, through recoding it in age (see Section 5). Marital status is usually considered as reliable but it is not public any longer. We will discuss alternative scenarios that respectively include it or exclude it as a key variable. We do not consider occupation and education reliable in this scenario, but recoding of these variables is suggested under the Nosy scenario. Summarizing, under the Individual Archive scenario we assume that the intruder would use place of residence, sex, age and possibly marital status as key variables.

3.3 The Household Archive scenario

The household archive scenario is based on the assumption that the External archive used by the intruder contains, for each household and for each household member, directly identifying variables, as well as some other variables.

As in the previous scenario, the intruder's strategy of attack would be matching the individual information provided by the household archive and the EU-SILC micro data. In this case, the intruder would use as matching variables not only individual characteristics, but also household characteristics. In particular, we assume that the intruder would use the household characteristics at individual level. We assume the external archive to have the same structure of the EU-SILC user data base. Particularly, each record in the file represents a single individual and a household identifier is associated to each record, allowing for household recognition.

The external archive we consider is the *Population Register*. In every municipality, a population register collects information about households, as well as each household member. These archives are not public, but a single individual might ask for information about one or a few households. Thus, under this scenario, the intruder's chance of access to the external archive is lower than in the previous case.

Information provided by the population register and present in the EU-SILC micro data, are the same individual variables provided by the electoral register, plus (i) the household size and (ii) the parental relationship. We assume that the intruder might consider as a reliable key variable the parental relationship recoded in six categories (grandparents, parents, partners, daughters or sons, granddaughters or grandsons, other relationships), coherently with the information provided by the survey, even if the classification used by the population register may be different. Summarizing, under the household archive scenario we assume that the intruder would use the same individual variables considered in the individual archive scenario, as well as parental relationship and household size.

3.3 A Longitudinal scenario

Longitudinal data structures provide a same set of identification variables several times, in a given period of observation (say four years in EU-SILC survey). An intruder might use the specific key variables pattern of change in order to identify individuals. Rare patterns of variables change might ease individual spontaneous recognition. Nevertheless, as in the Nosy scenario, we assume that once extremely identifying variables are properly recoded and the number of individuals residing in any geographical domain is high enough, the intruder should not be able to know whether the pattern of change is unique (or rare) in the population.

We consider the case of the Individual or Household archive scenario under a longitudinal prospective. In these situations, an intruder might try to link the external archive and the micro data on the base of key variables patterns of change. We have to consider that electoral registers provide key variables that are not expected to change, apart from the case of the place of residence and the marital status. As far as the population register is concerned, we assume that the intruder is not likely to have access to it several times, and at the same reference periods of the survey. If such a scenario take place, the re-identification risks could be extremely high and data protection may imply unacceptable levels of information loss.

Further analyses on re-identification risk may be conducted when future waves of the survey will be available. Suitable disclosure scenarios may be defined at least on the bases of the individual scenario. At this stage, we suggest that a certain level of risk is worth to be run, but the user should have access to data only under strict license.

4. The Individual and household risk of re-identification

The anonymisation process of a micro data file might be developed in two distinct phases: the first one consists on evaluating the disclosure risk; the second one concerns the application of protection method to the data, where the risk of re-identification is considered too high. As previously stated, we consider a measure of the disclosure risk based on a probabilistic estimation of the individual re-identification risk (see for details Franconi and Polettini, 2004). The individual approach allows to apply protection methods only to those records that present a risk higher than a pre-fixed threshold. Protection methods taken into account are mainly "global recoding" and "local suppression". Usually a preliminary step of global recoding is used in order to reduce the number of suppressions to an acceptable level. The aim is lowering the re-identification risk under a given threshold for all the records and, at the same time, minimising the information loss.

We firstly note that for social data, identification variables are mainly categorical or can be treated as categorical (the variable "Age" for example). Given the set of identification variables, the approach is based on the relation between the frequency of a certain combination of identification variables in the sample data (f_k, for k representing the combination under investigation) and the frequency of the same combination in the population (F_k). The idea behind the method is that a statistical unit, represented by a combination identification variables' values, is "at risk" if the same combination is "rare" in the population. The true value of F_k is often unknown, and consequently we estimate it using sampling information available in the file, namely the sample weights. Recently Istat tested the effectiveness of the method simulating sample survey data from the population census (see for details Di Consiglio *et al* , 2003).

