Italian Structure of Earnings Survey 2002

Microdata file for scientific purposes

File Description

SUMMARY

1. Introduction

When collecting data, Istat must guarantee that the statistical information would not be disseminated if confidential information about a single respondent could be too accurately inferred. A statistical disclosure control methodology was set-up in order to guarantee the confidentiality of respondents. The units at risk of re-identification were identified. Only the key or confidential variables were modified. Only the values of the units at risk of re-identification were perturbed.

The user of the microdata file for research must comply with the conditions of the signed agreement.

In order to foster the correct use of the data, the constraints induced by the statistical disclosure limitation methodology will be discussed in this document. In section 2 useful information about the status of each variable in the microdata file for scientific purposes is provided. The main effects of the statistical disclosure limitation methodology are shown in Section 3. Section 4 highlights some cautions that should be taken into account when performing very detailed statistical analyses using this microdata file for scientific research.

2. Variables description

This section summaries the changes of the variables with respect to the file that could be analyzed at Istat or Eurostat Safe Centers. Such changes are necessary in order to protect the confidentiality of respondents.

The labels and names of variables follow the Commission Regulation 1916/2000. The labels used in this document are also indicated in the document *MFR_ISTAT_SES2002_SurveyMethodology*.

The disseminated files are semi-colon (";") separated.
The decimal separator is ".".
The first row contains the names of the variables.
The file *MFR_ISTAT_SES2002FileA.csv* contains the information on enterprises.
The file *MFR_ISTAT_SES2002FileB.csv* contains the information on employees.

Table 1 – The status[1] of each variable in the Italian SES 2002 microdata file for scientific research.

Variable	Variable name	Type[2]	Status
A.1.1	Geographical location	Mandatory	Not changed
A.1.2	Size of enterprise	Mandatory	Changed
A.1.3	Principal economic activity	Mandatory	Changed
A.1.4	Form of economic and financial control	Mandatory	Not changed
A.1.5	Existence of collective pay agreements	Mandatory	Not changed
A.1.6	Total number of employees in the local unit	Optional	Removed
A.1.7	Principal market for the enterprise's products	Optional	Not changed
A.1.8	Size of the group of enterprises	Optional	Not changed
A.1.9	Country of residence of the entity controlling the group of enterprises	Optional	Not changed
A.4.1	Enterprise sample weights	Mandatory	Removed
B.2.1	Gender	Mandatory	Not changed
B.2.2	Employee's age	Mandatory	Changed
B.2.3	Occupation	Mandatory	Not changed
B.2.4	Management position or supervisory position	Optional	Not changed
B.2.5	Hightest completed level of education and training	Mandatory	Not changed
B.2.6	Length of service in the enterprise	Mandatory	Not changed
B.2.7	Full-time or part-time	Mandatory	Not changed
B.2.7.1	Share of a full-time	Mandatory	Not changed
B.2.8	Type of contract of employment	Mandatory	Not changed
B.2.9	Citizenship	Optional	Not changed
B.2.10	Coverage by a government scheme designed to promote employement	Optional	Not changed
B.2.11	Total period of career breaks - in months	Optional	Not changed

[1] The status refers **only** to the application of any statistical disclosure control methodology.
[2] The optional variables were not registered.

