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 could be easily adapted to different national settings. A strategy to achieve comparable dissemination at European level is also discussed.

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 (MFR). 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 MFR. The three main parts of the statistical disclosure control process: risk assessment, disclosure limitation method and the quality assessment 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.. A full survey description of CIS is given in Eurostat (2006).

The CIS 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. The stratification variables used for the
Italian CIS4, 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. In some strata a census was conducted. The average response rate turned out to be 49%. A calibration methodology, currently applied at ISTAT, was used for the estimation process. For CIS4, the number of enterprises and the 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).

3. Disclosure Scenarios
The evaluation of the individual re-identification risk is necessary to identify those survey records containing some information that could uniquely characterize a unit in the population. This information could be used by an intruder to increase his knowledge about a respondent.

The data protector generally 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, see Hundepool (2006). 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 MFR 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 the sampled units. It is hard to accept the hypothesis that the researcher could have some information obtained as an insider.

Once the MFR is released, the NSI has no more any control on the way the file is used. However, the signed contract legally impedes the researcher 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 MFR 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 sample because of their significant impact on the studied phenomenon. The greatest enterprises are also the most well-known ones. Consequently, a spontaneous identification might occur. Moreover, even the bona-fide 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. Below few common characteristics of these 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 Nace. Consequently, the dissemination of CIS data at a more aggregated level of NACE 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 or the share of turnover that is due to new or improved products (quantifying the economic relevance of innovations). Each registered component of \textit{RTOT} 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 Evanghelista (2006), Klomp (2001), Mastrostefano (2007).

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

As a result of this literature review, a possible list of statistics to be used as benchmarking purposes in data utility definition may be made. Ratios of innovation variables as a mean to analyse scaled quantities seem predominant. Also the in turnover variations seem relevant.

3.3. Harm
The disclosure scenarios should also assess the confidential content of the MFR. 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 its 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 file 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 scenario, the identifying variables are determined. For the Italian CIS4 it was considered that an enterprise could be identified using the following structural variables: \textit{Nace}, \textit{Nuts}, \textit{Size} and \textit{Turnover}, see Ichim (2007). The continuous variable \textit{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.

The main question of the risk assessment phase 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. 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 \textit{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 Sweeney (2002). Or, the mass of a unit 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 the 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 correctly 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.

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 $t$ of key variables may be written as $t = (t^n, t^c)$. Here $t^n$ denotes the vector of numerical variables and $t^c$ 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 units $t_1 = (t_1^n, t_1^c)$ and $t_2 = (t_2^n, t_2^c)$ could be defined by multiplying the Euclidean distance between the vectors of the numerical key variables, $t_1^n$ and $t_2^n$, and the inverse of the indicator function of the vectors of the categorical key variables, $t_1^c$ and $t_2^c$, as in equation (1).

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

If $\beta = 0$, no categorical variable is considered in the disclosure scenario. In the CIS framework, this is equivalent to the assumption that the intruder would try to identify a unit by comparing only the 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 approximate 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 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

![Figure 1. The risk-density concept for continuous variables.](image-url)
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 \( V \), then \( V \) 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 (2007).

<table>
<thead>
<tr>
<th>Key variables</th>
<th>( \beta_{\text{Nace}} )</th>
<th>( \beta_{\text{Size}} )</th>
<th>( \beta_{\text{Nuts}} )</th>
<th>Distance function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turnover</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( e(T_1, T_2) )</td>
</tr>
<tr>
<td>Nace</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>( I^{-1}(N_1, N_2) )</td>
</tr>
<tr>
<td>Turnover, Nace</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>( e(T_1, T_2) * I^{-1}(N_1, N_2) )</td>
</tr>
<tr>
<td>Turnover, Size</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>( e(T_1, T_2) * I^{-1}(S_1, S_2) )</td>
</tr>
<tr>
<td>Turnover, Nace, Size</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>( e(T_1, T_2) * I^{-1}(N_1, N_2) * I^{-1}(S_1, S_2) )</td>
</tr>
</tbody>
</table>

Table 1. 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. 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 Turnover independently on the Nace categories may be observed. 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, all the units represented by blue diamond points might be considered 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 as illustrated by the red squares symbols in figure 2. The situation worsens when 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 in the risk assessment phase is to define a value for the parameter \( k \) in the \( k \)-anonymity principle. Based on its own dissemination policy, the NSI may define the minimum number \( M^* \) of units to which a unit \( u \) should be confused to be considered safe. 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 conditions hold:

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

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 \( |N_{M^*}(u)| \) denotes the number of elements in \( N_{M^*}(u) \). \( LRD_{M^*} \) 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:

\[
LOF_{M^*}(u) = \frac{\sum_{u' \in N_{M^*}(u)} LRD_{M^*}(u')}{|N_{M^*}(u)|} \tag{4}
\]
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 because $LRD_{M^*}(u) = LRD_{M^*}(u), \forall u \in N_{M^*}(u)$. 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 $\alpha$ and define as being at risk of re-identification those units for which $LOF_{M^*}(u) > \alpha$. The threshold $\alpha$ may be set by simply using some quantile criteria. This is a very simple and robust approach, but, unfortunately, it would mean that, in each combination of the categorical key variables, there exists a fixed percentage of units at risk. In presence of isolated units, the ordered $LOF_{M^*}$ values present a sudden change in slope, as illustrated in figure 3. The value $\alpha$ corresponding to this abrupt change would give a reliable indication of the isolated units. $\alpha$ could be automatically determined by means of structural change models, see Zeileis (2003). An advantage of such setting of $\alpha$ is that the percentage of units at risk of re-identification would not be $a$-priori defined. Manually increasing/decreasing the value of $\alpha$ 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.

Moreover, it should also be noted that for extreme choices of $\alpha$, none or all the units would be considered at risk of re-identification. More details on the usage of the $LOF_{M^*}$ function in the SDC framework may be found in Ichim (2007).

4. Disclosure Limitation Methodology

Once the units at risk of re-identification are determined, if the risk is considered too high, a protection method should be applied in order to reduce it. 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, Nace and Size 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 Nuts. Based on the percentages of rare cases for different Nuts levels, see table 2, it was decided to release only the national level. The other two structural categorical variables were left almost unchanged, except for very few combinations of Nace where a 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 these units at risk was achieved also by perturbing the structural continuous variable, i.e. Turnover.

<table>
<thead>
<tr>
<th>Geographical location</th>
<th>Percentage of rare cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUTS3</td>
<td>16.62%</td>
</tr>
<tr>
<td>NUTS2</td>
<td>7.47%</td>
</tr>
<tr>
<td>NUTS1</td>
<td>1.48%</td>
</tr>
</tbody>
</table>

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

The perturbation methods should be assessed from two points of view. First the protection method should be adequate to the type of dissemination. For MFR, a selective protection should be used. Only the units at risk and only the key and confidential variables should be modified. The selectiveness property 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 it was 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. A simple method like the imputation from the nearest neighbour could be used. In the SDC 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. For the Italian CIS4 microdata file, for each combination of categorical key variables, an individual ranking was applied only 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. Consequently, all the units would be subject to a micro-aggregation process because of their location on the tail of the 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 the categorical key variables. On the contrary, if there is no categorical key variable, the micro-aggregation would be applied irrespective of any combination of categorical key variables. The drawbacks of this latter approach were discussed also in Leppälähti (2007) and Franconi (2007).

The micro-aggregation parameter should be equal to or greater than $M^*$. Only 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 SDC. In table 3, several particular cases of the discussed methodology are shown. The other possible extensions may be easily derived.

<table>
<thead>
<tr>
<th>Categorical keys</th>
<th>Continuous keys</th>
<th>$\alpha$</th>
<th>Perturbation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Turnover</td>
<td>Max</td>
<td>No</td>
</tr>
<tr>
<td>None</td>
<td>Turnover</td>
<td>(0, Max)</td>
<td>Imputation from the nearest safe neighbour and individual ranking only on tails, irrespective of any combination of categorical key variables</td>
</tr>
<tr>
<td>None</td>
<td>Turnover</td>
<td>0</td>
<td>Individual ranking on all the units, irrespective of any combination of categorical key variables</td>
</tr>
<tr>
<td>None</td>
<td>Turnover, X</td>
<td>Max</td>
<td>No</td>
</tr>
<tr>
<td>None</td>
<td>Turnover, X</td>
<td>(0, Max)</td>
<td>Imputation from the nearest safe neighbour and multivariate micro-aggregation only on tails, irrespective of any combination of categorical key variables</td>
</tr>
<tr>
<td>None</td>
<td>Turnover, X</td>
<td>0</td>
<td>Multivariate micro-aggregation on all the units, irrespective of any combination of categorical key variables</td>
</tr>
<tr>
<td>Nace</td>
<td>Turnover</td>
<td>Max</td>
<td>No</td>
</tr>
<tr>
<td>Nace</td>
<td>Turnover</td>
<td>(0, Max)</td>
<td>Imputation from the nearest safe neighbour and individual ranking only on tails, for each category of the categorical key variable</td>
</tr>
<tr>
<td>Nace</td>
<td>Turnover</td>
<td>0</td>
<td>Individual ranking on all the units, for each category of the categorical key variable</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover</td>
<td>Max</td>
<td>No</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover</td>
<td>(0, Max)</td>
<td>Imputation from the nearest safe neighbour and individual ranking only on tails, for each combination of the categorical key variables</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover</td>
<td>0</td>
<td>Individual ranking on all the units, for each combination of the categorical key variables</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover, X</td>
<td>Max</td>
<td>No</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover, X</td>
<td>(0, Max)</td>
<td>Imputation from the nearest safe neighbour and multivariate micro-aggregation only on tails, for each combination of the categorical key variables</td>
</tr>
<tr>
<td>Nace, Size</td>
<td>Turnover, X</td>
<td>0</td>
<td>Multivariate micro-aggregation on all the units, for each combination of the categorical key variables</td>
</tr>
</tbody>
</table>