The individual approach has been implemented in Argus (software and manual available at http://neon.vb.cbs.nl/casc/; see also Polettini and Seri, 2003), allowing for two alternative risk computations: *Base Individual Risk* (BIR) and *Base Household Risk* (BHR). The former is the individual risk just described, and it is used when each record in the file represents a single statistical unit and these are independent from each other. The latter is intended to be computed when data structure is characterised by hierarchy between statistical units as in the case of household data. Particularly, each record in the file represents a single individual, and a household identifier is associated to each record allowing for household recognition. We assume that an intruder tries to link individual records to an external archive with similar characteristics using both household and individual information (see Section 3.3). Alternatively we could consider an household as a statistical unit (a single record) joining the sets of identification variables of each household components into a single one. In such a case, considering a household as a single unit, a BIR approach might be used. Similarly, when dealing with longitudinal data, key variables recorded at different point in time might be simultaneously used as matching variables. Previous experiences (Benedetti and Franconi, 1998) have shown that when the household and longitudinal data structures are explicitly considered, the re-identification risk becomes extremely high. As we argued, data protection would imply a strong information loss, affecting the research interest in the data.

BHR estimate, as implemented in Argus, is based on the individual risk assuming that, if an individual is correctly linked and identified, all household components might be identified as well. The value of BHR is the same for each household member. Given a fixed threshold α, an individual is classified at risk if the estimate of BHR is higher than α. In order to apply local suppression to the records at risk we consider that: let hhs be the household size (number of members of the household), if BIR is lower than α/hhs for all the members of the household then BHR is lower than α. Thus, we only need to check for records having BIR higher then α/hhs and apply local suppression as in the case of independent records. In other words, the higher the household size, the lower is the threshold considered. This certainly represents and advantage because higher level of safety are asked for larger households. We consider the household risk (BHR) as a reasonable approximation of the hierarchical risk that occur when an household is re-identified, and consequently all the household members are re-identified.

5. Empirical results

So far, only the first wave (2004) of the EU-SILC survey has been carried out in Italy. Consequently, we cannot empirically address the disclosive features implied by the longitudinal data structure. The following analyses are based on provisional data, organized in 61750 individual records.

A first step to reduce the disclosure risk consists in dropping or recoding identifying variables (see Table 1).

Table 1 EU-SILC variables to be recoded or dropped

		Dropped	***Recoded***
Sample Variables	Primary Strata	X	
	Psu-1 (First Stage)	X	
	Psu-2 (Second Stage)	X	
	Order Of Selection Of Psu	X	
Individual Variables	Month Of Birth	X	
	Year Of Birth		X
	Month Moved Out The Household Or Died	X	
	Month Moved In The Household	X	
	Day Of The Personal Interview	X	
	Month Of The Personal Interview	X	
	Citizenship 1		X
	Citizenship 2	X	
	Highest Isced Level Attained		X
	Parental Relationship		X
Household Variables	Day Of Household Interview	X	
	Month Of Household Interview	X	
	Number Of Rooms Available To The Household		X
	Place Of Residence		X

In a second step, we estimate the individual and household risk of re-identification to evaluate the number of suppressions needed respectively under the Individual and Household Archive scenario. Provided that key variables might be recoded according to different levels of aggregation, we propose and discuss some alternative solutions.

It is worth noting that we do not consider the different scenarios independently. Particularly, in order to avoid spontaneous recognition defined in the Nosy scenario also local suppressions applied under the Individual and Household scenario are needed. The anonymisation strategy we propose is based on the simultaneous application of the protection methods suggested under the different scenarios. Empirical analyses are produced using the software μ-Argus.

The Individual Archive scenario has been defined through the following key variables: sex, age, and place of residence. Age is recoded according to two standards: (i) age is top coded at 85 years, and the key variable is called Age85; (ii) age is top coded at 85 years and simultaneously recoded, aggregating from 0 to 2 years, form 3 to 5 years, from 6 to 10 years (on the base of the first levels of the educational system), and the corresponding key variable is called Age85_edu. Similarly, the place of residence is recoded according to two standards: (i) regions in 19 modalities, Valle D'Aosta and Piemonte aggregated, and the key variable is named Region; (ii) regions in 11 modalities (according to NUTS nomenclature), and the key variable is called Macro Region. These standards of recoding produce four alternative solutions. They are also tested including marital status as a key variable, called Mar Stat.

