Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research

Daniela Ichim
Istituto Nazionale di Statistica
via Cesare Balbo, 16,
00184 Roma, Italia
ichim@istat.it

Abstract
This paper describes a methodology for the dissemination of microdata stemming from the Community Innovation Survey. Both risk assessment and disclosure limitation phases are introduced in a flexible parametric form. The methodology can be easily adapted to different national settings. A strategy to achieve comparable dissemination at European level will be indicated.

1. Introduction

The mission of the National Statistical Institutes (NSI) is to produce reliable, impartial, transparent, accessible and pertinent information. The dissemination of this information should be performed in full compliance with the regulations pertaining to the privacy of respondents.

Nowadays researchers increase their demand of analysis of microdata. A way to satisfy the users' needs is the dissemination of microdata files for research purposes. At a first glance, the dissemination of detailed information and the preservation of the confidentiality of respondents might seem two very conflicting objectives. Anyway, by a careful analysis of the product to be released, the right balance may be found.

Taking as an example the Community Innovation Survey, this paper illustrates the approach adopted by ISTAT (Italian National Statistical Institute) for the dissemination of microdata files for research. The three main parts of the statistical disclosure control process: risk assessment, disclosure limitation method and the assessment of the quality of the microdata file are presented.

2. Community Innovation Survey

The Community Innovation Survey (CIS) collects information on the innovation tendency at firm level. On each statistical unit, the enterprise, CIS registers information on the economic activity (*Nace*), geographical location (*Nuts*), number of employees (*Size*), *Turnover*, expenditure on innovation and research (*RTOT*), etc.. The latter is decomposed with respect to factors like intramural/extramural research, acquisition of machinery, acquisition of external knowledge, personnel training, etc.. Various facets of innovation are also investigated, e.g. factors that determine or hamper innovation, number of employees with higher education, number of registered patents, etc.. A full survey description of CIS is given in Eurostat (2006)].

The main statistical unit for CIS4 was the enterprise, as defined in the Council Regulation 696/1993 on statistical units or as defined in the national statistical business register. The target population was defined by the enterprises whose principal economic activity may be classified in one of the categories of NACE 10 to NACE 74. All enterprises included in the target population followed the minimum coverage which was defined by all enterprises with 10 employees or more.

The survey was based on a one stage stratified simple random sample. At least 6 enterprises in each stratum were selected. In the case of less than 6 enterprises in a stratum, a full census was conducted. The stratification variables used for the CIS 4, were: the economic activities according to NACE classification, enterprise size according to the number of employees and regional aspects at NUTS 2 level. A multi-variable and multi-domain sample allocation was used.

The official, up-to-date, statistical business register, called ASIA (Archivio Statistico delle Imprese Attive - statistical business register of active enterprises) was used. The Italian CIS4 sample included 44,571 enterprises out of a population of about 193,300 enterprises with 10 employees or more and potentially active in the year 2004 and the average response rate turned out to be 49%.

A calibration methodology, see Deville and Särndal (1992), currently applied at ISTAT, was used for the estimation process. The final weights were obtained by adopting the following procedure: an initial weight was assigned to each sampled unit with reference to the sampling plan as the reciprocal of the inclusion probability. Two correction factors for initial weights were then calculated: a first one was the unit non response factor; a second one was to satisfy equality between estimation of auxiliary variables and known totals from the Register. The final weights were thus obtained as the result of the product between initial weights and correction factors. For CIS4, as well as for most of the business surveys, number of enterprises and number of employees were used as auxiliary variables, according to the information provided by the Italian Official Business Register ASIA. More details on the Italian CIS4 may be found in ISTAT (2004).

The European Union Regulation CE 831/2002 establishes a list of business surveys for which access is granted for research purposes: the Community Innovation Survey, the Structure of Earnings Survey and the Continuing Vocational Training Survey. These surveys have undergone a complex process of harmonisation that inherently includes comparability as an important dimension of the quality framework. Comparability aims at measuring the impact of differences in applied statistical concepts and definitions on the comparison of statistics between geographical areas, non-geographical domains, or over time. The factors that may cause several statistical figures to lose comparability are attributes of the survey that produces them. Such features may be grouped into two broad categories: the first one relates to survey concepts and the second one relates to measurement and estimation methodologies. To address the problems deriving from the first type of attributes, the approach usually taken at European level is via a regulatory framework where all the concepts of the survey are clearly defined and harmonised. This common framework clearly defines the phenomenon under study, target population, statistical units to be surveyed and all possible metadata descriptions for all the variables involved so as to avoid “structural” non-comparability. As far as the second group of issues is concerned, indications on the suggested methodologies for every survey phase are

given: sampling design, data collection, weight calculation, imputation and so on. To improve standardisation on all phases, routines are provided by Eurostat for the use of member states. However, “member states are in general free to use whatever methods they prefer as long as some quality thresholds are met”, see Eurostat (2004). For the CIS in order to ensure what we have called “structural” comparability across countries, Eurostat, in close cooperation with the EU Member States, developed a standard core questionnaire, with an accompanying set of definitions and detailed metadata based on Eurostat (2006). To address the type of incompatibilities due to issues of measurement estimation, clear methodological recommendations were given at European level.

3. Disclosure Risk Scenarios

Microdata has many analytical advantages over aggregated data, but also poses more serious disclosure issues. For microdata, disclosure occurs when an individual can be re-identified by an intruder using information contained in the file, and when on the basis of that, the intruder could increase his knowledge about the identified individual, see Hundepool (2006).

To assess the disclosure risk, the data protector makes realistic assumptions about what an intruder might know about respondents and what information would be available to him to match against the microdata and potentially make an identification and disclosure. These assumptions are known as disclosure risk scenarios.

The disclosure scenario consists of the analysis of the users and their needs and the analysis of the file content: the key and confidential variables.

3.1. Users

For the microdata files for research the potential users are known in advance: researchers that sign an agreement with ISTAT in order to get the MFR for performing their analyses. A first characteristic of this type of release is that any nosy colleague scenario cannot be deemed realistic; obviously, the researchers are not colleagues and not even competitors of any sampled unit. It is hard to accept the hypothesis that the researcher could have some information obtained as an insider.

