CASC PROJECT

Computational Aspects of Statistical Confidentiality

31 December 2003

Proposal for the creation of a micro-data file for research for Business Surveys – Part 1

Giovanni Seri
Silvia Polettini

ISTAT, PSM/C
Via C. Balbo, 16
00184, Roma
Italy

Deliverable No: 5-D5

1. Introduction

Assessing the performance of a disclosure limitation method for business microdata is a difficult task widely discussed in the literature (see recent work of Brand, 2002, Yancey *et al*. , 2002, Sebé *et al*., 2002), and a standard "best" way to proceed is not yet available. At any rate, the performance of any SDC method is universally measured in terms of "information loss" and "risk of disclosure".
For business microdata most of the information collected takes the form of quantitative variables with skew distributions. Such variables are often representative of enterprise size and so can lead to enterprise re-identification. For example, information about turnover can lead to the identification of a very large and well-known enterprise in a particular class of the NACE classification of economic activity. This means that, even though they are not always publicly available, quantitative variables are "natural" key variables as they are extremely identifying. The practical consequence of this is that all units are unique (rare) with respect to a small set of quantitative variables. Moreover, in many cases, populations of enterprises are sparse and firms are easily identifiable simply by their economic activity and geographical position. Finally, other *a priori* information such as knowledge about the survey design can be used to identify an enterprise, see Cox (1995).
Because of these problems, available re-identification criteria based on the rareness (uniqueness) of a unit in the population (see for example, Skinner and Holmes, 1998, Fienberg and Makov, 1998, Benedetti and Franconi, 1998) are not always applicable and the protection method based on data reduction (such as global recoding and local suppression, see Willenborg and de Waal, 2001) used in the case of social microdata may be ineffective or cause too large an information loss.
As a consequence, many of the protection techniques specifically proposed for business microdata aim to perturb the original data in such a way that enterprises are not recognisable. Perturbative methods achieve data protection from a twofold perspective. On one hand, if the data are modified, re-identification by means of record linkage or matching algorithms is made harder and uncertain; on the other hand, even when an intruder is able to re-identify a firm, he/she cannot be sure that the data disclosed are consistent with the original data. As it is the case with data reduction, the level of data protection generally increases with perturbation of data. Of course, this latter aspect has to be balanced with the need to make the information content of perturbed data as similar as possible to that of the original data in order to preserve the quality of statistical results. The other aspect of the problem is to establish whether or not the modified data lead to breaches of confidentiality, i.e. re-identification of statistical units.
This work aims at verifying the effectiveness and applicability of the statistical disclosure control techniques proposed in Franconi and Stander (2002) and in Polettini *et al.* (2002). The proposed approach has been tested on real data from the Community Innovation Survey (CIS). CIS is a business survey having a Europe-wide perspective involving a mixture of continuous and categorical variables. In Polettini *et al*. (2002) the analytical validity of the method has been exploited. Therefore in this paper we mainly explore the assessment of the level of protection guaranteed by the method.
In the next Section we give a description of the experimental data used. In Section 3 we discuss the risk of disclosure for business microdata and describe the disclosure scenario. In Section 4 and 5 we outline the framework used for the risk assessment. Experimental results are presented in Section 6 and some conclusions are given in Section 7.

2. Experimental data

We consider the microdata set to be protected as a matrix A with n rows representing units and m+s columns representing the m key variables (x_j, j=1,…,m) and the s confidential variables (c_r, r=1,…,s) respectively:

$$A=(X,C), \text{ where } X=\{x_{i,j}, i=1,\ldots,n;\ j=1,\ldots,m\} \text{ and } C=\{c_{i,r}, i=1,\ldots,n;\ r=1,\ldots,s\}. \quad (1)$$

As C contains all the confidential variables that do not allow for re-identification, this matrix is generally released unchanged and it is not directly involved in the risk assessment. The application of protection methods consists in replacing X with a different matrix $Y=\{y_{i,j}, i=1,\ldots,n; j=1,\ldots,m\}$. For our purposes only X and Y are relevant and C will be ignored in the following. In some cases we can have s=0, which means that all the released variable may lead to re-identification of a unit and therefore need to be protected.

The data used in this work come from the Italian sample of the Community Innovation Survey (CIS). The variables of the CIS can be divided into two sets. The first contains all the general information about the enterprise such as its main Economic activity, Geographical area, Number of employees, Turnover, Export, and Group membership. The second contains survey specific information concerning "innovation" and "research". The only key variable we consider in this second group is the amount of Expenditure for research and innovation (R&I). The variables Turnover, Expenditure for R&I and Exports are measured in millions of Italian lire. We use the logarithmic transformation to reduce the skewness of the data. According to the strategy of disclosure limitation applied, we stratify by Economic activity (two digit NACE code) omitting enterprises with zero Turnovers or Exports (see Polettini *et al.*, 2002 for details). In this paper, we consider enterprises involved in the following main economic activities:

Economic activity (Manufacture of)	NACE code	No of enterprises
food products and beverages	15	222
wearing apparel and dressing	18	157
chemicals and chemical products	24	205
rubber and plastic products	25	214
other non-metallic mineral products	26	185
fabricated metal products, except machinery and equipment	28	338
machinery and equipment n.e.c	29	528
furniture; manufacturing n.e.c.	36	274
Total		2123

Besides the economic activity, recoded as just mentioned before, we only consider quantitative key variables. In this work, other categorical key variables as the Geographical area are omitted in order to maintain a congruous number of units in each sector of economic activity. Protection of variables other than those we discuss in the following requires special consideration. We present results of our risk assessment framework for the following sets of quantitative key variables:

Case 1. Turnover;

Case 2. Turnover and Number of employees;

Case 3. Turnover, Number of employees and Exports;

Case 4. Turnover, Number of employees, Exports and Expenditure for R&I.

3. Risk assessment for perturbed business microdata

The risk of disclosure is usually measured as risk of re-identification, the re-identification being the possibility of establishing a link between a microdata record released by a NSI and a target enterprise. Consequently, protection methods are applied in order to make statistical units not recognisable in the population. As regards perturbed economic microdata, the approach suggested in the literature to assess the risk of disclosure refers to record linkage procedures. Domingo-Ferrer and Torra (2001) report on "distance based record linkage" (Pagliuca and Seri, 1999), "probabilistic record linkage" (Winkler, 1998, Yancey *et al.*, 2002) and "interval disclosure" (Domingo-Ferrer and Torra, 2001), keeping into account different sets of key variables or parameters.

We define the disclosure scenario for a NSI assuming that the external archive available to the intruder coincides with the original file, X. We consider this disclosure scenario as the "worst" for a NSI because it is implicitly assumed that the intruder knows that the target enterprise is included in the released file and that there is no differences between the original data and the external archive due, for example, to classification errors. In other words, the intruder has the "best" representation of the original data to try re-identification.
The disclosure scenario and the linkage context are coherent because both assume the presence of two distinct archives (the released file and an external archive available to the intruder) with overlapping information (key variables are assumed to be in both the archives) useful to perform a link between records.
A linkage procedure is based on a comparison between a released record **y** and a record **x**, say d(**x**,**y**), and a decision rule to designate a pair (**x**,**y**) as a link or not. In the case under study, as the data are perturbed and variables are numerical, d(**x**,**y**) can be defined as a distance. The distance induces a variable Z defined as: z=d(**x**,**y**).
All the possible pairs (**x**,**y**) are in the product space X**x**Y. Let M be the set of pairs corresponding to correct links and let U be the set of nonlinks:

$$XxY=M\cup U.$$

The two distributions:

