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

# A1. Community Innovation Survey 4

## *1. Introduction*

This document describes a methodology for the anonymisation of the Community Innovation Survey microdata. The application in case of the Italian CIS4 is presented.

The anonymisaton methodology takes into account both economic features of the data and the dissemination policy of the National Statistical Institute. An important key point is the fact that the final protected data set would be released for research purposes, and hence subject to a signed contract. Consequently, a rigorous study of possible disclosure scenarios is carried out in order to define the identifying variables. Two spontaneous identification scenarios based on structural and non structural information are modelled . 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. In the disclosure scenario based on structural information, the basic idea 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 around this value is not deemed sufficient (below a certain threshold). 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 situation, considering both continuous and categorical key variables. Depending on the thresholds, the number of isolated (hence at risk of re-identification) units would tend to increase or decrease. The thresholds may be defined according to the observed phenomenon, assumed disclosure scenario and national dissemination policy, proving the great flexibility of the methodology. A disclosure scenario is also modelled in order to take into account some qualitative information a possible intruder might have. Justified by the research purposes of the microdata file release, **only** the key variables and **only** the records at risk of re-identification would be perturbed whereas the rest of the file would be released unchanged. Perturbation is mainly achieved by an imputation from the nearest safe unit and a particular microaggregation. For extreme thresholds choice in the identification phase, the methodology reduces to microaggregation. A deterministic adjustment procedure is performed in order to maintain the published totals. Modifying only the key variables and only for the records at risk of re-identification, many variables (including the sampling weights) would remain unchanged, hence coherence with many already published aggregate statistics would be naturally achieved.

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

1. Definition of the disclosure scenarios.
    a. spontaneous identification scenario based on structural information

    b. spontaneous identification scenarios based on non-structural information
        i. dominance scenario
        ii. uniqueness scenario
2. Preliminary work on variables.
    a. variable suppression
    b. global recoding
    c. preliminary rounding
3. Risk assessment: identification of units at risk.
    a. spontaneous identification scenario based on structural information (clustering)
    b. spontaneous identification scenarios based on non-structural information (uniqueness)
4. Microdata protection
    a. imputation from the nearest clustered unit
    b. micro-aggregation on tails
    c. multiplicative perturbation
    d. micro-aggregation of some records of variables related to the first year of the reference period
5. Adjustment to preserve published totals.
6. Audit strategies.
    a. negative values
    b. insufficient protection
    c. overprotection
7. Information loss assessment.
    a. variance comparison
    b. correlations comparison
    c. users perspective.
8. Description of the microdata file to be released.

This document gives a detailed description of these eight steps. Section 2 briefly summarizes the CIS4 survey. In section 3, the assumptions of the disclosure scenarios are discussed. In section 4, some preliminary work (variable suppression, coding, etc.) on variables is described. Based on the disclosure scenarios previously determined, section 5 defines the units at risk and  introduces a method for their automatic identification. It also presents the practical implementation of the spontaneous identification scenarios based on non-structural information. The microdata perturbation presented in section 6 is applied to the units considered at risk of re-identification. As discussed in section 7, in order to preserve some characteristics (e.g. weighted totals), a further adjustment is performed on the perturbed variables. In section 8, some possible approaches to the audit problem are discussed. Finally, in section 9 several measures for information loss assessment are discussed whereas in section 10 the choices made and the results in terms of released variables are summarized.

Throughout the document, the results obtained by applying this anonymisation methodology to the Italian CIS4 survey microdata are presented.

## 2. Community Innovation Survey: Brief Description of the Data

CIS provides information on the characteristics of innovation activity at enterprise level, see Eurostat (2005). Some of the main observed variables are: principal economic activity, geographical information, number of employees in 2002 and 2004, turnover in 2002 and 2004, expenditure in intramural RD (in line with the turnover in 2004), expenditure in extramural RD (in line with the turnover in 2004), expenditure in acquisition of machinery (in line with the turnover in 2004), expenditure in other external knowledge (in line with the turnover in 2004), expenditure in training, market (in line with the turnover in 2004), total innovation expenditure (in line with the turnover in 2004), number of persons involved in intra RD (in line with the number of employees in 2004).

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

*Principal economic activity*

The following industries are 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), architectural and engineering activities (NACE 74.2), technical testing and analysis (NACE 74.3).

"Non-core" industries that are covered in addition are: motor trade (NACE 50), retail trade (NACE 52), hotels and restaurants (NACE 55), real estate activities (NACE 70), renting of machinery and equipment without an operator (NACE 71), research and development (NACE 73), other business activities: legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling; business and management consultancy; holdings (NACE 74.1); advertising (NACE 74.4); labour recruitment and provision of personnel (NACE 74.5); investigation and security activities (NACE 74.6); industrial cleaning (NACE 74.7); miscellaneous business activities n.e.c. (NACE 74.8).

*Size of the enterprise*

All enterprises included in the target population follow the minimum coverage which is all enterprises with 10 employees or more.

For the Italian CIS4, the sampling frame was 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. A sampling survey was conducted, resulting in about 22000 respondent enterprises. The sampling weights were also recorded.

## *3. Disclosure Scenarios*

As the microdata file will be disseminated for research purposes, ”nosy colleague” and “external register” scenarios are not deemed realistic. Instead, only “spontaneous identification” scenarios are considered realistic and discussed in this section.

### 3.1 Spontaneous Identification Based on Structural Information

Since structural variables are generally publicly known, it is supposed that an intruder could use this information to identify an enterprise. Moreover, it is known that a business register was used as a sampling frame. Consequently, a possible intruder *a-priori* knows that an enterprise possibly belonging to the sample, it is surely included in the business register. As the possible intruder is a researcher, it may be assumed that he wouldn’t perform a **complete** record linkage for re-identification purposes. Nonetheless, the researcher may be curious about some units. For example, he might know in advance the most famous/dominant enterprises. Alternatively, some units may be highlighted during the analysis and the intruder may try to find some more information about such enterprises. In other words, it may be supposed that the intruder/researcher **might** use some structural information **only** for the re-identification of **several particular enterprises**.

Public business registers report general information on name, address, turnover (*TURN*), number of employees (*EMP*), etc. The survey observed variables that are also reported in a publicly available register could be considered as identifying variables. The survey information content (level of detail) of each identifying variable with a significant identifying power should be modified. Data utility criteria should be the main constraint: variables frequently used in statistical analyses should be less modified.

For the Italian CIS4 microdata file, the following variables are registered in an external register:

1. turnover (*TURN*)
2. principal economic activity (*NACE*)
3. number of employees (*EMP*)
4. region (*NUTS*)

Since *TURN* is a proxy of the enterprise dimension, it is frequently included in researcher’s analyses. Then, releasing *TURN* in its original form would be an important aim of the dissemination of the microdata file. The other publicly available identifying variables are either removed or modified.

Some other information on enterprises is recorded in the microdata file and public registers, with different reference dates, e.g. number of employees and turnover in the first year of the reference period of the survey. Being somehow obsolete information, the present disclosure scenario assumes that there is no available public business register that contains such historical information. Moreover, due to the quantity and quality of the resources needed for such re-identification, it is supposed that the researcher would not directly use the information referred to the first year of reference period. Hence, in this scenario based on structural information, these variables are not considered as identifying variables.

## 3.2 Spontaneous Identification Based on Non-structural Information

In the CIS4 survey data set 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 hardly published in an external register, but they can assume extreme values on some units. Mere additional information would then clearly identify an enterprise. Special attention should 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 identifying variables to be disseminated.

In order to evaluate this risk of re-identification in an automatic manner, four spontaneous identification subscenarios are considered realistic. These scenarios are based on several assumptions on the intruder a-priori knowledge.

### *3.2.1 Dominance Scenario Based on Investment in Research*

This scenario is based on the assumption that **large enterprises invest much more in innovation** than the other enterprises. Since large enterprises are the most famous ones, the values assumed by the categorical identifying variables on such units are generally known, or easily derived. A possible intruder, supposing (or knowing) that the target enterprise belongs to the sample (as it generally happens for the largest enterprises), could assume that this economic unit is the one with the maximum total innovation expenditure (*RTOT*), for example. In this scenario, for a given combination of the categorical identifying variables, an intruder could identify the enterprise with the maximum *RTOT* and then he would check whether this enterprise is dominant. The dominance should be assessed with respect to (some) continuous identifying variables to be disseminated, see below. If both these conditions hold, the intruder would probably associate this record to the target enterprise.

Of course, this dominance scenario would also apply to each *RTOT* component deemed at risk of re-identification (the intruder should have some previous knowledge on the particular *RTOT* component to be used as identifying variable).

#### *3.2.2 Dominance Scenario Based TURN at the First Year of the Reference Period*

This scenario is based on the assumption that **some large enterprises are always dominant**. It is supposed that an intruder a-priori could know which are the constantly dominant enterprises since such units are generally also well-known. Moreover, such enterprises generally belong to the sample.

#### *3.2.3 Dominance Scenario Based on Variations with Respect to the Beginning of the Reference Period*

This scenario is based on the assumption that **large enterprises** corresponding to extremely **large *TURN* (or *EMP*) variations** are known by the general public. A possible intruder could remember that a well-known enterprise had a boom or a crack during the reference period. Being a researcher, it might be supposed that the intruder indeed has this accurate knowledge of the phenomenon under study. He might look for such a huge variation between *TURN* and *TURN2002*. Then, supposing that the target enterprise belongs to the sample (this generally holds for the large enterprises), he would associate it to the record corresponding to this extreme variation. Having in mind the target enterprise and its economic characteristics, the intruder would probably look for the largest variation inside the correct combination of categorical key variables.

The same reasoning might hold for the variable "number of employees" (*EMP*).

#### *3.2.4 Uniqueness Scenario*

This scenario is based on the hypothesis that **large enterprises are generally included in the sample** (direct sampling weight equal to one). Consequently, when in a given combination of the categorical identifying variables there is an unique enterprise with a direct sampling weight equal to one, such enterprise could be at risk of re-identification. Supposing that the calibration would not alter such weight (or simply ignoring the existence of any calibration process), a possible intruder would search for those enterprises with weight one. In case there is an unique such unit in a given combination of the categorical identifying variables, such enterprise would then be associated to the target enterprise.

#### *3.2.5 Sampling Weights*

Providing information on the inclusion in the sample of a given unit (enterprise), the sampling weights information content may not always be negligible. The CIS4 survey microdata file contains also information on the direct and adjusted (e.g. for non-response) weights.

