Data Mining and Knowledge Discovery, 11, 181–193, 2005

DOI: 10.1007/s10618-005-0011-9

Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata

JOSEP M. MATEO-SANZ josepmaria.mateo@urv.net
JOSEP DOMINGO-FERRER josep.domingo@urv.net
FRANCESC SEBÉ francesc.sebe@urv.net
Department of Computer Engineering and Mathematics,Rovira i Virgili University of Tarragona, Av. Països Catalans 26, E-43007, Tarragona, Catalonia

Editor: Geoff Webb

Received October 27, 2004; Accepted March 31, 2005

Published online: 2 September 2005

Abstract. Inference control for protecting the privacy of microdata (individual data) should try to optimize the tradeoff between data utility (low information loss) and protection against disclosure (low disclosure risk). Whereas risk measures are bounded between 0 and 1, information loss measures proposed in the literature for continuous data are unbounded, which makes it awkward to trade off information loss for disclosure risk. We propose in this paper to use probabilities to define bounded information loss measures for continuous microdata.

Keywords: database security, privacy, statistical disclosure control, microdata protection, information loss measures

1. Introduction

The purpose of privacy protection in statistical databases, also known as Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL), is to allow publication of statistical data in such a way that they do not give away confidential information that can be linked to specific individuals or entities. This is a relevant issue in domains as diverse as official statistics, e-health and e-commerce.

The challenge for SDC is to modify data in such a way that the risk of disclosing information on specific respondents becomes low enough while keeping at a minimum the information loss, i.e. the loss of the accuracy sought by database users. Thus, it is of paramount importance for a data protector using SDC to be able to compare and trade off information loss and disclosure risk to reach a suitable balance.

SDC can be applied to information in several formats and flavors: tabular data, dynamic databases and also microdata (individual respondent records). This paper is related to microdata protection and, more specifically, to measuring the information loss caused by SDC methods, also known as masking methods, when applied to the protection of continuous (numerical) microdata.

1.1. Contribution and plan of this paper

Section 2 presents previous information loss measures defined in the literature for continuous data. Section 3 discusses the problems of combining these existing information loss measures with disclosure risk measures. Section 4 describes a new approach to measuring information loss for continuous data which, being based on probabilities, provides measures bounded between 0 and 1 that can readily be combined with risk measures. Section 5 presents empirical results showing the high correlation of the new measures with previous measures in the literature; it also presents an example application of the proposed measures to evaluate the information loss caused by several masking methods to a specific microdata set. Section 6 is a conclusion.

2. Background on information loss measures for continuous data

A strict evaluation of information loss must be based on the data uses to be supported by the protected data. The greater the differences between the results obtained on original and protected data for those uses, the higher the loss of information. However, very often microdata protection cannot be performed in a data use specific manner, for the following reasons:

- Potential data uses are very diverse and it may even be hard to identify them all at the moment of data release by the data protector.
- Even if all data uses can be identified, issuing several versions of the same original dataset so that the i-th version has an information loss optimized for the i-th data use may result in unexpected disclosure.

Since data must often be protected with no specific data use in mind, generic information loss measures are desirable to guide the data protector in assessing how much harm is being inflicted to the data by a particular masking method.

Defining what a generic information loss measure is can be a tricky issue. Roughly speaking, it should capture the amount of information loss for a reasonable range of data uses. We will say there is little information loss if the protected dataset is analytically valid and interesting according to the following definitions by Winkler (1999):

- A protected microdata set is *analytically valid* if it approximately preserves the following with respect to the original data (some conditions apply only to continuous attributes):
  1. Means and covariances on a small set of subdomains (subsets of records and/or attributes)
  2. Marginal values for a few tabulations of the data
  3. At least one distributional characteristic
- A microdata set is *analytically interesting* if six attributes on important subdomains are provided that can be validly analyzed.

More precise conditions of analytical validity and analytical interest cannot be stated without taking specific data uses into account. As imprecise as they may be, the above definitions suggest some possible measures:

- Compare raw records in the original and the protected dataset. The more similar the masking method to the identity function, the lesser the impact (but the higher the disclosure risk!). This requires pairing records in the original dataset and records in the protected dataset. For masking methods based on the original data, each record in the protected dataset is naturally paired to the record in the original dataset it originates from. For synthetic microdata preserving only some features of the original data, pairing is more artificial. In Dandekar et al. (2002) we proposed to pair a synthetic record to the nearest original record according to some distance.
- Compare some statistics computed on the original and the protected datasets. The above definitions list some statistics which should be preserved as much as possible by a masking method.