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

5. Data Quality
The last step of the disclosure limitation process is the quality assessment 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 SDC 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 RTOT would be
perfectly preserved if the individual ranking irrespective of the categorical key variables were 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, applying such a slight perturbation could not give sufficient protection. As it may be observed in the lower part of the figure 4, if this type of individual ranking were applied to the Turnover, there might be some units that exactly preserve their dominance characteristic. In other words, even without knowledge of the exact values, if the intruder knew that an enterprise was dominant before the MFR 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.

5.1 Record Linkage
A further experiment was performed to assess the ability of the $LOF_M$ function to detect the units at risk. A record linkage (RL) experiment was performed using the Chambers of Commerce database. In the RL experiment, the units correctly matched were determined. Then the number of neighbours in certain neighbourhoods of these units at risk was computed. Two sets of blocking variables, \{Nace\} and \{Nace, Size\}, were used. The units at risk in this framework were compared with the ones identified by the $LOF_M$ function. For each unit correctly matched in the RL experiment, $LOF_M$ was measured using two values for $M^*$, i.e. $M^* = 3$ and $M^* = 5$, respectively. To yield the two risk measures comparable, neighbourhoods of predefined width around the Turnover values were considered. For example, see table 4, among the units correctly matched in the RL experiment performed using \{Nace\} 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\} and having less than 5 units within 10% of their Turnover values were labelled at risk by the $LOF_M$ 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 were large enterprises, i.e. with more than 250 employees.

<table>
<thead>
<tr>
<th>$M^*$</th>
<th>1 unit within 10%</th>
<th>less than $M^*$ units within 10%</th>
<th>less than $M^*$ units within 20%</th>
<th>less than $M^*$ units within 30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nace</td>
<td>3</td>
<td>88%</td>
<td>84%</td>
<td>97%</td>
</tr>
<tr>
<td>Nace Size</td>
<td>3</td>
<td>63%</td>
<td>60%</td>
<td>74%</td>
</tr>
<tr>
<td>Nace</td>
<td>5</td>
<td>88%</td>
<td>73%</td>
<td>87%</td>
</tr>
<tr>
<td>Nace Size</td>
<td>5</td>
<td>63%</td>
<td>58%</td>
<td>70%</td>
</tr>
</tbody>
</table>

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. 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 RL experiments were performed using the original microdata file derived from the Italian CIS4 file, following the approach presented in 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, according to 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 RL experiment, then the number of units correctly identified 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 now 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 increase the number units correctly matched in the RL experiment using \{Nace, Size\} as blocking variables.

### 5.2 Information Content

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. Ratio of Turnover variances before and after the application of the perturbation.](image)

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 MFR 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 continuous 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.](image)
6. Harmonised dissemination

In its co-ordinating role Eurostat is working on 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: 1) development of general methodological guidelines, 2) definition of benchmarking statistics and 3) 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 unique statistical disclosure methodology. This strategy surely has the lowest costs in terms of implementation, testing and application. It might be thought 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 would imply, to start with, the indication of the methodological paradigm of SDC. Such paradigm would state 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 would 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 MFR could be achieved twofold: 1) developing flexible anonymisation methodologies and 2) constraining the output of the anonymisation methodology. In other words, the harmonisation concept could be defined by means of a set of minimal requirements on both input and output of the anonymisation process.

6.1 Input Harmonisation – disclosure limitation methodology

In principle, on the input phase, a significant improvement might be reached by using flexible SDL 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 SDL methodology could be more easily accepted by many Member States. Of course, each Member State should first agree with the underlying disclosure scenario and the risk assessment methodology.
As previously stated, an interesting feature of the anonymisation procedure outlined in section 4 is that for extreme choices of the parameters in the risk assessment phase the $LOF_M^*$ protection reduces to different protection methods. 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
Data utility/data quality is 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. In the SDC framework it is proposed to control 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. Careful definition and tuning of the benchmarking statistics coupled with clear threshold setting would allow comparability of analyses among different protection methods and different parameter choices in different Member States. Given the assurance of a pre-defined acceptable re-identification risk level, preservation of the 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.

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 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. Of course, the complete harmonisation of the dissemination of MFR remains to be achieved, but the flexibility and comparability seem a promising starting point.

Acknowledgement Istat is not responsible for any views or results presented in the paper. This work was supported in part by the ESSnet-SDC project.

References