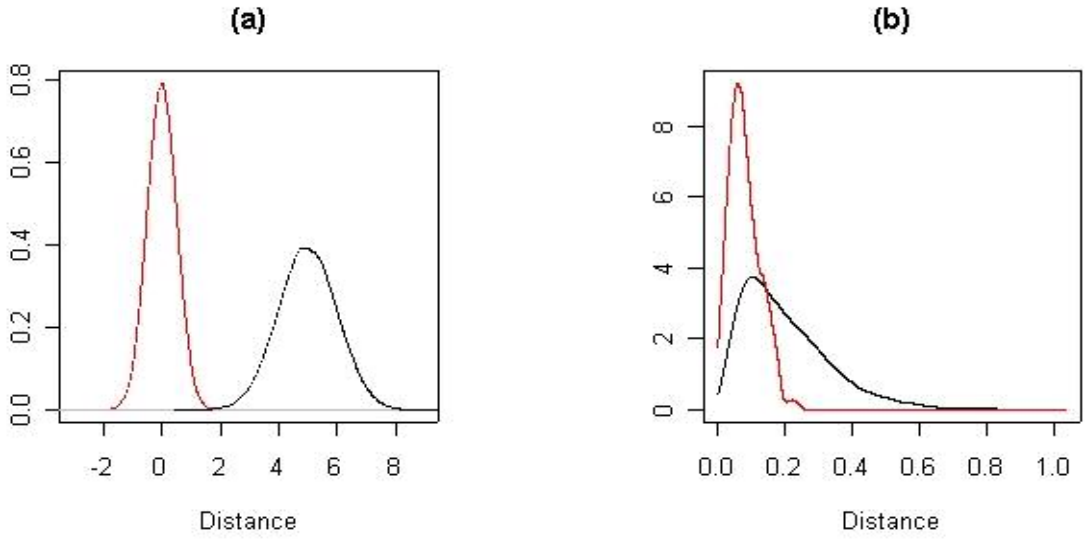
$$m(z) = P(Z = z \mid (\mathbf{x},\mathbf{y}) \in M)$$

and (**2**)

$$u(z) = P(Z = z \mid (\mathbf{x},\mathbf{y}) \in U)$$

are the basic ingredients of probabilistic record linkage and their estimation is the main issue in the record linkage literature. As in our framework the two archives involved in the procedure are completely known, we can assign without uncertainty each pair to link/nonlink. Therefore the disclosure scenario assumed in this work allows us to compute the two distributions in (2).

Figure 1. Example of densities for the comparison variable Z corresponding to links, red line on the left, and nonlinks, black line on the right



The "discrepancy" between m(z) and u(z) in (2) is indicative of the effectiveness of the linkage procedure. The farther apart the two distributions, the easier the designation of a pair as a link or nonlink. A theoretical example is shown in Figure 1 (a): the values of the variables Z used to compare records are strongly different for links and nonlinks and the two distributions m(z), red line

on the left, and u(z), black line on the right, do not overlap. Similarly, the closer the two distributions are, the harder is to make correct links. Following this reasoning, a synthetic index for the level of protection can be obtained by measuring the "discrepancy" between the two distributions. In Figure 1 (b) the estimated distribution of m(z) and u(z), for Case 1 (Turnover as the only key variable) and NACE code 18, are plotted. We used the Kolmogorov-Smirnov two-sample test of the null hypothesis that the observed values of m(z) and u(z) were drawn from the same continuous distribution: Table 1 shows the results for Case 1. The test reveals that the distributions of the distance for matches and non matches are not indistinguishable. Similar results were obtained for Case 2, 3 and 4. However we can still examine the extent of separation between the above distributions. The record linkage procedure described by Fellegi and Sunter (1969) indeed results in higher error rates, e.g. less effective matching algorithms, when these densities do overlap.

Table 1. Kolmogorov-Smirnov test

Nace code	No. of enterprises	ks-test	p-value
15	222	0.4897	2.20E-16
18	157	0.4877	2.20E-16
24	205	0.6487	2.20E-16
25	214	0.5693	2.20E-16
26	185	0.5598	2.20E-16
28	338	0.5477	2.20E-16
29	528	0.6328	2.20E-16
36	274	0.5058	2.20E-16

4. Criteria for risk assessment of perturbed microdata

Perturbation is the most widely adopted procedure to protect business microdata. Like all the protection methods, it faces the trade-off between information loss and safety. We do not address the first issue, that for our application is discussed in Franconi and Stander (2002) and Polettini *et al.* (2002). We will mainly deal with the safety of the file. To measure the extent of protection of the microdata file, we introduce two related concepts of protection. These are both considered at the individual level; this enables us to effectively perform a risk assessment of the data, offering the opportunity of add further protection to selected, high risk records that do not meet the desired protection level.

We consider a statistical unit "not recognisable" under two respects, representing two related aspects: the first concerns the *perturbation* of the data. In a sense, we can say that perturbative methods "bash the statistical unit's face in". If NSIs do not release confidential variables, e.g. s=0 in (1), a certain amount of perturbation can be considered to ensure enough protection to release a record. In the context of tabular data release this concept is applied to each cell of a table.

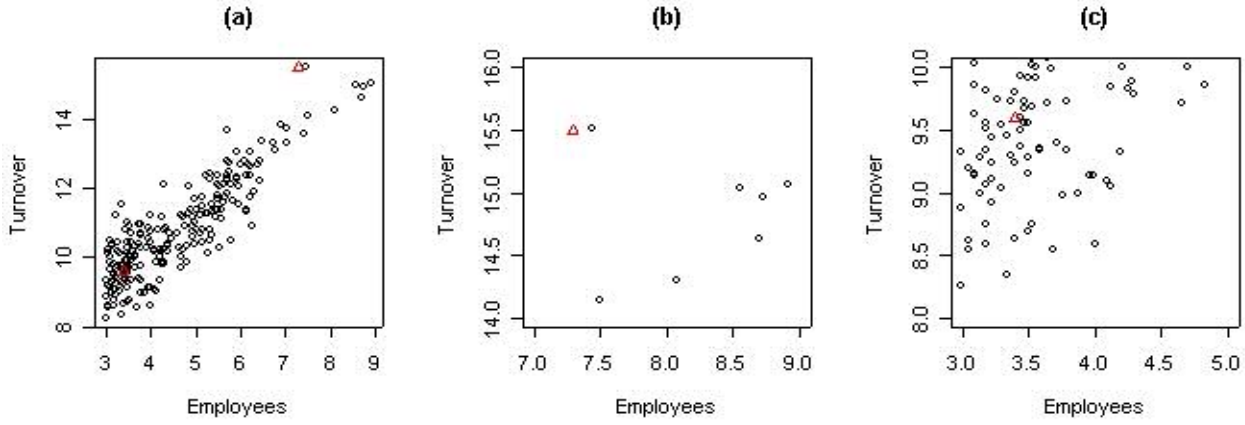
The second aspect refers directly to the concept of record re-identification. Roughly speaking, a statistical unit is not recognisable if it is "hidden" in the population: when a large number of units in the population share with the target record the same characteristics, re-identification of this record is more uncertain than it is when the record is a unique in the population. Since our data are perturbed, an intruder performing an exercise of record linkage will not find any record matching exactly the target. The intruder will therefore consider as possible links those records of the external archive which are similar to the target with respect to a set of key variables. This leads us to consider the concept of *neighbourhood* of a released record. For each record **y** in the released file, we denote as neighbour of **y** any unit **x** in the original sample (the external archive) that is "similar" to **y**. The level of protection ensured by each perturbed unit will depend on the *number of neighbours* we can attach to it.

Note that when releasing social microdata, the approach adopted is similar to the one we have just described. In that case the number of units in the population with identical scores on a set of categorical key variables is computed. In this case the neighbourhood of a unit *i* is the set of records that are indistinguishable from *i*. Likewise, both local suppression and global recoding aim to increase the number of units in the population likely to be confused with the one under investigation.

Of course, in order to define a neighbourhood, a measure of similarity between units is needed; in general we will refer to multivariate distance measures. Clearly, the definition of "similarity" and hence of neighbourhood is arbitrary as it depends on the intruder's strategy of attack, and this makes the procedure weaker. In addition, the intruder's strategy of attack is based on a measure of distance, but NSIs do not know which one; moreover, NSIs do not know which set of key variables is used by the intruder to try a re-identification. For this reason we consider 4 different sets of key variables, each corresponding to a strategy of attack.

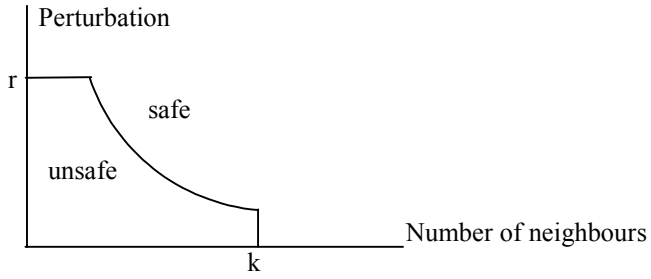
We consider as an example the case where Turnover and Number of employees are the only key variables. In Figure 2 (a) two perturbed units, triangles with coordinates (7.3,15.5) and (3.4,9.6), are plotted along with the real original values of the sample (logarithmic scale is used). In plot (b) and (c) the positions of those two perturbed units is zoomed in. Figure 2 (b) represents the situation when an outlier in the original data is weakly perturbed by the protection method. An intruder trying to compare the released record with the data in his/her archive will have great confidence that the link between two such points is a true link because the released record is very close (similar) just to a single isolated point. Figure 2 (c) shows the released record hidden in a crowded cloud of points, and of course this makes it harder to identify the correct link.

