**Case studies A: The release of Microdata Files for Research Purposes**

# A2. Structure of Earnings Survey 2002

## *1. Introduction*

The European 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. It is a 4-yearly survey conducted under Council Regulation 530/1999 and Commission Regulation 1916/2000. 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.

According to Council Regulation No. 322/97 and Commission Regulation 831/2002, microdata files could be disseminated for research purposes. This document describes a methodology for the dissemination of a microdata file for research purposes and its application to the Italian SES 2002 microdata.

The statistical disclosure control methodology takes into account both the economic features of the data and the special class of users: scientific researchers. A detailed study of different possible disclosure scenarios is carried out in order to define identifying variables and a careful risk assessment analysis of employees is performed to single out the records at risk of disclosure. In the microdata file to be disseminated, only the records at risk of disclosure are protected (i.e. modified) whereas the rest of the file is released unchanged. Enterprise risk assessment is performed considering both the economic classification, regional detail and size classes because these are considered identifying variables. Protection of the enterprises is achieved by means of global recoding of some categorical variables. Protection of employees is achieved through a constrained regression model, the constraints being a data utility criteria. Comparison with the already published tables is also performed.

The microdata anonymisation methodology is based on the following eight steps:

1. Definition of two disclosure scenarios: for enterprises and employees.
2. Preliminary work on variables.
3. Risk assessment: re-identification of units (enterprises) at risk.
4. Protection of enterprises at risk.
5. Risk assessment: re-identification of employees at risk.
6. Protection of employees.
7. Information loss assessment.
8. Description of the microdata file to be disseminated.

The enterprises at risk of re-identification are determined (step 3) using an approach based on rare cases (sample and population). Protection of these units (step 4) is achieved by aggregating the categorical variable, considering also the a-priori defined priority of such variables.

The employees at risk of re-identification are determined (step 5) with respect to a spontaneous re-identification scenario, simulating an intruder reasoning. Protection of these employees is achieved by means of a linear regression model, considering the perturbation method from an user perspective.

## *2. Structure of earnings survey: brief description of the data*

The Structure of Earnings Survey (SES) is a four-yearly survey whose objective is to provide accurate and harmonised data on earnings in the EU Member States, Acceding Countries and Candidate Countries for policy-making and research purposes. The SES 2002 gives detailed and comparable information on relationships between the level of remuneration, individual characteristics of employees and their employer (economic activity, dimension and location of the enterprise).

The collection of data for the SES 2002 was obtained from tailor-made questionnaire. The reference year was 2002 while the reference month is October. The Italian sample resulted in 8817 enterprises and 81975 employees.

The statistics of the SES 2002 refer to enterprises with at least 10 employees in the areas of economic activity defined by sections C-K of NACE Rev.1. The population of employees was represented by those employees having an employment contract in the reference month in the enterprise. Specifically, the employees covered in the SES 2002 were those who actually received remuneration during the reference month. For the sampled employees who had period(s) of unpaid absence during the reference month, their earnings should had been adjusted on to a full months basis. Where it was not feasible to adjust their monthly earnings, then such employees should had been excluded from the sample. More details on the European survey may be found in Eurostat (2004).

The following variables[1] were observed in the Italian SES 2002 and subject to disclosure control in the present microdata anonymisation procedure. In parentheses there are the names of these variables used in this document. The other variables mentioned in the Council Regulation 1916/2000 were optional variables and they were not observed in the Italian SES 2002.

**A. Variables related to enterprise**
**A.1.1** Geographical location of the observation unit[2]. [*Nuts*]
**A.1.2** Size of enterprise to which the enterprise belongs to[3].[*SizeEnt*]

[1]The labels strictly follow the Council Regulation 1916/2000.
[2]Nomenclature of territorial statistical units NUTS level 1.
[3]The number of employees is classified in: 10-49, 50-249, 250-499, 500-999, 1000 and more employees.

**A.1.3** Principal economic activity of the observation unit[4]. [*Nace*]
**A.1.4** Form of economic and financial control[5].
**A.1.5** Existence of collective pay agreements covering the majority of employees in the observation unit[6].
**A.1.6** (optional) Total number of employees in the enterprise[7]. [*Emp*]
**A.4.1.** Sample weights[8].

**B. Variables related to employee**
**B.2.1** Gender.[*Gender*]
**B.2.2** Employee's age. [*Age*]
**B.2.3** Occupation[9].[*Occup*]
**B.2.4** Management position or supervisory position. [*ManPos*]
**B.2.5** Highest completed level of education and training[10].
**B.2.6** Length of service in the enterprise.
**B.2.7** Is the employee full-time or part-time? [*FtPt*]
**B.2.7.1** Share of a full-timer's normal hours.
**B.2.8** Type of contract of employment.
**B.3.0** Average gross hourly earnings in the representative month.
**B.3.1** Total gross earnings for a representative month. [*MontlyEarnings*]
**B.3.1.1** Earnings related to overtime.
**B.3.1.2** Special payment for shift work.
**B.3.2** Total gross annual earnings in the reference year. [*AnnualEarnings*]
**B.3.2.1** Number of weeks to which the gross annual earnings relate.
**B.3.2.2** Total annual bonuses.
**B.3.2.2.2** (optional) Annual bonuses based on productivity.
**B.3.4** Number of paid hours during the representative month. [*PaidHrsMonth*]
**B.3.4.1** Number of overtime hours paid in the reference month.
**B.3.5** Annual days of absence.
**B.3.5.1** Annual days of holiday leave excluding days of sick leave[11].
**B.3.5.1.1** Holiday entitlement or number of holidays actually taken.
**B.4.2** sample weights[12].

---

[4]2-digit level of NACE Rev. 1 for sections C-K.
[5]Public Control, Private Control, Shared Control.
[6]National level or interconfederal agreement; industry agreement; agreement for individual industries in individual regions; enterprise agreement; single observation unit (local unit) agreement; other type of agreement; no collective agreement exists.
[7]A simple head count of the total number of employees in the reference month was required, covering all employees..
[8]Calibrated with respect to *Nace*, *Nuts* and *SizeEnt* categories.
[9]ISCO-88 (COM) classification at the two digit level.
[10]The information required concerns the level of general, professional or higher education, which the employee has completed. Completed implies the successful completion of the training and is normally (but not invariably) accompanied by an appropriate paper qualification. It is only necessary to code the highest level reached.
[11]Total number of paid annual holidays (expressed in days) actually taken by the employee, excluding sick leave and public holidays. Depending on B3.5.1.1, it can be also be the number of paid entitled annual holidays.
[12]Calibrated with respect to the total number of employees.