To be specific, assume a microdata set with n individuals (records) and p continuous attributes. Let X be the matrix representing the original microdata set (rows are records and columns are attributes). Let X' be the matrix representing the protected microdata set. The following tools are useful to characterize the information contained in X and X':

- Covariance matrices V (on X) and V' (on X').
- Correlation matrices R and R'.
- Correlation matrices RF and RF' between the p attributes and the p factors $PC_1, \ldots, PC_p$ obtained through principal components analysis.
- Commonality between each of the p attributes and the first principal component PC_1 (or other principal components PC_i's). Commonality is the percent of each attribute that is explained by PC_1 (or PC_i). Let C be the vector of commonalities for X and C' the corresponding vector for X'.
- Factor score coefficient matrices F and F'. Matrix F contains the factors that should multiply each attribute in X to obtain its projection on each principal component. F' is the corresponding matrix for X'.

There does not seem to be a single quantitative measure completely capturing the informational difference between X and X'. Therefore, it was proposed in Domingo-Ferrer et al. (2001) and Domingo-Ferrer and Torra (2001a) to measure information loss through the discrepancies between matrices X, V, R, RF, C and F obtained on the original data and the corresponding X', V', R', RF', C' and F' obtained on the protected dataset. In particular, discrepancy between correlations is related to the information loss for data uses such as regressions and cross tabulations.

In Domingo-Ferrer et al. (2001) and Domingo-Ferrer and Torra (2001a), matrix discrepancy was measured in three ways:

Mean square error: Sum of squared componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Table 1. Information loss measures for continuous microdata.

	Mean square error	Mean abs. error	Mean variation
$X - X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V - V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(v_{ij}-v'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$R - R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF - RF'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C - C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F - F'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Mean absolute error: Sum of absolute componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Mean variation: Sum of absolute percent variation of components in the matrix computed on protected data with respect to components in the matrix computed on original data, divided by the number of cells in either matrix. This approach has the advantage of not being affected by scale changes of attributes.

Table 1 summarizes the measures proposed in Domingo-Ferrer et al. (2001) and Domingo-Ferrer and Torra (2001a). In this table, p is the number of attributes, n the number of records, and components of matrices are represented by the corresponding lowercase letters (e.g. x_{ij} is a component of matrix X). Regarding $X - X'$ measures, it makes also sense to compute those on the averages of attributes rather than on all data (call this variant $\bar{X} - \bar{X}'$. Similarly, for $V - V'$ measures, it would also be sensible to use them to compare only the variances of the attributes, i.e. to compare the diagonals of the covariance matrices rather than the whole matrices (call this variant $S - S'$).

In Yancey et al. (2002), it was observed that dividing by x_{ij} causes the $X - X'$ mean variation to rise sharply when the original value x_{ij} is close to 0. This dependency on the particular original value being undesirable in an information loss measure, the authors of Yancey et al. (2002) proposed to replace the mean variation of $X - X'$ by the more stable measure

$$\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{\sqrt{2}S_j}$$

where S_j is the standard deviation of the j-th attribute in the original dataset.

3. Trading off information loss and disclosure risk

There is a broad choice of methods for continuous microdata protection (see Domingo-Ferrer and Torra (2001a) for an introduction). To increase the *embarras du choix*, most of such methods are parametric (e.g., in microaggregation, one parameter is the minimal number of records in a cluster), so the user must go through two choices rather than one: a primary choice to select a method and a secondary choice to select parameters for the method to be used.

The optimal method and parameterization will be the ones yielding an optimal tradeoff between information loss and disclosure risk. Thus, we need to be able to combine measures of information loss and measures of disclosure risk. Two approaches to do this are conceivable:

Explicit: A score (formula) is adopted which combines information loss and disclosure risk measures. This is the approach adopted in Domingo-Ferrer and Torra (2001b). Using a score permits to regard the selection of a masking method and its parameters as an optimization problem. This was exploited in Sebé et al. (2002): a masking method was applied to the original data file and then a post-masking optimization procedure was applied to decrease the score obtained.

