WP.12
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
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

DISCLOSURE SCENARIO AND RISK ASSESSMENT: STRUCTURE OF EARNINGS SURVEY

Supporting Paper

Prepared by Daniela Ichim and Luisa Franconi (Istat, Italy)

Disclosure scenario and risk assessment: structure of earnings survey

Daniela Ichim, Luisa Franconi

Istituto Nazionale di Statistica, via C. Balbo 16, Rome, Italy.
(ichim@istat.it, franconi@istat.it)

Abstract. The anonymisation of Structure of Earnings Survey microdata is addressed. A practical implementation of the statistical disclosure control paradigm is illustrated. Two realistic disclosure scenarios for enterprise and employee re-identification are presented. Both scenarios are based on a careful analysis of the microdata file information content. Global recoding and a constrained regression model are implemented to protect the units at risk of re-identification. Using this approach, only few variables and only the records at risk are modified. This leads to an important information preservation.

1 Introduction

The Structure of Earnings Survey (SES) provides detailed information on the level and structure of remuneration of employees, their individual characteristics and the enterprise or local unit to which they belong to. The SES outcome represents an uniquely rich data source on gross earnings in Europe which is increasingly important for evidence-based policy making, in particular for monitoring economic growth and social cohesion. Furthermore, the SES data are indispensable for employers and employees as regards the demand and supply of labour.

National Statistical Institutes (NSI) disseminate more and more microdata files for research purposes. Such information dissemination is always constrained by the preservation of confidentiality of respondents. Statistical disclosure limitation methods are commonly applied to any survey microdata file. Special characteristics of microdata should be taken into account by the protection methods. Given the particular characteristics and needs of the typical user, the microdata files for research purposes should leave the possibility to perform very reliable analysis. An optimal trade-off between risk of re-identification and information loss should always be looked for and found by the NSIs.

The theoretical strategy proposed in [1] was put into practice in order to anonymise the Italian SES microdata file. Both respondent and user requirements were considered.

The paper is organized as follows. Section 2 briefly presents the Italian structure of earnings survey (reference year 2002). The adopted disclosure scenarios are described in section 3. Some preliminary work on variables is presented in section 4. Risk estimation issues and protection of enterprises and employees are discussed in section 5 and 6, respectively. An information loss assessment is provided in section 7.

2 Italian Structure of Earnings Survey

SES is one of the business surveys specified in the European Commission Regulation 831/2002. In Italy, SES data was collected by means of a sampling survey. A two stage sampling scheme was followed. Especially for this wave, Italy had obtained a derogation allowing a stratified sampling of enterprises instead of local units, as would be required by the European Commission Regulation 1916/2000. The enterprises sampling frame was the most up-to-date version of (Archivio Statistico delle Imprese Attive - statistical business register of active enterprises). Employees were sampled through a proportional to size sampling scheme. The observed variables are those indicated in the Regulation 1916/2000. Generally, the optional variables were not observed in the Italian survey. At enterprise level, structural economic variables were registered: economic activity ($Nace$), number of employees ($SizeEnt$), geographical location ($Nuts$), form of economic and financial control and existence of collective pay agreements. On employees, besides $Gender$ and Age, variables related to education, profession and contractual position were surveyed. Annual and monthly earnings corresponding to the reference month (October 2002) were registered together with their various components. The working time was accounted for through variables like number of paid hours. More details on the Italian survey may be found in [4].

The sampling weights were computed in order to indicate each unit representativeness, see [2]. Generally, the weights should satisfy some restrictions, for example the preservation of some known population totals. The enterprise sampling weights were derived from the inverses of the inclusion probabilities by means of a multivariate calibration procedure, see [3, 4]. To compute the employees final weights, the procedure indicated in [2] was followed.

3 Disclosure scenario

Coherently to the SES hierarchical structure, two disclosure scenarios were adopted, taking into account the particular observed phenomenon: one for enterprise re-identification and one for employee re-identification. No nosy colleague or external register scenarios were deemed realistic.