Once the microdata file for research is released, the NSI has no more any control on the way the file is used. However, the signed contract legally impede the researcher to try to identify any unit. This means that the NSI generally takes the researchers on trust. Consequently, since any record linkage experiment involves a lot of resources (time, methodology, tools, etc.), the NSI presumes that the researcher wouldn’t deliberately try to match the microdata file with an external database containing direct identifiers.

Anyway, even if they are considered as bona-fide users, the researchers might unintentionally recognize some units. For example, in a business microdata framework, it is publicly known that the greatest enterprises are generally included in the microdata file because of their significant impact on the studied phenomenon. The greatest enterprises are also the most famous ones. Consequently, a spontaneous identification or recognition might occur. Moreover, even the researcher might be simply curious about some units revealed as “particular” from his analyses.

3.2. Research Potential

In order to gain some insights on possible statistical usages of CIS data a brief review on the scientific literature based on such data was carried out. An example of such review is provided in Arundel (2005). Below few common characteristics of several analyses based on CIS microdata are listed.

Analyses are commonly performed at NACE 2-digit level, using the data at national level. This proves the strategic importance of the economic variable. Consequently, the dissemination of CIS data at a more aggregated level of the economic activity would be almost useless.

A relationship between the economic performance of companies and their innovation attitude is commonly investigated. The economic performance may be modelled, for example, through turnover, employment and their variations. Examples of studied statistics are the innovation intensity (expenditure per employee on innovation linked to employment growth, by internal or external innovation) or the share of turnover that is due to new or improved products (quantifying the economic relevance of innovations). Each registered component of the expenditure on innovation is equally used to analyse the innovation phenomenon. Correlations and ratios involving these components and the ones expressing the economic performance seem to be particularly important. Such analyses may be found, for example, in Belderbos (2004), Evanghelista (2006), Klomp (2001), Loof (2002), Mastrostefano (2007).

As usual in survey statistics, weighted means are widely used. Besides being part of the already published tabular data, weighted means were found to be involved in the majority of analyses. For example, any share is expressed using the weighted totals. Consequently, preservation of such statistics seems crucial.

As a result of this overview, a possible list of statistics to be used for benchmarking purposes in data utility may be made. Ratios of innovation variables as a mean to analyse scaled quantities seems predominant. Also the change in turnover with respect to the first year of the reference period seems relevant.

3.3. Harm

The disclosure scenarios have also the role of assessing the confidential content of the microdata file.

For CIS the confidential content is mainly related to the expenditure on innovation, research and development. Variables like research in intra/extramural research and development, expenditure on acquisition of machinery, expenditure on external knowledge represent both the core of the survey and the confidential content.

3.4 Identification

It is here understood that the direct identifiers are completely removed from the microdata file. However, other variables in the microdata can be used as indirect identifying variables, e.g. gender, age, principal economic activity, enterprise size in terms of number of employees, etc.. Based on the disclosure risk scenario, the identifying variables are determined. The other variables in the file are confidential or sensitive variables and represent the data not to be disclosed.

The first step of the anonymisation process is the risk assessment. The main question is: when a unit cannot be identified? Intuitively, a unit cannot be identified when it could be confused with several/many other units. The difficulty is to express this simple concept using a sound statistical methodology.

Of course, not all the variables could be used in order to identify a unit u or, on the contrary, to assess if u could be confused with other units. The variables used for this task are called identifying variables. For the Italian CIS4 it was considered that an enterprise could be identified using the following structural variables: principal economic activity, geographical location, number of employees and turnover, see Ichim (2006). The continuous variable *Turnover* is the variable expressing the concept of dominance or magnitude of an enterprise. This disclosure scenario is a general one because most of the structural variables, both categorical and continuous, are considered identifying (key) variables. Of course, in other national settings, different subsets of these key variables might be considered so, but the corresponding disclosure scenarios would be only particular cases of the scenario adopted for the Italian CIS4.

As previously stated, a unit is not at risk if it cannot be singled it out from the rest. In presence of solely categorical identifying variables, the methodological solution is given by the k-anonymity principle: a unit cannot be identified with certainty when there are at least k units with the same values of the key variables, see Sweeny (2002). Or, the mass of a given point is greater than a given threshold.

By definition, a continuous variable takes on each unit almost a different value. That's why the ***exact*** k-anonymity principle is no more useful in this setting. But the k-anonymity expresses also the density concept for discrete variables. The extension of this density concept is by far much easier. If the density around a unit is very high, the unit should be safe. This is mainly due to the uncertainty that governs any measurement process. On the contrary, if a unit is very distant from its closest neighbours, the chance to identify it increases significantly, even if the measurements have some degree of uncertainty. This safety concept for continuous variables is illustrated in figure 1. The black dashed circles illustrate a group of units that could be confused one with another. From the group of red dotted circles, a unit is clearly distinguishable, hence at risk.

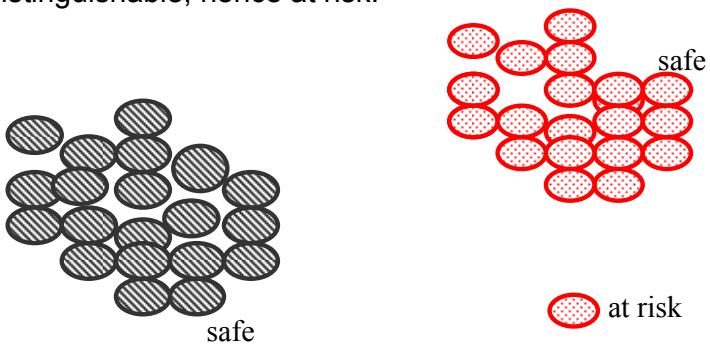

Figure 1. The density concept for continuous variables.

