DOCUMENTI ISTAT

n. 2/2007

Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment

D. Ichim

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

DOCUMENTI ISTAT

n. 2/2007

Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment

D. Ichim(*)

(*) ISTAT - Servizio Progettazione e supporto metodologico nei processi di produzione statistica

Contributi e Documenti Istat 2007

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

Abstract

This work presents a procedure for the anonymisation of a microdata file to be released solely for research purposes. The proposal is based on a detailed analysis of possible disclosure scenarios and a rigorous individual risk assessment. The latter is based on a clustering algorithm, in particular a density based aggregation method. Two protection methods are applied to the units considered at risk of identification: a nearest clustered imputation and microaggregation on the tails. Both information loss and data utility are evaluated from a researcher point of view.

1. Introduction

This document describes a proposal for the anonymisation of the microdata stemming from the Community Innovation Survey (CIS), see [1] and [2]. Results of its application to the Italian CIS3 microdata are reported. There already exist several proposals for the anonymisation of business microdata files, mainly based on the application of individual ranking, see [3], for example. In [3] the microaggregation is applied without taking into account any output of the risk assessment phase (disclosure scenario and risk assessment). Moreover, the microaggregation would be applied to the whole file, irrespective to peculiarities of the observed phenomenon and without taking into account possible relationships between variables (for example turnover needs to be greater than export or expenditures). This strategy in the application of the method would result in a series of considerable drawbacks. Firstly, in an economical system, e.g. the Italian one, where the number of large size enterprises is reduced and commonly one or two extremely large size companies dominate a NACE division, an individual ranking that does not consider stratification by NACE would result in insufficient protection for the most identifiable enterprises. Secondly, if the method is not applied per single NACE division and per size classes all the aggregates produced and published will not be maintained with a considerable loss for the users that will not be able to compare results. Moreover, individual ranking would result in a change of the relationships between variables and may lead to incoherent values inside records (i.e. single enterprises). In this document, a different, more flexible, strategy is discussed.

The presented methodology takes into account both economic features of the data and the National Statistical Institute dissemination policy. A key point is the fact that the final protected data set would be released for research purposes, and hence subject to a signed contract. Knowing the class of data users, a rigorous study of possible disclosure scenarios is carried out in order to define identifying variables, including a spontaneous identification scenario. Then a careful risk assessment analysis is performed to single out the records at risk. The risk assessment is performed considering both the economic classification and size classes, when these are considered identifying variables. The basic ingredient of this framework is that a value of an identifying variable (or indeed the values from a set of identifying variables) is considered at risk if it is isolated i.e. the “density” of the points close to this value is not deemed sufficient. The "density" concept is defined using both the distance between points and number of points in a neighbourhood. The distance used may be easily extended to a multivariate case, considering both continuous and categorical

variables. With respect to some a-priori defined thresholds, the number of isolated (hence at risk of identification) points would tend to increase or decrease. These thresholds may be defined according to the observed phenomenon, assumed disclosure scenario and national dissemination policy, proving the great flexibility of the method. Justified by the research purposes of the microdata file release, **only** the records at risk of identification are perturbed (i.e. modified) whereas the rest of the file is released unchanged. Protection is mainly achieved by a nearest clustered imputation method and a particular microaggregation, to which this method reduces for an extreme thresholds choice in the identification phase. Modifying only the key variables and only for the records at risk of identification, a lot of variables (including the calibrated weights) would remain unchanged, hence coherence with many already published aggregate statistics would be easily achieved.

The microdata anonymisation proposal is based on the following six steps:

1. Definition of the disclosure scenario.
2. Preliminary work on variables.
3. Risk assessment: identification of units at risk.
4. Microdata protection.
5. Information loss assessment.
6. Description of the microdata file to be released.

The rest of the document gives a detailed description of these six steps. Section 2 briefly summarizes the CIS3 information. In section 3, hypotheses for different disclosure scenarios are discussed. In section 4, some preliminary basic work (variable suppression, coding, etc.) on variables is described. Based on the disclosure scenario previously determined, in section 5 the isolated units are defined and a method for their automatic identification is introduced. The microdata perturbation method presented in section 6 is applied only to the isolated units. In order to preserve some statistical characteristics (e.g. totals), another adjustment is performed. Finally, in section 7 several measures for information loss and data utility assessment are proposed.

Throughout the document, the results obtained by applying this anonymisation procedure to the Italian CIS3 data are illustrated.

2. Community Innovation Survey: brief description of the data

CIS provides information on the characteristics of innovation activity at enterprise level. The surveyed variables are completely listed in [2]. Some of the main observed variables are:

1. principal economic activity
2. geographical information
3. number of employees in 1998 and 2000;
4. turnover in 1998 and 2000;
5. exports in 1998 and 2000;
6. gross investment in tangible goods: 2000;
7. number of valid patents at end of 2000;
8. number of employees with higher education (in line with the number of employees in 2000);
9. expenditure in intramural RD (in line with the turnover in 2000);
10. expenditure in extramural RD (in line with the turnover in 2000);
11. expenditure in acquisition of machinery (in line with the turnover in 2000);
12. expenditure in other external knowledge (in line with the turnover in 2000);

13. expenditure in training, market (in line with the turnover in 2000);
14. total innovation expenditure (in line with the turnover in 2000);
15. number of persons involved in intra RD (in line with the number of employees in 2000).

The CIS statistical population is determined by the size of the enterprise and its principal economic activity.

Principal economic activity

The following industries were included in the population:

- mining and quarrying (NACE 10 - 14)
- manufacturing (NACE 15 - 37)
- electricity, gas and water supply (NACE 40 - 41)
- wholesale trade (NACE 51)
- transport, storage and communication (NACE 60 - 64)
- financial intermediation (NACE 65-67)
- computer and related activities (NACE 72)
- research and development (NACE 73)
- architectural and engineering activities (NACE 74.2)
- technical testing and analysis (NACE 74.3)

Size of the enterprise

At least all enterprises with 10 or more employees in any of the specified economical classifications were included in the statistical population.

The sampling frame was in general the best available business register containing basic information such as names, addresses, NACE-division, size and region of all enterprises in the target population. Innovation data was collected both through census or sample survey. In most countries a sample survey was conducted, hence the weights (direct or calibrated) were also recorded. In Italy, a sampling survey was conducted, resulting in about 13000 respondent enterprises in the sample.

3. Disclosure scenarios

As the microdata file to be released is intended for research purposes, a "nosy colleague" scenario is not deemed realistic. Instead, other two scenarios, "external register" and "spontaneous identification" are possible and discussed in this section.

3.1 External register

Generally, business registers are publicly available. A possible intruder could use such information to identify an enterprise. Moreover, it is known that a business register was used as a sampling frame. Consequently, an intruder *a-priori* knows that an enterprise possibly belonging to the sample, it is surely included in the business register. Public business registers report general information on name, address, turnover (*TURN*), number of employees (*EMP*), principal activity of an enterprise (*NACE*), region (*NUTS*). Among these variables, the only one that could be released in its original form is *TURN*. The others are either removed or modified.