3.1 Enterprise scenario

Given the existence of publicly available business registers, it may be supposed that an intruder could use such information to re-identify an enterprise. Moreover, it is known that, for business surveys, large enterprises are always included in the sample. Public business registers report general information on name, address, number of employees, principal economic activity, geographical location, turnover, exports, etc. Among these variables, $Nace$, $Nuts$ and $SizeEnt$ were observed in SES. Consequently, they were the key variables of the adopted disclosure scenario. Of course, only enterprises belonging to combinations with very small frequencies should be considered at risk of re-identification.

The survey confidential information related to enterprises is represented by the economic/ fiscal/ social policies that could be deduced/inferred from the variables observed on employees. But the content to be protected against confidentiality breaches is not due to the variables directly observed on enterprises. In conclusion, a disclosure scenario based on confidential variables was not deemed realistic for enterprise re-identification.

3.2 Employee scenario

The observed variables related to employees may be classified in two categories. There are “social” and “fiscal” variables. The latter cannot be subject to an external disclosure scenario since they are not publicly available, at least in Italy. The “social” variables observed in the Italian 2002 SES (age, gender, etc.) may be available in external registers, but they exhibit very high frequencies. Irrespective of other variables, the "social" ones cannot be considered as identifying variables.

In absence of any other information, variables on earnings, number of paid hours and absence days cannot be subject to any spontaneous identification. Combined only with *Gender* and *Age*, these variables cannot either be used for spontaneous employee identification. The reason is that the observed social variables are not at all identifying (high frequencies). In other words, it was believed that an intruder could not identify an employee only by means of gender, age, number of paid hours and earnings variables, for example. Moreover, the variable Management position or supervisory position ($ManPos$) is hardly identifying because of the second option in its definition.

Instead, enterprise information, representing an employee activity, could be used to identify an employee. Assuming such knowledge, an employee could be identified by means of the "social" variables. In Italy, there does not exist any reliable external register containing information on occupation and/or education. Consequently, these variables could be hardly used by an intruder, except for spontaneous identification based on very detailed personal *a-priori* knowledge. But in such cases, the disclosure information content would probably be substantially diminished. The same reasoning applies to variables related to the type of contract and length of stay in service. In conclusion, it was considered that only *Gender* and *Age* could be combined with enterprise information for employee identification.

About the disclosure content, it was assumed that an intruder could be interested only in extremely high earnings. "Small-medium" earnings were not judged at risk since in Italy many occupational categories are subject to some kind of national contract, at least as a common basis. Hence such earnings should not be "appealing" for an intruder. Given the Italian economy structure, it was believed that small-medium size enterprises (and their employees, too) were hardly identifiable because of their high frequencies. As the microdata file is to be released for research purposes, such interest of an intruder in small-medium enterprises cannot be fully claimed. Hence, it was supposed that only high earnings corresponding to large (and generally well-known) enterprises could be considered interesting by an intruder.

In conclusion, the adopted disclosure scenario assumes that the employee identification is possible by means of: a) information on enterprise (*Nace* x *Nuts* x *SizeEnt*), b) social variables (*Gender* x *Age*) and c) extremely high earnings in large enterprises.

Since the anonymised microdata file would be disseminate for research purposes, modification of only key and confidential variables was deemed sufficient. The other variables could be released unchanged.

4 Recoding

For the Italian 2002 SES microdata file, several variables were recoded taking into account both confidentiality issues and user requirements.

1. Number of employees was recoded in 4 categories: $E10-49$, $E50-249$, $E250-999$, $E1000+$. The new variable was called $Size$.

2. Coherently with Istat dissemination policy, *Nace* divisions 10-14 were aggregated together, as well as the divisions 15 and 16.

3. *Age* was recoded in 14-19, 20-29, 30-39, 40-49, 50-59, 60+.

Rare case	**Frequency**
Sample uniques	70
Population uniques	78
Sample and population uniques	**28**
Sample doubles	50
Population doubles	54
Sample uniques and population doubles	**19**
Sample and population doubles	**15**

Table 1: Frequency of combinations $Nace$ x $Nuts$ x $Size$ with 1 or 2 units.

4. *Length of service in the enterprise* was recoded in intervals of 4 years. The original variable was kept, too. The recoded variable, *Len*, was used in the perturbation stage.