Figure 2. Quantifying the extent of protection by the number of neighbours (NACE 15). Key variables are Turnover and no. Employees



5. Confidentiality plot for perturbed microdata

The aspects discussed in the previous section, namely the amount of perturbation of the original datum and the number of units that are similar to the original record, can be jointly exploited to assess the protection of a record. In particular, a graphical tool connecting the above mentioned aspects can be introduced. We denote by "confidentiality plot" a graph in which the perturbed data can be represented with coordinates the "number of neighbours" (horizontal axis) and the "amount of perturbation" (vertical axis). We use this tool to check safety of each individual record in a file and the corresponding quality of representation of the original record.



The threshold "r" means that the released value is safe if it is distorted over the r% of the original value, whatever the risk of re-identification. On the other hand, the threshold "k" means that if the perturbed value is close to more than k units in the population, then no perturbation is required to protect this value. The curve represents the trade-off between these two aspects: the more the released value is hidden in the population, the less the perturbation that is required, and *vice versa*. The area under the curve is defined "unsafe" zone, because points in this area represent records that are not protected enough. Perturbed data can be plotted on this graph with coordinates the "number of neighbours" (horizontal axis) and the "amount of perturbation" (vertical axis).

If the released file contains confidential variables, re-identification has to be given priority and only the threshold "k" on the number of neighbours becomes relevant. Different graphs are generated by different scenarios or methods. For example, for microaggregated data (Domingo-Ferrer and Mateo-Sanz, 2002) the graph would consist of a vertical line at k because the method impose that at least k units are identical as regard the key variables.

The position of each point in the confidentiality plot with respect to the vertical axis can also be interpreted as an index of the quality of representation. In other words "information loss" and "perturbation" induced in a single record are equivalent labels for the vertical axis. As the CIS data consists also of confidential variables, the threshold for the re-identification risk can be properly represented in the graph as a vertical line for a fixed "k".

The connection between the "confidentiality plot" and the R-U confidentiality map of Duncan *et al.* (2001) is clear. R-U confidentiality map were introduced to compare the performances of different disclosure limitation methods. The main difference is that a point in the R-U confidentiality map represents a disclosure limitation techniques (e.g. a given data release), whereas a point in the confidentiality plot represents individual data (in tabular data release context it can be referred to a single value in a cell).

In this study for each individual record we propose to measure the amount of perturbation (information loss) as the relative error, e.g. the Euclidean distance, between the released and the original data. The aim is to measure the error that is to be accepted by a user accessing the released data in place of the original data. Denoting by **y** the vector of key variables for a generic record in the released file, and by **x** the corresponding vector of true values, we compute:

$$\text{Information loss} = \|\mathbf{y}-\mathbf{x}\|/\|\mathbf{x}\|. \qquad (\mathbf{3})$$

This definition can be extended to those protection methods that results in predictive intervals (in this case **y** is the interval midpoint) or in suppression of values (in this case it is often possible to evaluate a feasibility interval for the true value). From the intruder's point of view, **y** can be considered the estimate of **x** that minimizes the error in (3). Clearly, the measure of information loss in (3) is also a measure of the perturbation induced in the data, as it represents the *distance of the released value from the truth*. This approach suggest a generalization of the confidentiality plot: for any protection method (perturbative or not), it can be exploited to assess the individual risk of disclosure of each single data release (e.g. a cell of a table or a record in a file). We will to develop these aspects elsewhere.

In order to compute the number of neighbours of each record, according to the record linkage approach described in Section 3, we define the comparison variable Z as the relative euclidean distance between a released record **y** and each record **x** in the external archive:

$$z=\|\mathbf{y}-\mathbf{x}\|/\|\mathbf{y}\| \quad \forall\, \mathbf{y} \in Y;\ \forall\, \mathbf{x} \in X.$$

For each **y** to be released, the number of neighbours is computed as the number of original records **x** $\in$ X such that z is lower than a threshold δ:

$$\text{Number of neighbours of } \mathbf{y} = \#\{\, \mathbf{x} \in X : z < \delta\}. \qquad (4)$$

Next, we denote as "neighbourhood" of **y**:

$$N(\mathbf{y})=\{\, \mathbf{x} \in X : z < \delta\}.$$

Probabilistic record linkage procedures consider the probability of the type 1 error, i.e. the probability of designating a pair as a link when it is not. The two densities of the distributions m(z) and u(z) in (2) were estimated over the set of pairs $(\mathbf{x},\mathbf{y})\in M$, corresponding to correct links and $(\mathbf{x},\mathbf{y})\in U$, corresponding to nonlinks, respectively. We remark that, as stated before, the two sets M and U are completely known for the given disclosure scenario. Assuming that lower distances are likely to be measured in the occurrence of a true link, we have:

$$\Pr(z < \delta | (\mathbf{x},\mathbf{y}) \in U)=\alpha$$

where α is the acceptable level for the type 1 error and the δ, the critical distance, is fixed accordingly. In practice, the neighbourhood of **y** consists of all the original data $\mathbf{x}\in X$ that, for given α, are not rejected as nonlinks according to the probabilistic record linkage procedure. Choosing a smaller α turns into reducing the number of neighbours of the released record, which makes the linkage procedure more prudential.

Figure 3, (a) and (b), show the histograms of the distances measured for NACE 29 using Turnover as the only key variable (Case 1). In Figure 3 (c) the estimated densities m(z) and u(z) are plotted for α=0.05. α represents the area on the left of the vertical line at δ under the u(z) curve (the flat one).

Another approach that we investigated is to proceed according to the Fellegi-Sunter procedure (Fellegi and Sunter, 1969). The likelihood ratio:

$$t(z)=m(z)/u(z)$$

is used to perform a statistical test of the null hypothesis

$$H_0: \{(\mathbf{x},\mathbf{y}) \in U, \text{ the true distribution is } u(z)\}$$

versus

$$H_1: \{(\mathbf{x},\mathbf{y}) \in M, \text{ the true distribution is } m(z)\}.$$

Let τ, the critical value, be such that:

$$P(m(z)/u(z)> \tau | (\mathbf{x},\mathbf{y}) \in U)= \alpha;$$

H_1 is accepted if $z\in\{z : m(z)/u(z)> \tau\}$ for the given value of α. We then consider for α=0.05:

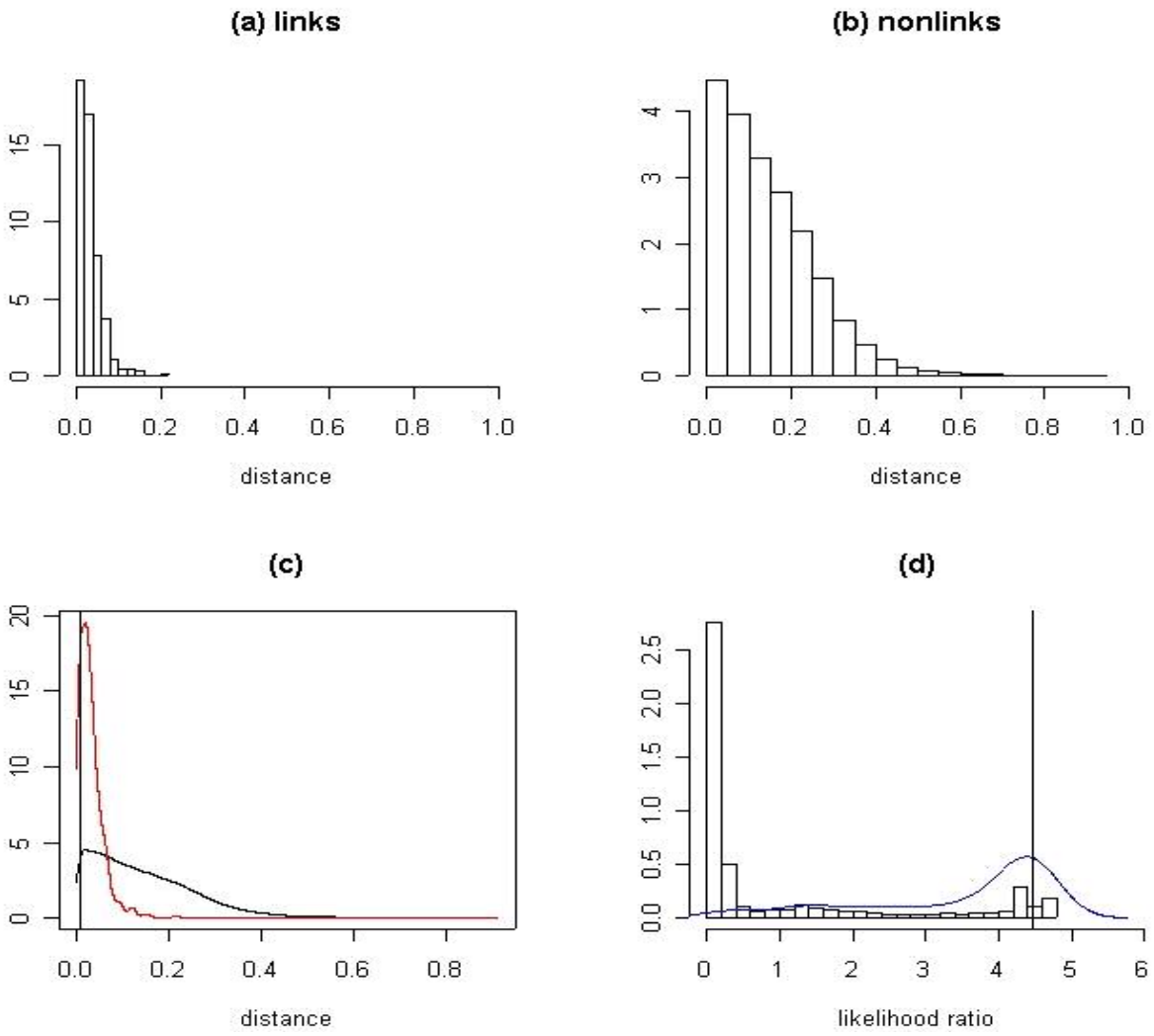
$$\text{Number of neighbours of } \mathbf{y} = \#\{\mathbf{x} \in X : m(z)/u(z) > \tau\}; \qquad (5)$$

the neighbourhood of **y** is defined as:

$$N(y)= \{\mathbf{x} \in X : m(z)/u(z) > \tau\}.$$

Figure 3 (d) shows, for NACE 29 and Turnover as the only key variable, the histogram of $t(z)|(\mathbf{x},\mathbf{y})\in U$, the estimated density of $t(z)|(\mathbf{x},\mathbf{y})\in M$ and the vertical line representing τ for α=0.05.

Figure 3. Case 1, NACE 29. (a) and (b), histograms and (c), densities of distances between links and nonlinks; (d) histogram of the likelihood ratio, t(z)



The two different ways to count the neighbours of a record **y** described in (4) and (5) led always to similar (often the same) results; for computational simplicity we preferred (4). Figure 4 shows the graphical approach denoted as "confidentiality plot", for Case 1 and NACE 29; this is the case that presents the biggest differences between the two approaches among the 4 cases under study. For each record **y** in the released file, a point is plotted on the graph with coordinates: the number of neighbours (horizontal axis) and the information loss (vertical axis). Squares identify cases for which the nearest-neighbour is the correct link, that is when for a given record **y** the pair $(\mathbf{x},\mathbf{y})\in M$ and **x** is the nearest neighbour of **y**. Crosses identify cases for which the correct link is in the neighbourhood, that is when for a given record **y** the pair $(\mathbf{x},\mathbf{y})\in M$ and $\mathbf{x}\in N(\mathbf{y})$. A filled square highlights the unit presenting the highest Turnover in the original data, which in most cases is the most easily re-identifiable. This latter consideration suggests another possible intruder's strategy of attack: the comparison of ranks for the biggest and therefore more easily identifiable enterprises.

6. Experimental results

In order to describe the test data used in this experiment, we show in Figure 5 the plots (on logarithmic scale) of the perturbed data *vs* the original data of the four eligible key variables separately. All figures in this section refers to NACE 18. Each observation is represented as little black circle. The straight line represents equality between perturbed and original data.

Figure 4. Confidentiality plot for business perturbed microdata: distance based

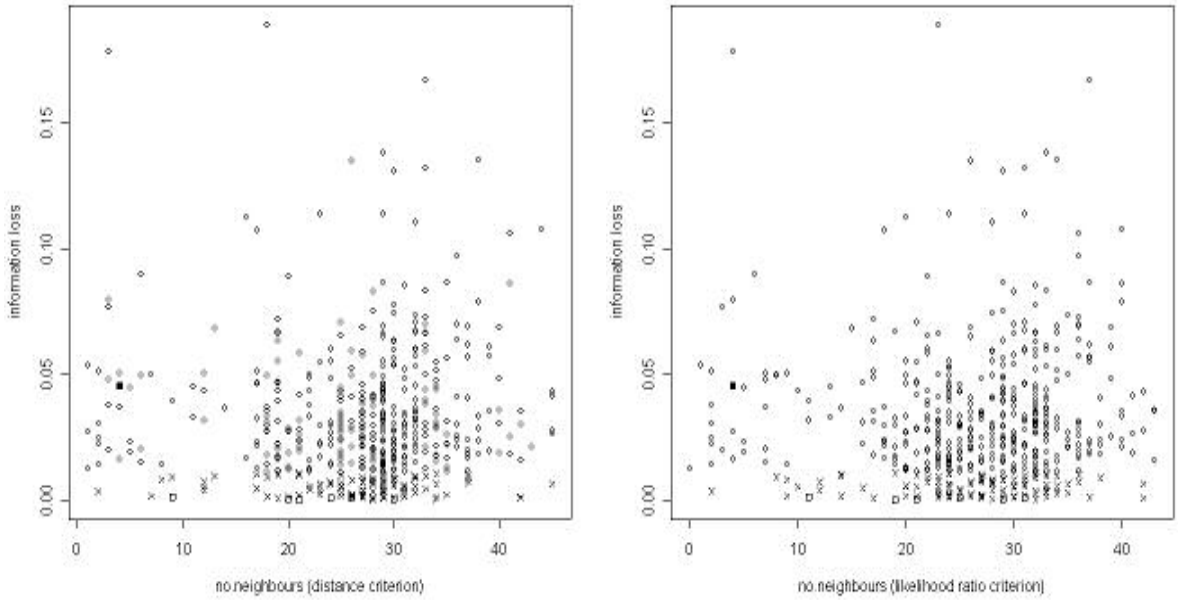
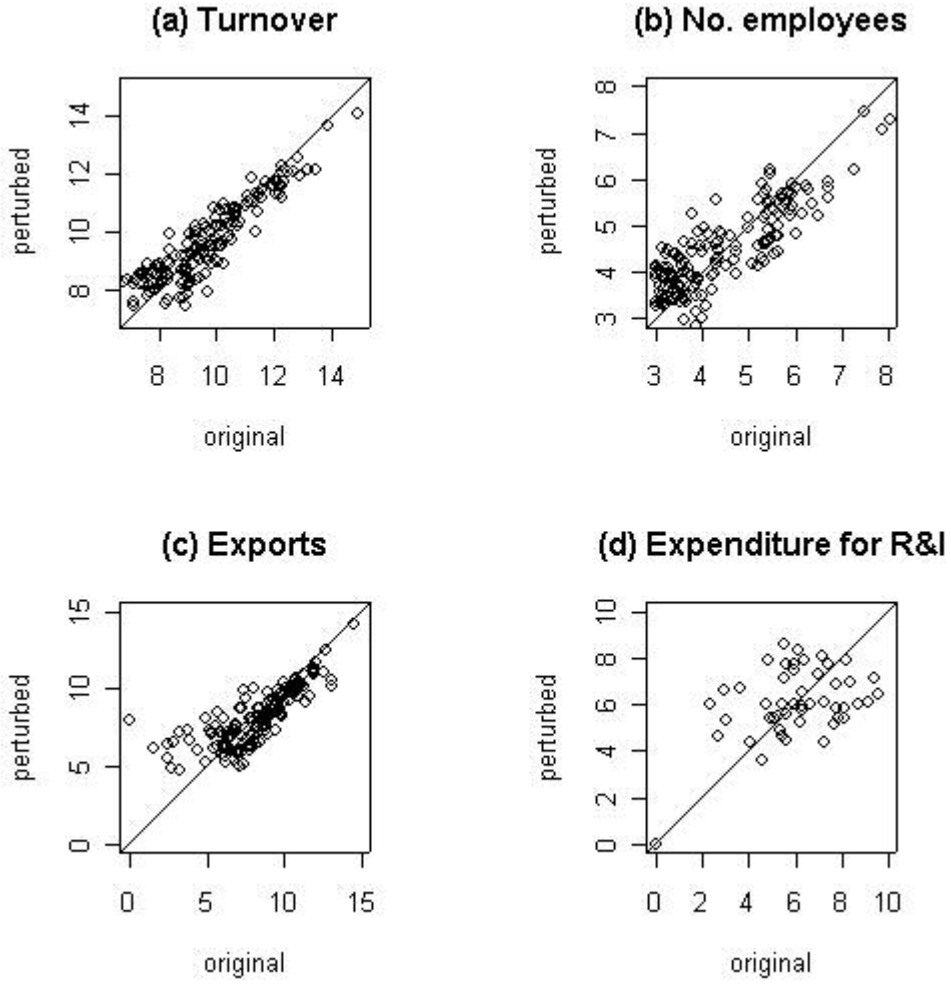