B.3.0	Average gross hourly earnings in the representative month	Mandatory	Changed
B.3.1	Total gross earnings for a representative month	Mandatory	Changed
B.3.1.1	Earnings related to overtime	Mandatory	Not changed
B.3.1.2	Special payment for shift work	Mandatory	Not changed
B.3.2	Total gross annual earnings in the reference year	Mandatory	Changed
B.3.2.1	Number of weeks to which the gross annual earnings relate	Mandatory	Not changed
B.3.2.2	Total annual bonuses	Mandatory	Not changed
B.3.2.2.1	Regular bonuses not paid at every period	Optional	Not changed
B.3.2.2.2	Annual bonuses based on productivity	Optional	Not changed
B.3.2.2.3	Annual premium related to profit sharing	Optional	Not changed
B.3.3	Employees' social security contrinutions and taxes paid by the employer on behalf of the employee to government authorities during the representative month	Optional	Not changed
B.3.3.1	Compulsory Social Security contributions	Optional	Not changed
B.3.3.2	Taxes	Optional	Not changed
B.3.4	Number of paid hours during the representative month	Mandatory	Not changed
B.3.4.1	Number of overtime hours paid in the reference month	Mandatory	Not changed
B.3.5	Annual days of absence	Mandatory	Not changed
B.3.5.1	Annual days of holiday leave	Mandatory	Not changed
B.3.5.1.1	Holiday entitlement or number of holidays	Mandatory	Not changed
B.3.5.2	Annual days of sick leave	Optional	Not changed
B.3.5.2.1	Annual days of sick leave paid by the employer	Optional	Not changed
B.3.5.2.2	Annual days of sick leave not paid by the employer	Optional	Not changed
B.3.5.3	Annual days of vocational training	Optional	Not changed
B.3.6	Annual estimation of payments in kind	Optional	Not changed
B.4.2	Employee sample weights	Mandatory	Not changed

In the next section, more details on the perturbation of variables are given.

3. Statistical disclosure control methodology

3.1 Disclosure scenarios

Following the hierarchical structure of the SES microdata file, two disclosure scenarios were deemed realistic: one for enterprises and one for employees.

3.2. Changes on variables

3.2.1 Variable suppression

The following types of variables were removed from the microdata file to be released.

1. Direct identifiers (name, address, etc).
2. Variables that were not observed in the Italian survey (majority of optional variables in the European survey; for example, citizenship)
3. Key variables considered too detailed:
    a. Number of employees of the local unit

In order to keep the microdata file structure, the removal operation consisted in application of "OPT" value for each value of each removed variable.

3.2.2 Global recoding

For the release of the Italian SES 2002 microdata file for scientific purposes, the following variables were recoded:

1. Number of employees of the enterprise (variable A1.2) was aggregated in 4 categories: *E10 – 49, E50 – 249, E250-999, E1000+*.
2. NACE divisions 10-14 were aggregated together into a new category called *R10*. NACE divisions 15-16 were aggregated into a new category called *R15*.
3. Age (variable B2.2) was recoded in 14-19, 20-29, 30-39, 40-49, 50-59, 60+. These categories were called *"20"),"* *[20-30)", "[30-40)","* *[40-50)","* *[50-60)"* and *"[60",* respectively.

3.3. Protection of enterprises

Protection of enterprises at risk of re-identification was achieved by means of a global recoding procedure. The global recoding procedure was applied to the variable *SizeEnt*. The resulting variable was called *Size* The global recoding was performed with respect to the population frequencies. This means that none of the population combinations of *NACE*, *NUTS* and *Size* contain a very reduced number of units.

3.4. Protection of employees

Only the variables *AnnualEarnings* and *MontllyEarnings* were modified. Only records of employees considered at risk of re-identification were modified. A controlled-model based protection method was applied to *AnnualEarnings*. Perturbation of employees at risk of re-identification was achieved by means of a constrained regression model. The *MonthlyEarmings* variable was proportionally modified.

4 Information content analysis

When applying whatever statistical disclosure limitation methodology, some information loss is unavoidable. (otherwise it wouldn't be possible to guarantee the confidentiality of the respondents).

4.1 Quality of Information on Enterprises

SizeEnt was initially recoded in 4 classes. Then, due to the enterprises protection, more aggregated size classes were created. Only size classes containing sampling and population uniques or doubles were aggregated. The next table presents the resulting sample frequencies of the *SizeEnt*.

Table 2 – Frequencies of size-classes in the Italian SES 2002 microdata file for scientific purposes.

SizeEnt	**Scientific purposes microdata file frequency**
E10-49	4794
E1000	213
E250-999	1112
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

4.2 Quality of Information on Employees

Only records of employees at risk of re-identification were modified. That is, only 0.39% of the employees records were modified. The next table presents summary statistics of the absolute relative perturbation (percentages) introduced on the records at risk of re-identification, on variable *AnnualEarnings*.