5. Total gross annual earnings in the reference year (*AnnualEarnings*) was recoded in categories of 10000 euro. The original variable was kept, too. The recoded variable, *AnnEarn*, was used to determine the employees at risk of re-identification.

5 Enterprises anonymisation

5.1 Enterprises at risk of re-identification

The key variables in the enterprises disclosure scenario were $Nace, Nuts$ and $Size$. These key variables are all categorical. With respect to the adopted disclosure scenario, the enterprises at risk were those belonging to combinations of key variables with frequencies below an *a-priori* given threshold. The drawback of this approach is that it does not consider the survey characteristics. As the Italian 2002 SES was a *sampling* survey, both sample frequencies and population frequencies were considered. When a population combination of key variables contains many enterprises, it would be more difficult to identify a sampled enterprise, even if it is a sample unique. Consequently, a sampled enterprise was considered at risk when both population and sample frequencies were simultaneously inferior to 3.

Considering the key variables recoded as in the previous section, the sample of enterprise contains 641 non-empty combinations. Instead, the population of enterprises contains 910 such combinations. The table 1 presents the frequencies of sample and population rare cases. With respect to the adopted disclosure scenario, 62 enterprises were considered at risk of re-identification.

5.2 Protection

Protection of enterprises at risk of re-identification was achieved by means of a dedicated global recoding procedure, see [5]. Since $Nace$ and $Nuts$ are the most important from the user point of view, only $Size$ was recoded.

As the enterprises at risk of re-identification were defined by means of both sample and population frequencies, the global recoding was applied controlling the population frequencies. Indeed, for each sample combination at risk, the corresponding population combination was identified. Two sample combinations were aggregated if the overall fre-

Original		**After recoding**	
Size	**Frequency**	**Size**	**Frequency**
$E10-49$	4852	$E10-49$	4794
$E1000$	247	$E1000$	213
$E250-999$	1257	$E250-999$	1112
$E50-249$	2461	$E50-249$	2322
		$E10-49\_E50-249$	53
		$E10-49\_E50-249\_E250-999\_E1000$	13
		$E250-999\_E1000$	152
		$E50-249\_E250-999$	152
		$E50-249\_E250-999\_E1000$	6

Table 2: Frequencies of *Size* before and after the application of global recoding.

quency of the population combinations was higher than the threshold. Obviously, this procedure can be applied only if the population frequencies are available.

As stated by the survey experts, it is preferable to aggregate a *Size* category with the category corresponding to immediately larger enterprises. When such recoding was not sufficient (the population frequency could still be smaller than the threshold), aggregation with the category corresponding to immediately smaller enterprises was investigated. If needed, aggregation of all *Size* categories for a given *Nace* x *Nuts* combination was performed. The sample frequencies before and after global recoding are presented in table 2.

6 Employees anonymisation

6.1 Employees at risk of re-identification

With respect to the adopted disclosure scenario, an employee could be identified only if information on enterprise is used. Considering the information content, only extreme earnings in large (and well-known) enterprises could be interesting for an intruder. For the anonymisation of the Italian 2002 SES microdata, enterprises with more than 250 employees were considered as large enterprises. In this sample there are 1504 enterprises with more than 250 employees (over a total of 8817 in the sample), while the population contains 3093 such enterprises (over a total of 193256). There are 40687 (over 81975) sampled employees belonging to a large enterprise. Due to the performed *SizeEnt* global recoding, further uncertainty is introduced. For example, an intruder wouldn't know whether an enterprise belonging to the category $E50-249\_E250-999$ belongs to $E50-249$ or $E250-999$.

Concerning the variable used for identification, several considerations hold. Firstly, as previously discussed, only a spontaneous identification scenario was adopted. This means that it was supposed that a possible intruder has no access to values of some variables possibly known only by a nosy colleague (for example, *paid hours* or *absence days*). Moreover, re-identification of an employee using his/her *number of absence days* or *number of working days* would be quite difficult. Therefore, an employee identification would be possible only by means of some previous "guesses" on ranges of earnings variables. Considering that the microdata file would be released for research purposes, the *Annual earnings* was deemed more adequate for spontaneous identification purposes. This was due to the “management” bonuses generally included in the annual earnings.