Implicit: No specific score can do justice to all methods for all data uses and all disclosure scenarios. Thus, another possibility is for the data protector to separately compute several information loss and disclosure risk measures and choose the most appropriate method based on a combination of the most relevant measures for the specific data use/disclosure scenario. This implicit approach is adopted in Yancey et al. (2002), for example.

Whether explicit or implicit, a combination of information loss and disclosure risk measures is best performed if both types of measures can be bounded within the same range. Unfortunately, whilst disclosure risk measures are bounded, current information loss measures for continuous data are not:

- Disclosure risk measures can normally be regarded as probabilities or proportions bounded between 0 and 1, e.g. the probability that a certain respondent is re-identified or the proportion of correctly re-identified records after a record linkage attack;
- However, being mean square errors, mean absolute errors and mean variations, the information loss measures discussed in Section 2 are unbounded. Moreover, mean variations may become huge when measured on magnitudes close to 0.

In Trottini (2003), the above mismatch was detected and a solution consisting of enforcing upper bounds on information loss measures was proposed. In practice, the proposal in Trottini (2003) was to limit those measures in Table 1 based on the mean variation to a predefined maximum value. Such an *unnatural* truncation clearly damages the accuracy of

the resulting measures. We propose in the rest of this paper an approach which *naturally* yields bounded information loss measures for continuous attributes.

Remark 3.1. Note that, for categorical attributes, information loss measures are naturally bounded, because of the finite range of such attributes. Thus, the problem of unbounded information loss measures appears only for continuous attributes.

4. Probabilistic information loss measures

In what follows, we view the original dataset X as a population with n records and the protected dataset X' as a sample with n' records.

4.1. A generic measure

Given a population parameter θ on X, we can compute the corresponding sample statistic $\hat{\Theta}$ on X'. Let us assume that $\hat{\theta}$ is the value taken by $\hat{\Theta}$ for a specific sample. The more different is $\hat{\theta}$ from θ, the more information is lost when publishing the sample X' instead of the population X. We show next how to express that loss of information through probability.

If the sample size n' is large, the distribution of $\hat{\Theta}$ tends to normality with mean θ and variance $\mathrm{Var}(\hat{\Theta})$. According to Kendall et al. (1994), values of n' greater than 100 are often large enough for normality of all sample statistics to be acceptable. Fortunately, most protected datasets released in official statistics consist of $n' > 100$ records, so that assuming normality is safe. Thus, the standardized sample discrepancy

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\Theta})}}$$

can be assumed to follow a $N(0,1)$ distribution.

Therefore, a probabilistic information loss measure $pil(\theta)$ referred to parameter θ is the probability that the absolute value of the discrepancy Z is less than or equal to the actual discrepancy we have got in our specific sample X', that is

$$pil(\hat{\Theta}) = 2 \cdot P(0 \leq Z \leq \frac{|\hat{\theta} - \theta|}{\sqrt{Var(\hat{\Theta})}}) \tag{1}$$

4.2. Notation for population parameters and sample statistics

In this section, we identify and denote several population parameters θ and corresponding sample statistics $\hat{\Theta}$ which can be relevant to measure information loss. We start with population parameters (on X) and then continue with sample statistics (on X').

Let the r-th moment about zero of the j-th attribute of X be denoted by:

$$\mu_r^0(j) = \frac{\sum_{i=1}^{n} x_{ij}^r}{n}$$

The r-th central moment of the j-th attribute of X is expressed as:

$$\mu_r(j) = \frac{\sum_{i=1}^{n} \left(x_{ij} - \mu_1^0(j)\right)^r}{n}$$

The (r, s)-th central moment of the j-th and j'-th attributes of X can be computed as:

$$\mu_{rs}(j, j') = \frac{\sum_{i=1}^{n} \left(x_{ij} - \mu_1^0(j)\right)^r \left(x_{ij'} - \mu_1^0(j')\right)^s}{n}$$

