Istat
Istituto Nazionale di Statistica

Farm Structure Survey

Reference year: 2005

File description

July 2009

CONTENTS

1. Introduction 3
2. Description of variables 4
3. Statistical disclosure control methodology 5
4. Information content analysis 6
5. References 10
6. Contacts 10

1. Introduction

When collecting data, Istat must guarantee that the statistical information would not be disseminated if confidential data about a single respondent could be too accurately inferred. For the Farm Structure Survey 2005, for the dissemination of a microdata file for research purposes (MFR), a statistical disclosure control methodology was set-up in order to guarantee the confidentiality of respondents.

The user of the MFR must comply with the conditions of the signed agreement.

Information about the status of each variable in the MFR is given in section 2. The statistical disclosure limitation methodology is briefly illustrated in Section 3. In section 4 the impact of the statistical disclosure limitation methodology is evaluated by means of some comparisons with the original data.

2. Description of variables

This section summarizes the changes of the variables with respect to the original microdata file. Such changes are necessary in order to protect the confidentiality of respondents.

The labels and names of variables follow the Commission Regulation 2139/2004/EC.

The decimal separator is ".".
In the tab delimited MFR, the first row contains the names of the variables.
The missing character is ":".
The MFR contains information on agricultural holdings.
The MFR contains 47780 records and 319 variables. The sum of the weights (variable A09) equals 1728532.

The following variables were suppressed: A04A, A04D, A05, A08, A09A.

Several numerical variables were suppressed, but the following relationships were maintained:

Suppressed Variables	**Released Variables**	**Relationship**
D09E1, D09F, D09G	D09	D09 = D09E1 + D09F + D09G
D18A, D18B, D18B1, D18B23	D18	D18 = D18A + D18B + D18B1 + D18B23
F01, F02	F	F = F01 + F02
G01A, G01B, G01C	G01	G01 = G01A + G01B + G01C
G03A, G03B	G03	G03 = G03A + G03B
G04C, G04D	G04A, G04B, G04	G04 = G04A + G04B + G04C + G04D
I08AD22, I08B, I08C, I08D, I08E	I08	I08 = I08AD22 + I08B + I08C + I08D + I08E
J09A, J09B	J09	J09 = J09A + J09B
J10A, J10B	J10	J10 = J10A + J10B
J16A, J16B, J16C, J16D	J16	J16 = J16A + J16B + J16C+J16D
D01, D02, D03, D04, D05, D06, D08	**D01**	**D01** = D01 + D02 + D03+ D04 + D05 + D06 + D08
D10, D11, D12	**D10**	**D10** = D10 + D11 + D12
D23, D24, D25	**D23**	**D23** = D23 + D24 + D25
D26, D27, D28, D29, D30	**D26**	**D26** = D26 + D27 + D28 + D29 + D30
D31, D32, D33	**D31**	**D31** = D31 + D32 + D33

Table 1. Suppressed numerical variables and maintained relationships. In bold the variables that were created as aggregations of observed variables.

The numerical variables were modified, except for A06, A09, A12, A13, A14, A15, A16, A17, A18 and the variables whose names start with letters "K" and "L". The variable A06 was recoded.

3. Statistical disclosure control methodology

3.1 Disclosure scenarios

With respect to the collected information, two disclosure scenarios were deemed realistic: one based on structural variables and the second one based on non-structural variables. Also the special kind of agreement between Istat and researcher was taken into consideration.

3.2. Changes on variables

3.2.1 Variable suppression

The following types of variables were suppressed from the MFR:

1. The direct identifiers (name, address, etc): e.g. A08, A09A;
2. The variables considered too detailed to be released: A04A, A04D;
3. A variable that generated too low sample and population frequencies: A05;
4. Several numerical variables that could allow the derivation of the farming type at a too detailed level: see first column of the table 1.

In order to keep the microdata file structure, each value of each suppressed alphanumeric variable was replaced by ":".

3.2.2 Global recoding

For the Italian FSS 2005 MFR, the variable A06 was recoded in order to give information only on the *particular types* (50 categories) of farms.