The direct weights are by default all equal to 1 when a census is conducted. But in such cases, releasing the same value (one) for all the weights would not increase the intruder knowledge. Moreover, the intruder already knows that the target enterprise was included in the sample (census). If a sampling survey is conducted, the direct weight of an enterprise is directly related to its probability of inclusion in the sample. The **direct weights** are considered too disclosive because the largest enterprises are surely included in the sample. Consequently, it would be sufficient to find a unique direct weight equal to 1 in the corresponding category of cross-classifying identifying variables to identify the enterprise without uncertainty. Considering also that these are not the weights to be used in the estimation/analysis phase, it is safer to remove the direct weights from the microdata file.

To account for non-response rate and for auxiliary information, a calibration is generally performed. Concerning the adjusted weights, the output of any calibration process, the above procedure is much more uncertain. Firstly because even if an intruder is informed on the auxiliary variables considered in the calibration process, it would be harder for him to achieve these auxiliary values, too. Secondly, to be able to compute back the direct weights (if possible, anyway), statistical expertise on the entire survey process is strictly necessary.

For the Italian CIS4 survey data, it was considered that the adjusted weights may be released unchanged.

### 3.3 Identifying Variables

The identifying variables of the hypothesized disclosure scenarios are: economic classification (*NACE*), geographical location (*NUTS*), number of employees at the end of the reference period (*EMP*), turnover at the end of the reference period (*TURN*), total innovation expenditure (*RTOT*) and its components deemed at risk of re-identification (section 3.2.1) and the sampling weights. Considering also the issues discussed in section 3.1, the categorical identifying variables considered are *NACE*, *EMP* and *NUTS*, while the variable with respect to which the dominance should be assessed is *TURN*. This is a very general disclosure scenario since it includes most of the structural variables. Depending on the phenomena under study (e.g. the economical phenomenon) and on the dissemination policy of the NSI, only a subset of these variables could be considered in the disclosure scenario. More details may be found in Ichim (2008).

## *4. Preliminary Work on Variables*

### 4.1 Variable Suppression

1. Some variables are removed because they may be invalidated from the quality point of view.
2. Direct identifiers are removed from the microdata file to be released. Some of these are:
    a. Name
    b. Address
3. Other variables
    a. Direct sampling weights
    b. Stratum A
    c. Stratum B

## 4.2 Global Recoding

Some variables are aggregated according to the structure of the surveyed economy. Based on feedback from the scientific community, the CIS microdata file should be released containing information on NACE at 2 digits and three enterprise size classes. Usually these minimum requirements could be easily fulfilled, but national characteristics of the data should be also considered. The dissemination policy of the National Statistical Institute is a natural constraint. For example, some NACE classifications 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 or dissemination problem, but rather a feature of the surveyed phenomenon. For example, when such phenomenon is not well represented, NACE divisions might be aggregated (preserving the NACE hierarchy).

For the Italian CIS4 microdata file, the following aggregations were used. These aggregations are routinely used by Istat. In other national settings, these agegations could change.

1. *Principal economic activity* (*NACE*). A new variable *NACE2* is then obtained:
    a. NACE is aggregated in NACE at 2 digits
    b. NACE is not recoded at NACE 2 digits for the categories NACE = 742 and NACE = 743.
    c. NACE 10, 11, 12, 13 and 14 are aggregated into a single class, *NACE2* = 10
    d. NACE 15 and 16 are aggregated into a single class. *NACE2* = 15
    e. NACE 40 and 41 are aggregated into a single class, *NACE2* = 40

2. *Number of employees* (*EMP*). *EMP* is recoded in three main classes. These classes can be combined every time the number of enterprises in a certain *NACE2* category is not considered sufficient . A new variable *EMPclass* is then created. For the Italian CIS4 microdata file, the following categories were then obtained:
    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. *Number of employees at the first year of the reference period* (*EMP2002*) is recoded in the same three main classes. These classes are:
    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"

Categories of identifying variables with too significant identifying power are commonly aggregated into a single category.

Maintaining *NUTS* at (macro)regional level, the number of combinations of *NACE2*, *EMPclass* and *NUTS* with less than three enterprises is too high. Consequently, at least in the Italian economic framework, the release of a geographical information at (macro)regional level is considered highly disclosive. In conclusion, *NUTS* categories are aggregated into a single one, releasing this information only at national level of detail. In other national settings, this recoding could not be necessary.

*Ho* is aggregated in the following classes: 1 (Home), 2 (Europe + Candidate Countries + EFTA countries) and 3 (rest of the world), 0 (no answer).

## 4.3 Significant Digits in Continuous Variables

The anonymisation methodology should avoid any overprotection because the microdata file would be disseminated for research purposes: if a unit may be confused (see below) with others, its re-identification would be difficult and its original *TURN* value is released unchanged. Since *TURN* is a continuous economic variable, it is a highly identifying one: it assumes almost unique values on each population 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 (more or less) light rounding to the variable *TURN*. In other words, one should choose the number of significant digits to be preserved (the rounding base). Otherwise, this step may be skipped. Rounding is a perturbation method because it avoids the re-identification based on exact disclosure. Since identification of units at risk should be also based on the released values, it is more natural to discuss it here. 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* variable should be considered as an input to the subsequent identification phase.

The reference year of CIS4 is 2004. The microdata file for research would be released in 2008. Hence, for the Italian CIS4 microdata file, an initial rounding of the variable *TURN*  was deemed necessary. Rounding is especially recommended when values on some units are imputed from some (external) register. The rounding base should vary between 1 and 10, depending on the register accuracy.

The same procedure should be applied to the *TURN* values corresponding to the first year of the reference period, namely 2002.

For the Italian CIS4, the rounding base was the unity.

# *5. Identification of Units at Risk*

A unit is considered at risk if it is "recognisable" in either spontaneous identification scenario. The scenario based on structural information will be addressed in section 5.1 while the scenario based on non-structural information will be discussed in section 5.2.

## 5.1 Spontaneous Identification Based on Structural Information

This section presents a distance-based approach for the evaluation of the identification risk in this scenario. More details and updated versions are discussed in Ichim (2008) and Ichim (2009).

### *5.1.1 Assumptions on identification of units at risk*

The method models an intruder uncertainty. The basic assumption is that a unit couldn't be identified by an intruder if can be confused with other units. The underlying idea is the *k*-anonymity principle, see Sweeny (2002). Both identification and/or confusion should be assessed with respect to the identifying variables. In particular, for each combination of categorical identifying variables (see section 3.3), identification/confusion should be evaluated with respect to the continuous identifying variables to be released. If the identifying variables were all categorical variables, a sample/population uniques approach might be considered, but this is hardly the case for the Community Innovation Survey.

Since *NUTS* was aggregated at national level, with respect to this spontaneous identification scenario (section 3.1), for each combination of *NACE2* and *EMPclass*, the turnover *TURN* is the unique remaining identifying variable. *TURN* being an economic continuous variable, it assumes almost unique values on each unit, even if it was previously rounded. Hence, the identification of the units at risk cannot be based on sample/population uniques approaches. Consequently, the degree of confusion of a given unit should be measured with respect to the distance to the other units.

The main idea of the method is that an intruder cannot distinguish between *TURN* values that are too close to each other. It should also be noted that an insufficient number of close units could anyway lead to an approximate disclosure. This should be avoided, too. Consequently, it is assumed that an intruder might 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.

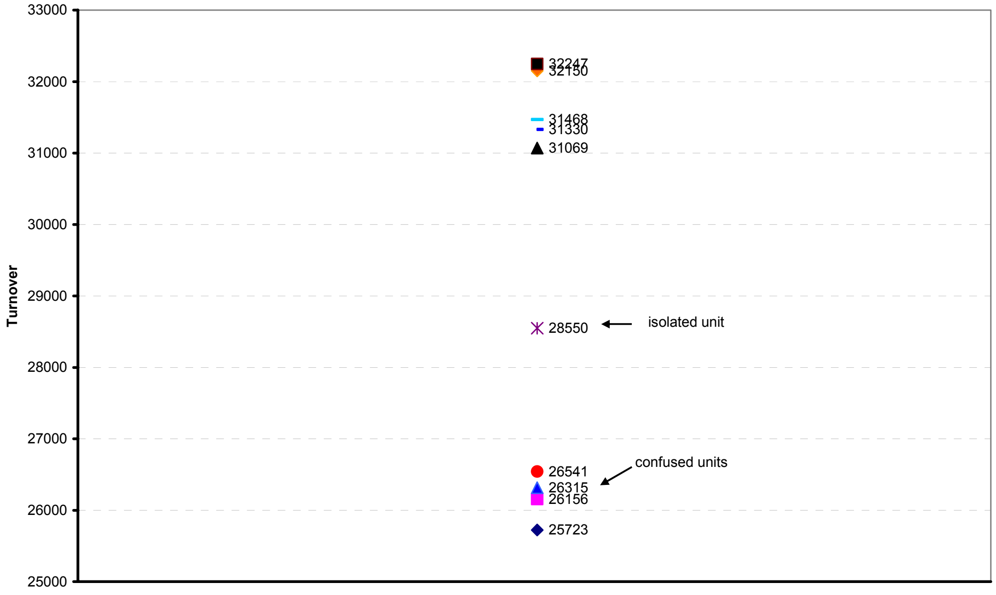


**Figure 1. Example of isolated and confused units.**

### *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 a cluster $C$ are considered not at risk of re-identification because they may be confused with the other units belonging to $C$. Instead, it is assumed that an intruder might distinguish between different clusters. Note that it is possible for clusters to contain a single unit. From the statistical disclosure control point of view, a cluster is considered at risk of re-identification if it does not contain a sufficient number of units.

Clustering algorithms are generally implemented in statistical software.

### *5.1.3 Density-Based Clustering Algorithm*

The identification of units at risk of re-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. In this framework, density is defined in terms of both distance and number of points. In literature, the algorithms taking into account these two features are called density based algorithms. The iterative algorithm used here for clusters identification is named DBSCAN, see Ester (1996). 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*. Clusters are constructed by means of the so-called "core" units, i.e. the units whose *Eps*-neighbourhoods contain at least a minimum number *MinPts* of units. Not all the units in a cluster have at least *MinPts* units in their *Eps*-neighbourhood. Such units, are still considered clustered because they belong to the *Eps*-neighbourhood of another unit in the cluster.

DBSCAN requires only the initialization of two parameters (thresholds): *Eps* (a positive real value) and *MinPts* (an integer number).