Table 3 – Summary statistics of the absolute relative perturbations of records at risk of re-identification (variable *AnnualEarmings*).

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

Coherence

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

Variable *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.

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

MonthlyEarnings values still resulted being in all cases greater than *Special payment for shift work*. Their correlation coefficient remained unchanged (0.03) after the perturbation.

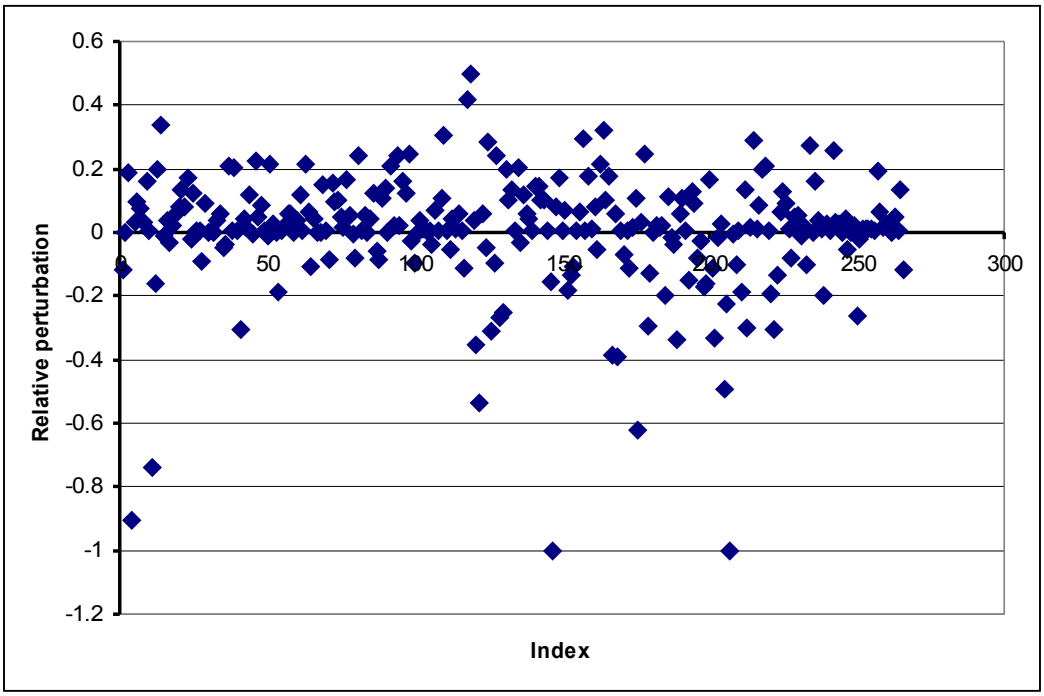
MonthlyEarnings values still resulted being in all cases greater than *Earnings related to overtime*. Their correlation coefficient remained unchanged (0.14) after the perturbation.

AnnualEarnings resulted still being greater than *Total annual bonuses*. Their correlation coefficient remained unchanged (0.59) after the perturbation.

AnnualEarnings resulted still being greater than *Annual bonuses based on productivity*. Their correlation coefficient changed from 0.41 (original value) to 0.40 (perturbed values).

Means comparison

The microdata file contains 26295 combinations of *Nace*, *Nuts*, *SizeEnt*, *Gender*, *Age*, *FtPt*, *ManPos*, *Occup* variables. Only 263 (1%) of the combinations of these variables were modified by the applied perturbation method. The graphic below illustrates the distribution of the relative changes in the weighted means of the *AnnualEarnings*[3] variable for all these 263 combinations that were modified by the perturbation procedure.



The mean of these relative perturbations is 0.005.

The weighted[4] 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*. The next tables present the relative perturbation (percentages) of the means with respect to the categories of these variables.