Of course, the problem is even more complicated when we deal with a mixture of categorical and numerical key variables, i.e. when the observed vector of key variables may be re-arranged into $\boldsymbol{t} = (\boldsymbol{t}^n, \boldsymbol{t}^c)$. Here $\boldsymbol{t}^n$ denotes the vector of numerical variables and $\boldsymbol{t}^c$ denotes the vector of categorical variables The measure of the density around a unit starts with the definition of a distance between points. The distance between two points $\boldsymbol{t}_1 = (\boldsymbol{t}_1^n, \boldsymbol{t}_1^c)$ and $\boldsymbol{t}_2 = (\boldsymbol{t}_2^n, \boldsymbol{t}_2^c)$ could be defined by

multiplying the Euclidean distance between the numerical key variables and the inverse of the indicator function for the categorical key variables, as in equation (1).

$$d(t_1, t_2) = e(t_1^n, t_2^n)(I^{-1}(t_1^c, t_2^c))^{\beta}, \quad \beta \in \{0, 1\} \tag{1}$$

If $\beta = 0$, no categorical variable is considered in the disclosure scenario. In the CIS framework, this is equivalent to supposing that the intruder would try to identify a unit **only** by comparing numerical variables like *Turnover*. It would mean that the intruder would completely ignore his knowledge about the structural categorical variables. In a national setting, this is not a realistic assumption since, at least for the largest enterprises, their principal economic activity, their size in terms of number of employees and even their geographical location are generally well-known. It follows that the inclusion of the structural categorical variables in the disclosure scenario is almost mandatory. The distance function between units should be accordingly modified. This could be simply done by setting $\beta = 1$ for all the structural categorical key variables. It is clear that this function incorporates many different scenarios which may reflect different national situations. In table 1, different distance functions for different scenarios are presented. The extensions to other cases are trivial. It's worthwhile noting that the distance function must reflect the choices made by the selection of the key variables, i.e. the disclosure scenario; if it is assumed that a unit might be identified by means of a variable, then this variable (information) should be part of the definition of the distance between units. For the Italian CIS4 microdata file, the key variables were *Nace, Size, Nuts* and *Turnover*, as described in Ichim (2006).

Key variables	**Distance function**
Turnover	$e(T_1, T_2)$
Nace	$I^{-1}(N_1, N_2)$
Turnover, Nace	$e(T_1, T_2) * I^{-1}(N_1, N_2)$
Turnover, Size	$e(T_1, T_2) * I^{-1}(S_1, S_2)$
Turnover, Nace, Size	$e(T_1, T_2) * I^{-1}(N_1, N_2) * I^{-1}(S_1, S_2)$

Table1. Examples of disclosure scenarios and corresponding distance functions between two units $u_1 = (T_1, N_1, S_1, G_1, \cdots)$ and $u_2 = (T_2, N_2, S_2, G_2, \cdots)$ where T stands for *Turnover*, N stands for *Nace*, S stands for *Size*, G stands for *Nuts* and so on.

The importance of the definition of a suitable disclosure scenario should be again stressed. As it can be noticed, the only subjective part in the disclosure scenario is the choice of the key variables. Unfortunately, there is no rule of thumb about this choice. Most depends on the assumptions made on the available external knowledge; surely the quantity and the quality of this information vary across Member States. The NSI should be anyway aware of the consequences of ignoring an important key variable. In figure 2, on the left, the scatterplot of the *Turnover* independently on the *Nace* categories may be seen. On the right, the same values were plotted, but this time for a single *Nace* category. If the data protector considers that *Nace* is not a key variable, it might consider that all the units represented by blue diamond points are safe, because there is always a sufficient number of close units. Instead, if an intruder uses anyway the structural information given by the *Nace* category when trying to identify a unit, the most dominant enterprise would be readily isolated. This situation only worsens if other structural categorical variables are included in the disclosure scenario. Consequently, the categorical structural variables cannot be completely eliminated from the disclosure scenario and the data protector should be aware about the consequences of such an extreme choice.

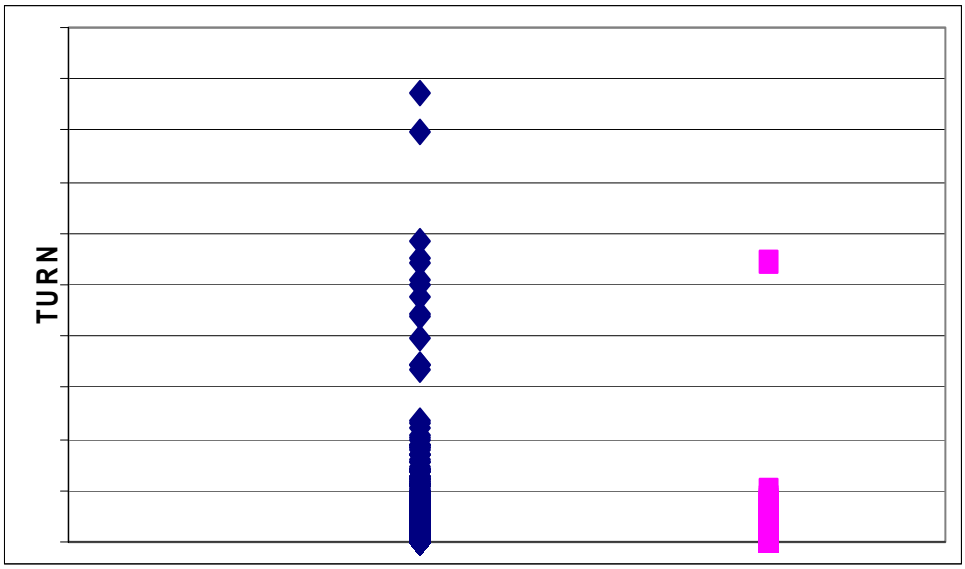

Figure 2. Scatterplot of *Turnover* values independently of *Nace* (on the left, blue diamond symbols) and for a single *Nace* category (on the right, red squares symbols).

The next step is to define a value for the parameter k in the k-anonymity principle. If a unit A may be confused only with unit B, is A at risk? If a unit A may be confused only with units the B and C, is A at risk of re-identification? How many confusing units are required for A in order to be considered safe?

Based on its own dissemination policy, a statistical agency may define the minimum number M^* of units to which u should be confused in order to be considered safe. An estimation of the uncertainty associated to the key variables should be accounted for when setting M^*. For the Italian CIS4 microdata file, M^* was set equal to 3.