Figure 5. Plots of perturbed data *vs* original data of Turnover, Number of employees, Exports and Expenditure for R & I



As expected, a few enterprises present high value of Turnover, Number of employees and Export, namely the more identifiable, while the Expenditure for R & I seems to be less reliable as proxy of the enterprise size mainly because of the presence of zeros. Anyway, the ranks of the highest enterprises is not the same across the key variables. Notice that the range of perturbed values is smaller than the one relative to the original values: this is a consequence of the protection method used.

Figures 6-9 shows confidentiality plots for Case 1 to 4, each representing a different set of key variables. Analogously, Tables 2-5 reports for each NACE code: the number of nearest neighbours that are correctly linked (number of records **y** such that the pair $(\mathbf{x},\mathbf{y})\in M$ and **x** is the nearest neighbour of **y**), the critical distance δ (maximum distance defining the neighbourhood), number of units correctly linked in their own neighbourhood (number of records **y** such that $(\mathbf{x},\mathbf{y})\in M$ and $\mathbf{x}\in N(\mathbf{y})$), critical value of the likelihood ratio τ and the corresponding number of neighbours.

Results should be interpreted in the light of the strongly prudential disclosure scenario we assumed, specifically: the hypothesis that the external archive coincides with the original data, the reduced number of records in the file (zero Turnovers or non-export firms are excluded) and the relatively small size of the neighbourhood, each contribute to lowering the number of neighbours. Moreover, as a consequence of the disclosure scenario, we count neighbours only on the basis of the sample data (we do not consider sampling weight). Anyway, for a given number k of neighbours, plots shows that there are always a few unsafe records.

For a few units the neighbourhood is empty. These are represented in the confidentiality plot with coordinate Number of neighbours = 0. In this case the perturbation applied to records **y** with empty $N(\mathbf{y})$ is higher than the critical distance defining the neighbourhood; moreover, no original record **x** is "close enough" to **y**.

If the intruder's strategy of attack is "nearest neighbour based", only an average percentage between 2% and 3% of correct links can be detected. However the fact that the enterprise with highest turnover is not always correctly linked using the nearest neighbour supports the idea of ineffectiveness of such a strategy of attack.

The average percentage of units **y** whose correct link lies in the neighbourhood $N(\mathbf{y})$ is between 18,4% and 31,6% depending on the set of key variables considered. It can also be observed that increasing the number of key variables the average critical distance increases as well, as expected.

Case 1 to 4 each define a strategy of data protection (see Section 2.1) but none of them is clearly superior. Of course, Case 1, i.e. an intruder using a single key variable, describes a situation that is easier to interpret.

In order to investigate the effect of different neighbourhood sizes, e.g. different critical distances defining the neighbourhood, we present the confidentiality plot relative to α=0.1 for Case 1 (see Figure 10). For NACE 18 this corresponds to a critical distance δ=0.0292. It means that a record in the external archive is considered neighbour of a released record if it differs of less than 2.92% from the released record in terms of Turnover. The two isolated points on the left of the plot highlight the presence of two "extreme" units, namely the two highest firms in the NACE group. It seems that, these easily identifiable units require a strong perturbation of data to make them similar at least to k other units in the external archive.

The average critical distance for the 8 NACE groups is 2.22% (to be compared with an average 1.1% in Table 2. We do not consider the possibility of directly selecting a threshold on the relative distance defining a neighbourhood as an alternative to setting a value for α. Of course, this is an alternative, perhaps more intuitive, way to proceed.

As another possible strategy of attack by the intruder we considered the comparison of ranks. For the biggest and therefore more easily identifiable enterprises, the intruder might adopt a different, perhaps more effective, strategy of attack. He/she might proceed by ranking units according to the variable that he/she considers proxy of enterprise size. Therefore we considered whether the ranks of the ten enterprises presenting the highest turnover are preserved. This rank based strategy seems to be more effective than the one based on the nearest neighbour, at least for the units we

investigated. We analysed only the highest 10 positions according to Turnover (see Table 6) as they are usually the most easily identifiable. We observed that 7 out of 8 times (8 are the NACE sectors investigated) the maximum rank is preserved. This means that the unit presenting highest Turnover in the perturbed data is correctly linked with the unit presenting the highest Turnover in the original data. For the second highest Turnover we observe 6 out of 8 occasions of coincidence.

7. Conclusions

Assessing the performance of a disclosure limitation method for business microdata is a difficult task. Particularly hard is to measure the risk of disclosure, as no widely accepted standard procedure is available.

Because of the high risk of disclosure of business microdata, perturbative protection methods are suggested in the literature. We have outlined a way to assess graphically the level of protection guaranteed by SDC methods to each record and applied this framework to the method presented in Franconi and Stander (2002) and in Polettini *et al.* (2002).

The disclosure scenario assumed is strongly prudential as the external archive available to the intruder for re-identification purpose is the original data file.

At a first stage we analysed the two distribution defined in (2) and we found them strongly overlapping. This induces a higher error rate in probabilistic record linkage procedures. As a consequence, the protection method applied is effective with regard to a global evaluation of the safety level of the file and against an intruder's strategy of attack based on record linkage. Note that this is the more frequently assumed strategy of attack in the context of perturbed microdata.

We then proposed a generalization of the criteria (namely, the amount of perturbation and the number of records that share the same or similar characteristics) usually employed to assess the re-identification risk of a single record in a microdata file. A graphical framework for joint evaluation of risk of disclosure and information loss has been suggested.

Experimental results have been presented relative to CIS data (for a partial set of enterprises engaged in economic activities corresponding to codes 15, 18, 24, 25, 26, 28, 29 and 36). Different strategies of attack by an intruder have been taken into account.

The presence of outliers relative to large, easily identifiable, enterprises, as highlighted by the confidentiality plot indicates the need for higher protection of these records. For this reason we could not claim safety of all data protected by the method under investigation. In order to protect these easily identifiable units a strong perturbation of data is required to make them similar to the other units in the file. In this case, the information loss can be extremely severe and hardly would the protected data meet the requirement of analytic validity of the data. It seems reasonable that we can have a satisfactory trade-off between protection and the quality of data included in the file only for a sample of small size enterprises. However, this means that we do not give a complete representation of the observed phenomena.

Acknowledgement

The authors would like to thank Luisa Franconi for helpful comments.

The views expressed are those of the authors and do not necessarily reflect the policies of Istat.