If $r = s = 1$, we get the covariance $\mu_{11}(j, j')$ between attributes j and j'. In this way, the correlation coefficient can be expressed as

$$\rho(j, j') = \frac{\mu_{11}(j, j')}{(\mu_{02}(j, j')\mu_{20}(j, j'))^{1/2}}$$

We now turn to the moments on a protected dataset X' with n' records corresponding to the original dataset X. The moments on X' are regarded as statistics. We denote the r-th moment about zero of the j-th attribute of X' by:

$$m_r^0(j) = \frac{\sum_{i=1}^{n'} (x'_{ij})^r}{n'}$$

The r-th central moment of the j-th attribute of X' is expressed as:

$$m_r(j) = \frac{\sum_{i=1}^{n'} \left(x'_{ij} - m_1^0(j)\right)^r}{n'}$$

The (r, s)-th central moment of the j-th and j'-th attributes of X' can be computed as:

$$m_{rs}(j, j') = \frac{\sum_{i=1}^{n'} \left(x'_{ij} - m_1^0(j)\right)^r \left(x'_{ij'} - m_1^0(j')\right)^s}{n'}$$

The correlation coefficient between two attributes in X' can be expressed as

$$r(j, j') = \frac{m_{11}(j, j')}{(m_{02}(j, j')m_{20}(j, j'))^{1/2}}$$

4.3. Selected specific measures

Expression (1) can be used to derive an information loss measure for any particular statistic. For the sake of concreteness and comparability, we will consider here the same statistics as Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001), with the slight

adaptation that direct comparison of data is replaced with quantile comparison. Given two attributes j and j', this yields the following measures:

- $pil(m_1^0(j))$ for the mean;
- $pil(m_2(j))$ for the variance;
- $pil(m_{11}(j, j'))$ for the covariance;
- $pil(r(j, j'))$ for Pearson's correlation;
- $pil(Q_q(j))$ for quantiles.

Being probabilities, all the above measures are naturally bounded within the [0,1] interval. In order to get dataset-wide measures, we must average over the various attributes (as in Table 1). This yields:

$$PIL\left(m_1^0\right) = \frac{\sum_{j=1}^{p} pil\left(m_1^0(j)\right)}{p} \tag{2}$$

$$PIL(m_2) = \frac{\sum_{j=1}^{p} pil(m_2(j))}{p} \tag{3}$$

$$PIL(m_{11}) = \frac{\sum_{1\le j<j'\le p} pil(m_{11}(j, j'))}{p(p-1)/2} \tag{4}$$

$$PIL(r) = \frac{\sum_{1\le j<j'\le p} pil(r(j, j'))}{p(p-1)/2} \tag{5}$$

Remark 4.1. The normality assumption for the $r(j, j')$ statistic between two attributes j and j' only holds when the population correlation $\rho(j, j')$ is sufficiently centered within the interval $[-1,1]$. For values of $\rho(j,j')$ close to -1 or 1, computing $pil(r(j,j'))$ using Expression (1) with a standard normal Z yields an overpessimistic information loss measure ($Var(r(j, j'))$ is very small). Still, when one takes the average $PIL(r)$ over all pairs of attributes, the result is usually coherent with the average $PIL(m_{11})$ for covariances, as one would expect.

Remark 4.2. Using the fact that the correlation coefficient is bounded in $[-1, 1]$, one might think of using as a non-parametric (and non-probabilistic) alternative to $PIL(r)$ the following one:

$$\frac{\sum_{1\le j<j'\le p} |r(j, j') - \rho(j, j')|}{p(p-1)} \tag{6}$$

Expression (6) is half the mean absolute error given in Table 1, which is bounded between 0 and 1. However, being non-probabilistic, this measure takes values often incoherent with the average information loss for covariances $PIL(m_{11})$. For example, one can easily get an $PIL(m_{11})$ in [0.75, 1] and an "optimistic" Expression (6) in [0,0.25]. This lack of coherence is clearly undesirable for statistics as related as the covariance and Pearson's correlation.

Remark 4.3. The loss measure $PIL(Q_q)$ for quantiles bears some resemblance to the information loss measure in Agrawal and Aggarwal (2001) consisting of half the expected

value of the L_1-norm between the densities of the original and protected attribute: both measures are bounded in the [0, 1] interval, but they are not equivalent because the measure in Agrawal and Aggarwal (2001) is not a probability.