## *3. Disclosure scenarios*

Following the hierarchical structure of the SES 2002 microdata file, two disclosure scenarios should be discussed. The first one should concern the enterprise re-identification, while the second disclosure scenario should concern the employee re-identification. The disclosure scenario(s) should take into account the particular national observed phenomenon. Throughout this document, the risk of disclosure is defined in terms of re-identification. This is a very general definition, widely accepted in many practical situations in the statistical disclosure control framework. This section describes the re-identification scenarios deemed realistic: one for enterprises and one for employees.

### 3.1 Enterprise scenarios

*3.1.1 Nosy colleague scenario; External register scenario*

The microdata file will be disseminated for research purposes. This is the main reason for discarding any "nosy colleague" and any "external register" scenarios. Indeed, it is hard to believe that the "bona-fide" researcher could be a "nosy colleague" or that he could act as an insider. Moreover, here it is supposed that the "bona-fide" researcher wouldn't perform any record linkage experiment for re-identification purposes (it is supposed that the scientific researcher is aware of the assumptions and costs of performing a record linkage experiment --- coherence of the involved databases, classifications, time, software, etc.).

*3.1.2 Disclosure scenario based on structural information*

The SES 2002 microdata file contains enterprise variables related to economical activity, geographical location and dimension of the enterprise[13]. It is obvious that these variables are structural variables.

Since the structural variables are generally publicly known [there are external databases that register such information with acceptable accuracy], it is supposed that an intruder could use this information to identify an enterprise. Moreover, it is known that a business register was used to construct the SES 2002 sampling frame. Consequently, a possible intruder *a-priori* knows that an enterprise possibly belonging to the sample, it is surely included in the business register. As the possible intruder is a researcher, it may be assumed that he wouldn't perform a complete[14] record linkage for re-identification purposes (see above). Nonetheless, the researcher may be curious about some units. For example, he might know in advance the most famous/dominant enterprises. Alternatively, some units may be highlighted during his analyses and the intruder may try to find some

[13]Expressed as of number of employees.
[14]For all records in the file.

more information about these enterprises. In other words, it may be supposed that the intruder-researcher **might** use some structural information **only** for the re-identification of **several particular enterprises**. Public business registers report general information on name, address, number of employees (*SizeEnt*), principal economic activity of an enterprise (*Nace*), region (*Nuts*) etc. Among these variables, *Nace*, *Nuts* and *SizeEnt* are registered also in the SES 2002 microdata file. Hence it may be assumed that an intruder could identify an enterprise using only information on *Nace*, *Nuts* and *SizeEnt*.

*3.1.3 Spontaneous re- identification scenario*

The confidential information on enterprises consists in the economic/fiscal/social policies that could be deduced/inferred from the variables observed on their employees. The information content that should be protected against confidentiality breaches is not exactly related to the variables observed directly on enterprises. Consequently, a spontaneous re-identification scenario for enterprise re-identification may not be deemed realistic.

## 3.2 Employee scenarios

*3.2.1 Nosy colleague scenario*

The microdata file will be disseminated for research purposes. This is the main reason for discarding any "nosy colleague" scenario.

*3.2.2 External disclosure scenario*

In the SES 2002, the employee variables may be classified in two categories. There are "social" and "fiscal" variables. The latter cannot be subject to an external disclosure scenario since they generally do not represent publicly available information. The social variables observed in the SES 2002 (*Gender*, *Age*, etc.) may be available in external registers, but the high (sample or population) frequencies of the combinations of their cross-classification show that they might not have an important identification power.

*3.2.3 Spontaneous re-identification scenario*

In absence of any additional information, observations on earnings, number of paid hours and absence days cannot be subject to a spontaneous re-identification scenario. Combined only with the observed social variables (*Gender*, *Age*), these variables cannot either be used for employee re-identification. The reason is that the social variables observed in the SES 2002 are not at all discriminating (because of the high frequencies of their combinations). In other words, since a nosy colleague scenario is not deemed feasible, it is hard to believe that an intruder could identify an employee only by means of her/his gender, age, number of paid hours and earnings variables, for example. Moreover, the variable *ManPos*, which could be highly identifying, should be understood as

information on the management or supervisory activities. Because of the second option in its definition, this variable almost loses its re-identification power.

Instead, some information on the enterprise could be used to identify an employee. Indeed, knowledge on the actual activity of an employee is represented in the SES 2002 microdata file by the information on the enterprise. Assuming that the intruder has such knowledge, an employee could be re-identified by means of the values of the social variables. In Italy, there does not exist any (publicly available and accurate) register containing information on occupation (ISCO 88) and education. Hence such variables could be hardly used by an intruder, except in cases of spontaneous re-identification based on very detailed personal *a-priori* knowledge. In such cases, the disclosure information content would probably be substantially diminished. The same reasoning applies to variables related to the type of contract of the employees and length of stay in service. Consequently, only variables *Gender* and *Age* could be combined with enterprise information to identify an employee.

About the information content of the re-identification, it is assumed that an intruder could be interested only in extremely high earnings. "Small-medium" earnings are not judged at risk of disclosure because many occupational categories are generally subject to some kind of national contract, as least as a common base. Hence such "small-medium" earnings should not be "interesting" for the possible intruder (the researcher). Moreover, it is supposed that only the high earnings corresponding to large (and generally well-known) enterprises could be interesting from an intruder point of view. The small-medium size enterprises[15] couldn't be identifiable (at risk of re-identification) because their frequency is far too high. Considering also the fact the microdata file will be disseminated for research purposes, the interest of an intruder in small-medium enterprises employees cannot be fully claimed. Consequently, the intruder could hardly be interested in the earnings of the employees of any small-medium enterprise. More details on the assumed disclosure scenarios may be found in Ichim (2007).