Figure 6. Case 1 for NACE 18: confidentiality plot

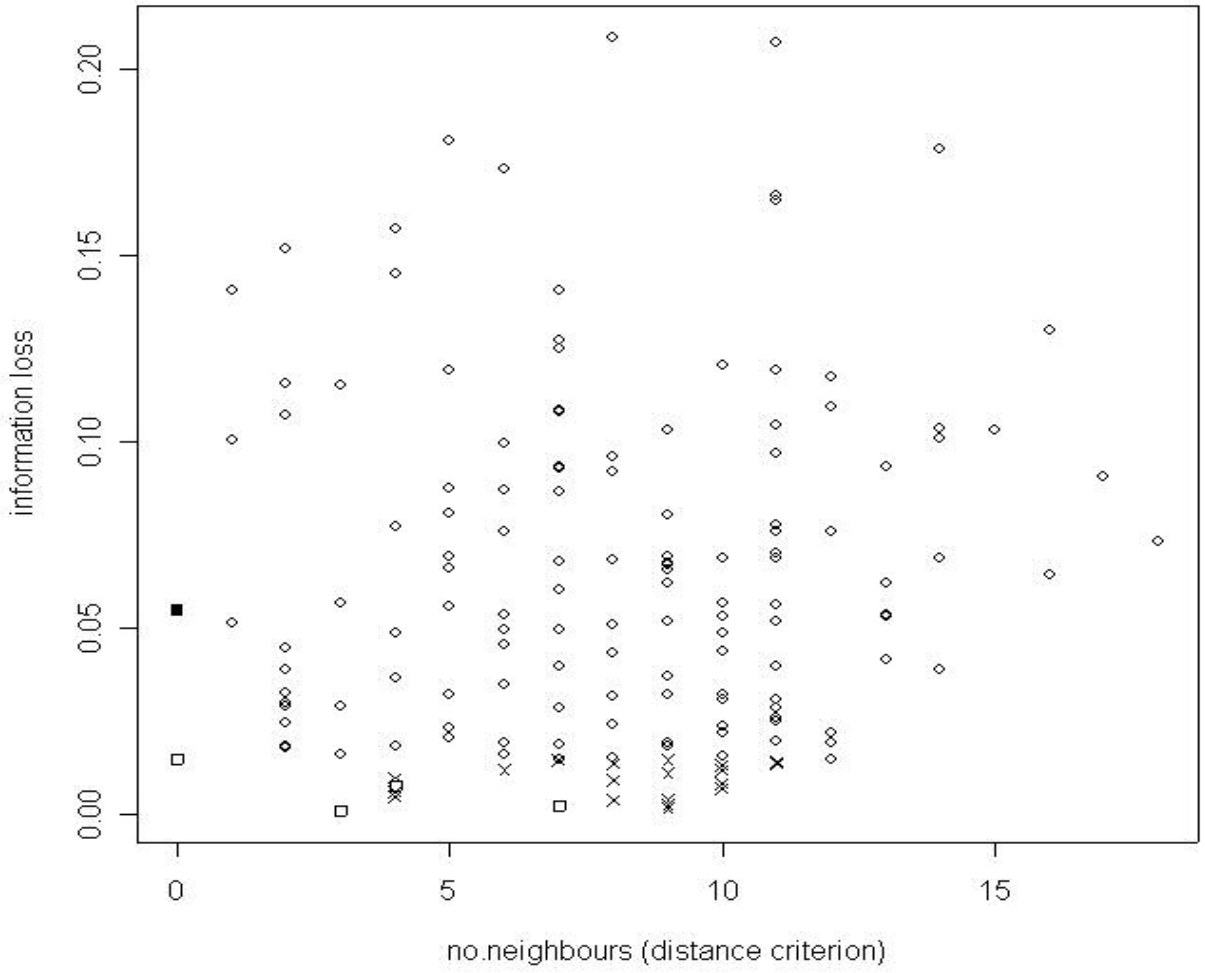


Table 2. Case 1., for each NACE: number of nearest neighbour correctly linked, critical distance δ, number of units linked correctly in their neighbourhood, critical value of the likelihood ratio τ and corresponding number of links in neighbourhood.

NACE	Correct nearest neighbour link (% no. record)	Critical distance δ*100	Correct link in neighbourhood – distance criterion (%)	Critical value for likelihood ratio=τ	Correct link in neighbourhood - likelihood ratio criterion
15	4 (1.8)	1.08	27 (12.2)	2.85	27
18	4 (2.5)	1.45	23 (14.6)	3.27	22
24	6 (2.9)	1.12	43 (21.0)	4.77	47
25	3 (1.4)	0.99	39 (18.2)	3.81	40
26	10 (5.4)	1.25	39 (21.1)	4.08	39
28	6 (1.8)	0.89	75 (22.2)	3.75	75
29	6 (1.1)	1.1	109 (20.6)	4.48	106
36	4 (1.5)	0.97	36 (13.1)	2.96	49
	Total 43 (2.0)	Average 1.1	Total 391 (18.4)		Total 405

Figure 7. Case 2 for NACE 18: confidentiality plot

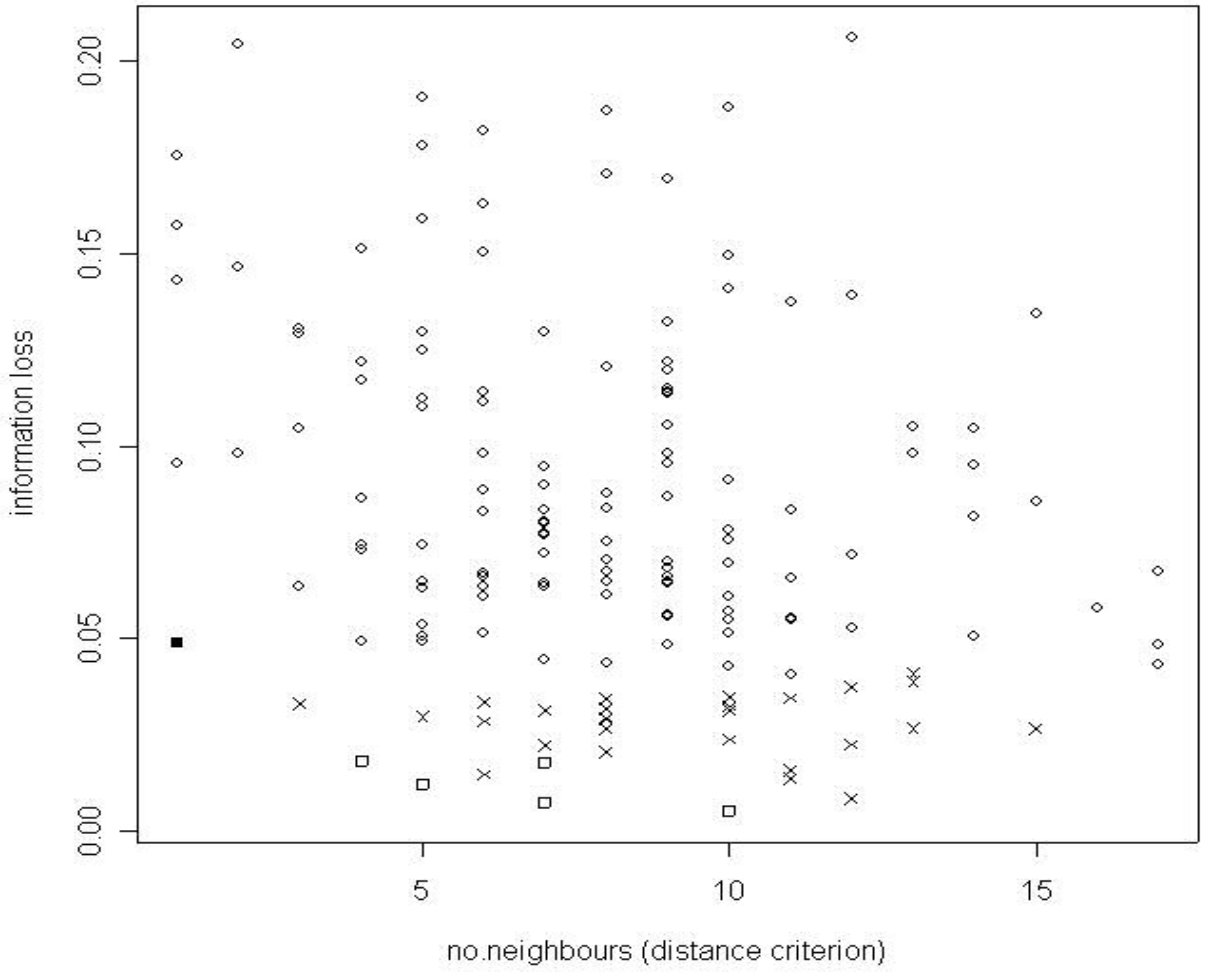


Table 3. Case 2., for each NACE: number of nearest neighbour correctly linked, critical distance δ, number of units linked correctly in their neighbourhood, critical value of the likelihood ratio τ and corresponding number of links in neighbourhood.