Some other information on enterprises is recorded in the microdata file, see [2], and public registers, e.g. number of employees and turnover in the first year of the reference period of the survey. Being obsolete information, this disclosure scenario assumes that no available public business register contains such historical information. Hence, the above mentioned variables are not considered as identifying variables.

3.2 Spontaneous identification

In the CIS3 dataset there are several confidential variables that may be subject to spontaneous identification. Some examples are total expenditure on innovation (*RTOT*), exports, number of persons involved in intra RD, etc. Such variables are never published in an external register, but they can assume extremely particular values on some units. Mere additional information would then clearly identify the enterprise. Special attention must be paid on these variables. A check performed by the **survey experts** is generally suggested. These validations are performed with respect to each combination of categorical key variables to be released.

Moreover, it must be also considered that the anonymised microdata file to be produced is released for research purposes, hence subject to the constraints of a signed contract.

3.3 Scenarios key variables

The key variables of the hypothesized disclosure scenarios are: economic classification (*NACE*), region (*NUTS*), number of employees (*EMP*) and turnover (*TURN*). The information content of these variables must be somehow reduced in order to increase the uncertainty of a possible intruder.

4. Preliminary work on variables

4.1 Variable suppression

Some variables are directly removed from the microdata file.

1. Direct identifiers are removed from the microdata file to be released. Some of these are:
    a. Name
    b. Address
2. Variables subject to spontaneous identification are generally suppressed. Some examples
    a. Country of head office (*Ho*)
    b. Export (at least for the Italian CIS3 microdata this is an identifying variable)
3. Other variables
    c. Turnover in the first year of the reference period
    d. Number of employees in the first year of the reference period

4.2 Global recoding

Some variables are aggregated according to several dissemination requirements. For example, Eurostat may ask for a minimum level of detail or the structure of the surveyed economy may constrain the release. Generally information on NACE at 2 digits and three enterprise size classes are the minimum requirements for the anonymised CIS microdata files to be released. Usually these minimum requirements are fulfilled, but national characteristics of the data must be also considered. The dissemination policy of the National Statistical Institute is a natural constraint. For example, some NACE divisions may never be released by their own, but always aggregated with others. Such a-priori aggregations generally depend on the economic structure of the country.

It is not a sampling problem, but rather a feature of the surveyed phenomenon. For example, when such phenomenon is not well represented, NACE divisions might be aggregated (with respect to the NACE hierarchy). Generally such aggregations are decided by survey experts. The following aggregations were used for the anonymisation of the Italian CIS3 microdata file:

1. *Principal economic activity* (*NACE*) is aggregated with respect to Eurostat requirements (item a) while the other items are performed with respect to the Italian dissemination policy and economic structure. A new variable *NACE2* is then obtained:
    a. NACE is aggregated in NACE at 2 digits
    b. NACE 10, 11, 13 and 14 are aggregated into a single class called NACE 10
    c. NACE 15 and 16 are aggregated into a single class called NACE 15
    d. NACE 40 and 41 are aggregated into a single class called NACE 40

2. *Number of employees* (*EMP*) is aggregated in five classes[1]. A new variable *EMPclass* is then created, with the following categories :
    a. small size: 10 – 49 employees, category "1"
    b. medium size: 50 – 249 employees, category "2"
    c. large size: more than 250 employees, category "3"
    d. only for NACE 20, 23, 30, 67 and 73, class 2 and class 3 are aggregated into a new one, category "2_3"
    e. only for NACE 37, 62 and 64 all number of employees classes are aggregated into a single one, category "1_2_3"

3. *NUTS* categories are aggregated into a single one (national level).

Once the *Number of employees* is recoded it loses most of its identifying properties; therefore the variable *EMPclass* is not considered an identifying variable.

4.3 Preliminary rounding

Considering that the anonymised CIS3 microdata file is intended for research purposes, the present method attempts to avoid overprotection: if a unit may be confused (see below) with others, its *TURN* value is released as it is. Since *TURN* is a continuous economic variable, it is a highly identifying one: it assumes almost unique values on each unit. If it is deemed that the release and reference survey dates are too close to each other, the first step would be the application of a dedicated controlled random rounding[2], see [4, 5], to the variable *TURN*. Otherwise, this step may be skipped. Rounding is a perturbation method, but since identification of units at risk should be based on released values, it is more natural to discuss it at this stage of the procedure. Otherwise, the identification of units at risk would be based on not rounded *TURN*, while the rounded *TURN* would be released. This is the reason for which (possibly) rounded *TURN* values must be considered as an input to the subsequent identification phase.

The reference period of CIS3 is 2000. The microdata file for research would be probably released in 2007. Hence, an initial rounding of the variable *TURN* in this particular situation was deemed unnecessary. For future surveys, anyway, initial rounding would be surely a procedure to be routinely implemented at least for the continuous key variables to be released. Rounding is especially recommended when values on some units are imputed from an external register.

[1] These aggregations are performed following the Eurostat CIS3 anonymisation proposal.
[2] Preserving also the means.

5. Identification of units at risk

A unit is considered at risk if it is “recognisable” either in the external register scenario or in the spontaneous identification scenario. The former will be addressed in section 5.1 and the latter in section 5.2.

5.1 External register

As the principal economic activity and number of employees variables are recoded as described in the paragraph 4.2, the only remaining key variable recorded in an external register is turnover. The method proposed in this document is flexible enough to incorporate other publicly available variables. For example, for future surveys, export might be released and hence considered in the identification phase. Moreover, the method is not limited to continuous variables, but (numeric) categorical ones can easily be taken into account.

5.1.1 Assumptions on identification of units at risk

The proposed method models an intruder uncertainty. Its main assumption is that a unit cannot be identified by an intruder when it is confused with other units. Both identification and/or confusion must be quantified with respect to the key variables to be released. In the external register disclosure scenario, for each combination of *NACE2* and *EMPclass*, the turnover *TURN* is the unique remaining key variable. Being the turnover an economic continuous variable, it assumes almost unique values on each unit, even if previously rounded. Hence, the identification of units at risk cannot be based on sample/population uniques approaches. Consequently, the degree of confusion of a given unit must be measured with respect to the distance to the other units. Here it is supposed that an intruder cannot distinguish between *TURN* values that are too close, with respect to a given distance. It is even worse (from an intruder point of view), when the "candidate" turnover values are all equal.

It is assumed that an intruder may confuse a unit U with others when there is a sufficient number of units in a well-defined (and not too large) neighbourhood of U. The units that cannot be confused with others are isolated units. Figure 1 shows examples of confused and isolated units.