In conclusion, the adopted disclosure scenario assumes that the employee re-identification is possible by means of:

a. information on the enterprise (*Nace* x *Nuts* x *SizeEnt*)
b. social variables (*Gender* x *Age*)
c. extremely high earnings related to large enterprises

# *4. Preliminary work on variables*

## 4.1 Variable suppression

The following categories of variables should be removed from the microdata file to be disseminated:

[15]Hence also their employees even if one (or more) has an extreme salary.

1. Direct identifiers (name, address, etc).
2. Variables that were not observed (for example, for the Italian SES 2002, the majority of optional variables [e.g. employee citizenship])
3. Key variables considered too detailed:
    a. Number of employees of the local unit

## 4.2 Global recoding

When the number of combinations with frequency 1 is extremely high, a possible method to reduce the information content of some variables is recoding, see Willenborg (2001). The global recoding may be applied to both continuous or categorical variables. Of course, the recoding, as all the other statistical disclosure limitation methods, should be applied only to key or confidential variables. Otherwise, an information loss is obtained without any gain from the confidentiality point of view For the SES 2002 microdata file, several variables should be recoded taking into account both confidentiality issues and the user requirements.

1. Number of employees of the enterprise (variable A1.2). Four[16] categories are suggested: *E10 – 49, E50 – 249, E250-999, E1000+*. In the rest of the paper, the new variable is called *Size.*
2. NACE divisions should be recoded according to the NSI dissemination policy. For example, Coherently with Istat dissemination policy, NACE divisions 10-14 were aggregated together into a new class called *R10*. Moreover, NACE divisions 15-16 were aggregated into a new class called *R15*.
3. Age (variable B2.2). Six categories are suggested: 14-19, 20-29, 30-39, 40-49, 50-59, 60+. In the rest of the paper, these classes are called *AGE1, AGE2, AGE3,AGE4,AGE5* and *AGE6* respectively.
4. Length of service in the enterprise (variable B2.6). Intervals of 4 years are suggested. The original variable should be preserved, too. See section 8 for the usage of this recoded variable, called *Len* in the rest of the paper.
5. Total gross annual earnings in the reference year (variable B3.2). Categories of 10000 euro are suggested. The original variable should be kept, too. See section 8 for the usage of the recoded variable, called *AnnEarn* in the rest of the paper.

# *5. Identification of enterprises at risk of disclosure*

## 5.1 Enterprises at risk

As discussed in section 3.1, the key variables in the enterprises disclosure scenario are *Nace*, *Nuts* and *Size.* These key variables are all categorical. With respect to the assumed disclosure scenario, the enterprises at risk of disclosure are the sampled enterprises that belong to combinations of key variables having frequencies below an *a-*

[16]Enterprises with less than 10 employees were not surveyed in the 2002 SES.

*priori* given threshold. Generally this threshold is set equal to 3. That is, an enterprise is considered at risk when it belongs to a combination of key variables whose frequency is 1 or 2. The drawback of this approach is that it does not consider the kind of survey it was conducted. Consideration of only the sample frequencies would result in an increased number of units at risk of re-identification; hence much more uncertainty should be introduced in the microdata file. This somehow contradicts the generally accepted basic assumptions of the dissemination of microdata files for research purposes. In presence of a signed contract, the NSIs generally trust as much as possible the "bona-fide" researchers. Consequently, the NSIs generally try to put the maximum possible emphasis on the data quality aspects. Having in mind the aim of preserving the main features of the dataset, the number of units at risk of disclosure should be very much controlled. No overestimation should be suitable for the dissemination of the microdata files for research purposes. For *sampling* surveys, e.g. the Italian SES 2002, both sample frequencies and population frequencies should be taken into account. When a combination of key variables contains many population enterprises, it would be more difficult to identify a sampled enterprise, even if it is a sample unique. Consequently, a sampled enterprise is considered at risk when both population and sample frequencies of the corresponding combination of key variables are simultaneously below the given threshold.

In the Italian SES 2002 data, with respect to the key variables recoded as described in section 4.2, the sample of enterprises contains 641 non-empty combinations of key variables. Instead, the population of enterprises contains 910 non-empty combinations. The table 1 presents the number of both sample and population rare combinations (frequency below the threshold 3):

| **Rarity** | # |
|---|---|
| Sample uniques | 70 |
| Population uniques | 78 |
| Sample that are population uniques | **28** |
| Population doubles | 54 |
| Sample doubles | 50 |
| Sample uniques that are population doubles | **19** |
| Sample doubles that are population doubles | **15** |

**Table 1. Number of sample and population unique and double cases in the Italian SES 2002 microdata file.**

In conclusion, the total number of enterprises that should be considered at risk is the sum of "sample and population uniques" (28), "sample uniques that are population doubles" (19) and "sample doubles that are population doubles" (15); there are 62 combinations at risk of disclosure.

## 6. Protection of enterprises

As generally accepted, the protection strategy should be chosen by considering also the user requirements, not only the data protection aim.

Protection of enterprises at risk is achieved by means of a free global recoding procedure, see Willenborg. Following the results of a brief review of the scientific literature on this topic, see Ichim (2008), variables *Nace* and *Nuts* are the most important from the user point of view. This is the main reason for recoding the categories of *Size* only within each combination of *Nace* and *Nuts* at risk of re-identification. The category of *Size* determined by the *Nace* x *Nuts* x *Size* combination at risk[17] is aggregated with a close category of *Size* variable, completely maintaining the *Nace* and *Nuts* categories.

Since the enterprises at risk of re-identification are defined by means of both sample and population frequencies, the aggregation is performed with respect to the population frequencies. In other words, for each sample combination at risk, the corresponding population combination should be identified. Then the new aggregated category of *Size* is defined with respect to the population frequencies and not with respect to the sample frequencies. Obviously, this step can be performed only if the population frequencies are fully available to the data protector. When this information is not readily available, the enterprise weights (grossing-up) factors could be used as well. Anyway, it should be reminded that the sum of weights is only an estimate of the population frequency.