Given this threshold M^* on units and the distance d between units, for each unit u, its M^*-th distance is computed. $M^*(u)$ is the distance between u and a unit u^* for which the following condition hold:

a) there are at least M^* units u' satisfying $d(u,u') \le d(u,u^*)$

b) there are at most M^*-1 units u' satisfying $d(u,u') < d(u,u^*)$

The neighbourhood $N_{M^*}(u)$ of u is defined as the subset of units closer than $M^*(u)$ with respect to u. The reachability distance of u with respect to a unit u' is defined in equation (2).

$$RD_{M^*}(u,u') = \max\{M^*(u'), d(u,u')\} \tag{2}$$

The inverse of the average of the reachability distances of the units $u' \in N_{M^*}(u)$ gives the local reachability density (LRD_{M^*}) of u, see equation 3.

$$LRD_{M^*}(u) = \left(\frac{\sum_{u' \in N_{M^*}(u)} RD_{M^*}(u,u')}{|N_{M^*}(u)|} \right)^{-1} \tag{3}$$

Here $|V|$ denotes the number of elements in V. LRD_{M^*} estimates the density around u using the M^*-th distances of the units in $N_{M^*}(u)$. Finally, the local outlier factor,

LOF_{M^*}, is defined as a measure of difference in density between a unit and its nearest neighbours:

$$LOF_{M^*}(u) = \frac{\sum_{u' \in N_{M^*}(u)} \frac{LRD_{M^*}(u')}{LRD_{M^*}(u)}}{|N_{M^*}(u)|} \quad (4)$$

The general properties of LOF_{M^*} are discussed in Breunig (2000). The LOF_{M^*} value of a unit u in a high density area is very close to 1, since $LRD_{M^*}(u') \approx LRD_{M^*}(u), \forall u' \in N_{M^*}(u)$. In such cases u can be confused at least with its M^* nearest neighbours, hence u is safe. On the contrary, if u is very distant from its nearest high density area, $LOF_{M^*}(u)$ would be very much greater than 1. Then u is an isolated unit; it is at risk of re-identification. The statistical agency may set a threshold α and define at risk of re-identification those units for which $LOF_{M^*}(u) > \alpha$. More details on the usage of the LOF_{M^*} function in the statistical disclosure control framework may be found in Ichim (2007). A cluster bsed methodology for the identification of units at risk was also described in Bacher (2002).

In practice, once the key variables were identified, two parameters have to be set. First, the threshold M^* should be derived from the dissemination policy of the NSI. For example, if the NSI uses the frequency rule of minimum 3 units, then M^* could be set equal to 3. The second parameter is a cut-off point for the LOF_{M^*}, the function that measures the density around a point. The LOF_{M^*} value is computed for each point. The units at risk could be identified by simply using some quantile criteria. This is a very simple and robust approach, but it would mean that a fixed percentage of units at risk exists in each combination of categorical key variables. In presence of isolated units, the ordered LOF_{M^*} values present a sudden change in slope, as illustrated in figure 3. The value α corresponding to this abrupt change would give a reliable indication of the isolated units. α could be automatically determined by means of structural change models, see Zeileis (2003). An advantage of such setting of α is that the percentage of units at risk of re-identification would not be *a-priori* defined. Increasing/decreasing the value of α would obviously decrease/increase the number of units at risk. This approach was actually implemented for the disclosure control of the Italian CIS4 microdata file.

A further advantage of the LOF_{M^*} measure is its independence on the location of the units at risk of re-identification. Indeed, such units may be found on both tails of the key variable distribution, as well as on its central part. This happens because the units at risk are determined using a threshold on the ***ordered*** LOF_{M^*} values. It should also be noted that for extreme choices of α, none or all the units would be considered at risk of re-identification.

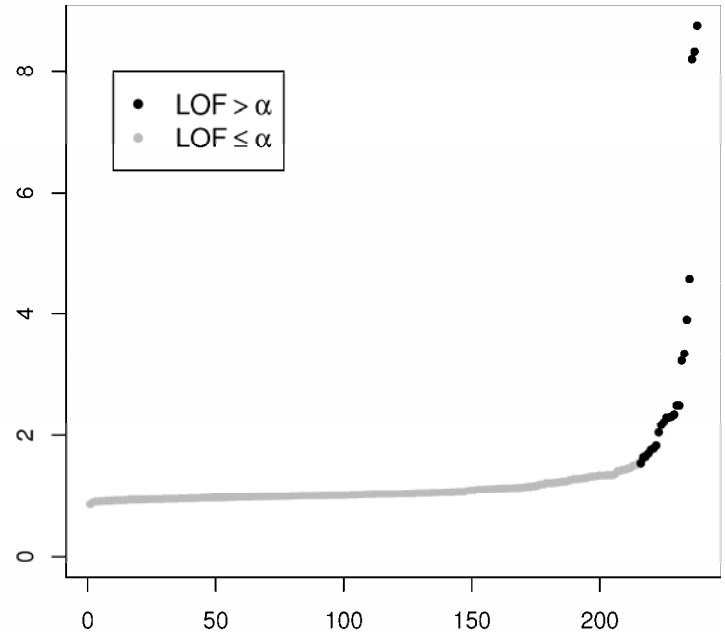

Figure 3. Abrupt change models for the selection of the parameter α.

4. Disclosure Limitation Methodology

Once the units at risk of re-identification were determined, if the risk is considered too high, a protection method should be applied in order to reduce the risk. Among the many protection methods, see, for example Willenborg (2001), each one with its own advantages and drawbacks, the data protector should choose the disclosure limitation method that solves the specific dissemination problem. As already discussed in section 3.2, the economic activity and the size of an enterprise seem to be, from a user point of view, the most important structural variables. That's why, in order to reduce the number of rare cases, for the Italian CIS4 microdata file, it was decided to recode the variable giving the geographical location of the enterprise. Based on the information given in table 2, the percentages of rare cases for different *Nuts* levels, it was decided to release only the national level. The other two structural categorical variables were left almost unchanged, except for very few combinations were a free local recoding of *Size* was performed. Then, for each combination of structural categorical key variables, the units at risk of re-identification with respect to *Turnover* were identified by means of the procedure described in section 3.4. For the Italian CIS4 microdata file, 8.25% of units were found at risk of re-identification using this methodology. The protection of units was achieved also by perturbing the structural continuous variable, i.e. *Turnover*.