5.1.2 Clustering algorithms

A cluster is a group of homogeneous units, with respect to some a-priori defined criteria. Based on clustering principles, units belonging to the same cluster are considered similar, indistinguishable. Units belonging to different clusters are considered different. Similarity and confusion both express the same concept, although in different frameworks. That is, units belonging to the same cluster are considered not at risk of identification because they may be confused with the other units belonging to the same cluster. Instead, it is assumed that an intruder might distinguish between different clusters. Moreover, the outlying units not belonging to any cluster are considered at risk of identification since they cannot be confused with any other unit. From the statistical disclosure control point of view, to guarantee unit confidentiality, for a unit U to be considered not at risk of identification, there must be a sufficient number of units that may be confused with U.

Clustering algorithms are generally implemented in statistical software.

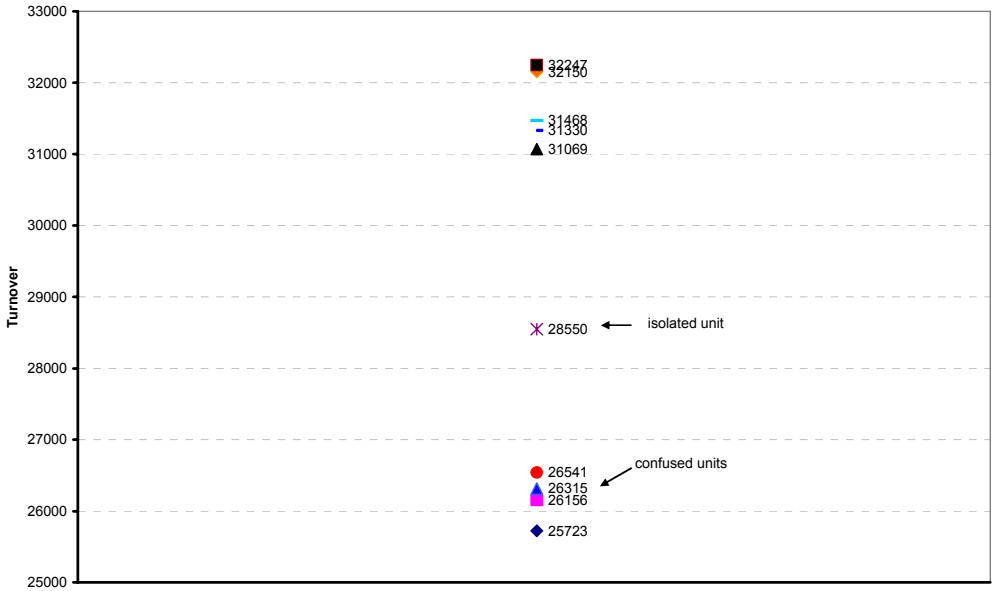


Figure 1. Example of isolated and confused units with respect to the variable turnover.

5.1.3 Density-based clustering algorithm

In this proposal, the identification of units at risk of identification is based on a density notion of clusters. Clusters may be viewed as subsets of data with high density. Furthermore, the density within the areas outside clusters is lower than the density inside any of the clusters. This is the main hypothesis of the proposal: when a unit belongs to a cluster, it belongs to a high density (sufficient number of close units) subset of data. Hence the unit may be considered as being confused with others. The clustering algorithms taking into account these two features (distance from each other and number of neighbours) are called density based algorithms. The algorithm used in this application, named DBSCAN, is completely described in [6]. The key idea is that for each unit of a cluster the neighbourhood of a given radius has to contain at least a minimum number *MinPts* of units, i.e. the density in the neighbourhood has to exceed some threshold. The algorithm is based on the *Eps*-neighbourhood of a unit U, i.e., the set of units for which the distance from U is less than *Eps*.

DBSCAN requires the initialization of only two parameters (thresholds): *Eps* and *MinPts*.

The DBSCAN clustering steps are briefly summarized below:

DB1. Start with a unit U. A cluster C_U containing U is initialized.

DB2. If the *Eps*-neighbourhood E_U of U contains at least *MinPts* units, the entire E_U is included in the cluster C_U. This step is repeated for each unit in E_U, increasing, if necessary, each time the dimension of the original cluster C_U.

DB3. if the *Eps*-neighbourhood E_U of U does not contain a sufficient number of units (< *MinPts*), the next unit of the data set is visited.

DB4. Clusters may be merged if the distance to each other is less than *Eps*. The distance between two clusters C_1 and C_2 is the smallest distance between two units A and B, $A \in C_1$, $B \in C_2$.

DB5. The units not belonging to any cluster are isolated units.

DBSCAN algorithm is implemented in R and available from www.r-project.org.

5.1.4 Parameter settings for DBSCAN

As stated by the survey experts, for each *NACE2* category, variable *EMPclass* is not an identifying variable for the Italian CIS3 dataset. As a consequence, the identification of isolated units by means of the DBSCAN algorithm was performed, for the Italian CIS3 dataset, only with respect to *NACE2*, without regard to *EMPclass*. For other economic structures/countries/surveys, consideration of *EMPclass* might be necessary. It must be stressed that the clustering algorithm has to be applied for each combination of categorical key variables.

Before applying the algorithm, a transformation might be applied on key variables. As *TURN* generally exhibits a very skewed distribution, a logarithmic transformation was used for the Italian CIS3 microdata file.

For continuous variables, the Euclidean distance function is suitable. It can also be easily extended to the multivariate case, when other key variables (e.g. export) could be released. If a variable transformation on *TURN* is not used, other distance functions might be thought. Usage of different distance functions for different combinations of categorical key variables is not suggested.

The choice of the number of units in a neighbourhood of a unit U, the parameter *MinPts*, depends both on the dissemination policy and on the accuracy of the public register considered in the disclosure scenario. A too low value would imply too many clusters with a sufficient number of units. Hence, the identification risk would not be really evaluated because most of the units would result as being confused with others. On the contrary, a too high value would be overprotective since too many units would result being isolated. A trade-off value should be found. Following various simulations, *MinPts* = 5 was deemed a valid value for Italian CIS3 dataset. Usage of the same value for all combinations of categorical key variables (only *NACE2* in the present case) is a coherent choice. When the economic structure of a country is very different from combination to combination, different values of *MinPts* might be anyway chosen.

The *Eps* value is determined with respect to the distances from the *MinPts*-th unit, for each combination of key variable. That is, the distances between units are computed. For each unit U, its corresponding distances are sorted in ascending order. Then, the *MinPts*-th element in this list is recorded. The vector of distances from the *MinPts*-th neighbour is considered and its third quartile Q3 is calculated. This third quartile is the threshold value *Eps*. This procedure for the threshold definition was used for the Italian CIS 3 data. Other criteria for the determination of the threshold *Eps* are possible, but again, for the sake of coherence, the same principles should be followed for each combination of categorical key variables.