As stated by the survey experts, it is preferable to aggregate a *Size* category with the category corresponding to immediately larger enterprises. To perform such aggregation, an order relationship among the *Size* categories is required. The most natural relationship could be used:

$$E10\text{-}49 < E50\text{-}249 < E250\text{-}999 < E1000+.$$

If such recoding is not sufficient (the number of population units could still be less than the predefined threshold 3), the aggregation with the category corresponding to immediately smaller enterprises could be investigated. If this aggregation is sufficient (the number of population units belonging to the new combination is greater than the threshold), this aggregation should be performed. Otherwise, aggregation of all *Size* categories in the current *Nace* x *Nuts* combination should be performed. In case this aggregation is not sufficient too, the regional detail (*Nuts*) would be released at national level, involving a huge information loss. Following the previously described procedure, for the Italian SES 2002 microdata, it was not necessary to recode the geographical detail at national level.

The enterprise protection procedure is summarized in the following steps:

1. List the combinations of key variables containing sample uniques or doubles that are also population uniques or doubles.

[17] Containing a sample unique or double, that is also a population unique or double.

2. For each sample combination $k$ of key variables containing enterprises at risk, identify the corresponding population combination and perform the next steps with respect to the population combinations:
3. Identify the *Size* category $c$ defining $k$
4. Find the immediately superior category of *Size*: $c_{sup}$
5. If the sum of population units in $c$ and $c_{sup}$ is greater than the 3, aggregate[18] $c$ and $c_{sup}$, otherwise perform step 6
6. Find the immediately inferior category of *Size*: $c_{inf}$
7. If the sum of population units in $c$ and $c_{inf}$ is greater than the threshold, aggregate $c$ and $c_{inf}$, otherwise perform step 8
8. If the number of population units in all *Size* categories is sufficient, aggregate all categories, otherwise, perform step 9
9. Aggregate all *Size* categories and aggregate all *Nuts* categories at national level[19].

If deemed necessary, it could be possible to use different versions of the above procedure. For example, if no aggregation with the immediately inferior *Size* class is acceptable, one could aggregate with the immediately superior *Size* class or to all the categories.

By applying the above aggregation procedure, the resulting data set (Italian SES2002 microdata file) contained 590 non-empty combinations of key variables**.** The table 2 presents the sample frequencies (*Size* categories) before and after recoding.

| Before recoding | | 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 |

[18]Following a natural ordering of the *Size* categories, *E1000+* can be aggregated only with the category *E250-999* and, if such aggregation is not sufficient, all *Size* categories are joint together. For the same reasoning, *E10-49* can be aggregated only with the category *E50-249* and, when this is not sufficient, all *Size* categories are aggregated.

[19]This should generally be sufficient, as it happened for the Italian microdata set. Otherwise (for example, for a highly dominating enterprise), probably a record suppression followed by a new recoding procedure would be more suitable.

**Table 2. Frequencies of the combinations in the Italian SES 2002 microdata file.**

Instead of a free global recoding, a fixed global recoding recoding could be performed, see Willenborg (2001). In this case, if aggregation of two *Size* categories is required in a *Nace* x *Nuts* combination.

## 7. Identification of employees at risk

When identifying the employees at risk of re-identification, the survey variables must be considered aggregated with respect to the previous steps of the anonymisation procedure dedicated to the enterprises protection (sections 4.2 and 6).

As discussed in section 3.2, with respect to the assumed disclosure scenario, an employee could be identified only when information on enterprise is used. Moreover, considering the possible information content, only extreme earnings in large (and well-known) enterprises could present some interest from an intruder point of view. For the anonymisation of the SES 2002 microdata file, only enterprises with more than 250 employees could be considered as large[20] enterprises. In the Italian SES 2002 microdata file there were 1504 enterprises with more than 250 employees (over a total of 8817 in the sample), while the population contained 3093 such enterprises (over a total of 193256). In the sample, there were 40687 (over 81975) employees belonging to a large enterprise. It should be noted that, due to the aggregations performed for the enterprises protection, further uncertainty was introduced. For example, a possible intruder wouldn't know whether an enterprise belonging to the category *E50-249_E250-999* belongs to *E50-249* or to *E250-999*.

With regard to which variable the re-identification of employees could be attempted, several considerations hold. Firstly, as previously stated, only the spontaneous re-identification scenario is assumed. This means that it is assumed that a possible intruder has no access to values of some variables that could be known only to a nosy colleague (for example, variables *paid hours* or *absence days*). Moreover, re-identification of an employee with respect to his/her *number of absence days* or *number of working days* would be quite difficult. Therefore it is considered that re-identification of employees would be possible only by means of some previous "ideas" on ranges of earnings variables. In the SES 2002 microdata file there are two such variables: *Monthly earnings* and *Annual earnings*. Considering that the microdata file will be disseminated for research purposes, the *Annual earnings* is more adequate for usage for re-identification purposes. This choice is mainly due to the fact that annual earnings generally include the "management" bonuses based on productivity results.

For identification of the employees at risk of re-identification, *AnnualEarnings* is aggregated into categories, resulting in a new variable *AnnEarn*, see section 4.2. The

[20]Consideration of national economical system is required when defining the threshold for "large" enterprises.

reason is that a possible intruder wouldn't be interested[21] in differences of few thousands of euro between two values. Moreover, a possible intruder wouldn't be even able to consider as different such two close values.

*AnnualEarnings* exceeding a certain threshold $T$, are considered as extremely high earnings, hence subject to spontaneous re-identification. The threshold value $T$ should be the same for all combinations of key variables. This statement holds because there is no reason to consider a certain earning value as being high in a combination $k_1$ while considering it as "ordinary" in another combination $k_2$. Moreover, a possible intruder would probably try to identify those employees with earnings greater than a certain $T_{intruder}$, an a-priori value imagined/thought by the intruder. This $T_{intruder}$ is the limit above which the intruder would consider all earnings as high (and interesting!) earnings. As $T_{intruder}$ probably depends on both personal experience and knowledge of the studied economical phenomenon, its value cannot be determined by the microdata protector. Based only on the observed data, the microdata protector estimates $T_{intruder}$ by means of a unique $T$. Taking into account the fact that the earnings probability distribution is very skewed, $T$ value is computed for the SES 2002 microdata as the 99% quantile of the distribution[22]. For the Italian SES 2002 microdata the $T$ value was about to 92.000 euro. Obviously, it is assumed that employees with extremely low earnings are not at risk of re-identification. Moreover, other quantiles of the earnings distribution could also be used.