4.4. *Variances of the selected statistics*

In order to use Expression (1) to construct $pil(m_1^0(j))$, $pil(m_2(j))$, $pil(m_{11}(j, j'))$, $pil(r(j, j'))$ and $p(Q_q(j))$ we need the variance of each statistic or at least an approximation to it. This is a technical but unavoidable issue.

Since we take the original dataset X as the population and the masked dataset X' as the sample, our sampling method is the particular masking method used to obtain X' from X. Therefore, to be strict, the variance of each sample statistic depends on the masking method. However, deriving the expression of the statistic variance for each masking method whose information loss is to be measured is a cumbersome and hardly feasible task. Our primary goal is to obtain information loss measures which can be easily applied to any masking method. In that spirit, we suggest to sacrifice accuracy to applicability and compute variances as if the sampling method were simple random sampling. Of course, some masking methods may substantially differ from simple random sampling: for example, if masking consists of replacing original values by their overall mean, one has zero variance for the sample mean and so on. However, the fact that the new measures using those simplified variances are highly correlated with previous information loss measures in the literature (see Section 5) shows that the above is a reasonable approximation.

We will drop attribute indexes j, j' in the remainder of this section to improve readability. Following the Chapter 10 of Kendall et al. (1994), we have that, under simple random sampling, the variance of the sample mean is

$$\mathrm{Var}\left(m_1^0\right) = \frac{\mu_2}{n'}$$

The variance of the sample variance is

$$\mathrm{Var}(m_2) = \frac{\mu_4 - \mu_2^2}{n'}$$

The variance of the sample covariance is

$$\mathrm{Var}(m_{11}) = \frac{\mu_{22} - \mu_{11}^2}{n'}$$

The variance of the sample Pearson's correlation coefficient is

$$\mathrm{Var}(r) = \frac{\rho^2}{n'}\left\{\frac{\mu_{22}}{\mu_{11}^2} + \frac{1}{4}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}}\right) - \left(\frac{\mu_{31}}{\mu_{11}\mu_{20}} + \frac{\mu_{13}}{\mu_{11}\mu_{02}}\right)\right\}$$

Finally, if $q \in [0, 1]$, the variance of the sample q-quantile Q_q is

$$\mathrm{Var}(Q_q) = \frac{q(1-q)}{n' f_{Q_q}^2}$$

where f_{Q_q} is the value of the attribute's density function for the abscissa Q_q. If we take the dataset X as our population, it is unlikely that we know the analytical expression of the attribute density functions. A simple method to estimate f_{Q_q} is to approximate it by counting the proportion of records included in an interval around Q_q for the specific attribute being considered and then dividing by the interval width. It remains to decide what interval should be taken. A possible (and arbitrary) option is $(Q_q - \varepsilon, Q_q + \varepsilon)$, where ε is the range of the attribute divided by 1000.

Remark 4.4. Kernel methods (Rosenblatt, 1956; Parzen, 1962; Silverman, 1982) are an alternative for density estimation based on histogram smoothing. They may yield better density estimates than the simple approach sketched above, but they usually require more computation. See Härdle (1991) for a comprehensive discussion on kernel density estimation.

5. Empirical results

The "Census" dataset used in Domingo-Ferrer and Torra (2001b) was taken. This dataset was extracted from the U. S. Current Population Survey 1995 and consists of 1080 records with 13 continuous attributes.

For that dataset and for all 109 masking methods considered in Domingo-Ferrer and Torra (2001b), the behavior of the new measures defined here was compared to the behavior of measures IL_1 to IL_5 defined in Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001). Specifically:

- $PIL(Q)$ was constructed as the average of $PIL(Q_5)$ through $PIL(Q_{95})$ in 5% increments, so that it represents the average impact on quantiles from 5% to 95% for all attributes; $PIL(Q)$ was compared to IL_1 (mean variation of original and masked individual attribute values, that is, of $X - X'$ in Table 1);
- $PIL(m_1^0)$ was compared to IL_2 (mean variation of attribute means);
- $PIL(m_2)$ was compared to IL_3 (mean variation of attribute variances);
- $PIL(m_{11})$ was compared to IL_4 (mean variation of attribute covariances, that is, of $V - V'$ in Table 1);
- $PIL(r)$ was compared to IL_5 (mean absolute error of Pearson's correlations, that is, of $R - R'$ in Table 1);
- The average

$$PIL = 100 * (PIL(Q) + PIL(m_1^0) + PIL(m_2) + PIL(m_{11}) + PIL(r))/5$$

was compared to the IL reported in Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer and Torra (2001a)

$$IL = 100 * (IL_1 + IL_2 + IL_3 + IL_4 + IL_5)/5$$

Table 2. Pearson's and Spearman's correlations between probabilistic measures and measures in Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001).