Summarizing, the identification of units at risk, for each combination of key variables, is the following:

a) if necessary, transform the data
b) choose the distance function
c) define the value *MinPts* and compute the threshold *Eps*
d) apply the density-based clustering algorithm DBSCAN with parameters *MinPts* and *Eps*. (apply steps **DB1** to **DB4** and iterate to merge clusters)
e) list the isolated units

In case of Italian CIS3 microdata, *TURN* has a very skewed distribution, for each combination of categorical key variables. When applied to CIS3 data, DBSCAN identified isolated units on both tails of the distribution. Moreover, DBSCAN detects also the isolated units on the central part of the distribution. In the next table the percentages of isolated points identified when applying the

DBSCAN to Italian CIS3 data are shown. The percentages are computed with respect to the total number of observations in each category.

		Percentage of isolated points					Percentage of isolated points		
Nace2	Number of Observations	Left tail	Right tail	Total	*Nace2*	Number of Observations	Left tail	Right tail	Total
17	489	0.61	7.98	13.91	*33*	323	2.79	11.46	16.72
15	628	2.39	4.14	14.81	*34*	289	2.08	9.00	17.99
10	232	5.17	12.50	18.53	*35*	238	2.94	11.76	15.13
18	631	1.58	5.71	15.85	*36*	523	1.34	9.94	17.59
19	373	4.02	5.36	15.28	*37*	101	3.96	9.90	14.85
20	422	3.55	4.74	16.35	*40*	212	6.13	9.91	19.81
21	353	6.80	9.92	19.55	*51*	655	6.26	6.72	18.78
22	420	3.33	9.76	18.10	*60*	605	4.13	6.45	17.02
23	94	3.19	9.57	17.02	*61*	81	6.17	1.23	12.35
24	523	4.97	7.27	14.91	*62*	26	3.85	11.54	23.08
25	498	0.20	3.01	14.26	*63*	546	5.13	9.71	19.41
26	573	1.22	6.63	15.53	*64*	62	0.00	16.13	16.13
27	347	3.46	5.76	18.16	*65*	516	3.29	7.17	18.80
28	712	1.69	4.92	16.85	*66*	101	1.98	0.99	8.91
29	697	2.15	1.87	14.20	*67*	153	9.15	13.07	22.22
30	85	0.00	8.24	10.59	*72*	411	1.70	9.98	18.00
31	466	5.58	6.87	16.95	*73*	61	3.28	13.11	16.39
32	250	2.40	10.40	16.80	*74*	268	4.10	14.93	20.90

5.2 Spontaneous identification

The spontaneous identification scenario is based on personal or highly specialized knowledge. Therefore, the identification risk may be assessed only based on expert's opinion, simulating the behaviour of very specialized intruder. With respect to this scenario, a check performed by (survey) experts is required.

The first step consists in enumerating the confidential variables subject to spontaneous identification risk. Only variables to be released must be taken into account. All combinations of categorical key variables must be considered. For the CIS3 survey the confidential variables that have to be checked as they may lead to spontaneous identification of a unit are *RTOT* and its various components (*RRINDX, RREXX*, etc) within *NACE2* categories.

In a second phase, for each combination of key variables, (survey) experts check the values the confidential variables assume on units. A list of units at risk is produced at the end of this anonymisation phase. These particular units will be further subject to a dedicated protection method.

6. Microdata perturbation

Microdata perturbation is achieved in several steps described in this section. Two kinds of methods are used. Firstly, methods applied to variables are considered, as also discussed in the "preliminary work" section. Secondly, two perturbation methods dedicated to the isolated units listed in the identification phase are presented. An adjustment is proposed in order to preserve turnover totals for each combination of the categorical key variables to be released.

6.1 Global recoding

The global recoding applied (paragraph 4.2) to variables *Principal economic activity* and *Number of employees*, transforming them into *NACE2* and *EMPclass* respectively, is a protection method since it reduces the information content of the variables, hence increasing an intruder uncertainty.

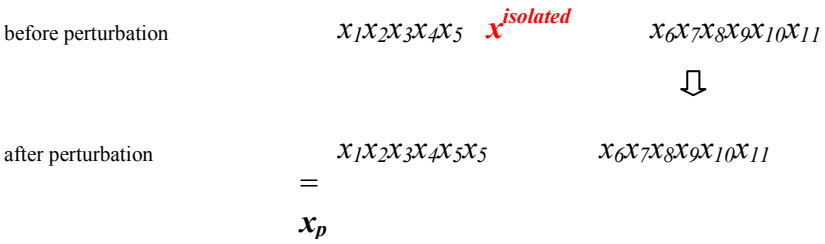
6.2 Rounding

The rounding (possibly) applied to *TURN* if deemed necessary (paragraph 4.3) is a protection applied to turnover. The protective power of this type of rounding strongly depends on the used rounding base.

6.3 Nearest clustered unit imputation

To avoid overprotection, **only** the isolated units identified by the clustering algorithm are modified. The perturbation of these units is performed with respect to each combination of key variables to which DBSCAN was applied. That is, perturbation is applied with respect to each *NACE2* category.

Isolated *TURN* values (units at risk) are to be perturbed in order to avoid identification. The perturbation proposed is a nearest clustered unit imputation: an isolated *TURN* value is replaced by the *TURN* value assumed on the closest[3] clustered unit. The procedure is presented below. "x_i" stands for the *i*-th clustered unit, "$\boldsymbol{x}^{isolated}$" for an isolated unit while "$\boldsymbol{x_p}$" represents the perturbed value of "$\boldsymbol{x}^{isolated}$".



The nearest clustered unit imputation is formulated below:

1. let $x^{isolated}$ be the value to be perturbed
2. find the closest clustered unit x_p:

x_p *s.t.* $d\left(x^{isolated}, x_p\right) = \min_{x_c \in \mathrm{C}}\left(d\left(x^{isolated}, x_c\right)\right)$ where C is the set of all clustered units

3. let the protected *TURN* value be $x_*^{isolated} = x_p$

The nearest clustered unit imputation may be easily extended to the multivariate case because the closest clustered unit is found with respect to a distance function that can be defined for the multivariate case.

[3] with respect to the same distance function used in the clustering algorithm

6.4 Microaggregation on tails

Since *TURN* variable generally has a very skewed distribution, protection of the smallest and largest isolated *TURN* values by nearest clustered unit imputation may be very expensive in terms of information. This is the reason for which on the left and right tails of *TURN* a microaggregation, see [3], is proposed for perturbing the isolated units. Here microaggregation is equivalent to individual ranking because there is only one variable (*TURN*). If necessary, microaggregation is suggested for multivariate variables. Microaggregation is performed with respect to the same combinations where the clustering algorithm was applied, here *NACE2* categories. Starting from the original values of the isolated units on each tail, groups of minimum k isolated units are identified. The first or the last group may have a slightly different number of units. The average of each group is then computed and each group member *TURN* value is replaced by this mean. In case the number of isolated units on the tail is between k and $2k$, all the units are averaged and their *TURN* values are replaced by the mean. In case the number of isolated units on the tail is even lower than k, the values of all these units are replaced by the value of the closest clustered unit.