The DBSCAN clustering steps are briefly summarized below:

**DB1.** Start with a unit U. A cluster $C$ 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 $C$. This step is repeated consecutively for each unit in $E_U$, increasing, if the case, the dimension of $C$.

**DB3.** If the *Eps*-neighbourhood $E_U$ of U does not contain a sufficient number of units (< *MinPts*), the next unit in $C$ is checked. If the inspection of all units in $C$ is finished, check the next unit in the dataset. Repeat until no units should be checked anymore.

**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 Implementation of DBSCAN*

**DBSCAN should be applied inside each combination of categorical identifying variables.**

For the Italian CIS4 survey data, the identification of isolated units by means of the DBSCAN algorithm was performed for each combination of *NACE2* and *EMPclass*. *TURN* was the clustering variable. In other national settings, consideration of more (or less) identifying variables could be necessary. It should be anyway stressed that the anonymisation procedure should be applied with respect to the entire set of identifying variables.

Before applying the algorithm, it could be necessary to apply a transformation on the continuous identifying variables. As *TURN* generally exhibits a very skewed distribution, a logarithmic transformation was used for the Italian CIS4 microdata file.

For continuous variables, the Euclidean distance function is suitable. It can also be easily extended to the multivariate case, when other (continuous) identifying 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 identifying variables is not reccomended.

The choice of the number of units in a neighbourhood of a unit U, the parameter *MinPts*, depends on the dissemination policy and on the accuracy of the external knowledge considered in the disclosure scenario. A too low value would imply too many clusters (with a sufficient number of units). Hence, the re-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. Consequently, the information loss would increase. A trade-off should be found, taking into account both the dissemination policy of the NSI and the accuracy of the structural information considered in the disclosure scenario.

Following various simulations, *MinPts* = 3 was deemed a reasonable value for Italian CIS4 microdata.

Usage of the same value for all combinations of categorical identifying variables (here *NACE2* and *EMPclass*) is a coherent choice. When the continuous identifying variable distribution (or, equivalently, the phenomenon under study) depends too much on the combination of categorical identifying variables, different values of *MinPts* might be chosen.

As discussed in the Ester (1996), the *Eps* value is determined with respect to the distances from the *MinPts*-th unit, for each combination of categorical identifying variables. That is, the matrix of distances between units are computed. For each unit U, this matrix contains a row with the distances from the other units to U. Then, for each unit U, the elements of the corresponding row are sorted in ascending order. Then, the *MinPts*-th column of this newly obtained matrix is considered and the value corresponding to the most abrupt change in the corresponding sorted vector is automatically detected. This value is the threshold value *Eps.* An easier approach could be the threshold setting by means of a quantile of the *MinPts*-th column, for example the third quantile.

For the Italian CIS 4 survey data, the threshold *Eps* was computed by means of the automatic criteria. In this way, the number of units at risk was not a-priori defined.

Other criteria for the determination of the threshold *Eps* are possible, but the same principles should be followed for each combination of categorical identifying variables.

Summarizing, the identification of units at risk procedure, for each combination of the categorical identifying variables, is the following:

a) if the case, transform the data
b) choose the distance function
c) choose 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

It should be noted that when the number of units in a given combination of categorical identifying variables is too small (for example smaller than *MinPts*), all these units would be considered isolated points. Anyway, due to the recoding procedure in section 4.2, such situation hardly happens in CIS microdata file. Moreover, application of any clustering algorithm on a very reduced number of units is meaningless.

When applied to the Italian CIS4 survey data, DBSCAN identified isolated units on both tails of the distribution, because *TURN* has a very skewed distribution inside each combination of *NACE2* and *EMPclass*. Moreover, DBSCAN detected also isolated units on the central part of the distribution. In Table 1 the percentages of isolated points identified when applying the DBSCAN to the Italian CIS4 survey data are shown. The percentages are computed with respect to the total number of observations.

The *TURN* values corresponding to the units at risk of re-identification should be perturbed (see section 6). Denote by $TURN^*$ the perturbed variable.

### *5.1.5 Possible Extensions*

The method is flexible enough to incorporate other identifying variables, see Ichim (2008). It is believed that, in other national settings, this approach could only be extended to include other identifying variables. Firstly, it should be noted that *NACE2*, *EMPclass* and *TURN* are among the minimal requirements. The number of categorical identifying variables could increase or decrease, but this generalisation presents no particular problem (only the number of categories would increase or decrease). Secondly, the *TURN* importance from a data utility point of view was already stressed in section 3.1. The number of variables used for identification/confusion evaluation could only increase. For example, for future surveys, other variables might be released and hence considered in the identification phase. In such cases, the distance function should be adequately updated.

| *Nace2* | *EMPclass* | Observations | Left Tail | Right Tail | Total | *Nace2* | *EMPclass* | Observations | Left Tail | Right Tail | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *10* | *1* | 297 | 2.69 | 1.35 | 7.41 | *35* | *3* | 16 | 0.00 | 6.25 | 6.25 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *10* | *5* | 38 | 5.26 | 2.63 | 7.89 | *36* | *1* | 212 | 2.36 | 3.30 | 7.08 |
| *15* | *1* | 428 | 1.87 | 2.57 | 7.71 | *36* | *2* | 100 | 2.00 | 8.00 | 10.00 |
| *15* | *2* | 160 | 4.38 | 1.88 | 8.75 | *36* | *3* | 25 | 4.00 | 8.00 | 20.00 |
| *15* | *3* | 63 | 1.59 | 6.35 | 12.70 | *37* | *6* | 106 | 0.94 | 5.66 | 7.55 |
| *17* | *1* | 299 | 1.67 | 1.67 | 5.35 | *40* | *1* | 167 | 3.59 | 3.59 | 7.78 |
| *17* | *2* | 94 | 5.32 | 2.13 | 7.45 | *40* | *2* | 82 | 3.66 | 4.88 | 8.54 |
| *17* | *3* | 43 | 0.00 | 6.98 | 6.98 | *40* | *3* | 38 | 0.00 | 2.63 | 2.63 |
| *18* | *1* | 336 | 2.38 | 1.19 | 6.85 | *45* | *1* | 4756 | 1.20 | 0.88 | 8.35 |
| *18* | *2* | 97 | 0.00 | 2.06 | 7.22 | *45* | *2* | 514 | 0.97 | 0.39 | 7.98 |
| *18* | *3* | 26 | 3.85 | 11.54 | 15.38 | *45* | *3* | 47 | 2.13 | 4.26 | 6.38 |
| *19* | *1* | 209 | 2.87 | 4.78 | 7.66 | *50* | *1* | 642 | 1.25 | 3.74 | 7.63 |
| *19* | *5* | 91 | 2.20 | 1.10 | 5.49 | *50* | *5* | 91 | 5.49 | 1.10 | 7.69 |
| *20* | *1* | 259 | 1.93 | 1.54 | 6.95 | *51* | *1* | 724 | 1.93 | 0.83 | 7.32 |
| *20* | *5* | 67 | 1.49 | 10.45 | 26.87 | *51* | *2* | 351 | 3.99 | 1.99 | 8.26 |
| *21* | *1* | 205 | 3.41 | 2.93 | 8.78 | *51* | *3* | 67 | 0.00 | 4.48 | 4.48 |
| *21* | *2* | 60 | 5.00 | 5.00 | 15.00 | *52* | *1* | 423 | 3.07 | 1.65 | 8.04 |
| *21* | *3* | 17 | 5.88 | 5.88 | 11.76 | *52* | *2* | 131 | 2.29 | 1.53 | 7.63 |
| *22* | *1* | 236 | 2.54 | 3.39 | 8.47 | *52* | *3* | 80 | 3.75 | 0.00 | 3.75 |
| *22* | *2* | 72 | 5.56 | 4.17 | 11.11 | *55* | *1* | 901 | 2.66 | 0.78 | 7.21 |
| *22* | *3* | 18 | 0.00 | 11.11 | 22.22 | *55* | *2* | 109 | 5.50 | 2.75 | 9.17 |
| *23* | *1* | 77 | 5.19 | 0.00 | 7.79 | *55* | *3* | 37 | 2.70 | 2.70 | 5.41 |
| *23* | *5* | 30 | 3.33 | 10.00 | 13.33 | *60* | *1* | 488 | 1.43 | 2.05 | 6.97 |
| *24* | *1* | 156 | 1.28 | 1.92 | 6.41 | *60* | *2* | 148 | 1.35 | 4.05 | 7.43 |
| *24* | *2* | 145 | 2.07 | 4.83 | 8.28 | *60* | *3* | 78 | 1.28 | 2.56 | 6.41 |
| *24* | *3* | 64 | 4.69 | 1.56 | 7.81 | *61* | *1* | 30 | 6.67 | 6.67 | 16.67 |
| *25* | *1* | 204 | 2.45 | 1.96 | 6.86 | *61* | *5* | 37 | 2.70 | 2.70 | 5.41 |
| *25* | *2* | 88 | 2.27 | 2.27 | 4.55 | *62* | *6* | 31 | 0.00 | 6.45 | 6.45 |
| *25* | *3* | 32 | 3.13 | 3.13 | 6.25 | *63* | *1* | 310 | 2.90 | 1.61 | 8.06 |
| *26* | *1* | 365 | 3.84 | 0.82 | 7.95 | *63* | *2* | 185 | 0.54 | 2.16 | 7.03 |
| *26* | *2* | 111 | 0.90 | 1.80 | 4.50 | *63* | *3* | 68 | 0.00 | 1.47 | 2.94 |
| *26* | *3* | 39 | 0.00 | 2.56 | 2.56 | *64* | *1* | 41 | 2.44 | 4.88 | 7.32 |
| *27* | *1* | 131 | 3.82 | 2.29 | 6.87 | *64* | *5* | 15 | 0.00 | 0.00 | 100.00 |
| *27* | *2* | 135 | 0.74 | 3.70 | 6.67 | *65* | *1* | 198 | 5.56 | 0.51 | 8.08 |
| *27* | *3* | 36 | 0.00 | 8.33 | 8.33 | *65* | *2* | 224 | 6.70 | 1.79 | 8.93 |
| *28* | *1* | 692 | 3.18 | 2.75 | 8.09 | *65* | *3* | 125 | 4.00 | 5.60 | 9.60 |
| *28* | *2* | 290 | 2.07 | 1.03 | 6.55 | *66* | *1* | 28 | 21.43 | 10.71 | 35.71 |
| *28* | *3* | 46 | 2.17 | 6.52 | 13.04 | *66* | *2* | 43 | 6.98 | 0.00 | 6.98 |
| *29* | *1* | 293 | 0.34 | 3.75 | 7.17 | *66* | *3* | 17 | 5.88 | 5.88 | 11.76 |
| *29* | *2* | 200 | 3.50 | 1.50 | 8.50 | *67* | *1* | 165 | 2.42 | 6.06 | 9.09 |
| *29* | *3* | 107 | 2.80 | 2.80 | 7.48 | *67* | *5* | 26 | 0.00 | 3.85 | 15.38 |
| *30* | *6* | 83 | 2.41 | 6.02 | 8.43 | *70* | *1* | 113 | 5.31 | 0.88 | 7.08 |
| *31* | *1* | 195 | 3.08 | 3.08 | 6.67 | *70* | *5* | 19 | 0.00 | 5.26 | 5.26 |
| *31* | *2* | 97 | 3.09 | 1.03 | 6.19 | *71* | *6* | 91 | 4.40 | 10.99 | 15.38 |
| *31* | *3* | 40 | 2.50 | 2.50 | 5.00 | *72* | *1* | 341 | 1.47 | 2.64 | 7.33 |
| *32* | *1* | 78 | 2.56 | 2.56 | 5.13 | *72* | *2* | 137 | 3.65 | 4.38 | 8.03 |
| *32* | *2* | 45 | 4.44 | 0.00 | 8.89 | *72* | *3* | 50 | 2.00 | 2.00 | 4.00 |
| *32* | *3* | 14 | 0.00 | 0.00 | 100.00 | *73* | *1* | 65 | 6.15 | 4.62 | 10.77 |
| *33* | *1* | 108 | 3.70 | 4.63 | 9.26 | *73* | *5* | 29 | 0.00 | 3.45 | 3.45 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *33* | *2* | 64 | 1.56 | 1.56 | 6.25 | *74* | *1* | 673 | 1.93 | 1.34 | 7.43 |
| *33* | *3* | 18 | 0.00 | 0.00 | 100.00 | *74* | *2* | 424 | 2.12 | 0.47 | 8.49 |
| *34* | *1* | 89 | 2.25 | 4.49 | 6.74 | *74* | *3* | 186 | 2.15 | 3.23 | 6.99 |
| *34* | *2* | 79 | 6.33 | 0.00 | 6.33 | *742* | *1* | 211 | 4.27 | 1.90 | 9.00 |
| *34* | *3* | 49 | 2.04 | 8.16 | 10.20 | *742* | *5* | 54 | 3.70 | 3.70 | 14.81 |
| *35* | *1* | 104 | 0.00 | 0.96 | 4.81 | *743* | *1* | 103 | 3.88 | 1.94 | 8.74 |
| *35* | *2* | 43 | 9.30 | 0.00 | 9.30 | *743* | *5* | 20 | 0.00 | 20.00 | 20.00 |