Then, for each combination of *Nace*, *Nuts*, *Size*, *Gender*, *Age*, *AnnEarn* the number of sampled employees with earnings greater than $T$ is computed. If there is a single employee with such characteristics, it is 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 SES 2002 microdata file, using this procedure, 317 employees were identified as being unique cases with respect to the combinations of key variables mentioned above. In the tables 3 – 9, the distribution of these 317 statistical units with respect to the considered key variables is presented. Moreover, the same distribution with respect to *Occup* is also shown. The categories that are missing in these tables contain no unique employees at risk of re-identification.

| *Nace* | # | *Nace* | # |
|---|---|---|---|
| *R15* | 24 | *N35* | 6 |
| *R17* | 15 | *N36* | 3 |
| *R18* | 3 | *N40* | 2 |
| *R19* | 1 | *N41* | 2 |
| *R20* | 1 | *N45* | 4 |
| *R21* | 7 | *N50* | 3 |

[21]The microdata file will be disseminated for research purposes.
[22]The threshold was computed only for large enterprises (with more than 250 employees).

| | | | |
|---|---|---|---|
| *R22* | 13 | *N51* | 16 |
| *R23* | 6 | *N52* | 8 |
| *R24* | 23 | *N55* | 5 |
| *R25* | 8 | *N60* | 2 |
| *R26* | 13 | *N61* | 1 |
| *R27* | 4 | *N63* | 4 |
| *R28* | 10 | *N64* | 3 |
| *R29* | 19 | *N65* | 42 |
| *R31* | 6 | *N66* | 15 |
| *R32* | 4 | *N72* | 7 |
| *R33* | 4 | *N74* | 24 |
| *R34* | 9 | | |

**Table 3. Distribution of the employees at risk by *Nace*.**

| *AnnEarn* | # |
|---|---|
| *AE9* | 27 |
| *AE10* | 74 |
| *AE11* | 60 |
| *AE12* | 47 |
| *AE13* | 41 |
| *AE14* | 24 |
| *AE15* | 18 |
| *AE16* | 6 |
| *AE17* | 5 |
| *AE18* | 4 |
| *AE19* | 3 |
| *AE20* | 2 |
| *AE21* | 2 |
| *AE22* | *2* |
| *AE24* | *1* |
| *AE27* | *1* |
| *AE9* | *27* |

**Table 4. Distribution of the employees at risk by *AnnEarn*.**

| *Nuts* | # |
|---|---|
| *ITC* | 157 |
| *ITD* | 82 |
| *ITE* | 61 |
| *ITF* | 15 |
| *ITG* | 2 |

**Table 5. Distribution of the employees at risk by *Nuts*.**

| *Age* | # |
|---|---|
| *AGE2* | 1 |
| *AGE3* | 44 |
| *AGE4* | 105 |
| *AGE5* | 137 |
| *AGE6* | 30 |

**Table 6. Distribution of the employees at risk by *Age*.**

| *Size* | # |
|---|---|
| *E1000* | 95 |
| *E250-999* | 176 |
| *E250-999 _ E1000* | 46 |

**Table 7. Distribution of the employees at risk by *Size*.**

| *Gender* | # |
|---|---|
| *F* | 19 |
| *M* | 298 |

**Table 8. Distribution of the employees at risk by *Gender*.**

| *Occup* | # |
|---|---|
| *I11* | 13 |
| *I12* | 179 |
| *I13* | 28 |
| *I21* | 28 |
| *I22* | 1 |
| *I23* | 5 |
| *I24* | 8 |
| *I31* | 7 |
| *I33* | 24 |
| *I34* | 2 |
| *I41* | 19 |
| *I52* | 2 |
| *I82* | 1 |

**Table 9. Distribution of the employees at risk by *Occup*.**

## *8. Employee protection*

Since the microdata file is disseminated for research purposes, modification of only records of employees considered at risk of re-identification is considered sufficient,. The statisticaò disclosure limitation methodology should be applied taking into account also the possible usages of the microdata file. As stated by survey experts, regression models are most frequently used on this kind of data in order to estimate the possible differences between different categories. For example, the researchers are generally interested in estimating the difference of *AnnualEarnings* between two categories of the regional detail (estimating differences between regional politics). The proposed protection method is based on a perturbation method based on a regression analysis: model selection, parameter estimation and computation of the fitted values.

### 8.1 Model selection

The class of models to be discussed is the output of a careful analysis of the user needs. For SES 2002 microdata file, only parametric linear models were considered.

The response variable should be the one with respect to which the perturbation should be applied. If necessary, for example for goodness-of-fit reasons, transformation of this variable could be investigated. For the SES 2002 microdata file, the response variable of the regression model is *AnnualEarnings*.

The choice of the explanatory variables is the most crucial step for the protection procedure. Firstly, the variables that could have a significant impact on the *AnnualEarnings* trend should be included among the explanatory variables. For example, if it is believed that *Nace* categories explain in a significant way this trend, *Nace* should be considered as an explanatory variable. Secondly, the assumed model should simulate somehow an user analysis (see also section 8.3).

For the SES 2002 microdata file, it was supposed that *AnnualEarnings* could be modelled as a linear combination of *Size*, *Gender, Age, ManPos, Occup,FtPt*, *Len, MontlyEarnings* and *PaidHrsMonth,* respectively. Such model was used for each combination of *Nace* and *Nuts*.

In conclusion, the mathematical formulation of the linear model used for employees protection is the following:

$$AnnualEarnings \sim \alpha_1 Size + \alpha_2 Gender + \alpha_3 Age + \alpha_4 Occup + \alpha_5 ManPos + \alpha_6 Len + \alpha_7 FtPt + \alpha_8 MonthlyEarnings + \alpha_9 PaidHrsMonth \quad (1)$$

## 8.2 Parameter estimation and computation of the fitted values

The weighted least squares method, see Draper (1998), is applied for the estimation of the parameters in equation (1), for each combination of *Nace* and *Nuts*. The least squares weights are the sampling weights (variable B4.2).