The information on the geographical location of the farms is given only at regional level (NUTS2), variable A07.

3.3. Perturbation of numerical variables

Protection of agricultural holdings at risk of disclosure was achieved by means of variable suppression, variable aggregation and individual ranking, see Defays *et. al*. The suppressed variables are indicated in the first column of Table 1. The aggregated variables are indicated in bold in the second column of Table 1. A particular version of microaggregation was applied to each numerical variable except for A06, A09, A12, A13, A14, A15, A16, A17, A18 and the variables whose names start with letters "K" and "L". The microaggregation was applied independently on each numerical variable (individual ranking). The individual ranking was applied using $k = 3$ as microaggregation parameter. The individual ranking was applied using A07 as blocking variable. A weighted version of individual ranking was applied, using the variable A09 (extrapolation factor) as weighting variable. A further control was implemented in order to take into account the sparsity of the phenomenon under study.

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).

Obviously, the information content of the non-modified variables was not modified. This means that the information content of about 64% of the variables remained unchanged.

Weighted Means Variations

For each perturbed numerical variable, since A07 was used as blocking variable, the weighted means were preserved at regional level (NUTS2 level).

The quartiles of the weighted means variations were computed. Namely, for each numerical variable and for each combination of A07 and *general types* of farming (A06), the weighted means were computed both for the original and perturbed data. These quantities were denoted WM_{orig} and WM_{pert} respectively. Then, the percentage variation was computed as $v = \frac{WM_{orig} - WM_{pert}}{WM_{orig}} * 100$. Finally, in Table 2, for each numerical variable, the quartiles of v over the 186 combinations of A07 and *general types* of farming (A06) are shown. The skew distributions are a consequence of the sparsity of the represented phenomenon. Several variables were deleted from Table 2 because of their extremely high rate of zero values.

Variable	Q1	Q2	Q3	Variable	Q1	Q2	Q3
A11	-0.07	0.00	0.08	*G02*	-1.48	0.04	0.84
A13	0.00	0.00	0.00	*G03*	-0.20	-0.02	0.18
A14	0.00	0.00	0.00	*G04*	-0.26	0.00	0.21
A15	0.00	0.00	0.00	*G04A*	-0.65	0.03	0.87
A16	0.00	0.00	0.00	*G04B*	-0.45	0.00	0.35
A17	0.00	0.00	0.00	*G05*	-5.59	0.01	4.42
A18	0.00	0.00	0.00	*G06*	-18.26	1.19	13.89
CC01	-0.12	-0.01	0.07	*G07*	-6.62	-0.01	0.15
CC02	-0.28	0.00	0.25	*H01*	-0.79	0.00	0.66
CC03	-1.87	0.01	1.55	*H02*	-0.41	0.00	0.32
CC05A	-1.79	0.07	1.28	*H03*	-0.17	0.00	0.15
CC05D	-14.91	-0.27	15.31	*I01*	-1.47	0.00	1.31
D01	-0.10	-0.01	0.08	*I03A*	-0.24	-0.01	0.16
D07	-0.92	-0.01	1.02	*I03B*	-0.27	0.00	0.20
D09	-2.35	-0.17	1.88	*I08*	-0.63	0.01	0.55
D10	-1.40	0.00	1.36	*J01*	-0.71	0.00	0.57
D14	-0.85	0.00	0.90	*J02*	-0.19	0.00	0.10
D14A	-1.04	-0.12	0.35	*J03*	-0.76	0.00	1.19
D14B	-6.95	-0.03	7.04	*J04*	-0.45	0.00	0.31
D15	-3.35	-0.06	2.09	*J05*	-0.48	0.00	1.10
D16	-1.28	0.00	7.96	*J06*	-0.75	0.00	0.71
D17	-13.50	-0.04	3.87	*J07*	-0.85	0.00	0.26
D18	-0.30	-0.01	0.14	*J08*	-0.49	0.01	1.06
D19	-10.19	-0.23	11.50	*J09*	-0.50	-0.01	0.49