**Table 1. Percentages of isolated units.**

## 5.2 Spontaneous Identification Based on Non-structural Information

These spontaneous identification scenarios are based on personal or highly specialized knowledge. Therefore, the re-identification risk may be assessed based on experts opinion, simulating the behaviour of specialized intruders. With respect to this scenario, a check performed by (survey) experts is highly recommended.

The first step consists in enumerating the confidential variables subject to spontaneous re-identification risk within well-defined combinations of categorical identifying variables. Only variables to be released should be taken into account.

For the Italian CIS4 survey the confidential variables that have to be checked are *RTOT* and its various components (*RRDINDX, RRDEXX, RMACX*, etc.) within each combination of *NACE2* and *EMPclass.* Also some information related to the first year of the reference period, *TURN2002* or *EMP2002*, might allow identification of some units.

In a second phase, for each combination of categorical identifying variables, a list of units at risk is produced. Also these particular units will be further subject to a selective protection method.

### *5.2.1 Dominance Scenario on Investment in Research*

For each combination of categorical identifying variables, the enterprises having the maximum *RTOT* value and being dominant with respect to a variable “V” at the same time should be identified. It is supposed that only large enterprise could have such characteristics. Consequently, only enterprises having more than 250 employees are studied in this scenario, but this definition could also depend on the economic structure of each member state. With respect to a single dominance variable “V”, an enterprise is considered to be dominant if the boxplot graphic of “V” identifies it as an extreme outlier. An upper extreme outlier is any data point observation which is 3*IQR higher than the third quartile (IQR is the interquartile range, the difference between the thrid and the first quantile). Moreover, a unit is dominant if it is classified as “isolated on the right tail” by the clustering algorithm. When the dominance is to be assessed with respect to several variables, an enterprise is defined as dominant when it is dominant with respect to at least one of the variables. The dominance variable(s) can be chosen only among the ones present in the files to be released because they do represent the information content achieved by a possible intruder from the anonymized data file. For example, considering a record in the file, an intruder could only link the maximum *RTOT* value with the perturbed *TURN* value and not with the original *TURN* value.

If it is deemed realistic to identify an enterprise only based on some *RTOT* component (*RMACX, RRDINX*, etc.), the same procedure should be applied for this component.

For the Italian CIS4 survey, as discussed at the end of the section 3.3, the dominance was assessed only with respect to *TURN* perturbed value. For this particular dataset, only *RTOT, RRDINX* and *RMARX* were considered as identifying. Table 2 shows the number of enterprises that could be identified in this scenario, over 112 non-empty combinations of *NACE2* and *EMPclass* (only enterprises with more than 250 employees were considered in this scenario). It should be specified that the total number of units considered at risk was 25.

| Variable | Number of units at risk |
|---|---|
| *RTOT* | 13 |
| *RRDINX* | 11 |
| *RMARX* | *14* |

**Table 2. Number of units at risk of re-identification in the dominance scenario on investment in research.**

### *5.2.2 Dominance Scenario Based on TURN at the First Year of the Reference Period of the Reference Period*

For each combination of categorical key variables, a large enterprise is considered at risk of re-identification if it is an upper extreme outlier and if, at the same time, it is the unique enterprise with this property. The upper extreme outliers were determined by means of the *TURN2002* values. An upper extreme outlier is any data point observation which is 3*IQR higher than the third quartile.

For the Italian CIS4 microdata file, this assessment was performed for each combination of *NACE2* and *EMPclass* variables. Only enterprises with more than 250 employees were taken into consideration. 12 enterprises were found to be at risk of re-identification in this scenario. The *TURN2002* values of these 12 enterprises should be perturbed (see section 6). Denote by *TURN2002*$^*$ the perturbed variable.

### *5.2.3 Dominance Scenario Based on Variations with Respect to the Beginning of the Reference Period*

For each combination of categorical key variables , the ratio between the perturbed *TURN* and *TURN2002*$^*$ is computed. The extreme outliers of this ratio are identified. In contrast with the other dominance scenarios, in this scenario, both lower and upper extreme outliers should be identified. If an unique lower extreme outlier is found, the corresponding enterprise is considered at risk of re-identification. If an unique upper extreme outlier is found, the corresponding enterprise is considered at risk of re-identification.

For the Italian CIS4 survey, 18 enterprises were considered as extreme outliers from the point of view of their variations *TURN*/TURN2002**, for a given combination of *NACE2* and *EMPclass*. This assessment was performed for each combination of categorical key variable with respect to large enterprises only (more than 250 employees).

Since the variables “number of employees” and “number of employees at the first year of the reference period” are both recoded, no variations could be computed based on their values.

*5.2.4 Uniqueness Scenario*

The risk of re-identification evaluation in this scenario is performed by counting, for each combination of the categorical identifying variables, those units having a direct sampling weight inferior to 1.5. If there is a single unit with this property, it is considered at risk of re-identification. Only large enterprises may be considered at risk of re-identification in this scenario. To perform any check, the direct weights should be available for the data protector.

Considering only enterprises with more than 250 employees, for the Italian CIS4 survey, no unit at risk of re-identification was found.

# *6. Microdata Protection*

Protection of microdata is achieved in several steps described in this section. Firstly, preliminary work on variables methods applied are considered, as also discussed in the previous sections. Secondly, two confusion based methods are applied to the *TURN* values of isolated units identified in the spontaneous identification scenario based on structural information. Finally, the *RTOT (RMACX, RRDINX, etc.)* values of the units considered at risk in the spontaneous identification scenario based on non-structural information are modified.

## 6.1 Global Recoding

The global recoding applied (section 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 Significant Digits

The rounding applied to *TURN* and *TURN2002* (paragraph 4.3) is a perturbation method.

## 6.3 Protection Against Spontaneous Identification Based on Structural Information

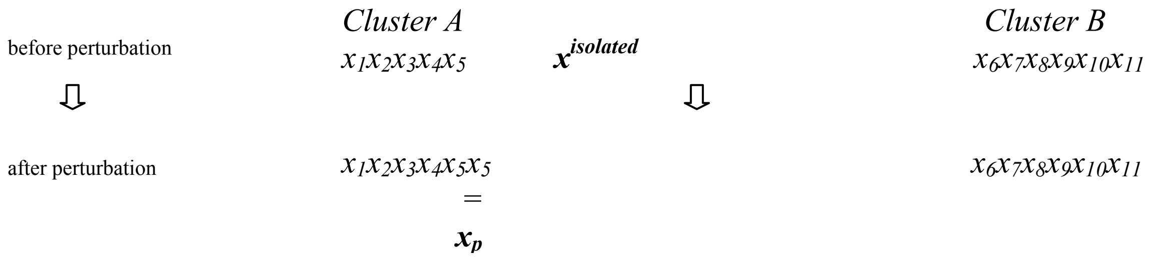
*6.3.1 Imputation from the Nearest Clustered Unit*

The isolated units identified by the clustering algorithm should be modified. Protection of these units is performed with respect to each combination of categorical identifying variables where DBSCAN is applied.

For the Italian CIS4 survey data, this perturbation was applied inside each combination of *NACE2* and *EMPclass*.

The perturbation, named "imputation from the nearest clustered unit", is an imputation with donor. That is, an isolated *TURN* value is replaced by the *TURN* value assumed by the closest (with respect to the same distance function used in the clustering algorithm) clustered unit. In this manner, the new *TURN* value will be confused with other values. It will be equal to at least another clustered value. Hence this unit would be considered as being confused with all the other units belonging to the same cluster (see section 5.1).

The imputation from the nearest clustered unit method is graphically presented below. "$x_i$" stands for the *i*-th clustered value, "$\boldsymbol{x}^{\boldsymbol{isolated}}$" for the value of an isolated unit while "$\boldsymbol{x_p}$" represents the perturbed value of "$\boldsymbol{x}^{\boldsymbol{isolated}}$".