NACE	Correct nearest neighbour link (% no. record)	Critical distance δ*100	Correct link in neighbourhood – distance criterion (%)	Critical value for likelihood ratio=τ	Correct link in neighbourhood – likelihood ratio criterion
15	5 (2.3)	3.2692	55 (24.8)	3.55	55
18	5 (3.2)	4.0838	33 (21.0)	3.68	33
24	6 (2.9)	2.8516	63 (30.7)	5.2	65
25	3 (1.4)	2.6919	58 (27.1)	4.8	59
26	13 (7.0)	3.181	58 (31.4)	4.6	58
28	8 (2.4)	2.4126	80 (23.7)	4.13	80
29	5 (0.9)	2.7472	158 (29.9)	4.71	162
36	4 (1.5)	2.7077	56 (20.4)	3.91	60
	Total 49 (2.3)	Average 3.0	Total 561 (26.4)		Total 572

Figure 8. Case 3 for NACE 18: confidentiality plot

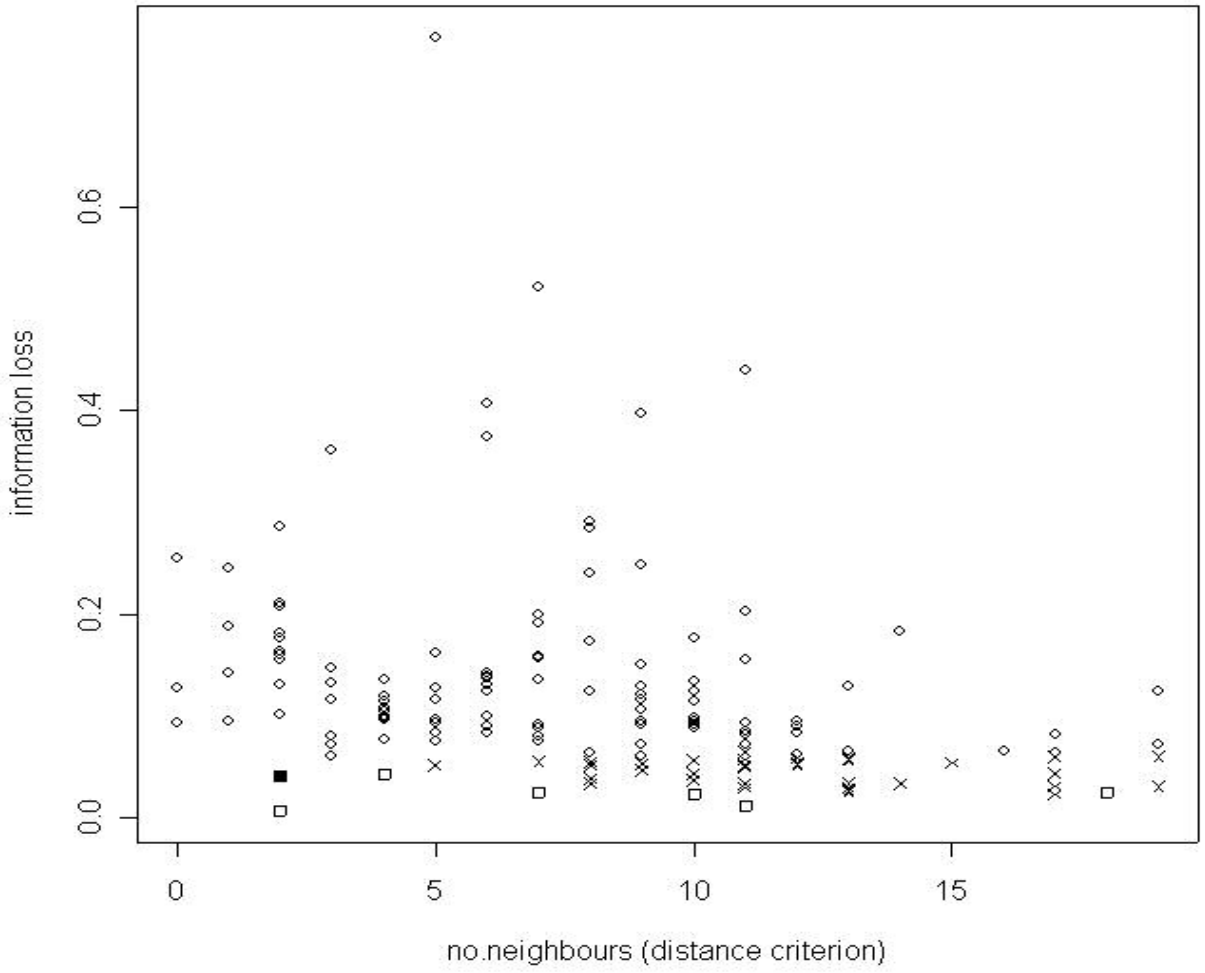


Table 4. Case 3., for each NACE: number of nearest neighbour correctly linked, critical distance δ, number of units linked correctly in their neighbourhood, critical value of the likelihood ratio τ and corresponding number of links in neighbourhood.

NACE	Correct nearest neighbour link (% no. record)	Critical distance δ*100	Correct link in neighbourhood – distance criterion (%)	Critical value for likelihood ratio=τ	Correct link in neighbourhood - likelihood ratio criterion
15	4 (1.8)	6.558	45 (20.3)	2.78	45
18	7 (4.5)	6.1898	44 (28.0)	3.13	44
24	8 (3.9)	4.5219	57 (27.8)	3.57	57
25	5 (2.3)	5.3055	65 (30.4)	3.05	65
26	6 (3.2)	5.4856	44 (23.8)	3.71	45
28	6 (1.8)	4.9268	67 (19.8)	2.9	67
29	6 (1.1)	4.2614	169 (32.0)	4.71	169
36	5 (1.8)	5.4675	52 (19.0)	3.26	52
	Total 47 (2.2)	Average 5.3	Total 543 (25.6)		Total 544

Figure 9. Case 4 for NACE 18: confidentiality plot

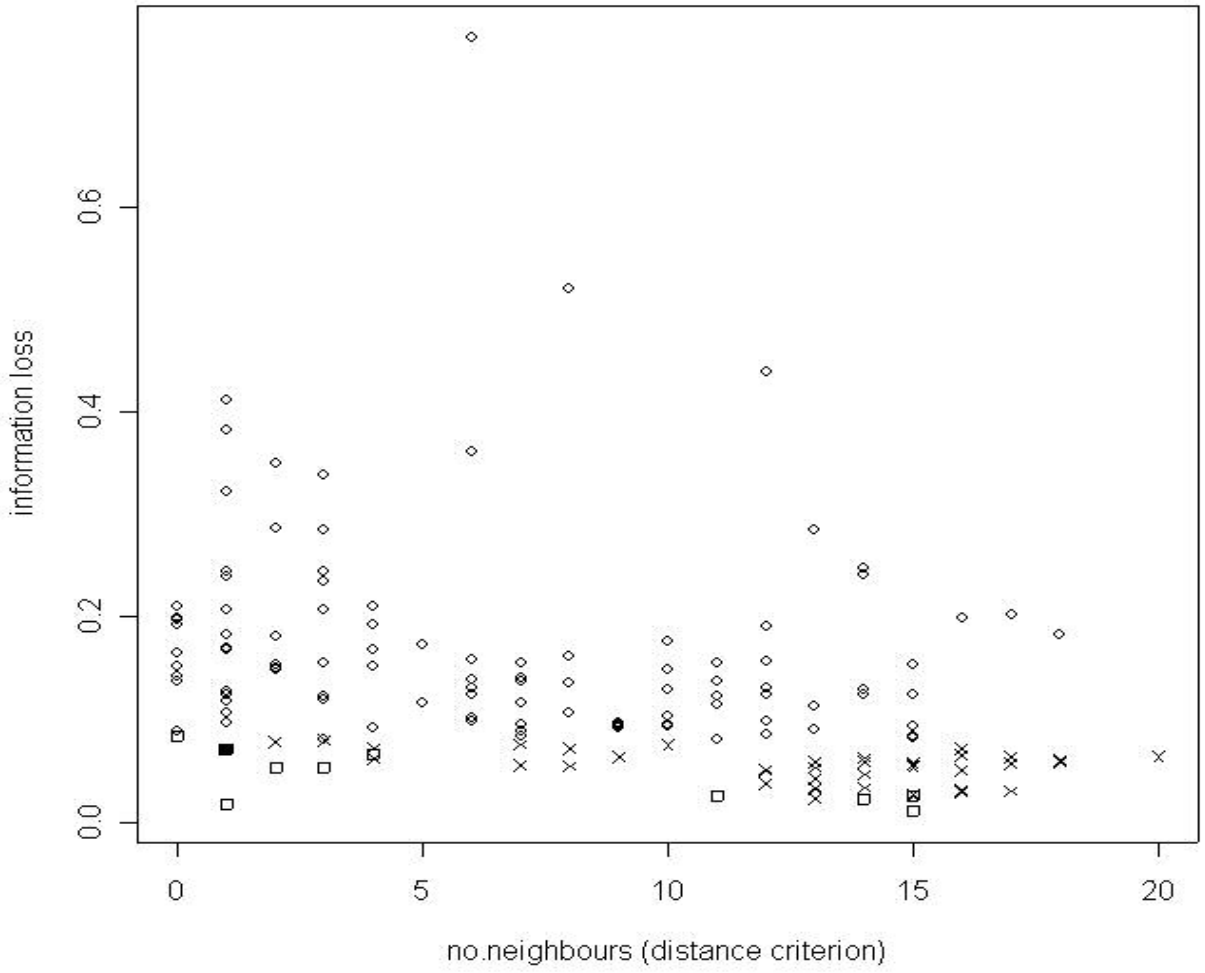


Table 5. Case 4., for each NACE: number of nearest neighbour correctly linked, critical distance δ, number of units linked correctly in their neighbourhood, critical value of the likelihood ratio τ and corresponding number of links in neighbourhood.