To ensure the coherence with the already published totals, a constrained minimization problem is solved, using dummy variables instead of the categorical ones in equation (1). The main constraint is given in terms of published totals: for each combination of *Nace* and *Nuts*, the relative difference between the original and perturbed value should not be greater than 0.5%, as required by the survey experts. Actually this constraint is transposed in terms of weighted totals of units at risk of re-identification, since these are the only records to be modified. Moreover, each perturbed value is restricted to belong to the interval (0.5*OriginalValue, 2*OriginalValue). Additionally, each perturbed value is restricted to be greater than the threshold used for the extreme earning re-identification. These last two constraints actually control the perturbation introduced in each record, on the key variable considered.

Fitted values are computed by replacing the estimates $\hat{\alpha}$ in equation (1).

For the Italian SES 2002 microdata file, because of a particular combination of the factors involved in equation (1), this procedure couldn't be applied for *Nace* = R64 and *Nuts* = ITC. For this particular combination of key variables, the *AnnualEarnings* values of the units at risk of re-identification were microaggregated (individual ranking with parameter 3).

## 8.3 Perturbation of records of employees at risk of re-identification

For the employees considered at risk of re-identification, only the values of the two earnings variables (annual and monthly) involved in this disclosure scenario are perturbed.

The values of the variable *AnnualEarnings* of the employees at risk of re-identification, are substituted by the corresponding fitted values obtained in the previous step. That is, for each combination of *Nace* and *Nuts*, if $A_i$ is the original value of *AnnualEarnings* of an employee at risk, the corresponding perturbed value $A_i^p$ is given by:

$$A_i^p \sim \hat{\alpha}_1 Size_i + \hat{\alpha}_2 Gender_i + \hat{\alpha}_3 Age_i + + \hat{\alpha}_4 Occup_i + \hat{\alpha}_5 ManPos_i$$
$$+ \hat{\alpha}_6 Len_i + \hat{\alpha}_7 FtPt_i + \hat{\alpha}_8 MonthlyEarnings_i + \hat{\alpha}_9 PaidHrsMonth_i$$

Then, the corresponding *Monthly earnings* $M_i$ value is proportionally modified:

$$M_i^p = \frac{M_i * A_i^p}{A_i}$$

By perturbing only the values of the employees at risk of re-identification, surely the means (with respect to combinations of categorical key variables) of *AnnualEarnings* and *MonthlyEarnings* would not be exactly preserved. But, the main trend of these variables would be preserved. When substituting the original value by the fitted value, the original value is moved towards the mean. This is the reason for which the trend is approximately preserved, even if slightly lowered (because only high earnings may be at risk of re-identification). In doing so, it is implicitly assumed that a user would study the trend of such variables, which is generally true. By leaving unchanged all the values corresponding to employees not at risk of re-identification, the information loss is very much controlled. More details on the applied perturbation method and its properties may be found in Ichim (2008).

## 8.4 Published totals

When perturbing the microdata, special attention must be paid on the already published information. Generally, any microdata release is anticipated by the publication of a set of tables containing information on the survey variables.

When disseminating anonymized microdata files, it is not always possible to exactly preserve the already published totals without a significant information loss from the point of view of 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 anonymized microdata file is necessary. The applied protection method controls by definition weighted totals for each combination of *Nace* and *Nuts*. That is, the variation between the original weighted totals and the "perturbed" weighted totals is, by construction of the perturbation method, smaller than 0.5%.

For the SES 2002 microdata file, the perturbation of the records of employees at risk of re-identification is achieved by replacing the original *AnnualEarnings* by the fitted value (see previous section). Since this perturbation is applied only to high earnings, a decrease of the weighted totals should generally be expected. Anyway, it must be noted that not all high earnings are decreased, but some of them could also be increased.

For the Italian SES 2002 microdata, several tables were already published, involving mainly the following variables[23]: *Nace*, *Nuts*, *Size*, *Gender*, *Age*, *FtPt*, *ManPos*, *Occup*. The microdata file contains 26295 combinations of these variables. Only 263 (1%) of these combinations were modified by the applied perturbation method. In figure 1, there is illustrated the distribution of the relative changes in the weighted means of the *AnnualEarnings*[24] variable for all these 263 combinations that were modified by the perturbation procedure.

[23]These variables were indicated as being the most important in the survey.

[24]Obviously, the means of the same combinations were modified also considering the *MonthlyEarnings* variable.

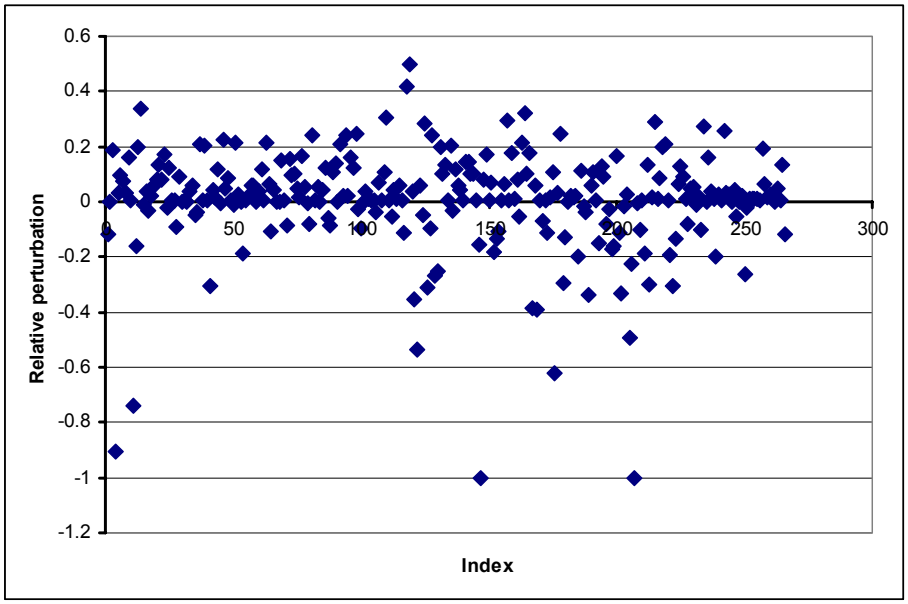


**Figure 1. Relative variations of the weighted means.**

The mean of these relative perturbations was 0.005. As these evaluations were performed at a much higher level of detail than the published tables, it is possible that, at higher hierarchy levels, these relative perturbations on the means would tend to compensate one another. These were the main reasons for not applying any further adjustment. When the distribution of the relative perturbations would be significantly different from zero, or in case the number of modified combinations means would be significantly higher, an adjustment for the preservation of the totals could be recommended.