The threshold is fixed at 0.01, that is an individual is considered as “safe” when in the population there are at least other 100 individuals showing the same combination of key variables. Records at risk are treated through local suppression. Table 2 shows the distribution of suppressions and the maximum of individual risk, by “solution” and key variable.

Results show that when considering Sex, Age85 and Macro Region as key variables (solution (1) and (3)), all individuals have a risk of identification lower than 0.01 (i.e. no suppressions have to be applied). In contrast, when marital status is added to this first set of key variables (solution (2) and (4)), 192 individuals have a risk of identification higher than the threshold. Most of the suppressions are applied to widow and divorced individuals, or to never married individuals but older age. That is, individuals showing less frequent combinations of marital status and age.

When the region is considered as a key variable, instead of macro region, we notice that the number of suppressions is still low if marital status is disregarded (solution (5) and (7)) and the maximum risk is not extremely high (0.036). The information loss due to the use of age85_edu instead of age85 does not produce a worth reduction of the suppressions. In both cases, if marital status is added as a key variable (solution (6) and (8)), the number of suppressions increases as well as the maximum risk.

Table 2 Individual Archive scenario: distribution of suppressions by solution and key variable (threshold fixed at 0.01).

Solutions	*Sex*	*Age85*	*Age85_edu*	*Region*	*Macro Region*	*Mar Stat*	*Ind. at risk*	*Suppressions*	*Max Ind.Risk*
(1)	0	0	----	----	0	----	0	0	0.008
(2)	0	0	----	----	0	192	192	192	0.078
(3)	0	----	0	----	0	----	0	0	0.008
(4)	0	----	0	----	0	192	192	192	0.078
(5)	7	1	----	0	----	----	8	8	0.036
(6)	7	1	----	0	----	598	606	606	0.093
(7)	7	----	0	0	----	----	7	7	0.036
(8)	7	----	0	0	----	598	605	605	0.093

We consider solution (1) as satisfactory, because the information loss due to recoding the place of residence in Macro Regions instead of Regions allows for not applying any suppression to the original micro data. Moreover, we disregard marital status at this stage, because this variable is not likely to be available to the intruder under the Individual Archive scenario, and in any case we consider it as a key variable under the Household Archive scenario.

As far as the Household Archive scenario is concerned, the same variables as in the previous scenario are considered, and parental relationship and the household size are included (named respectively Rel Par and HHsize). Under this scenario, the threshold is fixed at 0.04. It is higher than in the previous case because the population register (i.e. the intruder external archive) is not public. Thus, the intruder is not likely to have access to it for a region (or macro region).

Results of 3 different combinations of key variables are shown in Table 3. We firstly consider sex, age85, macro region, marital status, household size and parental relationship as key variables. Households estimated as at risk are 531, consequently 3041 suppressions are applied. Clearly, in some unsafe households there are more than one individual showing a risk higher than the threshold divided by household size. Thus, more than one suppression per household is applied. The second solution (2) shows that substituting age85 with age85_edu the information loss due to the aggregation of some ages in classes is not compensated by a significant reduction of suppressions. In the third solution (3) we use age85 and region instead of macro region, increasing the geographical information. As a consequence, the number of households at risk and of suppressions increases. Comparing these solutions, we notice that the information loss due to the use of macro region instead of region actually strongly reduces the risk of identification, and the number of suppressions. Thus we consider solution (1) as satisfactory.

Table 3 Household Archive scenario: distribution of suppressions by solution and key variable (threshold fixed at 0.04).

Solutions	*Sex*	*Age85*	*Age85_edu*	*Region*	*Macro Region*	*Mar Stat*	*Hous. Size*	*Rel Par*	*Hous. at risk*	*Suppr.*	*Max Hous.Risk*
(1)	259	1131	----	----	0	720	0	931	531	3041	0.51
(2)	217	----	1038	----	0	720	0	864	490	2839	0.51
(3)	485	2039	----	2	----	745	0	1162	828	4433	0.51

Finally, households of bigger size result to be more protected (see Table 4). Thus, disclosure problems due to the hierarchical data structure are solved to some extent by the anonymisation method proposed.

Table 4 Suppressions by household size under the Household Archive scenario, Solution 1.