An intruder wouldn’t be interested in differences of few thousands of euro between two values. Moreover, he wouldn’t be even able to consider as different such two close values. Consequently, to determine the employees at risk of identification, *AnnualEarnings* was recoded into categories, resulting in a new variable *AnnEarn*, see section 4.

AnnualEarnings values exceeding a certain threshold T were considered as extremely high earnings, hence subject to spontaneous identification. T should be the same for all combinations of categorical key variables. An intruder would probably try to identify those employees with earnings greater than a certain $T_{intruder}$, an *a-priori* value imagined/thought by him. This $T_{intruder}$ is the limit above which the intruder would consider all earnings as high and interesting. As $T_{intruder}$ probably depends on both personal experience and knowledge of the studied economical phenomenon, its value cannot be determined by the data protector. Based on the observed data only, $T_{intruder}$ was estimated by T. Considering the skew probability distribution of *AnnualEarnings*, for the Italian 2002 SES, T was computed as the 99% quantile of the distribution. Obviously, it was assumed that employees with extremely low earnings were not at risk of re-identification.

Then, for each $Nace, Nuts, Size, Gender, Age, AnnEarn$ combination, the sampled employees with earnings greater than T were counted. If there was a single employee with such characteristics, it was considered at risk of re-identification. Note that, in this way, the number of employees at risk of re-identification is not *a-priori* defined. In the Italian 2002 SES file, using this procedure, 317 employees were considered at risk.

6.2 Protection

Since the microdata would be disseminated for research purposes, perturbation of only records at risk of re-identification was deemed sufficient. Protection was applied taking into account also some probable usages of the microdata file. As stated by survey experts, regression models are frequently used to estimate the possible differences between different categories. For example, a researcher could be interested in estimating the difference on *AnnualEarnings* between two regions (estimating differences between regional politics). The employees protection was achieved by a perturbation method based on a regression model.

For the Italian 2002 SES, only parametric linear models were considered. The response variable was *AnnualEarnings.*

The explanatory variables choice is the most crucial step for the protection procedure. Firstly, the variables having a significant impact on the *AnnualEarnings* behaviour should be considered. For example, if it is believed that $Nace$ significantly explains the *Annual Earnings* variation, $Nace$ should be an explanatory variable. Secondly, the assumed model should simulate an user analysis. It was supposed that *AnnualEarnings* can be modelled as a linear combination of *Size, Gender, Age, ManPos, Occupation,FullTimePartTime, Len, MonthlyEarnings* and *PaidHoursMonth*, respectively. The model was used for each combination of $Nace$ and $Nuts$.

Taking into account the already published totals, a constrained minimization problem was solved, see [6]. The main constraint was: for each combination of $Nace$ and $Nuts$, the relative difference between the original and perturbed value should not be higher than 0.5%, as required by the survey experts. Moreover, each perturbed value was restricted to belong to the interval (0.5*OriginalValue, 2*OriginalValue). Additionally, each perturbed value was required to be higher than the threshold T. These last two constraints actually controlled the perturbation introduced in each record. The *AnnualEarnings* values of the employees at risk of re-identification, were replaced by the corresponding fitted values. Then, the *MonthlyEarnings* values of the records at risk were proportionally modified.

For the Italian 2002 SES microdata file, because of particular values of the factors involved in the regression model, the above methodology couldn't be applied for one combination of $Nace$ and $Nuts$. For this particular combination, the *AnnualEarnings* values of the units at risk of re-identification were micro-aggregated, see [7].

By perturbing only the records at risk of re-identification, surely the *AnnualEarnings* and *MonthlyEarnings* weighted means would not be exactly preserved. But, the trend behaviour of these variables would be actually preserved. This behaviour was slightly lowered because only high earnings may be at risk of re-identification.

7 Information loss assessment

Only 0.39% of employees records were modified. The extreme earnings were generally decreased. Over the 317 units at risk of re-identification, the *AnnualEarnings* values were increased for 88 units. The summary statistics of the records at risk perturbation are indicated in table 3.