Geographical location	Percentage of rare cases
NUTS3	16.62%
NUTS2	7.47%
NUTS1	1.48%

Table 2. Percentages of rare cases for different *Nuts* hierarchical details. The rare cases were determined based on the population frequencies, as given by the weights.

There are two general properties of the perturbation methods. First the protection method should be adequate to the type of dissemination. For microdata files for research, a selective protection should be used. **Only** the units at risk of re-identification and **only** the key and confidential variables should be modified. The

selectiveness guarantees that the information loss is very much reduced, or, at least, very much controlled.

Second, the disclosure limitation method should protect with respect to the assumed disclosure scenario. Here we assumed that a unit is safe if the k-anonymity principle is satisfied. Consequently, the perturbation should aim at achieving the k-anonymity. Of course, the k-anonymity principle should be simultaneously satisfied with respect to all key variables. The increase of uncertainty in the identification of a unit at risk could be done by a simple method like the imputation from the nearest neighbour. In the statistical disclosure control framework, the imputation should be performed using the value of the nearest **safe** neighbour; otherwise the increase in uncertainty could not be sufficient. This type of perturbation might produce a significant information loss on the tails. That's why a micro-aggregation, see Defays (1998), could perform better in such situations. The micro-aggregation firstly determines groups of M^* and then replaces the values of the units in each group by the mean of the group. For the Italian CIS4 microdata file, for each combination of categorical key variables, an individual ranking was applied on the tails of the *Turnover* distribution.

It is important to notice that, if $\alpha = 0$, all the units would be considered at risk of re-identification. The same result could be obtained if a quantile criteria is used, by setting the quantile threshold equal to zero. Consequently, all units would be subject to a micro-aggregation process because of their location on the tail of the distribution which in such cases is the entire distribution. If some categorical variables are included in the disclosure scenario, i.e. are considered key variables, the micro-aggregation would be applied with respect to each combination of key variables. On the contrary, if there is no categorical key variable, the micro-aggregation would be applied irrespective of any combination of key variables. The drawbacks of this latter approach were discussed also in Leppälahti (2007).

The last observation concerns the parameters of the micro-aggregation process. The minimum number of units belonging to the same group should be equal to or greater than M^*. This condition would ensure that the aimed k-anonymity criteria is achieved. If there are several continuous key variables, a multivariate micro-aggregation process should be applied, for each combination of categorical key variables. This is the only way to ensure that the k-anonymity criteria is satisfied with respect to all the key variables. Instead, if there is a unique continuous key variable, like for the Italian CIS4 microdata, the micro-aggregation reduces to individual ranking.

The disclosure limitation methodology presented above has the enormous advantage that it is a very flexible one. Indeed, for different choices of key variables and for different threshold settings, the methodology reduces to several well-known approaches to statistical disclosure control. In table 3, several particular cases of the discussed methodology are shown. The other possible extensions may be easily derived.

Categorical keys	**Continuous keys**	α	**Perturbation**
None	*Turnover*	Max	No
None	*Turnover*	(0, max)	Imputation from the nearest safe neighbour and individual ranking only on

			tails, irrespective of any combination of categorical variables
None	*Turnover*	0	Individual ranking on all the units, irrespective of any combination of categorical variables
None	*Turnover, X*	Max	No
None	*Turnover, X*	(0, max)	Imputation from the nearest safe neighbour and multivariate micro-aggregation only on tails, irrespective of any combination of categorical variables
None	*Turnover, X*	0	Multivariate micro-aggregation on all the units, irrespective of any combination of categorical variables
Nace	*Turnover*	Max	No
Nace	*Turnover*	(0, max)	Imputation from the nearest safe neighbour and individual ranking only on tails, for each category of the categorical key variable
Nace	*Turnover*	0	Individual ranking on all the units, for each category of the categorical key variable
Nace, Size	*Turnover*	Max	No
Nace, Size	*Turnover*	(0, max)	Imputation from the nearest safe neighbour and individual ranking only on tails, for each combination of the categorical key variables
Nace, Size	*Turnover*	0	Individual ranking on all the units, for each combination of the categorical key variables
Nace, Size	*Turnover, X*	Max	No
Nace, Size	*Turnover, X*	(0, max)	Imputation from the nearest safe neighbour and multivariate micro-aggregation only on tails, for each combination of the categorical key variables
Nace, Size	*Turnover, X*	0	Multivariate micro-aggregation on all the units, for each combination of the categorical key variables

Table 3. Particular cases of the proposed disclosure limitation methodology.

5. Data Quality

The last step of the disclosure limitation process is the assessment of the quality of the microdata file to be released. Whatever protection method has some impact on both data utility and degree of confidentiality. That's why in the statistical disclosure control framework, these two aspects of data quality, utility and confidentiality, should be simultaneously considered. For example, if only data utility were taken into account, the individual ranking could be applied independently of the categorical key variables. In such cases, many statistics would be almost exactly preserved, as shown in the upper part of the figure 4. Indeed, the correlation between *Turnover* and the total expenditure in innovation would be perfectly preserved in the individual ranking irrespective of the categorical key variables was applied. The same

phenomenon was observed for other statistics, too. Consequently, based only on a data utility criteria, the data protector could be satisfied with this protection method. Nevertheless, the same data protector should be aware of the confidentiality promise made to its respondents during the data collection phase. Applying such a slight perturbation as the one given by the application of the individual ranking irrespective of the categorical key variables, could not give sufficient protection. As it may be observed in the lower part of the figure 4, if this type of individual ranking was applied to the *Turnover*, there might be some units that exactly preserve their dominance in magnitude. In other words, even without knowledge of the exact values, if the intruder knew that an enterprise was dominant before the microdata file dissemination, he would be able to identify this enterprise if its *Turnover* value was perturbed without changing its dominance status. In a business framework, this drawback of the individual ranking applied irrespective of the categorical key variables is mainly due to the skewness of the continuous variables. It should be noticed that the selective protection method illustrated in section 4 eliminated this drawback because it was applied with respect to the stratifying/structural categorical key variables. Moreover, because of its selectiveness, this flexible protection method obviously preserves more information than a stratified micro-aggregation.