## *9. Information loss and information preservation*

The general consequences of the applied statistical disclosure limitation methodology were discussed in the previous sections. The impact of the protection method on the Italian SES 2002 microdata file was assessed by means of various statistical indicators and it will be presented in this section.

The table 10 presents a summary of the protection achieved by each variable.

| Code | Variable | Status | Code | Variable | Status |
|---|---|---|---|---|---|
| A.1.1 | Geographical location | not changed | | | |
| A.1.2 | Size of enterprise | changed | B.3.0 | Average gross hourly earnings in the representative month | changed |
| A.1.3 | Principal economic activity | changed | B.3.1 | Total gross earnings for a representative month | changed |
| A.1.4 | Form of economic and financial control | not changed | B.3.1.1 | Earnings related to overtime | not changed |
| A.1.5 | Existence of collective pay agreements | not changed | B.3.1.2 | Special payment for shift work | not changed |
| A.1.6 | Total number of employees | removed | B.3.2 | Total gross annual earnings in the reference year | changed |
| A.4.1 | Enterprise sample weights | not changed | B.3.2.1 | Number of weeks to which the gross annual earnings relate | not changed |
| B.2.1 | Gender | not changed | B.3.2.2 | Total annual bonuses | not changed |
| B.2.2 | Employee's age | changed | B.3.2.2.2 | Annual bonuses based on productivity | |
| B.2.3 | Occupation | changed | B.3.4 | Number of paid hours during the representative month | not changed |
| B.2.4 | Management position or supervisory position | not changed | B.3.4.1 | Number of overtime hours paid in the reference month | not changed |
| B.2.5 | Education | not changed | B.3.5 | Annual days of absence | not changed |
| B.2.6 | Length of service in the enterprise | not changed | B.3.5.1 | Annual days of holiday leave | not changed |
| B.2.7 | Full-time or part-time | not changed | B.3.5.1.1 | Holiday entitlement or number of holidays | not changed |
| B.2.7.1 | Share of a full-time | not changed | B.4.2 | Employee sample weights | not changed |
| B.2.8 | Type of contract of employment | not changed | | | not changed |

**Table 10. SDC status of each variable; Italian SES 2002 microdata file.**

### 9.1 Enterprises information loss

*SizenEnt* was initially recoded in 4 classes. Then, due to the enterprises protection (free global recoding), more aggregated size classes were created. Only *Size* classes containing sampling and population uniques or doubles were aggregated. The distribution of the enterprises before and after aggregation was shown in Table 2 (section 6).

## 9.2 Employees information loss

Only records of employees at risk of re-identification were modified. That is, only 0.39% of the employees records were modified. These extreme earnings were generally decreased. Over the 317 cases at risk of re-identification, only in 103 cases the *AnnualEarnings* values were increased. The next table presents summary statistics of the absolute relative perturbation (percentages) introduced on the records at risk of re-identification. It must be stressed that the smallest perturbations generally correspond to lower (but above the threshold) earnings.

| Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|
| 0.12 | 3.99 | 10.82 | 14.91 | 20.73 | 100 |

### *Consistency checks*

Since the *MonthlyEarnings* is modified proportionally with respect to *AnnualEarnings*, their consistency is automatically preserved.

Variable *Average gross hourly earnings in the representative month* is still computed as a ratio with respect to the perturbed *MonthlyEarnings*. Consequently, their consistency is maintained.

Variables related to time (number of worked hours, absence days, etc) are not at all modified, hence their internal consistency is preserved.

*MonthlyEarnings* values still resulted being in all cases greater than *Special payment for shift work*. Their correlation coefficient remained unchanged (0.03).
*MonthlyEarnings* values still resulted being in all cases greater than *Earnings related to overtime*. Their correlation coefficient remained unchanged (0.14).

*AnnualEarnings* resulted still being greater than *Total annual bonuses*. Their correlation coefficient remained unchanged (0.59).
*AnnualEarnings* resulted still being greater than *Annual bonuses based on productivity*. Their correlation coefficient changed from 0.41 to 0.40.

### *Means comparison*

The problem of comparison with the already published totals was already addressed in section 8.4 with respect to all combinations of the variables indicated there and with respect to *AnnualEarnings* variable only.

The weighted[25] means modifications of the *AnnualEarnings* and *MonthlyEarnings* variables were compared with respect to all combinations of each of the variables *Nace*, *Nuts*, *Size*, *Gender*, *Age* and *Occup*. In tables 11 – 15, the relative perturbation (percentages) of the means with respect to the categories of these variables are presented.

[25]The sampling weights were used.

| *Nace* | *AnnualEarnings* | *MonthlyEarnings* | *Nace* | *AnnualEarnings* | *MonthlyEarnings* |
|---|---|---|---|---|---|
| *R10* | 0.00 | 0.00 | *R37* | 0.00 | 0.00 |
| *R15* | 0.01 | 0.01 | *R40* | 0.00 | 0.00 |
| *R17* | 0.01 | 0.01 | *R41* | 0.01 | 0.01 |
| *R18* | 0.00 | 0.00 | *R45* | 0.00 | 0.00 |
| *R19* | 0.00 | 0.00 | *R50* | 0.00 | 0.00 |
| *R20* | 0.00 | 0.00 | *R51* | 0.01 | 0.01 |
| *R21* | 0.01 | 0.00 | *R52* | 0.00 | 0.00 |
| *R22* | 0.01 | 0.03 | *R55* | 0.01 | 0.00 |
| *R23* | 0.02 | 0.02 | *R60* | 0.00 | 0.00 |
| *R24* | 0.01 | 0.00 | *R61* | 0.01 | 0.01 |
| *R25* | 0.00 | 0.00 | *R62* | 0.00 | 0.00 |
| *R26* | 0.00 | 0.00 | *R63* | 0.00 | 0.00 |
| *R27* | 0.00 | -0.01 | *R64* | 0.00 | -0.01 |
| *R28* | 0.00 | 0.00 | *R65* | 0.01 | 0.00 |
| *R29* | 0.00 | -0.01 | *R66* | 0.02 | -0.06 |
| *R30* | 0.00 | 0.00 | *R67* | 0.00 | 0.00 |
| *R31* | 0.00 | 0.00 | *R70* | 0.00 | 0.00 |
| *R32* | 0.01 | 0.00 | *R71* | 0.00 | 0.00 |
| *R33* | 0.01 | 0.02 | *R72* | 0.01 | 0.02 |
| *R34* | 0.01 | 0.01 | *R73* | 0.00 | 0.00 |
| *R35* | 0.01 | 0.00 | *R74* | 0.01 | -0.01 |
| *R36* | 0.00 | 0.00 | | | |