Min	**Q1**	**Median**	**Mean**	**Q3**	**Max**
-50	-1.70	4.75	8.69	19.07	100

Table 3: Percentages of the absolute relative perturbation of *AnnualEarnings*.

Since the *MonthlyEarnings* was proportionally modified with respect to *AnnualEarnings*, their consistency was automatically preserved. *Average gross hourly earnings in the representative month* was still computed as a ratio with respect to the perturbed *MonthlyEarnings*. Consequently, their consistency was maintained. Time related variables (number of worked hours, absence days, etc.) were not at all modified. The order relationships between *MonthlyEarnings*, *Special payment for shift work* and *Earnings related to overtime* were maintained, as well as their correlation coefficients, 0.03 and 0.14, respectively. A similar conclusion may be stated for the relationships between *AnnualEarnings*, *Total annual bonuses* and *Annual bonuses based on productivity*.

Generally, any microdata release is anticipated by the publication of a set of tables containing information on the survey variables. It is not always possible to exactly preserve the already published totals without a significant information loss with respect to other statistics. In such cases, at least an assessment of the difference one might obtain between the published totals and the ones computed using the microdata file is necessary. By definition, the applied protection method controls weighted totals for each combination of $Nace$ and $Nuts$.

For the Italian 2002 SES microdata, several tables were already published, involving mainly the following variables: $Nace$, $Nuts$, *Size*, *Gender*, *Age*, *FullTimePartTime*, $ManPos$, *Occupation*. The microdata file contains 26295 combinations of these variables. Only 263 (1%) of these combinations were modified by the applied perturbation. The absolute relative changes in the weighted means of *AnnualEarnings* were all inferior to 0.3. The mean of these relative perturbations was 0.06. Considering that the already published tables do not have the same maximum level of detail with respect to which these evaluations were performed, it was supposed that, at higher hierarchical levels, these relative perturbations on the means would tend to compensate one another. These were the main reasons for not applying any further adjustment.

8 Conclusions

A detailed analysis of possible disclosure scenarios and an accurate definition of related identifying variables are the key points of this anonymisation procedure. Considering that the microdata file would be released for research purposes, the identification of units at risk is based on individual risk measures. This approach of a tailored definition of units at risk allows for dedicated protection methods that can save more information content.

Consideration of different scenarios is a key issue. For the Italian 2002 SES, for the enterprises an spontaneous identification scenario was adopted. Frequencies in the population of enterprises were considered for rare cases identification. Instead, for employees it was adopted a spontaneous identification scenario based on both an estimated threshold earning value and a rarity concept.

For the Italian 2002 SES microdata file, enterprise protection was achieved by aggregating categories of enterprise size. Consequently, two of the most important survey variables ($Nace$ and $Nuts$) remained completely unchanged. A perturbation was applied only to records of employees considered at risk of re-identification, resulting in a significant reduction of the information loss. The perturbation method was derived from an analysis of user requirements. A set of constraints derived from a data utility criteria were used, too. The sampling weights were unchanged. An evaluation of the information loss was also performed.

The sampling weights, as well as the number of respondents, could also contribute to the enterprise identification. Other issues related to the information content analysis will be subject to further research.

References

[1] Hundepool, A. *et. al.* (2006) "Handbook on Statistical disclosure control", available at www.cenex.org.

[2] "Structure of Earnings Survey 2002 - Eurostat's arrangements for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000", Eurostat, 6 April 2004.

[3] Deville, J.C., Sarndal, C.E. (1992) "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, **87**, 376-382.

[4] ISTAT (2005) "Dipendenti, ore lavorate e retribuzioni nelle imprese dell'industria e dei servizi Anno 2002", Statistiche in breve, available at www.istat.it.

[5] Willenborg, L. & De Waal, T. (2001) "Elements of statistical disclosure control", Lecture Notes in Statistics, New York: Springer.

[6] Draper N.R., Smith H. (1998) *Applied regression analysis*, Wiley-Interscience.

[7] Defays, D., Anwar, M.N. (1998) "Masking Microdata Using Micro-Aggregation", *Journal of Official Statistics*, **14 (4)**, 449–461.