The imputation from the nearest clustered unit is mathematically formulated below:

1. let $x^{isolated}$ be the value to be perturbed
2. find the closest clustered unit $x_p$:
   $x_p \; s.t. \quad d\left(x^{isolated}, x_p\right) = \min_{x_c \in \mathrm{C}}\left(d\left(x^{isolated}, x_c\right)\right)$ where $\mathrm{C}$ denotes the set of all clustered units

3. $x^{isolated}$ is replaced by $x_p$ in the microdata file to be released.

The imputation from the nearest clustered unit may be easily extended to a multivariate setting because the closest clustered unit is searched for with respect to a distance that can be generalized. It should be observed that values of donors may be higher or lower than the original values.

### *6.3.2 Microaggregation on Tails*

Since *TURN* variable has a skewed distribution for each combination of identifying variables, perturbation of the isolated *TURN* values on the tails by the imputation in section 6.3.1 may be very expensive in terms of information. Consequently, a microaggregation is performed for perturbing the isolated units on the tails. Since there is a single continuous identifying variable, microaggregation indeed reduces to individual ranking. This perturbation method was chosen because, like the clustering algorithm, it is based on the confusion principle, too (the *k*-anonymity principle). Microaggegation replaces each value by the average value of a group of units, creating groups of equal values: an intruder would not be able anymore to distinguish among the units.

Microaggregation should be performed on the original values of the isolated units on the tails. Microaggregation should be applied on each tail of the *TURN* distribution inside each combination of the categorical identifying variables where the clustering algorithm was used.

For the Italian CIS4 survey, it was applied inside each combination of *NACE2* and *EMPclass*.

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 greater number of units. The average of each group is then computed and each group member *TURN* value is replaced by the group mean. If the number of isolated units on the tail is between $k$ and $2k$, all these units are averaged and their *TURN* values are replaced by the mean. If 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 was deemed a sufficient microaggregation parameter value for the Italian CIS4 survey data.

This microaggregation on tails step has the double aim to protect the isolated tails units and, at the same time, to avoid overprotection (avoiding the imputation from the nearest clustered unit when such cluster is too distant from the isolated units). On each tail, it may be formulated as:

1. Count the number $N_t$ of isolated units on the tail
2. If $N_t \geq 2k$
   i. Determine groups of minimum $k$ isolated units. Considering only the *TURN* variable, simply sort the units and take groups of $k$ units. The last group may have a greater number of units.
   ii. Let $x_1, x_2, \ldots, x_g$ be the original *TURN* values in a group. Compute their arithmetic mean $\bar{x} = \frac{1}{g}\sum_{i=1}^{g} x_i$ .

iii. The perturbed *TURN* values are then $x_g^* = \bar{x}, i = 1,\ldots,g$

3. If $k \le N_t < 2k$

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. The perturbed *TURN* values are then $x_i^* = \bar{x}, i = 1,\ldots,N_t$.

4. If $N_t < k$

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 $c$ and its corresponding value $x_c$.

viii. The perturbed *TURN* values are then $x_i^* = x_c, i = 1,\ldots,N_t$ .

*TURN* perturbed values should be rounded to the same number of decimal digits as the original variable.

## 6.4 Protection Against Spontaneous Identification Based on Non-structural Information

Variables to be released which may be subject to spontaneous identification should be protected, too.

As these protections are based on the perturbed *TURN* values, it should be applied after the application of the adjustment described in section 7. Since a dominant unit is an isolated one, its *TURN* value might be modified by the adjustment procedure.

### *6.4.1 Dominance scenario based on investment in research*

*RTOT* and its components should be protected against this dominance scenario. As previously discussed, a check performed by survey experts is recommended. Generally, the number of units to be protected against spontaneous identification is very reduced. Hence a selective protection should be applied to all units considered at risk in this scenario. For all these units, all the components of *RTOT*, and *RTOT* itself, should be modified simultaneously in order to preserve the relationships among these variables. To preserve as much as possible the relationship between *RTOT* and *TURN*, these *RTOT* values are changed proportionally to the modification introduced in the corresponding *TURN* values. That is, the new value *RTOT*$^*$ is calculated as $RTOT^* = \frac{TURN^*}{TURN} RTOT$ , where *TURN*$^*$ denotes the perturbed value of *TURN.*

### *6.4.2 Dominance Scenario Based on TURN at the First Year of the Reference Period of the Reference Period*

Variable *TURN2002* should be perturbed to guarantee that no enterprise could be identified using the turnover values with respect to the information on the beginning of the reference period. *TURN2002* is more obsolete than *TURN*, hence less perturbation should be given to it.

In order to preserve the distribution of the ratio *TURN2002/TURN*, in a first step, *TURN2002* is perturbed. The perturbation is proportional to the perturbation of the ratio *TURN2002/TURN*. That is:

$$TURN2002^{\circ} = \frac{TURN2002}{TURN} * TURN^{*}$$

$$TURN2002^{*} = TURN2002^{\circ}$$

Then, for each combination of the categorical key variable containing enterprises at risk in this scenario, an individual ranking is performed on the right tail of the distribution. In order to reduce the information loss, only the group of the largest *TURN2002* values should be micro-aggregated. Denote by *TURN2002*$^*$ the perturbed *TURN2002* variable.

For the Italian CIS4 microdata file, for each combination of *NACE2* and *EMPclass* containing enterprises at risk in this dominance scenario, the largest $t = 3$ *TURN2002* values were micro-aggregated.

### *6.4.3 Dominance Scenario Based on Variations with Respect to the First Year of the Reference Period*

Variable *TURN2002* should be perturbed to guarantee that no enterprise could be identified using the turnover values with respect to the information on the beginning of the reference period. *TURN2002* is more obsolete than *TURN*, hence less perturbation should be given to it.

For each combination of categorical key variables,

a) Compute the ratio $R = TURN2002^{*}/TURN^{*}$.

b) For $R$, microaggregate:

   a. The smallest $p$ values and, independently,

   b. The largest $p$ values

Denote by $R^*$ the new variable.

c) Finally, compute the perturbed *TURN2002*$^{**}$ variable derives as:

$$TURN2002^{**} = R^{*} * TURN^{*}$$

For the Italian CIS4 survey, these steps were applied using $p = 3$ for each combination of *NACE2* and *EMPclass* containing enterprises at risk in this scenario.

*TURN2002*$^{**}$ values should be rounded to the same number of decimal digits as the original variable.

### *6.4.4 Uniqueness Scenario*

Since the uniqueness scenario is not based on “isolated” points, the perturbation of the *RTOT* values cannot be derived only from the perturbation introduced in the corresponding *TURN* value. This happens because this *TURN* value might not be an isolated or modified value. Considering that the number of the units at risk identified in the uniqueness scenario should be extremely reduced,  ,these *RTOT* values are changed with respect to the mean perturbation introduced in the *TURN* values.

1. let $RTOT_i$ be the value assumed by *RTOT* on an unit identified at risk of re-identification in the uniqueness scenario
2. if the *TURN*$_i$ corresponding value is changed into $TURN_i^*$ by the perturbation on *TURN*, then $RTOT_i^* = \frac{TURN_i^*}{TURN_i} RTOT_i$
3. otherwise, find the combination of categorical identifying variables to which the unit belongs to. Compute the average $A$ of the ratios $\frac{TURN^*}{TURN}$ in this category. If $A \neq 1$, then $RTOT^* = A * RTOT$ ; otherwise go to step 4.
4. compute the ratios average, but with respect to the categories of the most important categorical identifying variable (*NACE2* in case of CIS4 Italy).

The values of the *RTOT* components considered in any spontaneous identification scenario should be modified in the same manner, on the same enterprises.

The perturbed values of *RTOT* and its components should be rounded to the same number of decimal digits as the original variables.

## *7. Adjustment to Preserve the Published 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 the categorical identifying variables, *NACE2* and *EMPclass* for the Italian CIS4 microdata file. 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 the identifying variables, let $x_i, i = 1, \ldots, n$ and $w_i, i = 1, \ldots, n$ be the original (non perturbed) *TURN* values and weights respectively. Then $T = \sum_{i=1}^{n} x_i w_i$ is the total to be preserved. Let $x_i^*, i = 1, \ldots, n$ be the *TURN* perturbed values, according to the previously described perturbation steps. It should be observed that the applied perturbation method does not change the sampling 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 tail is not sufficient, as follows:

$x^{**}_{n-i+1} = x^{*}_{n-i+1} + \frac{D}{\sum_{j=1}^{k_1} w_{n-j+1}}, i = 1, \ldots, k_1$. Replacing $x^{*}_{n-i+1}$ by $x^{**}_{n-i+1}, i = 1, \ldots, k_1$, one can easily check that $T = \sum_{i=1}^{n} x_i w_i = \sum_{i=1}^{n} x_i^{**} w_i$ .

The value of $k_1$ might be chosen independently on the microaggregation parameter $k$ and on clustering parameter *MinPts*. It should depend on both number of points in the studied cross-classification cell and on the number of identified isolated units, but an unique value for all the cross-classification cells is suggested.

For the Italian CIS4 survey microdata anonymized file, $k_1 = 3$ was used.

Using the perturbation procedure described in sectio 6.5, the weighted totals of *RTOT* (and its components) with respect to the combinations of categorical identifying key variables are not preserved. Since the anonymized file is released for research purposes, it might be considered that the missed preservation of these totals would not involve any disclosure. If considered really necessary, the above presented adjustment procedure could be easily applied to *RTOT* and its identifying components. One should be aware that *RTOT* and its components are the most important from a user point of view and that the adjustment could modify the most interesting part of the studied phenomenon.

For the Italian CIS4 survey microdata anonymized file, the adjustment to preserve the *RTOT* totals was not performed.

## *8. Audit Strategies*

Protection being achieved by means of perturbation methods, an audit is required. This section gives some hints on the ways to check for negative values, insufficient protection and overprotection.

### 8.1 Check for Negative Values

Due to the adjustment applied for totals preservation (section 7), negative *TURN* values might result. Since the largest *TURN* values are generally nonnegative, for sake of coherence, care should be paid in order not to obtain negative values. Before perturbing the RTOT values, a check should be performed to reveal those negative units. If in a given combination of categorical identifying variables there are some negative *TURN* values, its total adjustment should be undone and restarted, considering additional groups (the same parameter $k_1$) of units on which distribute the difference $D$. In case it is still not possible to preserve the weighted total obtaining at the same time nonnegative values, preservation of only some aggregated weighted total with nonnegative values is preferred. If this trial fails too, the total is not preserved and the protection of isolated units stops before the microaggregation step (the isolated units on the right tail being perturbed only by the imputation from the nearest clustered unit).