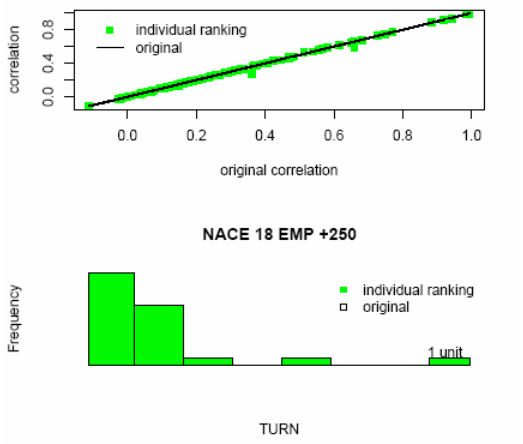

Figure 4. Comparison of two disclosure limitation methodologies in both data utility and confidentiality perspectives.

5.1 Record Linkage

A further experiment was performed in order to assess the ability of the LOF_{M^*} function to detect the units at risk. For this step a record linkage experiment was performed using the original microdata file and an external database. The Chambers of Commerce database was used because this register was involved also in the data correction phase in order to determine some of the imputation units. Firstly the quality in terms of completeness of this register was analysed. It was found that the three structural variables, *Nace, Size, Turnover*, were simultaneously registered only in 33% of cases, without considering the hierarchy level of detail. Considering only the part of the register containing complete information, in the record linkage experiment, the units correctly matched were determined. Then the number of neighbours in certain neighbourhoods of these units at risk of re-identification was computed. Two sets of blocking variables, {*Nace*} and {*Nace, Size*}, were used. The isolated units in this framework were compared with the ones

identified by the LOF_{M^*} function. For each unit correctly matched in this record linkage experiment, LOF_{M^*} was measured using two values for the threshold M^*, i.e. $M^* = 3$ and $M^* = 5$, respectively. The structural change model was always used in order to determine the units at risk. To yield the two risk measures comparable, neighbourhoods of a predefined width around the *Turnover* values were taken into account. For example, as it may be observed in table 4, among the units correctly matched in the record linkage experiment performed using only *Nace* as blocking variable and having 1 unit within 10% of their *Turnover* value, 88% were labelled at risk by the LOF_{M^*} measure with $M^* = 3$. At the same manner, 58% of the units correctly matched using {*Nace, Size*} as blocking variables and having less than 5 units within 10% of their *Turnover* values were labelled at risk by the LOF_{M^*} function with $M^* = 5$. The other entries in table 4 should be interpreted using the same reasoning. Generally it may be observed a good agreement between the two risk measures. It should be anyway mentioned that the agreement was perfect when the units at risk identified by either function were large enterprises, i.e. with more than 250 employees.

	M^*	**1 unit within 10%**	**less than M^* units within 10%**	**less than M^* units within 20%**	**less than M^* units within 30%**
Nace	3	88%	84%	97%	100%
Nace Size	3	63%	60%	74%	87%
Nace	5	88%	73%	87%	96%
Nace Size	5	63%	58%	70%	80%

Table 4. Comparison of LOF_{M^*} and record linkage.

No really significant difference was observed between the two thresholds on M^*, 3 and 5. At a first glance it might seem strange that higher agreement percentages were obtained when using only *Nace* as blocking variable. This is due to the fact that when the matching unit is looked for inside the *Nace* category, ignoring the size information, there are much more units, so the probability of an incorrect match increases. In this setting, the units considered at risk by either method are the really isolated units. Such units would probably be correctly labelled at risk by many measures based on a density concept. In table 4, by keeping constant the number of neighbours and by increasing the width of the neighbourhood, different degrees of isolation were simulated. That's why the agreement percentage generally increases: if the degree of isolation increases, it is easier for the risk measures to detect it.

Similar record linkage experiments were performed using the original microdata file derived from the Italian CIS4 file, following the approaches presented in Winkler (2004) and Domingo-Ferrer (2003). Since the disclosure limitation method was applied in order to satisfy the k-anonymity criteria, the probability of a correct match obviously decreases when the k-anonymity criteria is satisfied. This decrease is proportional to M^*. The only issue that could be highlighted from these experiments is again the importance of the definition of the disclosure scenario. For example, suppose that only *Nace* is considered a key variable. Consequently, following the procedure described in section 4, the perturbation method is applied only with respect to the *Nace* categories. If *Nace* and *Size* are used as blocking variables in

the record linkage experiment, then the number of units correctly identified significantly increases. This drawback is due to the fact that the LOF_{M^*} function is applied with respect to the *Nace* categories. Hence many units that could be at risk if the *Nace* category were split in several subsets (according to the *Size* categories) locate in high density areas. Consequently, they are not detected by the LOF_M function and they are not modified by the selective protection method. These non-perturbed units obviously increase the number units correctly matched in the record linkage experiment using {*Nace, Size*} as blocking variables.

5.2 Information Content

Many other considerations on the information content of the microdata file hold. Using a selective masking, a lot of statistics and statistical indicators were maintained. This is due to the fact that only the key and confidential variables were modified. Moreover, since the weights were not modified, the coherence with the already published statistics is guaranteed by default. Of course, not all the statistics were exactly preserved. For example, the changes induced in the variances of *Turnover* in each combination of categorical key variables are shown in figure 5. In general, since only the units at risk were perturbed, the modification of the statistical indicators was not significant.