Household size	*Number of records with suppressions*	*Number of records in the sample*	*% of records with suppression*
1	32	6342	0.5
2	192	13644	1.4
3	414	15612	2.7
4	621	18143	3.4
5	613	5840	10.5
6	559	1501	37.2
7	302	483	62.5
8	60	88	68.2
9	39	45	86.7
10	16	30	53.3
11	22	22	100.0
Total	2870	61750	

According to the analyses carried out, a proposal to protect the EU-SILC micro data can be exploited as follows: (i) variables in Table 1 are dropped or recoded; (ii) age is top-coded at 85 years, and place of residence in 11 Macro Regions according to

NUTS nomenclature; (iii) BIR is estimated, and no suppressions have to be applied (solution 1 in Table 2); (iv) BHR is estimated and key variables are suppressed for individuals belonging to households at risk (solution 1, Table 3).

5. Conclusions

In this work we propose an approach to define a MFR from provisional EU-SILC micro data. Particularly we highlight problems of statistical disclosure control when dealing with data presenting both hierarchical and longitudinal structure, as is the case of the EU-SILC micro data. We described the individual approach to the risk of disclosure based on a probabilistic estimate of the re-identification risk. The risk of disclosure has been analysed taken into account some disclosure scenarios in the Italian context. In particular, we consider (i) a Nosy scenario (Section 3.1) where disclosure is possible by spontaneous recognition, and (ii) two scenarios where re-identification may arise by record linkage techniques (Individual and Household archive scenarios, respectively in Section 3.2 and Section 3.3). In this last cases re-identification risks can be estimated and a threshold can be fixed in order to classify record "at risk" or "safe". Consequently, protection method can be applied in order to minimise information loss, guaranteeing the respect of the fixed acceptable level of risk.

We argue that high levels of risk are estimated when both hierarchical and longitudinal structure of data are taken into account. On the other hand, we observe that the disclosure scenario for such a situation may occur rarely, because of the low chance to access reliable external archive with household information and in different point in time, coherently with the observation period of the survey. Nevertheless, analyses on re-identification risk may be conducted when future waves of the survey will be available. Suitable disclosure scenarios taking into account longitudinal structure of the data may be defined at least on the bases of the individual scenario. At this stage, we suggest to consider these aspects mainly under the Nosy scenario. A MFR can be proposed on the basis of the experimental results presented in Section 5, provided that recoding and dropping of variables reported in Table 1 are applied. Anyway, we recommend, when releasing the EU-SILC user data base, to ask the researcher for signing an agreement, and consequently guarantee the data protection on legal basis too.

Acknowledgements

We thank Luisa Franconi e Maurizio Lucarelli for providing helpful comments and suggestions.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Istat.

References

Abowd, J. M. & Woodcock, S. D. (2000). *Disclosure Limitation in Longitudinal Linked Data*. In P. Doyle, J. I. Lane, J. J. Theeuwes and L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies 215-278. North-Holland.

Benedetti, R., Franconi, L. & Capobianchi, A. (2003). Individual Risk of Disclosure Using Sampling Design Information. Istat Contributi n. 14/2003. Available at http://www.istat.it/dati/pubbsci/contributi/Contr_anno2003.htm.

Benedetti, R.& Franconi, L. (1998). *Applied Issues on Disclosure Avoidance in Complex Microdata Files.* SDC-Statistical Disclosure Control Project (Esprit-20426) Deliverable No: MI2-D2.

Di Consiglio, L., Franconi, L. & Seri, G. (2003). *Assessing individual Risk of Disclosure: an Experiment.* In Proceedings of the Joint ECE/Eurostat Work Session of Statistical Data Confidentiality, Luxembourg, April 7-9 2003. Available at http://neon.vb.cbs.nl/casc/.

Domingo-Ferrer, J. & Torra, V. (2001). *Disclosure Control Methods and Information Loss for Microdata.* In P. Doyle, J. I. Lane, J. J. Theeuwes and L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies 91-110. North-Holland.

Franconi, L. & Polettini, S. (2004). Individual Risk Estimation in μ-Argus: A Review. In J. Domingo-Ferrer and V. Torra (Eds), Privacy in Statistical Database. *Springer-Verlag*, Berlin Heidelberg 2004.

Polettini, S.& Seri, G. (2003). *Guidelines for the Protection of Social Micro-data Using Individual Risk Methodology: Application with μ-Argus Version 3.2.* CASC-Computational Aspects of Statistical Confidentiality Deliverable No: 1.2-D3. Available at http://neon.vb.cbs.nl/CASC/deliv/12D3_guidelines.pdf

Willenborg, L. & de Waal, T. (2001). Elements of Statistical Disclosure Control. Springer, New York.