For example, in the Italian CIS4 survey, this situation happen for *NACE2* = 29 and *EMPclass* = 1. The difference was distributed on more groups of units and the total was preserved also with respect to the *NACE2* = 29 and *EMPclass*=1.

### 8.2 Check for Insufficient Perturbation

If the number of isolated units is very small, the units on the right tail might achieve an insufficient perturbation. This is more likely to occur for the isolated units on the right tail assuming unique values with respect to the already protected *TURN* variable. The number of isolated units per combination of categorical variables should anyway be monitored, see for example the Table 1. Moreover, the minimum absolute relative perturbation should give a good indication on the presence of units that received insufficient perturbation (extreme  values are immediately observed). If this situation occurs, the protection procedure should be applied in a different manner inside the indicated combination of categorical identifying variables. The microaggregation step is changed: instead of applying it only to isolated points, the microaggregation should be applied on the highest $k$ *TURN* values, eventually increasing the number of groups. Since these situations are not likely to occur, an automatic data protection and audit procedures is probably the best choice.

For *NACE2* = 25, *EMPclass* = 1 (Italian CIS4 survey data), there were only four isolated units on the right tail. Due to the microaggregation step and to the adjustment for weighted totals preservation, one of this values was not sufficiently perturbed.The microaggregation was applied also on non-isolated units.

### 8.3 Check for Overprotection

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 the adjustment used for the preservation of the weighted totals. 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 is more likely to occur on the isolated points on the left tail assuming unique values on the already protected *TURN* variable. To monitor these situations, the maximum absolute relative perturbation should give a good indication on the presence of units that changed too much their values. In this case, the adjustment is applied only on units on the right tail, even if their number is lower than $k_1$. Note that only the number of isolated units can be lower than $k_1$. For the Italian CIS4 survey, this situation did not occur.

## *9. Information Loss and Information Preservation*

Protection methods unavoidably change the informative content of a microdata file. In this section some hints on how to evaluate the information loss in case of CIS microdata are given. There is no universal definition of data quality, but the impact of the perturbation method on several statistical indicators could be assessed.

The weighted totals (of *TURN*) are preserved for each combination of *NACE2* and *EMPclass* variables, if possible (see section 6.5). Otherwise, they are preserved for each *NACE2* category only.

For the Italian CIS4, all weighted totals of *TURN* were exactly preserved, for all combinations of the categorical identifying variables.

### 9.1 Number of Modified *TURN* Values

With respect to the spontaneous identification scenario based on structural information, the only perturbed variable is *TURN*. The perturbation is applied only to the isolated units. All clustered units are to be released with their original values if not used in the adjustment procedure (section 7). By imputing an isolated value by the nearest clustered *TURN* observation, the information loss is not too large and protection is guaranteed. Moreover, the microaggregation applied on the distribution tails has mainly the same effects. Table 3 presents, for each combination of *NACE2* and *EMPclass*, the number of modified *TURN* values, for the Italian CIS4 survey.

| ***Nace2*** | ***EMPclass*** | **Observations** | **Percentage of modified units** | ***Nace2*** | ***EMPclass*** | **Observations** | **Percentage of modified units** |
|---:|---:|:---:|:---:|---:|---:|:---:|:---:|
| *10* | *1* | 297 | 7.41 | *35* | *3* | 16 | 18.75 |
| *10* | *5* | 38 | 13.16 | *36* | *1* | 212 | 7.08 |
| *15* | *1* | 428 | 7.71 | *36* | *2* | 100 | 10.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *15* | *2* | 160 | 8.75 | *36* | *3* | 25 | 24.00 |
| *15* | *3* | 63 | 12.70 | *37* | *6* | 106 | 7.55 |
| *17* | *1* | 299 | 5.35 | *40* | *1* | 167 | 7.78 |
| *17* | *2* | 94 | 8.51 | *40* | *2* | 82 | 8.54 |
| *17* | *3* | 43 | 6.98 | *40* | *3* | 38 | 7.89 |
| *18* | *1* | 336 | 6.55 | *45* | *1* | 4756 | 8.31 |
| *18* | *2* | 97 | 8.25 | *45* | *2* | 514 | 8.17 |
| *18* | *3* | 26 | 15.38 | *45* | *3* | 47 | 8.51 |
| *19* | *1* | 209 | 7.66 | *50* | *1* | 642 | 7.63 |
| *19* | *5* | 91 | 7.69 | *50* | *5* | 91 | 9.89 |
| *20* | *1* | 259 | 6.95 | *51* | *1* | 724 | 7.32 |
| *20* | *5* | 67 | 26.87 | *51* | *2* | 351 | 8.26 |
| *21* | *1* | 205 | 8.78 | *51* | *3* | 67 | 4.48 |
| *21* | *2* | 60 | 15.00 | *52* | *1* | 423 | 8.04 |
| *21* | *3* | 17 | 23.53 | *52* | *2* | 131 | 8.40 |
| *22* | *1* | 236 | 8.47 | *52* | *3* | 80 | 7.50 |
| *22* | *2* | 72 | 11.11 | *55* | *1* | 901 | 7.21 |
| *22* | *3* | 18 | 27.78 | *55* | *2* | 109 | 9.17 |
| *23* | *1* | 77 | 11.69 | *55* | *3* | 37 | 10.81 |
| *23* | *5* | 30 | 13.33 | *60* | *1* | 488 | 6.97 |
| *24* | *1* | 156 | 6.41 | *60* | *2* | 148 | 7.43 |
| *24* | *2* | 145 | 8.28 | *60* | *3* | 78 | 7.69 |
| *24* | *3* | 64 | 10.94 | *61* | *1* | 30 | 20.00 |
| *25* | *1* | 204 | 7.84 | *61* | *5* | 37 | 10.81 |
| *25* | *2* | 88 | 5.68 | *62* | *6* | 31 | 9.68 |
| *25* | *3* | 32 | 12.50 | *63* | *1* | 310 | 8.06 |
| *26* | *1* | 365 | 7.95 | *63* | *2* | 185 | 7.03 |
| *26* | *2* | 111 | 5.41 | *63* | *3* | 68 | 5.88 |
| *26* | *3* | 39 | 7.69 | *64* | *1* | 41 | 9.76 |
| *27* | *1* | 131 | 6.87 | *64* | *5* | 15 | 100.00 |
| *27* | *2* | 135 | 6.67 | *65* | *1* | 198 | 9.09 |
| *27* | *3* | 36 | 8.33 | *65* | *2* | 224 | 8.93 |
| *28* | *1* | 692 | 8.09 | *65* | *3* | 125 | 9.60 |
| *28* | *2* | 290 | 6.55 | *66* | *1* | 28 | 35.71 |
| *28* | *3* | 46 | 13.04 | *66* | *2* | 43 | 13.95 |
| *29* | *1* | 293 | 7.17 | *66* | *3* | 17 | 23.53 |
| *29* | *2* | 200 | 8.50 | *67* | *1* | 165 | 9.09 |
| *29* | *3* | 107 | 7.48 | *67* | *5* | 26 | 23.08 |
| *30* | *6* | 83 | 8.43 | *70* | *1* | 113 | 8.85 |
| *31* | *1* | 195 | 6.67 | *70* | *5* | 19 | 15.79 |
| *31* | *2* | 97 | 8.25 | *71* | *6* | 91 | 15.38 |
| *31* | *3* | 40 | 10.00 | *72* | *1* | 341 | 7.33 |
| *32* | *1* | 78 | 6.41 | *72* | *2* | 137 | 8.03 |
| *32* | *2* | 45 | 15.56 | *72* | *3* | 50 | 8.00 |
| *32* | *3* | 14 | 100.00 | *73* | *1* | 65 | 10.77 |
| *33* | *1* | 108 | 9.26 | *73* | *5* | 29 | 10.34 |
| *33* | *2* | 64 | 9.38 | *74* | *1* | 673 | 7.43 |
| *33* | *3* | 18 | 0.00 | *74* | *2* | 424 | 8.73 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *34* | *1* | 89 | 6.74 | *74* | *3* | 186 | 6.99 |
| *34* | *2* | 79 | 10.13 | *742* | *1* | 211 | 8.53 |
| *34* | *3* | 49 | 10.20 | *742* | *5* | 54 | 16.67 |
| *35* | *1* | 104 | 6.73 | *743* | *1* | 103 | 9.71 |
| 35 | 2 | 43 | 16.28 | *743* | *5* | 20 | 20.00 |

**Table 3. Number of modified units.**

## 9.2 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 is not necessarily observed. A comparison between the variances of the original and perturbed variables is suggested. This comparison should be performed for each combination of categorical identifying variables where microaggregation is applied, in this case, for each combination of *NACE2* and *EMPclass*. The ratios between the *TURN* variance after, $\sigma^*$, and before, $\sigma$, protection are shown in Table 4 and in Figure 2.