D21	-1.02	-0.01	1.24	*J10*	-0.63	0.00	0.51
D23	-7.31	-0.12	0.32	*J11*	-8.94	-0.06	5.56
D26	-1.09	-0.02	0.84	*J12*	-1.85	0.00	3.24
D31	-17.70	0.00	53.40	*J13*	-1.41	0.00	0.39
D34	-78.06	0.08	10.80	*J14*	-6.35	0.00	1.17
D35	-52.00	0.00	2.28	*J15*	-4.84	0.00	1.55
E	-0.09	0.00	0.04	*J16*	-24.69	-1.58	1.69
F	-0.32	0.00	0.20	*J17*	-11.30	0.00	3.25
G01	-0.75	-0.04	0.43	*J18*	-11.94	-0.36	4.32

Table 2. The quartiles of the weighted means variations over the combinations of A07 and A06. Q1 = first quartile, Q2 = second quartile (median) and Q3 = third quartile.

Variances Variations

The quartiles of the variances variations were computed. For each numerical variable and for each combination of A07 and *general types* of farming (A06), the variances were computed both for the original and perturbed data. These quantities were denoted V_{orig} and V_{pert} respectively. Then, the percentage variation was computed as $v = \frac{V_{orig} - V_{pert}}{V_{orig}} * 100$.

Finally, in table 3, for each numerical variable, the quartiles of v over the 186 combinations of A07 and *general types* of farming (A06) are shown. The skew distributions are a consequence of the sparsity of the represented phenomenon. Several variables were deleted from Table 3 because of their extremely high rate of zero values.

Variable	**Q1**	**Q2**	**Q3**	**Variable**	**Q1**	**Q2**	**Q3**
A11	-1.11	0.00	1.67	*G02*	-5.63	0.00	3.26
A13	0.00	0.00	0.00	*G03*	-4.84	0.00	1.99
A14	0.00	0.00	0.00	*G04*	-1.98	0.00	2.78
A15	0.00	0.00	0.00	*G04A*	-2.53	0.10	6.78
A16	0.00	0.00	0.00	*G04B*	-5.55	0.00	3.47
A17	0.00	0.00	0.00	*G05*	-12.95	1.80	19.87
A18	0.00	0.00	0.00	*G06*	-57.24	5.18	35.90
CC01	-1.06	0.07	1.92	*G07*	-77.76	-7.10	46.46
CC02	-1.69	0.13	2.75	*H01*	-4.03	0.16	7.39
CC03	-9.11	0.18	8.91	*H02*	-3.04	0.25	5.12
CC05A	-6.40	0.38	6.12	*H03*	-4.39	0.00	2.93
CC05D	-45.09	2.95	52.85	*I01*	-7.30	0.32	9.54
D01	-1.78	-0.15	1.01	*I03A*	-1.35	0.02	2.41
D07	-1.54	0.51	10.04	*I03B*	-1.79	0.01	2.30
D09	-7.86	-0.36	12.23	*I08*	-4.42	0.27	4.06
D10	-3.90	0.70	6.83	*J01*	-9.04	0.00	8.95
D14	-4.37	0.09	6.89	*J02*	-0.81	0.00	2.28
D14A	-4.52	0.00	5.38	*J03*	-4.64	0.00	9.97
D14B	-10.34	0.49	23.38	*J04*	-2.43	0.00	3.25
D15	-9.86	0.00	13.25	*J05*	-3.99	0.00	13.05
D16	-11.11	1.12	36.45	*J06*	-2.43	0.00	5.76
D17	-28.82	2.29	18.17	*J07*	-1.35	0.00	4.07
D18	-4.18	-0.02	1.96	*J08*	-2.59	0.47	8.89

D19	-36.15	-0.24	33.86	*J09*	-2.67	0.00	3.64
D21	-5.96	0.25	8.11	*J10*	-2.98	0.00	7.56
D23	-18.03	-1.50	43.98	*J11*	-29.96	0.10	23.69
D26	-2.26	0.00	3.12	*J12*	-9.62	0.00	18.49
D31	-171.16	70.61	90.59	*J13*	-14.95	0.00	3.26
D34	-98.81	7.66	57.41	*J14*	-25.72	0.00	12.21
D35	-152.35	-109.14	84.67	*J15*	-12.48	-0.13	8.91
E	-0.38	0.00	0.39	*J16*	-95.04	-3.57	8.76
F	-1.73	0.32	2.88	*J17*	-42.57	0.00	22.31
G01	-5.53	-0.02	3.40	*J18*	-45.92	-4.78	25.45

Table 3. The quartiles of the variances variations over the combinations of A07 and A06. Q1 = first quartile, Q2 = second quartile (median) and Q3 = third quartile.