[3] obviously, the means of the same combinations were modified also considering the *MonthlyEarnings* variable.
[4] The sampling weights were used.

Table 4 – Relative perturbations for each *Nace* category.

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 5 - Relative perturbations for each *Nuts* category.

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 6 - Relative perturbations for each *Gender* category.

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

Table 7 - Relative perturbations for each *SizeEnt* category.

SizeEnt	*AnnualEarnings*	*MonthlyEarnings*
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 8 - Relative perturbations for each *Age* category.

AGE	*AnnualEarnings*	*MonthlyRarnings*
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 9 - Relative perturbations for each *Occup* category.

Occup	*AnnualEarnings*	*MonthlyEarnings*	*Occup*	*AnnualEarnings*	*MonthlyEarnings*
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

5 Recommendations on the use of the microdata file for scientific purposes[5]

At European level, Structure of Earnings Survey is particularly relevant as it provides specific information of earnings differentials by demographic and professional characteristics of employees on one side and the production unit on the other side. As a consequence even slight variations of the indicators, may be crucial.

In the previous section the main effects of the statistical disclosure limitation methodology aimed to the protection of both enterprises and employees have been presented. As highlighted, if individual categorical variables are considered, the effects of the modifications on *AnnualEarnings* and *MonthlyEarnings* result to be negligible: the difference of *AnnualEarnings* and *MonthlyEarnings* in the microdata file for scientific purposes to the non-perturbed microdata file, is mimimum.

Compared to the original stratification, the attribution of the enterprise to the size class mentioned in the Implementation Regulation for SES is not possible for 376 enterprises (over a total of 8817 enterprises in the sample). The new size classes allow only a partial comparison between large size enterprises (with more than 250 employees) and medium and small size enterprises (with less than 250 employees). As a matter of fact for 171 enterprises is not possible to distinguish between large and small-medium enterprises.

As a consequence the user has to take into account the effects of the statistical disclosure limitation methodology in the cases listed in Table 11. If the analysis is performed at a too detailed combination of stratification variables the resulting indicators may significantly differ from those published at national level or even in the New Cronos data base. Anyway, these differences between the microdata file for scientific purposes and the non-perturbed microdata file might occur only when variables *AnnualEarnings* and *MonthlyEarnings* are involved in the analysis. Moreover, these differences may occur only when the corresponding few NACE and NUTS categories are involved in the analysis.

Table 11 contains the list of combinations of NACE codes and NUTS 1 for whom the statistical indicators for all original size classes (and in particular for the classes "enterprises with less than 250 employees" and "enterprises with more than 250 employees") cannot be constructed.

Table 11 - List of combinations of NACE 2 digit and NUTS 1 for whom indicators relative to size class "more than 250 employees" cannot be built

Nace 2 digit	NUTS1
30	ITC
10	ITG
10	TOTAL
17	ITF
20	ITE
22	ITF
23	ITG
24	ITF
26	ITF
26	ITG
27	ITG
29	ITF
29	ITG

Nace 2 digit	NUTS1
30	ITF
33	ITF
35	ITG
37	ITE
37	TOTAL
40	ITG
50	ITD
50	ITG
61	ITD
62	ITD
62	ITF
67	ITD

Note: the first row (combination NACE 30 and NUTS1 ITC) refers to size class "less than 250 employees"

[5] The section has been written by S.Cardinaleschi

6 Contacts

The responsible for the statistical disclosure limitation methodology may be contacted at ichim@istat.it.

7 References

1. Ichim, D. (2008) *Controlled model-based disclosure limitation of business microdata*. XLIV Riunione Scientifica della Società Italiana di Statistica, Università della Calabria, 25-27 giugno 2008.
2. Ichim, D., Franconi, L. (2007) Disclosure scenario and risk assessment: Structure of Earnings Survey, *Eurostat Monographs on statistical data confidentiality*, available at http://epp.eurostat.ec.europa.eu/portal/page?_pageid=3154,70730193,3154_70730647&_dad=portal&_schema=PORTAL