| **Nace2** | **EMPclass** | $\sigma^*/\sigma$ | **Nace2** | **EMPclass** | $\sigma^*/\sigma$ | **Nace2** | **EMPclass** | $\sigma^*/\sigma$ | **Nace2** | **EMPclass** | $\sigma^*/\sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *10* | *1* | 1.03 | *25* | *3* | 1.10 | *35* | *3* | 0.91 | *63* | *1* | 1.01 |
| *10* | *5* | 0.37 | *26* | *1* | 1.00 | *36* | *1* | 0.93 | *63* | *2* | 0.82 |
| *15* | *1* | 1.04 | *26* | *2* | 1.08 | *36* | *2* | 0.95 | *63* | *3* | 0.99 |
| *15* | *2* | 1.13 | *26* | *3* | 0.79 | *36* | *3* | 0.29 | *64* | *1* | 1.09 |
| *15* | *3* | 0.77 | *27* | *1* | 1.03 | *37* | *6* | 0.93 | *64* | *5* | 1.04 |
| *17* | *1* | 1.16 | *27* | *2* | 0.76 | *40* | *1* | 1.02 | *65* | *1* | 1.00 |
| *17* | *2* | 1.03 | *27* | *3* | 0.79 | *40* | *2* | 0.54 | *65* | *2* | 0.89 |
| *17* | *3* | 0.93 | *28* | *1* | 1.01 | *40* | *3* | 0.25 | *65* | *3* | 0.92 |
| *18* | *1* | 1.05 | *28* | *2* | 1.23 | *45* | *1* | 1.00 | *66* | *1* | 0.71 |
| *18* | *2* | 0.87 | *28* | *3* | 1.02 | *45* | *2* | 1.01 | *66* | *2* | 1.00 |
| *18* | *3* | 0.56 | *29* | *1* | 0.63 | *45* | *3* | 0.93 | *66* | *3* | 0.35 |
| *19* | *1* | 0.92 | *29* | *2* | 1.11 | *50* | *1* | 1.01 | *67* | *1* | 1.03 |
| *19* | *5* | 0.60 | *29* | *3* | 1.03 | *50* | *5* | 1.00 | *67* | *5* | 0.87 |
| *20* | *1* | 0.99 | *30* | *6* | 0.29 | *51* | *1* | 0.89 | *70* | *1* | 0.80 |
| *20* | *5* | 0.92 | *31* | *1* | 1.02 | *51* | *2* | 0.88 | *70* | *5* | 0.75 |
| *21* | *1* | 0.98 | *31* | *2* | 0.96 | *51* | *3* | 0.74 | *71* | *6* | 0.54 |
| *21* | *2* | 0.96 | *31* | *3* | 0.83 | *52* | *1* | 1.10 | *72* | *1* | 1.08 |
| *21* | *3* | 0.51 | *32* | *1* | 1.10 | *52* | *2* | 1.01 | *72* | *2* | 0.99 |
| *22* | *1* | 0.98 | *32* | *2* | 0.96 | *52* | *3* | 1.00 | *72* | *3* | 0.96 |
| *22* | *2* | 1.07 | *32* | *3* | 0.35 | *55* | *1* | 0.95 | *73* | *1* | 1.05 |
| *22* | *3* | 0.87 | *33* | *1* | 0.88 | *55* | *2* | 0.78 | *73* | *5* | 0.99 |
| *23* | *1* | 1.01 | *33* | *2* | 1.06 | *55* | *3* | 0.53 | *74* | *1* | 1.07 |
| *23* | *5* | 0.97 | *33* | *3* | 1.00 | *60* | *1* | 0.99 | *74* | *2* | 1.05 |
| *24* | *1* | 0.95 | *34* | *1* | 0.92 | *60* | *2* | 0.76 | *74* | *3* | 0.80 |
| *24* | *2* | 0.98 | *34* | *2* | 1.02 | *60* | *3* | 0.44 | *742* | *1* | 0.76 |
| *24* | *3* | 0.39 | *34* | *3* | 0.37 | *61* | *1* | 0.62 | *742* | *5* | 0.86 |
| *25* | *1* | 0.97 | *35* | *1* | 1.13 | *61* | *5* | 0.48 | *743* | *1* | 1.04 |
| *25* | *2* | 0.84 | *35* | *2* | 1.00 | *62* | *6* | 0.29 | *743* | *5* | 0.84 |

**Table 4. Ratios between the *TURN* variances.**

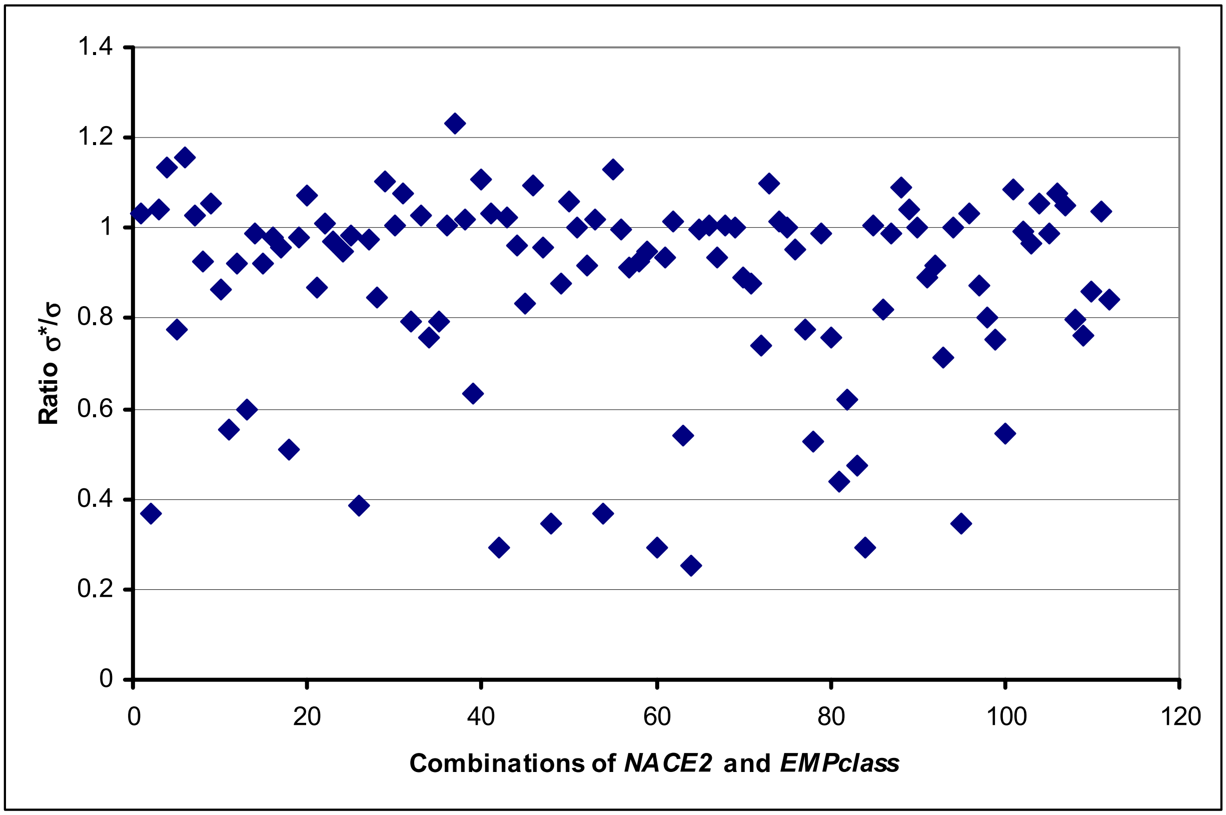


**Figure 2. Variance comparison of *TURN* before and after protection.**

## 9.3 Data Utility: Correlations Comparison

Correlations may also be compared to assess the degree of information loss. In Table 5 the correlations between the original and perturbed *TURN* values, between original *TURN* and *RTOT* and between perturbed *TURN* and *RTOT* are shown.