Measure pair	Pearson's r	Spearman's r
$(PIL(Q),IL_1)$	0.693	0.902
$(PIL(m_1^0), IL_2)$	0.918	1.000
$(PIL(m_2),IL_3)$	0.531	0.977
$(PIL(m_{11}),IL_4)$	0.592	0.950
$(PIL(r),IL_5)$	0.781	0.995
(PIL, IL)	0.824	0.955

Comparison was performed by computing Pearson's and Spearman's (rank) correlations for each of the above six pairs of measures taken over the 109 masking methods. The results are given in Table 2. It can be seen that both types of measures are highly correlated, regardless of whether Pearson's or Spearman's coefficient is used: the pair of average measures (*PIL*, *IL*) has coefficients 0.824 and 0.955. The fact that Spearman's rank correlation is so high is very interesting, because it means that both types of measures rank the 109 masking methods in much the same way.

We next give the values for the new measures for some masking methods and parameterizations which had been found in Domingo-Ferrer and Torra (2001b) to yield a good score, i.e. a good tradeoff between information loss and disclosure risk. These were:

- *Rankswap7*. Rank swapping (Moore, 1996) with parameter $p = 7$. Rank swapping consists of: first, ranking the values of an attribute in ascending order; second, swapping each value of the attribute with another value randomly chosen so that the rank of the two swapped values does not differ by more than $p\%$ of the total number of records.
- *Mic3mul7*. Microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002) taking three attributes at a time and group size $k = 7$. Microaggregation consists of clustering records of at least k records; rather than publishing attribute values for each individual, the average of the attribute values over the group to which the individual belongs is published. Groups can be formed independently for each attribute, considering two attributes at a time, three attributes at a time, all attributes at once, etc.
- *Micmul3*. Microaggregation taking all attributes at once and group size $k = 3$.
- *Noise16*. Additive noise with parameter $p = 0.16$. Gaussian noise is added to the original data to get the masked data. If the standard deviation of the original attribute is s, noise is generated using a $N(0, ps)$ distribution.

Table 3 gives the probabilistic information loss measures corresponding to versions of the "Census" dataset masked using the above four masking methods. For microaggregated data, a simple rescaling transformation can bring the impact $PIL(m_2)$ on variances down to 0 without significant increase for the other measures. The same transformation brings $PIL(m_1^0)$ and $PIL(m_2)$ to 0 for noise-added data. The rescaling transformation for the j-th

Table 3. Probabilistic information loss measures on the "Census" dataset.

Method	$PIL(Q)$	$PIL(m_1^0)$	$PIL(m_2)$	$PIL(m_{11})$	$PIL(r)$
Rankswap7	0	0	0	0.4540	0.5933
Mic3mul7	0.3640	0	0.4403	0.3022	0.3309
Micmul3	0.4987	0	0.7004	0.1846	0.5341
Noise16	0.4524	0.1608	0.4447	0.1342	0.4929
Scal_Mic3mul7	0.3760	0	0	0.2138	0.3309
Scal_Micmul3	0.4979	0	0	0.3631	0.5341
Scal_Noise16	0.4467	0	0	0.3271	0.4929

masked attribute is

$$\left(\frac{(x_{ij} - m_1^0(j))\sqrt{\mu_2(j)}}{\sqrt{m_2(j)}}\right) + \mu_1^0(j)$$

In Table 3, *Scal_Mic3mul7*, *Scal_Micmul3* and *Scal_Noise16* are microaggregation and noise addition methods with rescaling. It can be seen that:

- *Scal_Mic3mul 7* is the method among those considered in Table 3 that performs best for the overall set of probabilistic measures considered;
- However, *Rankswap7* might be preferable if quantile preservation is critical;
- Additive noise considered here is uncorrelated, that is, it is applied independently to each attribute. Correlated noise addition, with a covariance matrix Σ equal to the covariance matrix of original data, would probably yield lower $PIL(m_{11})$ and $PIL(r)$. However, our objective is to present new information loss metrics rather than the best masking methods; thus, we have used uncorrelated noise for comparability with Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001) where that was the kind of noise considered.