Considering that k equal *TURN* values would be released, $k = 3$ is considered a sufficient parameter value. This microaggregation has the double aim to protect the isolated units on tails and, at the same time, to avoid overprotection (avoiding the nearest clustered imputation when such cluster is too distant from the isolated units). The microaggregation step on each tail is formulated in the following:

1. count the number N_t of isolated units on the tail
2. If $N_t \geq 2k$, then
   i. Determine groups of minimum k isolated units. Considering only the *TURN* variable, simply sort the units and take groups of k units.
   ii. Let $x_1, x_2, \ldots, x_k$ be the original *TURN* values in a group. Compute their arithmetic mean $\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i$.
   iii. Let the group perturbed *TURN* values be $x_i^* = \bar{x}, i = 1, \ldots, k$
3. If $k \leq N_t < 2k$, then
   iv. Let $x_1, x_2, \ldots, x_{N_t}$ be the original *TURN* values. Compute their arithmetic mean $\bar{x} = \frac{1}{N_t}\sum_{i=1}^{N_t} x_i$.
   v. Let the perturbed *TURN* values be $x_i^* = \bar{x}, i = 1, \ldots, N_t$.
4. If $N_t < k$, then
   vi. Let $x_1, x_2, \ldots, x_{N_t}$ be the original *TURN* values.
   vii. Find the nearest (with respect to the distance used) clustered unit c. In the univariate case, this point is the smallest clustered unit (for the left tail) or the greatest clustered unit (for the right tail)
   viii. Let the perturbed *TURN* values be $x_i^* = c, i = 1, \ldots, N_t$.

6.5 Adjustment to preserve the totals

For coherence with the already published tables, the k_1 largest isolated units on the right tail are next adjusted in order to preserve *TURN* weighted totals for each combination of *NACE2* and *EMPclass* variables. When the number of isolated units on the right tail is less than k_1, the largest k_1 isolated units are adjusted to preserve the totals. For a given combination of *NACE2* and *EMPclass*, let $x_i, i = 1, \ldots, n$ and $w_i, i = 1, \ldots, n$ be the original (non perturbed) *TURN* values and calibration weights respectively. Denote by $T = \sum_{i=1}^{n} x_i w_i$ the total to be preserved. Let $x_i^*, i = 1, \ldots, n$ be the *TURN* perturbed values (sections 6.3 and 6.4). Note that the applied perturbation methods do not change the calibration weights. The difference $D = T - \sum_{i=1}^{n} x_i^* w_i$ is distributed on $x_{n-k_1+1}^*, \ldots, x_{n-1}^*, x_n^*$, the largest k_1 isolated units on the right tail or the largest isolated units if the number of isolated units on the right tail is not sufficient, as follows: $x_{n-i+1}^{**} = x_{n-i+1}^* + \frac{D}{\sum_{j=1}^{k_1} w_{n-j+1}}, i = 1, \ldots, k_1$ and $x_j^{**} = x_j^*, j = 1, \ldots, n - k_1$. One can easily check that $T = \sum_{i=1}^{n} x_i w_i = \sum_{i=1}^{n} x_i^{**} w_i$. For coherence with the microaggregation step, $k_1 = k$ could be considered. The value $k_1 = 3$ was used for the Italian CIS3 microdata file.

Since the (largest) *TURN* values are generally nonnegative, for sake of coherence, care must be paid in order not to obtain negative values. In such cases, additional groups of k_1 units on which distribute the difference D might be considered. When it is not possible to preserve the weighted totals for each combination of *NACE2* and *EMPclass* obtaining at the same time nonnegative values, preservation of only *NACE2* weighted total with nonnegative values is preferred. This situation might happen when there is a reduced number of units on the right tail. If this approach fails too, the weighted total should not be preserved and the protection of isolated units should stop before the microaggegation step (the isolated units on the right tail being perturbed by the nearest clustered unit imputation). For example, in the Italian CIS3 survey, this happened for *NACE2*=73 and *EMPclass*=2_3 where there were only 14 sampled units, half of which were identified as units at risk (isolated units). The microaggregation step was skipped and the total was preserved only with respect to the *NACE2* = 73 category.

It might happen that the number of isolated points on the right tail and on the central part of the distribution is not sufficient to apply any adjustment. Then, the adjustment would be performed by considering together isolated units on the left and right tail. Consequently, the isolated unit(s) on the left tail would be too much modified. This situation happened in case of Italian CIS3 dataset for *NACE2* = 23 when the modification of a left tail isolated point was more that 300%. In this case, the adjustment was applied only on right tail units.

In case of *NACE2* = 66 (Italian CIS3 dataset), in one size class, there were only two isolated points. Following the treatment previously described, the perturbation of the higher isolated unit was deemed insufficient. Hence microaggregation was applied on the highest k *TURN* values.

The latter considerations on the special cases show that an audit procedure is necessary at the end of the protection phase, as in whatever perturbation method.

6.6 Protection against spontaneous identification

Variables to be released which may be subject to spontaneous identification must be protected too. For example, in case of the Italian CIS3 data, *RTOT* is such a variable. As previously discussed, a check performed by survey experts is required. Generally, the number of units to be protected against spontaneous identification is very reduced. Hence a dedicated protection must be applied. For example, such *RTOT* values might be changed as follows: $RTOT^* = \frac{TURN^*}{TURN} RTOT$. In this particular survey data, Italian CIS3 dataset, there was not necessary to perturb *RTOT* values in any manner, as suggested by the survey experts.

7. Information loss and information preservation

Protection methods unavoidably reduce the informative content of a microdata file. In this paragraph some hints on how to evaluate the information loss in case of CIS data are given.

In the present proposal, the only changed variable is *TURN*. The perturbation is applied only to the isolated units. All clustered units are to be released with their original values. By imputing an isolated value by the nearest clustered *TURN* observation, the information loss is not too large and protection is enforced at the same time. Other imputation methods could be considered, e.g. imputation by the nearest cluster mean, but the difference between the original and protected datasets would surely increase.

The weighted totals are preserved for each combination of *NACE2* and *EMPclass* variables, if possible. Otherwise, they are preserved for each *NACE2* combination only.

Checking the relationships between variables, only 4 *RTOT* (over 12964 records) values resulted being higher than the perturbed *TURN* values, for the Italian CIS3 microdata file.

7.1 Variance comparison

Microaggregation generally decreases the variance of the involved variables. Due to the final adjustment made on the last group(s) of 3 (isolated) units to preserve weighted totals and to the applied imputation, this effect should not necessarily be observed. A comparison between the variances of the original and perturbed variables is suggested. This comparison must be performed for each combination of categorical key variables where microaggregation was applied, in this case, for each combination of *NACE2*. The ratios between the *TURN* variance after, σ^*, and before, σ, protection are shown in the next table and in figure 2.