| *Nace2* | *EMPclass* | cor(Turn,TurnProtected) | cor(RTot,RTotProtected) | cor(Turn,RTot) | cor(TurnProtected,RTotProtected) | *Nace2* | *EMPclass* | cor(Turn,TurnProtected) | cor(RTot,RTotProtected) | cor(Turn,RTot) | cor(TurnProtected,RTotProtected) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *10* | *1* | 1 | 1 | 0.2 | 0.19 | *35* | *3* | 0.85 | 1 | 0.51 | 0.49 |
| *10* | *5* | 0.56 | 0.98 | 0.99 | 0.56 | *36* | *1* | 0.97 | 1 | 0.26 | 0.26 |
| *15* | *1* | 0.98 | 1 | 0.1 | 0.09 | *36* | *2* | 0.98 | 1 | 0.39 | 0.43 |
| *15* | *2* | 0.98 | 1 | 0.16 | 0.18 | *36* | *3* | 0.81 | 1 | 0.88 | 0.48 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 3 | 0.94 | 0.99 | 0.66 | 0.48 | 37 | 6 | 1 | 1 | 0.14 | 0.16 |
| 17 | 1 | 0.96 | 1 | 0.18 | 0.19 | 40 | 1 | 0.99 | 1 | 0.11 | 0.11 |
| 17 | 2 | 0.99 | 1 | 0.29 | 0.31 | 40 | 2 | 0.62 | 1 | -0.03 | -0.05 |
| 17 | 3 | 0.88 | 1 | 0.11 | 0.17 | 40 | 3 | 0.76 | 0.94 | 0.94 | 0.55 |
| 18 | 1 | 0.95 | 1 | 0.16 | 0.15 | 45 | 1 | 1 | 1 | 0.09 | 0.09 |
| 18 | 2 | 0.99 | 1 | 0.48 | 0.47 | 45 | 2 | 0.97 | 1 | 0.02 | 0.02 |
| 18 | 3 | 0.88 | 1 | 0.13 | 0.2 | 45 | 3 | 0.98 | 1 | 0.04 | 0.03 |
| 19 | 1 | 0.98 | 1 | 0.15 | 0.15 | 50 | 1 | 0.99 | 1 | 0.01 | 0.02 |
| 19 | 5 | 0.89 | 1 | 0.54 | 0.65 | 50 | 5 | 1 | 1 | 0.12 | 0.13 |
| 20 | 1 | 0.99 | 1 | 0.19 | 0.19 | 51 | 1 | 0.98 | 1 | 0.04 | 0.04 |
| 20 | 5 | 0.97 | 1 | 0.24 | 0.25 | 51 | 2 | 0.97 | 1 | 0.04 | 0.04 |
| 21 | 1 | 1 | 1 | 0.24 | 0.25 | 51 | 3 | 0.9 | 1 | 0.07 | 0.09 |
| 21 | 2 | 0.99 | 1 | 0.28 | 0.29 | 52 | 1 | 0.99 | 1 | 0.23 | 0.21 |
| 21 | 3 | 0.77 | 0.97 | 0.33 | 0.04 | 52 | 2 | 1 | 1 | 0.1 | 0.1 |
| 22 | 1 | 0.93 | 1 | 0.1 | 0.09 | 52 | 3 | 1 | 1 | 0.17 | 0.17 |
| 22 | 2 | 1 | 1 | 0.26 | 0.24 | 55 | 1 | 0.99 | 1 | 0.06 | 0.07 |
| 22 | 3 | 0.93 | 1 | 0.47 | 0.58 | 55 | 2 | 0.94 | 1 | -0.01 | -0.01 |
| 23 | 1 | 1 | 1 | 0.48 | 0.47 | 55 | 3 | 0.87 | 1 | 0.21 | 0.48 |
| 23 | 5 | 0.91 | 0.98 | 0.15 | 0.54 | 60 | 1 | 0.99 | 1 | 0.12 | 0.13 |
| 24 | 1 | 0.99 | 1 | 0.26 | 0.28 | 60 | 2 | 0.93 | 1 | -0.02 | -0.02 |
| 24 | 2 | 0.98 | 1 | 0.62 | 0.56 | 60 | 3 | 0.79 | 0.99 | 0.37 | 0.39 |
| 24 | 3 | 0.74 | 0.98 | 0.31 | 0.12 | 61 | 1 | 0.94 | 1 | -0.11 | -0.12 |
| 25 | 1 | 1 | 1 | 0.1 | 0.1 | 61 | 5 | 0.84 | 1 | 0.07 | 0.21 |
| 25 | 2 | 0.99 | 1 | 0.41 | 0.28 | 62 | 6 | 0.58 | 0.98 | 0.99 | 0.52 |
| 25 | 3 | 0.75 | 1 | 0.06 | 0.26 | 63 | 1 | 1 | 1 | 0.19 | 0.19 |
| 26 | 1 | 1 | 1 | 0.28 | 0.28 | 63 | 2 | 0.95 | 1 | 0.04 | 0.05 |
| 26 | 2 | 0.98 | 1 | 0.19 | 0.26 | 63 | 3 | 0.99 | 1 | 0.29 | 0.32 |
| 26 | 3 | 0.93 | 1 | 0.1 | 0.07 | 64 | 1 | 0.9 | 1 | 0.75 | 0.48 |
| 27 | 1 | 1 | 1 | 0.32 | 0.31 | 64 | 5 | 0.93 | 0.98 | 0.66 | 0.63 |
| 27 | 2 | 0.89 | 1 | 0.27 | 0.18 | 65 | 1 | 0.99 | 1 | 0.2 | 0.19 |
| 27 | 3 | 0.9 | 1 | 0.07 | 0.07 | 65 | 2 | 0.97 | 1 | 0.04 | 0.06 |
| 28 | 1 | 1 | 1 | 0.26 | 0.26 | 65 | 3 | 0.96 | 1 | 0.44 | 0.43 |
| 28 | 2 | 0.95 | 1 | 0.26 | 0.3 | 66 | 1 | 0.94 | 1 | 0.06 | 0.07 |
| 28 | 3 | 1 | 1 | 0.22 | 0.22 | 66 | 2 | 1 | 1 | 0.3 | 0.3 |
| 29 | 1 | 0.95 | 1 | 0.07 | 0.1 | 66 | 3 | 0.73 | 1 | 0.65 | 0.27 |
| 29 | 2 | 0.99 | 1 | 0.11 | 0.11 | 67 | 1 | 1 | 1 | 0.29 | 0.28 |
| 29 | 3 | 0.98 | 1 | 0.73 | 0.74 | 67 | 5 | 0.96 | 1 | 0.58 | 0.47 |
| 30 | 6 | 0.55 | 0.68 | 0.42 | 0.46 | 70 | 1 | 0.93 | 1 | 0.1 | 0.14 |
| 31 | 1 | 0.99 | 1 | 0.36 | 0.35 | 70 | 5 | 0.93 | 1 | 0.77 | 0.55 |
| 31 | 2 | 1 | 1 | 0.38 | 0.38 | 71 | 6 | 0.88 | 0.96 | 0.47 | 0.28 |
| 31 | 3 | 0.94 | 0.98 | 0.68 | 0.47 | 72 | 1 | 0.99 | 1 | 0.18 | 0.17 |
| 32 | 1 | 0.96 | 1 | 0.56 | 0.48 | 72 | 2 | 0.99 | 1 | 0.31 | 0.33 |
| 32 | 2 | 1 | 1 | 0.42 | 0.43 | 72 | 3 | 0.99 | 1 | 0.57 | 0.63 |
| 32 | 3 | 0.64 | 1 | 0.99 | 0.55 | 73 | 1 | 0.97 | 1 | 0.48 | 0.43 |
| 33 | 1 | 0.95 | 1 | 0.48 | 0.64 | 73 | 5 | 0.98 | 1 | 0.92 | 0.88 |
| 33 | 2 | 0.97 | 1 | 0.47 | 0.58 | 74 | 1 | 0.98 | 1 | 0.1 | 0.1 |
| 33 | 3 | 1 | 1 | 0.34 | 0.34 | 74 | 2 | 0.99 | 1 | 0.12 | 0.11 |
| 34 | 1 | 0.97 | 1 | 0.17 | 0.18 | 74 | 3 | 0.95 | 0.96 | 0.37 | 0.55 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *34* | *2* | 1 | 1 | 0.18 | 0.17 | *742* | *1* | 0.84 | 1 | 0.16 | 0.16 |
| *34* | *3* | 0.67 | 1 | 0.36 | 0.48 | *742* | *5* | 0.93 | 1 | 0.19 | 0.22 |
| *35* | *1* | 0.98 | 1 | 0.02 | 0.02 | *743* | *1* | 0.98 | 1 | 0.01 | 0 |
| *35* | *2* | 1 | 1 | 0.23 | 0.23 | *743* | *5* | 0.91 | 1 | 0.57 | 0.57 |

**Table 5. Coefficients of correlations.**

## 9.4 Variations with Respect to the First Year of the Reference Period

The impact of the perturbation applied to *TURN2002* was assessed by means of the ratios between *TURN2002* and *TURN*, since this is the main usage of the variable *TURN2002*. For the Italian CIS4 microdata file, the Table 6 presents the comparison between some statistical indicators assessed on original and perturbed data. Only the combinations of *NACE2* and *EMPclass* that were changed are listed. The other 82% of combinations was unchanged from the point of view of the distribution of *TURN2002*/*TURN*.

| ***Nace2*** | ***Empclass*** | **NbObs** | **MeanPerturbedOverMeanOrig** | **VariancePerturbedOverVarianceOrig** | **CorrPerturbedOrig** |
|---|---|---|---|---|---|
| 33 | 3 | 18 | 1.02 | 1.53 | 0.84 |
| 66 | 3 | 17 | 1.06 | 1.32 | 0.97 |
| 22 | 3 | 18 | 0.97 | 1.17 | 0.84 |
| 67 | 5 | 26 | 0.98 | 1.13 | 0.98 |
| 742 | 5 | 54 | 0.99 | 1.10 | 0.99 |
| 24 | 3 | 64 | 1.00 | 1.05 | 0.99 |
| 20 | 5 | 67 | 1.00 | 1.02 | 0.99 |
| 55 | 3 | 37 | 1.00 | 1.01 | 0.98 |
| 35 | 3 | 16 | 0.99 | 1.01 | 1.00 |
| 36 | 3 | 25 | 1.00 | 1.00 | 0.97 |
| 18 | 3 | 26 | 1.00 | 1.00 | 0.99 |
| 45 | 3 | 47 | 1.00 | 0.96 | 0.98 |
| 32 | 3 | 14 | 1.00 | 0.86 | 0.93 |
| 15 | 3 | 63 | 1.00 | 0.86 | 0.96 |
| 28 | 3 | 46 | 1.01 | 0.73 | 0.56 |
| 21 | 3 | 17 | 0.99 | 0.48 | 0.78 |
| 17 | 3 | 43 | 0.98 | 0.44 | 0.81 |

| | | | | | |
|---|---|---|---|---|---|
| 34 | 3 | 49 | 0.89 | 0.09 | 0.66 |
| 50 | 5 | 91 | 0.47 | 0.00 | 0.48 |

**Table 6. Statistical indicators on TURN2002/TURN perturbation**

## 9.5 Data Utility: Users Perspective

The data utility is measured with respect to the *RTOT* and some of its components, for example, *RRDINDX, RRDEXX, RMACX*. Since *TURN*, *RTOT* and its components are the only perturbed variables, 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* and *RTOT*, but experts suggestions are very useful in such simulations. For the Italian CIS4 survey data, the quantiles of the above mentioned ratio variables were compared. Generally, a very good agreement was observed. Usually, the maximum (over the combinations of *NACE2* and *EMPclass*) absolute difference was lower than 0.4. Except for *NACE2* = 72, *EMPclass* = 2, 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 2.8. For *NACE2* = 72, *EMPclass* = 2, the maximum absolute differences between quantiles was above 30.

## *10. Concluding Remarks*

In this paper it was proved that a detailed analysis of possible disclosure scenario and the definition of related identifying variables coupled with a careful risk assessment to detect real units at risk is suitable to address the real risks of disclosure of a microdata file. This strategy of selective identification of risk allows for selective protection methods that can save more information content of the data than a generalised application of any perturbation method. The inclusion of different scenarios is a key issue.

The 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.

### 10.1. Annex

According to the statistical disclosure control methodology, the modified variables are presented in Table 7.

| **Description** | **Variable** | **Anonymisation summary** |
|---|---|---|
| Name of the enterprise | Id | Removed |
| Address | Nuts | Changed |

| | | |
|---|---|---|
| Main activity | Nace | Changed |
| Country of head office | Ho | Changed |
| Total turnover in 2002 | Turn2002 | Changed |
| Total turnover in 2004 | Turn | Changed |
| Total number of employees in 2002 | Emp2002 | Changed |
| Total number of employees in 2004 | Emp | Changed |
| Expenditure in intramural RD | RRdInX | Changed |
| Expenditure in extramural RD | RRdExX | Changed |
| Expenditure in acquisition of machinery | RMacX | Changed |
| Expenditure in other external knowledge | ROekX | Changed |
| Expenditure in marketing | RMarX | Changed |
| Expenditure in training | RTrX | Changed |
| Expenditure in preliminary activities | RPreX | Changed |
| Total innovation expenditure | RTot | Changed |
| Stratum to which ent before when sampled | StrB | Removed |
| Stratum to which ent belong acc to quest | StrA | Removed |
| Weighting factors | Weight | Removed |

**Table 7. Variables changed by the statistical disclosure control methodology.**

## 11. References

1. Eurostat (2005), *The Third Community Innovation Survey CIS3. Methodology of Anonymisation.* 11-07-2005
2. Ichim, D. (2008), *Community Innovation Survey: A Flexible Approach to the Dissemination of Microdata Files for Research Purposes,* Proceedings of the European Conference on Quality in Official Statistics, Rome, 8-11 July 2008, available at http://q2008.istat.it.
3. Ichim (2009), Disclosure Control of Business Microdata: a Density-Based Approach, International Statistical Review, to appear.
4. Ester, M., Kriegel, H. P., Sander, J., Xu, X (1996), *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

5. Sweeny, L. (2002), *k-anonymity: a Model for Protecting Privacy*. International Journal of Uncertainty, Fuzziness and Knwoledge Based Systems, 10(5), 557-570.