6. Conclusions

Statistical Disclosure Control is about optimizing the tradeoff between disclosure risk and the information loss inflicted to data. We have shown a way to obtain probabilistic information loss measures for assessing the impact of masking methods on continuous microdata sets. Previous information loss measures were unbounded and compared only awkwardly with disclosure risk, which is bounded between 0 and 1. Being probabilistic, the measures presented in this paper are also bounded between 0 and 1, so they make it easier for data protectors to find an optimal balance between information loss and disclosure risk.

Remark 6.1. A web form and the C source code for computing the probabilistic information measures proposed in this paper on any pair of original and masked datasets are available at http://vneumann.etse.urv.es/SDC/measures.

Acknowledgments

Thanks go to Jordi Castellà for his help in preparing the web form `http://vneumann.etse.urv.es/SDC/measures`. Also, comments by William Winkler greatly helped improving the presentation of this paper. This work was partly funded by the Spanish Ministry of Science and Technology and the European FEDER Fund under project TIC2001-0633-C03-01 "STREAMOBILE" and also by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-01 "PROPRIETAS".

References

Agrawal, D. and Aggarwal, C.C. 2001. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara CA: ACM.

Dandekar, R., Domingo-Ferrer, J., and Sebé, F. 2002. Lhs-based hybrid microdata vs. rank swapping and microaggregation for numeric microdata protection. In Inference Control in Statistical Databases, J. Domingo-Ferrer (Ed.), volume 2316 of LNCS, Berlin, Heidelberg: Springer, pp. 153–162

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(1):189–201.

Domingo-Ferrer, J., Mateo-Sanz, J.M., and Torra, V. 2001. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In Pre-proceedings of ETK-NTTS'2001 vol. 2, Luxemburg: Eurostat, pp. 807–826

Domingo-Ferrer, J. and Torra, V. 2001a. Disclosure protection methods and information loss for microdata. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz (Eds.), North-Holland: Amsterdam, pp. 91–110, http://vneumann.etse.urv.es/publications/bcpi

Domingo-Ferrer, J. and Torra, V. 2001b. A quantitative comparison of disclosure control methods for microdata. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz (Eds.), North-Holland: Amsterdam, pp. 111–134, http://vneumann.etse.urv.es/publications/bcpi

Härdle, W. 1991. Smoothing Techniques with Implementation in S. New York: Springer-Verlag

Kendall, M.G., Stuart, A., J.K. Ord, S.F.A., and O'Hagan, A. 1994. Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory (6th Edition). London: Arnold

Moore, R. 1996. Controlled data swapping techniques for masking public use microdata sets. U.S. Bureau of the Census, Washington, DC (unpublished manuscript).

Parzen, E. 1962. On estimation of a probability density and mode. Annals of Mathematical Statistics, 35:1065–1076.

Rosenblatt, M. 1956. Remarks on some non-parametric estimates of a density function. Annals of Mathematical Statistics, 27:642–669.

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 Inference Control in Statistical Databases, J. Domingo-Ferrer (Ed.), volume 2316 of LNCS, Berlin, Heidelberg: Springer, pp. 163–171

Silverman, B.W. 1982. Kernel density estimation using the fast fourier transformation. Applied Statistics, 31:93–97.

Trottini, M. 2003. Decision models for data disclosure limitation. PhD thesis, Carnegie Mellon University. http://www.niss.org/dgii/TR/Thesis-Trottini-final.pdf

Winkler, W.E. 1999. Re-identification methods for evaluating the confidentiality of analytically valid microdata. In Statistical Data Protection, J. Domingo-Ferrer (Ed.), Luxemburg: Office for Official Publications of the European Communities. (Journal version in Research in Official Statistics, vol. 1, no. 2, pp. 50–69, 1998).

Yancey, W.E., Winkler, W.E., and Creecy, R.H. 2002. Disclosure risk assessment in perturbative microdata protection. In Inference Control in Statistical Databases, J. Domingo-Ferrer (Ed.), volume 2316 of LNCS, Berlin, Heidelberg: Springer, pp. 135–152