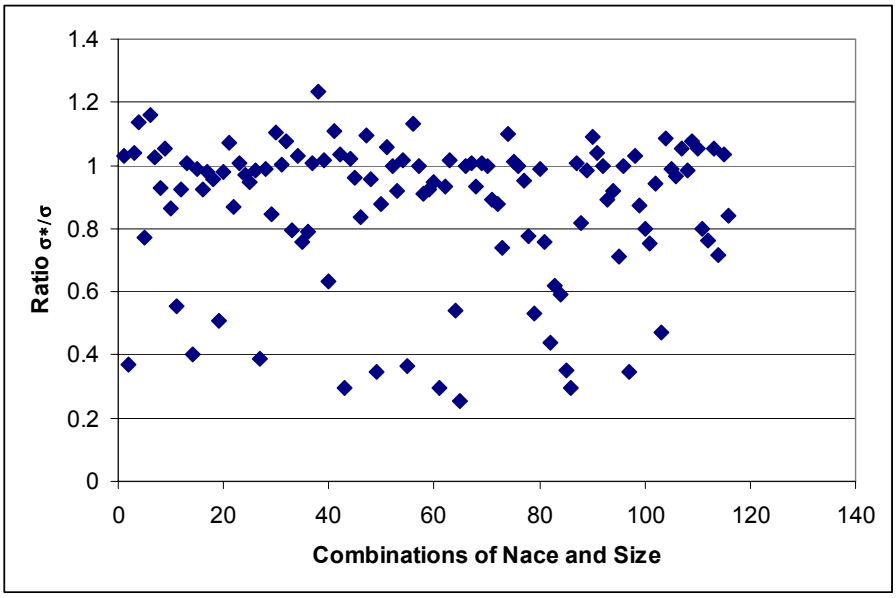

Figure 5. Comparison of *Turnover* variances before and after the application of the statistical disclosure methodology.

In absence of a sound statistical definition of data utility, the least it can be done is to assess the impact of the perturbation method on the variables most used by researchers. Or, otherwise stated, the research potential of the microdata file to be released should be assessed after the application of the disclosure limitation method. As already described in section 3.2, for CIS, the share of innovation seems one of the most used variables. In figure 6, a comparison between the selective LOF_{M^*}-based masking method and the stratified individual ranking is illustrated. Similar results were obtained for other combinations of categorical key variables and for

other variables. As expected, the stratified individual ranking reduces the variability of the data, especially on the tails of the distributions.

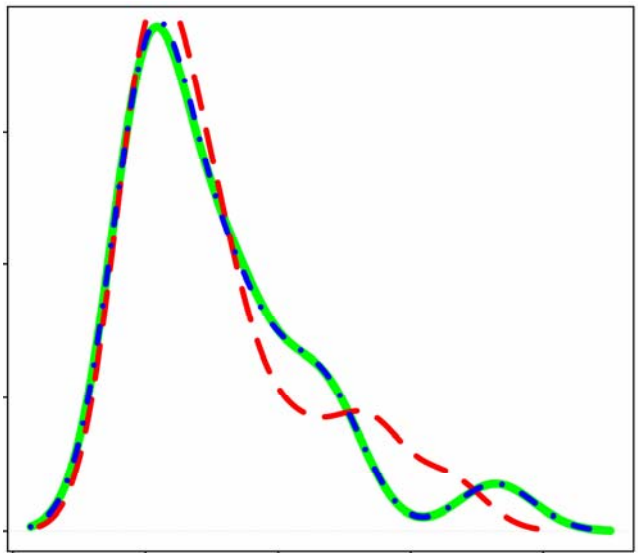

Figure 6. Comparison in LOF_{M^*} - based and individual ranking applied with respect to the categorical key variables *Nace* and *Size*. The original data are represented by the green solid line; the individual ranking is represented by the red dashed line; the LOF_{M^*} is represented by the blue dotdashed line.

6. Harmonised dissemination

In its co-ordinating role Eurostat is working in harmonising surveys throughout Europe providing guidance to Member States in collecting and processing data using comparable methods. The harmonisation process undergone by the CIS can be summarised in three steps: development of general methodological guidelines, definition of benchmarking statistics and assessment of the effects of different practices on such statistics and, finally, the definition of a threshold for determining when an action is necessary.

Currently the dissemination procedure at European level foresees the application of a single statistical disclosure methodology. This strategy surely has the lowest costs in terms of implementation, testing and application. It might be believed that this strategy also produces highly harmonized results. Nonetheless, the application of the same statistical disclosure limitation to two different data sets might produce very different qualitative and quantitative results.

The application the same harmonisation strategy at European level in the context of anonymisation procedures for the release of microdata files for research would imply, to start with, the indication of the methodological paradigm of statistical disclosure control. Such paradigm states the definition of a disclosure scenario, subsequent definition of risk, a measure to assess it, procedures to reduce the risk and finally, but absolutely crucial for the whole process, measures of data utility allowing the final users to judge how poor/good the results of his analysis on the anonymised microdata would be. Such utility measures represent the benchmarking statistics for comparability. In fact, in the anonymisation phase one main goal should be the production of anonymised data sets sharing certain statistics with the original microdata. The key of the whole process should then be the definition of protection

methods that maintain such statistics or the customisation of existing procedures to guarantee pre-selected characteristics.

Since the organisational heterogeneity of Member States is, for the moment, a fixed constraint at European level, a harmonized European dissemination of microdata files for research could be achieved twofold: 1) developing flexible anonymisation methodologies and 2) constraining the output of the anonymisation methodology. In other words, the harmonisation concept is defined by means of a set of minimal requirements on both input and output of the anonymisation process.

6.1 Input Harmonisation

In principle, on the input phase, a significant improvement might be reached by using flexible statistical disclosure control methods. Different variants of the same statistical disclosure limitation methodology could be easily implemented and tested. For example, the implementation of the individual ranking could depend on the microaggregation parameter p; then, each Member State should select its most appropriate value for this parameter p, e.g. 3 or 5 or some other value. The implementation of the same statistical disclosure limitation methodology with respect to different stratification domains is another simple example of flexibility. For instance, the methodology could be applied to the entire microdata file or with respect to the domains defined by the categorical key variables (generally the structural categorical variables). In other words, by simply changing the values of some parameters, the statistical disclosure methodology could be more easily accepted by many Member States. Of course, the Member State should previously accept the underlying disclosure scenario and the corresponding risk assessment methodology.