NACE	Correct nearest neighbour link (% no. record)	Critical distance δ*100	Correct link in neighbourhood – distance criterion (%)	Critical value for likelihood ratio=τ	Correct link in neighbourhood - likelihood ratio criterion
15	11 (5.0)	9.4117	64 (28.8)	3.74	64
18	11 (7.0)	8.1874	50 (31.8)	4.01	50
24	7 (3.4)	7.2591	69 (33.7)	4.35	69
25	14 (6.5)	8.4581	69 (32.2)	3.97	68
26	11 (5.9)	8.1834	59 (31.9)	4.65	60
28	12 (3.6)	7.6494	100 (29.6)	3.82	100
29	9 (1.7)	7.409	183 (34.7)	4.46	183
36	8 (2.9)	7.771	76 (27.7)	4.54	76
	Total 8 (2.9)	Average 8.0	Total 670 (31.6)		Total 670

Figure 10. Case 1 for NACE 18: confidentiality plot for α=0.1

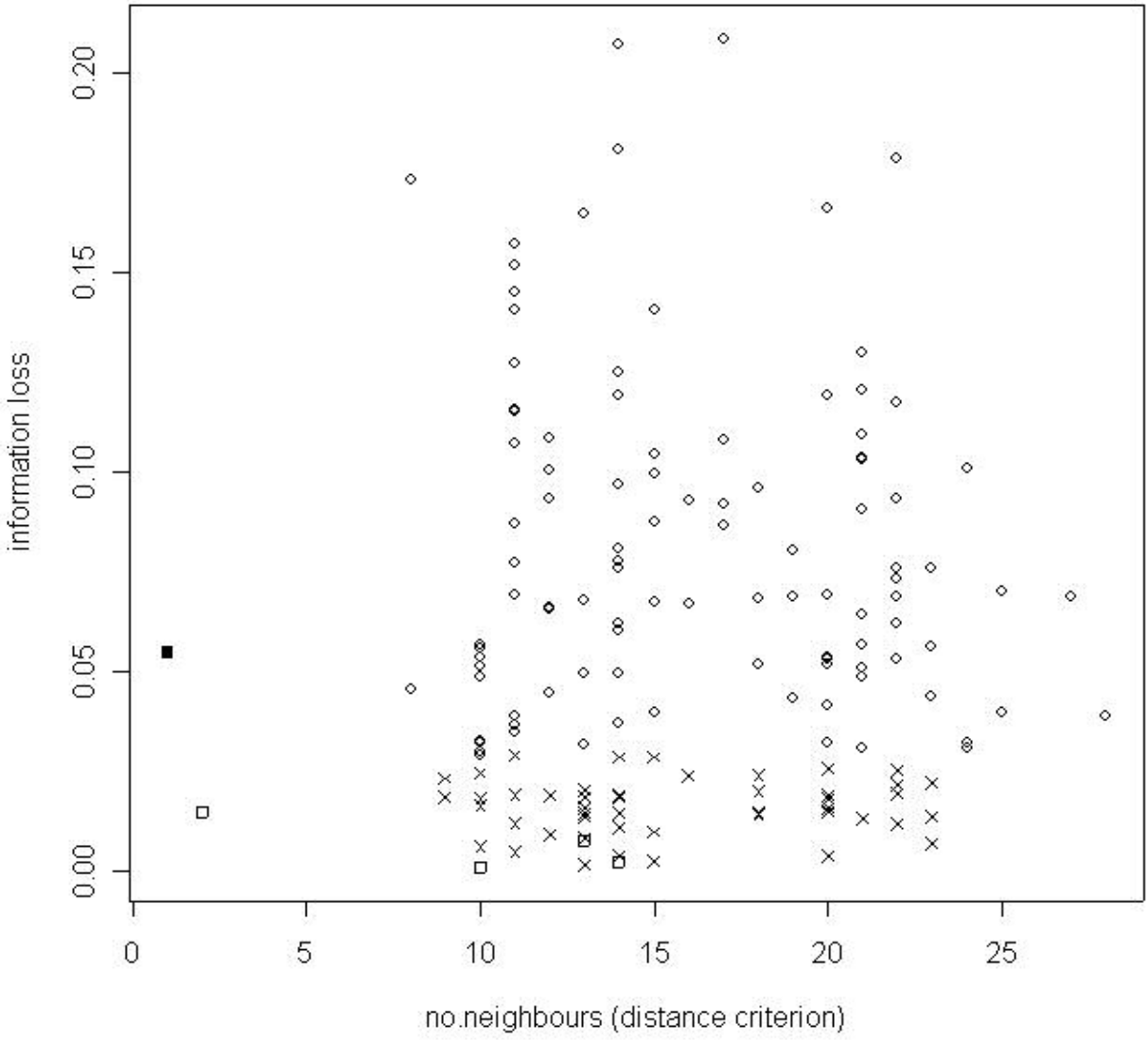


Table 6. Number and position of ranks preserved among the ten highest scores of turnover of the original data

NACE	Nr. of ranks preserved in the first ten positions	Antiranks									
		1[1]	2	3	4	5	6	7	8	9	10
15	3	-	-	X	X	X	-	-	-	-	-
18	3	X	X	-	-	-	-	-	X	-	-
24	3	X	X	X	-	-	-	-	-	-	-
25	4	X	X	-	-	-	X	-	X	-	-
26	3	X	X	-	-	-	X	-	-	-	-
28	2	X	X	-	-	-	-	-	-	-	-
29	2	X	-	-	-	-	-	X	-	-	-
36	4	X	X	X	X	-	-	-	-	-	-

[1] Cells in this column present "X" if the maximum rank is preserved and "-" otherwise. Similarly for the other columns.

References

Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. *Pre-proceedings NTTS '98, New Techniques and Technologies for Statistics*, Sorrento, 1, 225-232.

Brand, R. (2002). Microdata protection through noise addition. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 97-116.

Cox, L.H. (1995). Protecting confidentiality in business surveys. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (Eds.), New-York: Wiley, 443-476.

Domingo-Ferrer, J., and Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering,* 14, 189-201.

Domingo-Ferrer, J., and Torra, V., (2001). A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 111-133, 2001.

Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L. (2001). Disclosure risk vs. data utility: the R-U confidentiality map. *Technical Report LA-UR-01-6428*, Los Alamos National Laboratory.

Fellegi, I.P. and Sunter, A.B., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, (1969), 1183-1210.

Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14, 385-397.

Franconi, L. and Stander, J. (2002). A model based method for disclosure limitation of business microdata. *Journal of the Royal Statistical Society, D*, 51, 1-11.

Pagliuca, D. and Seri, G. (1999). Some results of individual ranking method on the System of Enterprise Accounts Annual Survey. *Esprit SDC Project, Deliverable MI-3/D2*.

Polettini, S. Franconi, L. and Stander, J. (2002). Model Based Disclosure Protection. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 83-96.

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J.M. and Torra, V. (2002). Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 163-171.

Skinner, C.J. and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

Willenborg, L. and de Waal, T. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Winkler, W.E. (1998). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics*, 1, 50-69.

Yancey, W.E., Winkler, W.E. and Creecy, R.H. (2002). Disclosure risk assessment in perturbative microdata protection via record linkage. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 135-152.