NACE2	σ^*/σ	*NACE2*	σ^*/σ
17	0.72	*33*	1.07
15	0.97	*34*	1.26
10	0.58	*35*	0.90
18	0.73	*36*	0.83
19	1.01	*37*	1.00
20	0.84	*40*	1.62
21	1.03	*51*	0.47
22	0.92	*60*	1.08
23	0.94	*61*	0.97
24	0.52	*62*	0.55
25	0.55	*63*	1.51
26	1.03	*64*	0.26
27	0.46	*65*	0.97
28	0.81	*66*	1.00
29	0.72	*67*	0.78
30	0.96	*72*	0.92
31	0.63	*73*	1.02
32	0.89	*74*	1.05

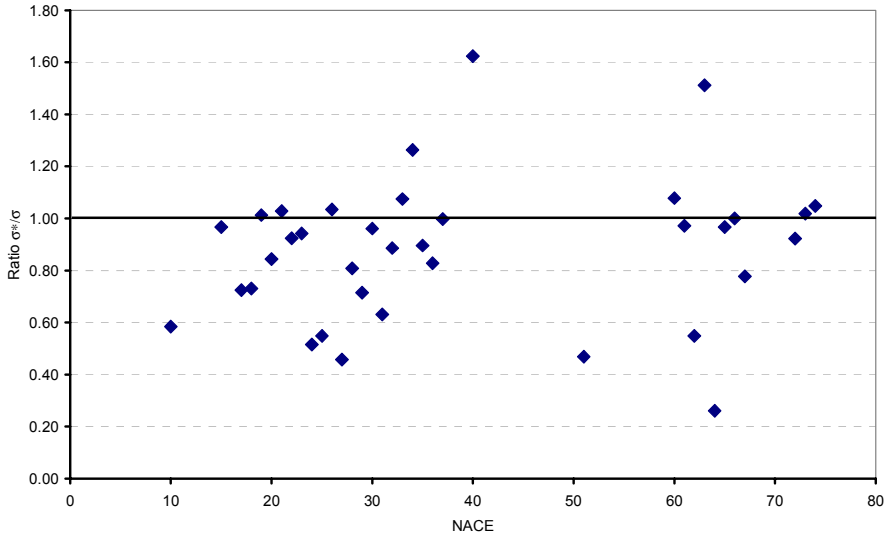


Figure 2. Comparison of σ^* and σ.

7.2 Data utility: correlations comparison

Correlations may also be compared to assess the degree of information loss. In the next table the correlations between the original and perturbed *TURN* values, between original *TURN* and *RTOT* and between perturbed *TURN* and *RTOT* are shown. *TURN** denotes the *TURN* perturbed variable.

NACE2	Corr(*TURN*,*TURN**)	corr(*TURN*,RTOT)	corr(*TURN**,RTOT)	*NACE2*	Corr(*TURN*,*TURN**)	corr(*TURN*,RTOT)	corr(*TURN**,RTOT)
17	0.93	0.15	0.17	***33***	0.96	0.72	0.63
15	0.98	0.71	0.77	***34***	0.9	0.9	0.66
10	0.85	0.12	0.18	***35***	0.9	0.77	0.67
18	0.92	0.34	0.5	***36***	0.96	0.41	0.38
19	0.81	0.15	0.17	***37***	0.97	0.18	0.26
20	0.93	0.28	0.33	***40***	0.87	0.3	0.33
21	0.98	0.4	0.43	***51***	0.76	0.19	0.15
22	0.96	0.72	0.64	***60***	0.98	0.47	0.4
23	0.93	0.04	0.06	***61***	0.99	0.31	0.27
24	0.74	0.55	0.49	***62***	0.65	0.1	0.12
25	0.77	0.22	0.4	***63***	0.9	0.43	0.35
26	0.96	0.39	0.39	***64***	0.58	0.51	0.54
27	0.72	0.39	0.46	***65***	0.99	0.48	0.47
28	0.95	0.27	0.29	***66***	1	0.32	0.32
29	0.93	0.62	0.61	***67***	0.93	0.56	0.62
30	0.86	0.52	0.46	***72***	0.95	0.55	0.6
31	0.81	0.73	0.73	***73***	0.94	0.94	0.91
32	0.75	0.95	0.56	***74***	0.99	0.24	0.22

7.3 Data utility: users perspective

The data utility is measured with respect to the variables *RTOT,* and its components *RRDINDX, RREXX, RMACX,* etc. Since *TURN* is the only perturbed variable, only ways in which researchers could use ratios *RTOT/TURN, RRDINDX/TURN, RREXX/TURN, RMACX/TURN* are taken into account. Of course, one cannot imagine all possible usages of *TURN*, but experts suggestions are very useful in such simulations. For the Italian CIS3 data, the quantiles of the above mentioned ratio variables were compared. Generally, a very good agreement was observed. Usually, the difference was lower than 1e-2. Except for *NACE2* = 73, the maximum absolute difference between the quantiles of the ratios computed using the original *TURN* and quantiles of the ratios computed with perturbed *TURN* was 0.58. For *NACE2* = 73, the differences between quantiles of the above listed variables were 1.14, 0.27, 0.27, 0.01, respectively.

8. Concluding remarks

It is believed that a detailed analysis of possible disclosure scenarios and the definition of related identifying variables coupled with a careful risk assessment to detect real units at risk are suitable to evaluate the real risks of disclosure of a microdata file. This strategy of selective identification of risk allows for dedicated protection methods that can save more information content of the data than a generalised application of a perturbation method. Consideration of different scenarios is a key issue of this proposal. However for the risk evaluation in the spontaneous recognition scenario one should rely on an expert opinion.

The proposed strategy is extremely flexible allowing for the selection of different parameters and the inclusion of several identifying variables. Weights are unchanged and the users may obtain the same published values for many aggregated statistics. Most of the variables are released in their original form.

9. Summary of changes

According to the anonymisation procedure here proposed, the microdata file released for research purposes should contain the following variables ("removed" = variable not to be released, "changed"= a protection method was applied, "unchanged"= no treatment):

Variable	Code	Note	Variable	Code	Note
Name of the enterprise	Id	removed	Total turnover in 2000	Turn	changed
Address	Nuts	changed	Exports in 1998	Exp98	removed
Main activity	Nace	changed	Exports in 2000	Exp	removed
Enterprise part of a group	Gp	unchanged	Gross investment in tangible goods: 2000	InvTa	unchanged
Country of head office	Ho	removed	Total number of employees in 1998	Emp98	removed
Enterprise established	Est	unchanged	Total number of employees in 2000	Emp	changed
Turnover increased by 10 pct	TurnInc	unchanged	Stratum A	StraA	removed
Turnover decreased by 10 pct	TurnDec	unchanged	Stratum B	StraB	removed
Enterprise most significant market	SigMar	unchanged	Direct weights	weight	removed
Total turnover in 1998	Turn98	removed			

All the other variables remained unchanged.