Modified Zero Values

For the Farm Structure Survey 2005, the zero values represent a missing phenomenon. By applying the individual ranking, some zero values might be transformed in non-zero values. In this way, an artificial phenomenon could be created. In Table 4, for each numerical variable, the number N of zero values that were transformed in non-zero values is shown. It should be observed that the greatest values of N correspond to variables with the largest variations of the weighted means/variances.

Variabile	**N**	**Variabile**	**N**
A11	0	*G02*	2
A13	0	*G03*	3
A14	0	*G04*	0
A15	0	*G04A*	0
A16	0	*G04B*	0
A17	0	*G05*	3
A18	0	*G06*	0
CC01	0	*G07*	13
CC02	0	*H01*	0
CC03	0	*H02*	0
CC05A	0	*H03*	0
CC05D	7	*I01*	0
D01	0	*I03A*	0
D07	2	*I03B*	0
D09	2	*I08*	4
D10	0	*J01*	0
D14	0	*J02*	0
D14A	0	*J03*	0
D14B	0	*J04*	0
D15	1	*J05*	0
D16	7	*J06*	0
D17	5	*J07*	0
D18	0	*J08*	0
D19	6	*J09*	0
D21	1	*J10*	0
D23	7	*J11*	2
D26	4	*J12*	2

D31	6	*J13*	0
D34	11	*J14*	3
D35	16	*J15*	0
E	0	*J16*	2
F	0	*J17*	0
G01	0	*J18*	6

Table 4. Number N of modified zero values.

Individual Perturbations

The impact of the perturbation method on the information loss might be evaluated using the individual perturbations, too. In Table 5, for each numerical variable, the median of the individual perturbations is presented.

Variable	Median	Variable	Median
A11	-0.001	*G04A*	-0.002
CC01	-0.001	*G04B*	-0.001
CC02	-0.001	*G05*	0.002
CC03	-0.001	*G06*	0.002
CC05A	0.001	*G07*	-0.028
CC05D	-0.030	*H01*	-0.001
D01	-0.001	*H02*	-0.001
D07	0.017	*H03*	-0.001
D09	0.001	*I01*	0.001
D10	0.001	*I03A*	-0.001
D14	-0.001	*I03B*	-0.001
D14A	-0.001	*I08*	0.001
D14B	-0.001	*J01*	-1.000
D15	-0.001	*J02*	-1.000
D16	-0.002	*J03*	-1.000
D17	0.001	*J04*	-1.000
D18	0.001	*J05*	-1.000
D19	-0.002	*J06*	-1.000
D21	-0.001	*J07*	-1.000
D23	-0.001	*J08*	1.000
D26	-0.001	*J09*	1.000
D31	-0.075	*J10*	-1.000
D34	-0.013	*J11*	1.000
D35	-0.303	*J12*	1.000
E	-0.001	*J13*	-1.000
F	0.001	*J14*	-1.000
G01	-0.001	*J15*	-1.000
G02	-0.001	*J16*	-1.000
G03	-0.001	*J17*	-1.000
G04	-0.001	*J18*	-1.000

Table 5. Median of the individual perturbations.

5. References

D. Ichim (2009), "*Italian Farm Structure Survey: SDC Methodology for an MFR Dissemination*", Documenti Istat (2009), to appear.

D. Defays and M.N. Anwar (1998) , "*Masking Microdata Using Micro-Aggregation*", Journal of Official Statistics, Vol.14, No.4, 1998. pp. 449-461.

6. Contacts

For further details on the statistical disclosure control methodology, please contact rilascio.microdati@istat.it.