**Table 11. Relative perturbations of the means with respect to *Nace*.**

| *Nuts* | *AnnualEarnings* | *MonthlyEarnings* |
|---|---|---|
| *ITC* | 0.01 | 0.00 |
| *ITD* | 0.01 | 0.00 |
| *ITE* | 0.00 | 0.00 |
| *ITF* | 0.00 | 0.00 |
| *ITG* | 0.01 | 0.01 |

**Table 11. Relative perturbations of the means with respect to *Nuts*.**

| *Gender* | *Annual earnings* | *Monthly earnings* |
|---|---|---|
| *F* | -0.01 | -0.01 |
| *M* | 0.01 | 0.01 |

**Table 12. Relative perturbations of the means with respect to *Gender*.**

| *Size* | *Annual earnings* | *Monthly earnings* |
|---|---|---|
| *E10-49* | 0.00 | 0.00 |
| *E10-49_E50-249* | 0.00 | 0.00 |
| *E10-49_E50-249_E250-999_E1000* | 0.00 | 0.00 |
| *E1000* | -0.04 | -0.05 |
| *E250-999* | 0.10 | 0.08 |
| *E250-999_E1000* | 0.02 | 0.02 |
| *E50-249* | 0.00 | 0.00 |
| *E50-249_E250-999* | 0.00 | 0.00 |
| *E50-249_E250-999_E1000* | 0.00 | 0.00 |

**Table 13. Relative perturbations of the means with respect to *Size*.**

| *Age* | *Annual earnings* | *Monthly earnings* |
|---|---|---|
| *AGE1* | 0.00 | 0.00 |
| *AGE2* | 0.01 | 0.00 |
| *AGE3* | -0.02 | -0.02 |
| *AGE4* | 0.01 | 0.00 |
| *AGE5* | 0.13 | 0.10 |
| *AGE6* | -1.22 | -1.14 |

**Table 14. Relative perturbations of the means with respect to *Age*.**

| *Occup* | *Annual earnings* | *Monthly earnings* | *Occup* | *Annual earnings* | *Monthly earnings* |
|---|---|---|---|---|---|
| *I11* | -1.73 | -1.77 | *I51* | 0.00 | 0.00 |
| *I12* | 0.92 | 0.83 | *I52* | 0.01 | 0.01 |
| *I13* | 0.24 | 0.18 | *I61* | 0.00 | 0.00 |
| *I21* | -0.89 | -1.00 | *I71* | 0.00 | 0.00 |
| *I22* | 0.25 | 0.24 | *I72* | 0.00 | 0.00 |
| *I23* | -7.26 | -6.96 | *I73* | 0.00 | 0.00 |
| *I24* | -2.86 | -2.46 | *I74* | 0.00 | 0.00 |
| *I31* | 0.05 | 0.04 | *I81* | 0.00 | 0.00 |
| *I32* | 0.00 | 0.00 | *I82* | -0.01 | -0.01 |
| *I33* | -0.06 | -0.08 | *I91* | 0.00 | 0.00 |
| *I34* | 0.01 | 0.01 | *I92* | 0.00 | 0.00 |
| *I41* | 0.01 | 0.00 | *I93* | 0.00 | 0.00 |

**Table 15. Relative perturbations of the means with respect to *Occup*.**

## *10. Concluding remarks*

A detailed analysis of possible disclosure scenarios and an accurate definition of related identifying variables are the key points of the proposed statistical disclosure control

methodology. Considering that the microdata file would be disseminated for research purposes, the re-identification of units at risk is based on individual risk measures. This approach of selective re-identification of units at risk allowed for selective protection methods that could save more information content of the data.

Consideration of different scenarios is a key issue. For the SES 2002 microdata file, for enterprises, a disclosure scenario based on structural information is deemed realistic. Moreover, the population frequencies are considered for the identification of the rare cases. Instead, for employees it is adopted a spontaneous re-identification scenario based on both a dominance criteria (threshold on the earnings variable) and a rarity concept. Only some earnings variables are considered subject to a re-identification process.

For the SES 2002 microdata file, the protection of enterprises is achieved by aggregating categories of the enterprise size. Consequently, two of the most important survey variables (*Nace* and *Nuts*) are completely unchanged. A perturbation is applied only to records of employees considered at risk of re-identification, resulting in a significant reduction of the information loss. The perturbation method is derived from an analysis of user requirements. Moreover, a set of constraints derived from a data utility criteria is used. Sampling weights are unchanged. Therefore, users may obtain the same published values for many aggregated statistics.

## *11. References*

**1.** D. Defays, M.N. Anwar, *Masking Microdata Using Micro-Aggregation*, Journal of Official Statistics, Vol.14, No.4, 1998. pp. 449-461
**2.** L. Willenborg, T. De Waal (2001), *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, New York: Springer.
**3.** N.R. Draper, H. Smith (1998), *Applied regression analysis*, Wiley-Interscience; 3rd edition.
**4.** D. Ichim, L. Franconi (2007), *Disclosure Scenario and Risk Assessment: Structure of Earnings Survey*, Joint Unece/Eurostat work session on statistical data confidentiality. Manchester 2007, available at http://www.unece.org/stats/documents/2007/12/confidentiality/wp.12.e.pdf.
**5.** D. Ichim (2008), *Controlled Model-Based Disclosure Limitation of Business Microdata*, Atti della XLIV Riunione Scientifica della Società Italiana di Statistica, Cleup sc, pp. 293-300.
**6.** Eurostat (2004), *Structure of Earnings Survey 2002 - Eurostat's arrangements for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000*.