As previously stated, an interesting feature of the anonymisation procedure outlined in section 4 is that for extreme choice of the parameters in the risk assessment phase the protection process reduces to individual ranking. Indeed, if a degenerate distance is considered for the categorical key variables, the risk assessment and the protection method would be applied irrespective of *Nace* and *Size*. Additionally, if $\alpha = 0$, all units would be considered at risk of re-identification. According to the procedure described in this paper, all these units would be protected using micro-aggregation. An evolution of the current European situation could see the LOF_{M^*} -based selective masking as a possible framework for choosing different degrees of anonymisation.

6.2 Output Harmonisation

On the output phase, the definition of a battery of quality criteria could be used to put in practice the comparability concept. Microdata files are disseminated for research purposes, therefore data utility/data quality are one of the most important characteristics of the output of the European dissemination flow. Timeliness, consistency, efficacy and comparability are only some of the dimensions of data quality who are of interest to the users. Data utility is neither easy to define nor easy to quantify. Here it is proposed to assess it through the definition of relevant statistics for the type of data under analysis. Then, quality criteria or thresholds on these relevant statistics should be set. Moreover, possible remedies should be indicated for the cases when the quality criteria are not met. Careful definition and tuning of

benchmarking statistics coupled with clear threshold setting would allow comparability of analyses among different methods and different parameter choices in different Member States. Given the assurance of a pre-defined acceptable re-identification risk level, preservation of benchmarking statistics should then be the primary objective, independently on the anonymisation methodology.

This framework implies an initial investment in identifying the relevant statistics and relative thresholds but, then, the whole procedure is expected to become part of the production process. Also this initial stage can be performed with the help of Member States who have gained already experience in this field. The flexibility allowed by the anonymisation process should increase the number of Member States adhering to the dissemination project and therefore the number of data sets available to users.

6. Conclusions

In this work two quality dimensions were discussed: confidentiality and data utility. Both risk of re-identification measure and protection were inspired from the k-anonymity principle. The properties of these tools were discussed.

The risk measure is a flexible one, which could be easily adapted to any mixture of continuous and categorical variables. The flexibility of the risk measure and the selectiveness of the protection method allow us to choose from different degrees of anonymisation. The possibility to obtain these degrees of anonymisation allows us to focus on the users needs. The CIS experts and users could define a set of benchmarking statistics useful to measure relevant data utility aspects and to set thresholds to guarantee a common baseline quality for anonymous microdata.

In this way the comparability of analyses could be achieved in a multinational setting. Of course, the complete harmonisation of the dissemination of microdata files for research remains to be achieved, but the flexibility seems a promising starting point.

Acknowledgement Istat is not responsible for any views or results presented in the paper. The author was partially supported by the European project ESSnet-SDC.

References

Arundel, A. and Bordoy, C. (2005) “The 4th Community Innovation Survey: Final Questionnaire, Supporting Documentation, and the State-of-Art for the Design of the CIS”, working paper, available on request.

Bacher, J., Brand, R. and Bender, S. (2002), “Re-identifying Register Data by Survey Data Using Cluster Analysis: an Empirical Study”, *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5), 589-607.

Belderbos, R., Carree, M. and Lokshin, B. (2004), “Cooperative R&D and Firm Performance”, paper presented at Conference on Industrial Dynamics, Innovation and Development.

Breunig, M., Kriegel, H., Ng, R. T. & Sander, J. (2000), “LOF: identifying density-based local outliers”, *ACM SIGMOD Int. Conf. on Management of Data, Dallas, TX*, 93-104.

Defays, D. and Anwar, M.N. (1998), "Masking Microdata Using Micro-aggregation", *Journal of Official Statistics*, 14(4), 449-461.

Deville, J.C., Särndal, C.E. (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, 87, 376-382.

Domingo-Ferrer, J. and Torra, V. (2003), "Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage" *Statistics and Computing*, 13(4), 343-354.

Eurostat (2004), "Innovation in Europe: Results for the EU, Iceland and Norway", Panorama of the European Union, Theme 9 Sciences and technologies, European Communities.

Eurostat (2006), "The Fourth Community Innovation Survey (CIS4). Methodological Recommendations", Doc. Eurostat/F4/STI/CIS/2b.

Evangelista, R. and Mastrostefano, V. (2006), "Firm Size, Sectors and Countries as Sources of Variety in Innovation", *Economics of Innovation and New Technology*, 15 (3), 247–270.

Hundepool, A. et. al. (2006), "Handbook on Statistical Disclosure Control", available at http://neon.vb.cbs.nl/casc/.

Ichim, D. (2007), "Microdata Anonymisation of the Community Innovation Survey Data: a Density Based Clustering Approach for Risk Assessment", *Documenti Istat*, 2, available at www.istat.it.

Ichim, D. (2007), "Disclosure Control for Business Microdata: a Density-Based Approach", submitted.

ISTAT (2004), "Statistiche sull'innovazione delle imprese. Anni 1998-2000". *Informazioni* n. 12.

Klomp, L. and van Leeuwen, G. (2001), "Linking Innovation and Firm Performance: a New Approach", *International Journal of the Economics of Business*, 8(3), 343–364.

Leppälahti, A. and Teikari, I. (2007), "Problems with Micro-data from Small Countries", paper presented at the 32nd CEIES Seminar Innovation Indicators - more than technology.

Loof, H. and Heshmati, A. (2002), "Knowledge Capital and Performance Heterogeneity: a Firm Level Innovation Study*", International Journal of Production Economics*, 76(1), 61–85.

Mastrostefano, V. and Pianta, M. (2007), "Innovation Dynamics and Employment Effects", submitted.

Sweeny, L. (2002), "k-anonimity: A Model for Protecting Privacy", *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5), 557--570.

Willenborg, L. and De Waal, T. (2001), *Elements of statistical disclosure control*, Lecture Notes in Statistics. New York: Springer.

Winkler, W. (2004), "Re-identification Methods for Masked Microdata", *Privacy in Statistical Databases*, Eds. J. Domingo-Ferrer and V. Torra, 216-230. Berlin: Springer-Verlag.

Zeileis, A., Kleiber, C., Kramer, W. and Hornik, K. (2003), "Testing and Dating of Structural Changes in Practice", *Computational Statistics and Data Analysis*, 44, 109--123.