10. References

1. Doc.Eurostat/F4/STI/CIS/M2/8.
2. Doc. Eurostat *The Third Community Innovation Survey CIS3. Methodology of Anonymisation*. 11-07-2005.
3. D. Defays, M.N. Anwar, *Masking Microdata Using Micro-Aggregation*, Journal of Official Statistics, Vol.14, No.4, 1998. pp. 449-461.
4. Cox, L.H., Ernst, L.R.: *Controlled Rounding*. INFOR 20 (1982) 423-432.
5. I. P. Fellegi, *Controlled Random Rounding*, Survey methodology, 123-133, 1975.
6. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996), *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

Contributi ISTAT(*)

1/2002 - Francesca Biancani, Andrea Carone, Rita Pistacchio e Giuseppina Ruocco - *Analisi delle imprese individuali*

2/2002 - Massimiliano Borgese .-.*Proposte metodologiche per un progetto d'indagine sul trasporto aereo alla luce della recente normativa comunitaria sul settore*

3/2002 - Nadia Di Veroli e Roberta Rizzi - *Proposta di classificazione dei rapporti di lavoro subordinato e delle attività di lavoro autonomo: analisi del quadro normativo*

4/2002 – Roberto Gismondi – *Uno stimatore ottimale in presenza di non risposte*

5/2002 – Maria Anna Pennucci – *Le strategie europee per l'occupazione dal Libro bianco di Delors al Consiglio Europeo di Cardiff*

1/2003 – Giovanni Maria Merola – *Safety Rules in Statistical Disclosure Control for Tabular Data*

2/2003 – Fabio Bacchini, Pietro Gennari e Roberto Iannaccone – *A new index of production for the construction sector based on input data*

3/2003 – Fulvia Cerroni e Enrica Morganti – *La metodologia e il potenziale informativo dell'archivio sui gruppi di impresa: primi risultati*

4/2003 – Sara Mastrovita e Isabella Siciliani – *Effetti dei trasferimenti sociali sulla distribuzione del reddito nei Paesi dell'Unione europea: un'analisi dal Panel europeo sule famiglie*

5/2003 – Patrizia Cella, Giuseppe Garofalo, Adriano Paggiaro, Nicola Torelli e Caterina Viviano – *Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità*

6/2003 – Enrico Grande e Orietta Luzi – *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat*

7/2003 – Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Indagine sperimentale sui posti di lavoro vacanti*

8/2003 – Mario Adua – *L'agricoltura di montagna: le aziende delle donne, caratteristiche agricole e socio-rurali*

9/2003 – Franco Mostacci e Roberto Sabbatini – *L'euro ha creato inflazione? Changeover e arrotondamenti dei prezzi al consumo in Italia nel 2002*

10/2003 – Leonello Tronti – *Problemi e prospettive di riforma del sistema pensionistico*

11/2003 – Roberto Gismondi – *Tecniche di stima e condizioni di coerenza per indagini infraannuali ripetute nel tempo*

12/2003 – Antonio Frenda – *Analisi delle legislazioni e delle prassi contabili relative ai gruppi di imprese nei paesi dell'Unione Europea*

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys):*

un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi **–** *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo **–** *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni **–** *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 **–** Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo **–** *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

(*) ultimi cinque anni

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

Documenti ISTAT(*)

1/2002 – Paolo Consolini e Rita De Carli - *Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati*

2/2002 – Stefania Macchia - *Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività economiche*

3/2002 – Maria De Lucia - *Applicabilità della disciplina in materia di festività nel pubblico impiego*

4/2002 – Roberto Gismondi, Massimo Marciani e Mauro Giorgetti - *The italian contribution towards the implementantion of an european transport information system: main results of the MESUDEMO project*

5/2002 – Olimpio Cianfarani e Sauro Angeletti - *Misure di risultato e indicatori di processo: l'esperienza progettuale dell'Istat*

6/2002 – Riccardo Carbini e Valerio De Santis – *Programma statistico nazionale: specifiche e note metodologiche per la compilazione delle schede identificative dei progetti*

7/2002 – Maria De Lucia – *Il CCNL del personale dirigente dell'area 1 e la valutazione delle prestazioni dei dirigenti*

8/2002 – Giuseppe Garofalo e Enrica Morganti – *Gruppo di lavoro per la progettazione di un archivio statistico sui gruppi d'impresa*

1/2003 – Francesca Ceccato, Massimiliano Tancioni e Donatella Tuzi – *MODSIM-P:Il nuovo modello dinamico di previsione della spesa pensionistica*

2/2003 – Anna Pia Mirto – *Definizioni e classificazioni delle strutture ricettive nelle rilevazioni statistiche ufficiali sull'offerta turistica*

3/2003 – Simona Spirito – *Le prestazioni assistenziali monetarie non pensionistiche*

4/2003 – Maria De Lucia – *Approfondimenti di alcune tematiche inerenti la gestione del personale*

5/2003 – Rosalia Coniglio, Marialuisa Cugno, Maria Filmeno e Alberto Vitalini – *Mappatura della criminalità nel distretto di Milano*

6/2003 – Maria Letizia D'Autilia – *I provvedimenti di riforma della pubblica amministrazione per l'identificazione delle "Amministrazioni pubbliche" secondo il Sec95: analisi istituzionale e organizzativa per l'anno 2000*

7/2003 – Francesca Gallo, Pierpaolo Massoli, Sara Mastrovita, Roberto Merluzzi, Claudio Pauselli, Isabella Siciliani e Alessandra Sorrentino – *La procedura di controllo e correzione dei dati Panel Europeo sulle famiglie*

8/2003 – Cinzia Castagnaro, Martina Lo Conte, Stefania Macchia e Manuela Murgia – *Una soluzione in-house per le indagini CATI: il caso della Indagine Campionaria sulle Nascite*

9/2003 – Anna Pia Maria Mirto e Norina Salamone – *La classificazione delle strutture ricettive turistiche nella normativa delle regioni italiane*

10/2003 – Roberto Gismondi e Anna Pia Maria Mirto – *Le fonti statistiche per l'analisi della congiuntura turistica: il mosaico italiano*

11/2003 – Loredana Di Consiglio e Stefano Falorsi – *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento Generale della Popolazione 2001*

12/2003 – Roberto Gismondi e Anna Rita Giorgi – *Struttura e dinamica evolutiva del comparto commerciale al dettaglio: le tendenze recenti e gli effetti della riforma "Bersani"*

13/2003 – Donatella Cangialosi e Rosario Milazzo – *Fabbisogni formativi degli Uffici comunali di statistica: indagine rapida in Sicilia*

14/2003 – Agostino Buratti e Giovanni Salzano – *Il sistema automatizzato integrato per la gestione delle rilevazioni dei documenti di bilancio degli enti locali*

1/2004 – Giovanna Brancato e Giorgia Simeoni – *Tesauri del Sistema Informativo di Documentazione delle Indagini (SIDI)*

2/2004 – Corrado Peperoni – *Indagine sui bilanci consuntivi degli Enti previdenziali: rilevazione, gestione e procedure di controllo dei dati*

3/2004 – Marzia Angelucci, Giovanna Brancato, Dario Camol, Alessio Cardacino, Sandra Maresca e Concetta Pellegrini – *Il sistema ASIMET per la gestione delle Note Metodologiche dell'Annuario Statistico Italiano*

4/2004 – Francesca Gallo, Sara Mastrovita, Isabella Siciliani e Giovanni Battista Arcieri – *Il processo di produzione dell'Indagine ECHP*

5/2004 – Natale Renato Fazio e Carmela Pascucci – *Gli operatori non identificati nelle statistiche del commercio con l'estero: metodologia di identificazione nelle spedizioni "groupage" e miglioramento nella qualità dei dati*

6/2004 – Diego Moretti e Claudia Rinaldelli – *Una valutazione dettagliata dell'errore campionario della spesa media mensile familiare*

7/2004 – Franco Mostacci – *Aspetti Teorico-pratici per la Costruzione di Indici dei Prezzi al Consumo*

8/2004 – Maria Frustaci – *Glossario economico-statistico multilingua*

9/2004 – Giovanni Seri e Maurizio Lucarelli – *"Il Laboratorio per l'analisi dei dati elementari (ADELE): monitoraggio dell'attività dal 1999 al 2004"*

10/2004 – Alessandra Nuccitelli, Francesco Bosio e Luciano Fioriti – *L'applicazione RECLINK per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*

1/2005 – Francesco Cuccia, Simone De Angelis, Antonio Laureti Palma, Stefania Macchia, Simona Mastroluca e Domenico Perrone – *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*

2/2005 – Marina Peci – *La statistica per i Comuni: sviluppo e prospettive del progetto Sisco.T (Servizio Informativo Statistico Comunale. Tavole)*

3/2005 – Massimiliano Renzetti e Annamaria Urbano – *Sistema Informativo sulla Giustizia: strumenti di gestione e manutenzione*

4/2005 – Marco Broccoli, Roberto Di Giuseppe e Daniela Pagliuca – *Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo Microstrat di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat*

5/2005 – Mauro Albani e Francesca Pagliara – *La ristrutturazione della rilevazione Istat sulla criminalità minorile*

6/2005 – Francesco Altarocca e Gaetano Sberno – *Progettazione e sviluppo di un "Catalogo dei File Grezzi con meta-dati di base" (CFG) in tecnologia Web*

* ultimi cinque anni

7/2005 – Salvatore F. Allegra e Barbara Baldazzi – *Data editing and quality of daily diaries in the Italian Time Use Survey*

8/2005 – Alessandra Capobianchi – *Alcune esperienze in ambito internazionale per l'accesso ai dati elementari*

9/2005 – Francesco Rizzo, Laura Vignola, Dario Camol e Mauro Bianchi – *Il progetto "banca dati della diffusione congiunturale"*

10/2005 – Ennio Fortunato e Nadia Mignolli – *I sistemi informativi Istat per la diffusione via web*

11/2005 – Ennio Fortunato e Nadia Mignolli – *Sistemi di indicatori per l'attività di governo: l'offerta informativa dell'Istat*

12/2005 – Carlo De Gregorio e Stefania Fatello – *L'indice dei prezzi al consumo dei testi scolastici nel 2004*

13/2005 – Francesco Rizzo e Laura Vignola – *RSS: uno standard per diffondere informazioni*

14/2005 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Launching and implementing the job vacancy statistics*

15/2005 – Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni e Leonardo Tininini – *La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT*

16/2005 – Ennio Fortunato e Nadia Mignolli – *Verso il Sistema di Indicatori Territoriali: rilevazione e analisi della produzione Istat*

17/2005 – Raffaella Cianchetta e Daniela Pagliuca – *Soluzioni Open Source per il software generalizzato in Istat: il caso di PHPSurveyor*

18/2005 – Gianluca Giuliani e Barbara Boschetto – *Gli indicatori di qualità dell'Indagine continua sulle Forze di Lavoro dell'Istat*

19/2005 – Rossana Balestrino, Franco Garritano, Carlo Cipriano e Luciano Fanfoni – *Metodi e aspetti tecnologici di raccolta dei dati sulle imprese*

1/2006 – Roberta Roncati – *www.istat.it (versione 3.0) Il nuovo piano di navigazione*

2/2006 – Maura Seri e Annamaria Urbano – *Sistema Informativo Territoriale sulla Giustizia: la sezione sui confronti internazionali*

3/2006 – Giovanna Brancato, Riccardo Carbini e Concetta Pellegrini – *SIQual: il sistema informativo sulla qualità per gli utenti esterni*

4/2006 – Concetta Pellegrini – *Soluzioni tecnologiche a supporto dello sviluppo di sistemi informativi sulla qualità: l'esperienza SIDI*

5/2006 – Maurizio Lucarelli – *Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica*

6/2006 – Natale Renato Fazio – *La ricostruzione storica delle statistiche del commercio con l'estero per gli anni 1970-1990*

7/2006 – Emilia D'Acunto – *L'evoluzione delle statistiche ufficiali sugli indici dei prezzi al consumo*

8/2006 – Ugo Guarnera, Orietta Luzi e Stefano Salvi – *Indagine struttura e produzioni delle aziende agricole: la nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti*

9/2006 – Maurizio Lucarelli – *La regionalizzazione del Laboratorio ADELE: un'ipotesi di sistema distribuito per l'accesso ai dati elementari*

10/2006 – Alessandra Bugio, Claudia De Vitiis, Stefano Falorsi, Lidia Gargiulo, Emilio Gianicolo e Alessandro Pallara – *La stima di indicatori per domini sub-regionali con i dati dell'indagine: condizioni di salute e ricorso ai servizi sanitari*

11/2006 – Sonia Vittozzi, Paola Giacchè, Achille Zuchegna, Piero Crivelli, Patrizia Collesi, Valerio Tiberi, Alexia Sasso, Maurizio Bonsignori, Giuseppe Stassi e Giovanni A. Barbieri – *Progetto di articolazione della produzione editoriale in collane e settori*

12/2006 – Alessandra Coli, Francesca Tartamella, Giuseppe Sacco, Ivan Faiella, Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Isabella Siciliani, Sara Colombini e Alessandra Masi – *La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*

13/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Intrastat*

14/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Extrastat*

15/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: comparazione tra rilevazione Intrastat ed Extrastat*

16/2006 – Fabio M. Rapiti – *Short term statistics quality Reporting: the LCI National Quality Report 2004*

17/2006 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zeli – *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI*

18/2006 – Mauro Albani – *La nuova procedura per il trattamento dei dati dell'indagine Istat sulla criminalità*

19/2006 – Alessandra Capobianchi – *Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa*

20/2006 – Francesco Altarocca – *Gli strumenti informatici nella raccolta dei dati di indagini statistiche: il caso della Rilevazione sperimentale delle tecnologie informatiche e della comunicazione nelle Pubbliche Amministrazioni locali*

1/2007 – Giuseppe Stassi – *La politica editoriale dell'Istat nel periodo 1996-2004: collane, settori, modalità di diffusione*

2/2007 – Daniela Ichim – *Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment*