Chapter 5

Disclosure Control Methods and Information Loss for Microdata

Josep Domingo-Ferrer*
Universitat Rovira i Virgili

Vicenç Torra
*Institut d'Investigació en Intelligència Artificial-CSIC**[spell out CSIC?]***

1. Introduction

Statistical disclosure control (SDC) seeks to modify statistical data so that they can be published without giving away confidential information that can be linked to specific respondents. The challenge for SDC is to achieve this modification with minimum loss of the detail and accuracy sought by database users. SDC methods for microdata are usually known as *masking methods,* of which there is a wide range. From the point of view of their operational principles, current masking methods fall into the following two categories (Willenborg and De Waal 2001):

- *Perturbative.* The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset.
- *Nonperturbative.* Nonperturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail on the original dataset. Global recoding, local suppression, and sampling are examples of nonperturbative masking.

* Some of the work reported in this chapter was funded in part by the U.S. Bureau of the Census under Contracts No. OBLIG-2000-29158-0-0 and OBLIG-2000-29144-0-0. Thanks go to Laura Zayatz for providing information on rank swapping. We would also like to thank Josep M. Mateo-Sanz and Francesc Sebé for their help in working out some of the examples in this chapter. The comments of the editors and several reviewers are gratefully acknowledged as well.

From the point of view of the data on which they are used, a different two-part classification applies:

- *Continuous.* A variable is considered continuous if it is numerical and arithmetic operations can be performed with it. Examples are income and age. Note that a numerical variable does not necessarily have an infinite range, as in the case of age (alas!).
- *Categorical.* A variable is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Examples are day of the week and eye color.

Structure of This Chapter

Section 2 introduces the notation used throughout the chapter. Section 3 describes masking methods that have been proposed for microdata protection. Section 4 proposes several methods to measure the information loss attributable to masking in both continuous and categorical cases. Section 5 concludes the chapter in a way that leads into Chapter 6, which compares SDC alternatives in terms of the information loss/confidentiality tradeoff.

2. Notation and Variable Types

We assume that the information in a microdata file can be represented as a two-dimensional table where one dimension is the set of records (*i.e.,* elements, individuals, persons) and the other is the set of variables.

The microdata file contains a value for each record-variable pair, so that it can be modeled as a function

$$\boldsymbol{V}:\boldsymbol{O} \rightarrow D(V_1) \times D(V_2) \times \ldots \times D(V_m)$$

where $\boldsymbol{O}$ denotes the set of records, $V_1, V_2, \ldots, V_m$ denote the variables, and $D(V_i)$ refers to the range of variable V_i.

Without loss of generality, the m-dimensional function V can be assumed to be of the form:

$$\boldsymbol{V}(O) = (V_1(O), V_2(O), \ldots, V_m(O))$$

where $V_i(\bullet):\boldsymbol{O} \rightarrow D(V_i)$ is a one-dimensional function assigning a value for variable V_i to a given record.

3. Masking Methods for Microdata Protection

Using the notation of Section 2, we can state the purpose of SDC more formally by saying that its goal is to supply the user with a masked microdata file V' similar to the original V (*i.e.,* with low information loss) in such a way that

1) Disclosure risk (*i.e.,* risk of identification of an individual) is low.
2) User analyses (regressions, means, *etc.*) on V' and on V yield the same or at least similar results.

This section describes masking methods that can be used to produce V' from V. Subsections discuss perturbative methods and nonperturbative methods. Complementary reviews of the literature on masking methods can be found in Adam and Wortmann (1989) and Willenborg and De Waal (2001).

Perturbative Methods

Perturbative methods allow for release of the entire microdata set, although perturbed values rather than exact values are given. Not all perturbative methods are designed for continuous data, a distinction that is addressed further below for each method.

Most perturbative methods reviewed below (including additive noise, data swapping, microaggregation, and post-randomization) are special cases of matrix masking. If the original microdata set is V, then the masked microdata set V' is computed as

$$V' = AVB + C$$

where A is a record-transforming mask, B is a variable-transforming mask, and C is a displacing mask (noise) (Duncan and Pearson 1991).

Table 1 lists the perturbative methods described below. For each method, the table indicates whether it is suitable for continuous and/or categorical data.

Table 1. Perturbative Methods Versus Data Types

Method	Continuous data	Categorical data
Additive noise	X	
Data distort. by probability distribution	X	X
Microaggregation	X	
Resampling	X	
Lossy compression	X	
Multiple imputation	X	
Camouflage	X	
PRAM		X
Rank swapping	X	X
Rounding	X	

Additive Noise. Additive noise (Kim 1986; Little 1993; Sullivan and Fuller 1989, 1990) consists of adding random noise with the same correlation structure as the original unmasked data. It is currently the only method that can preserve correlations.

Let $v_{ij} = V_i(o_j)$ be the unmasked value of variable V_i for individual o_j. Let $e_{ij} = E_i(o_j)$ be the noise added to v_{ij}, and let $v'_{ij} = v_{ij} + e_{ij}$. Further, let $V = \{V_{ij}\}$ be the matrix having v_{ij} as elements, and similarly $E = \{e_{ij}\}$ and $V' = \{v'_{ij}\}$. It is assumed that the expected value of the noise is $E(E) = 0$ and its variance is $Var(E) = cVar(V)$ for some constant c. The variance of the masked data is $Var(V') = (1 + c)Var(V)$. The variance of unmasked variables can be recovered as $Var(V')/(1 + c)$.

White (*i.e.,* Gaussian) noise is most frequently used, even though it may be subject to the *bias problem* (Matloff 1986): if V_i is a continuous positive variable with a strictly decreasing density function (*e.g.,* the exponential density) to which a perturbation that is symmetrical around 0 has been added (*e.g.,* Gaussian noise), Matloff shows that

$$E(V_i | V'_i = w) < w$$

where V'_i is the perturbed version of V_i.

Thus, the constant c is the only parameter that can be tuned. It alters (increases or decreases) the random noise being inoculated.

The nature of additive noise makes it unsuitable for categorical data. But it is well suited for continuous data, for the following reasons:

- It makes no assumptions about the range of possible values for V_i (which may be infinite).
- The noise being added is typically continuous and with mean zero, which suits continuous original data well.
- No exact matching is possible with external files. Depending on the amount of noise added, approximate (interval) matching might be possible.

Data Distortion by Probability Distribution. Data distortion by probability distribution (see Liew *et al.* 1985 on probability distortion) is a method suitable for both categorical and continuous variables. Three steps are needed to compute the distorted version of a confidential original dataset:

1) Identification of the underlying density function of each of the confidential variables in the dataset and estimation of the parameters associated with the density function.

2) Generation of a distorted series for each confidential variable from the estimated density function.

3) Mapping and replacement of the distorted series in place of the confidential series.

In the identification and estimation stage, the original series of the confidential variable (*e.g.,* salary) is screened to determine which of a set of predetermined den-

sity functions fits the data best. Goodness of fit can be tested with the Kolmogorov-Smirnov test. An example set of predetermined density functions could include Poisson, exponential, normal, gamma, Weibull, log-normal, uniform, triangular, chi-square. If several density functions are acceptable at a given significance level, selecting the one yielding the smallest value for the Kolmogorov-Smirnov statistics is recommended. If no density in the predetermined set fits the data, the frequency imposed distortion method can be used. With the latter method, the original series is divided into several intervals (somewhere between 8 and 20). The frequencies within the interval are counted for the original series and become a guideline to generate the distorted series. By using a uniform random number generating subroutine, a distorted series is generated until its frequencies become the same as the frequencies of the original series. If the frequencies in some intervals overflow, they are simply discarded.

Once the best-fit density function has been selected, the generation stage feeds the estimated distribution parameters to a random value-generating routine to produce the distorted series.

The final stage, mapping and replacement, is needed only if the distorted variables are to be used jointly with other nondistorted variables. Mapping consists of ranking the distorted series and the original series in the same order and replacing each element of the original series with the corresponding distorted element.

It must be stressed here that the approach described in Liew *et al.* (1985) is for one variable at a time. One could imagine a generalization of the method using multivariate density functions. However, this is not a trivial undertaking. It requires multivariate ranking/mapping and can lead to very poor fits.

The example below shows how distribution fitting can be used to construct a masked dataset from an original dataset.

Example 1

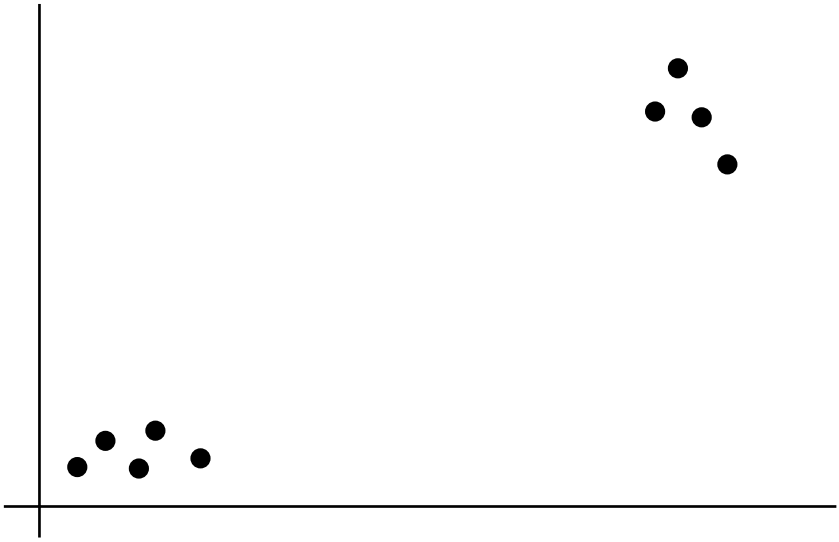
A distribution fitting software (Crystal Ball 2001) was used on the original (ranked) dataset 186, 693, 830, 1177, 1219, 1428, 1902, 1903, 2496, 3406. Continuous distributions tried were normal, triangular, exponential, log-normal, Weibull, uniform, beta, gamma, logistic, Pareto, and extreme value; discrete distributions tried were binomial, Poisson, geometric, and hypergeometric. The software allowed for three fitting criteria to be used: Kolmogorov-Smirnov, χ^2, and Anderson-Darling. According to the first criterion, the best fit occurred for the extreme value distribution with modal and scale parameters 1105.78 and 732.43, respectively; the Kolmogorov statistic for this fit was 0.1138. Using the fitted distribution, the following masked dataset was generated and used to replace the original one: 425.60, 660.97, 843.43, 855.76, 880.68, 895.73, 1086.25, 1102.57, 1485.37, 2035.34. To assess the disclosure risk associated with this masking, reidentification experiments such as those described in Chapter 6 can be conducted.

[authors put the symbol ¨ at the end of every Example paragraph and it is now deleted because it doesn't match any other formats. If that changes the meaning or the intent, please reinsert. Also please coordinate the appearance of this set of numbered examples with the set in Chapter 9.]

Microaggregation. Records are clustered into small aggregates or groups of size at least k. Rather than publishing an original variable V_i for a given record, the average of the values of V_i over the group to which the record belongs is published. The rationale behind microaggregation is that confidentiality rules permit publication of microdata sets if the records correspond to groups of *k* or more individuals, where no individual dominates (*i. e.,* contributes too much to) the group and *k* is a threshold value. To minimize information loss, groups should be as homogeneous as possible.

Classical microaggregation (Defays and Nanopoulos 1993) required that all groups except perhaps one be of size k; allowing groups to be of size $\geq k$ depending on the structure of data can be termed *data-oriented microaggregation* (Domingo-Ferrer and Mateo-Sanz 2002; Mateo-Sanz and Domingo-Ferrer 1999). Figure 1 illustrates the advantages of variable-sized groups. If classical fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. If, in contrast, variable-sized groups are allowed, then the five pieces of data on the left can be kept in a single group and the four on the right in another group; such a variable-size grouping yields more homogeneous groups, which implies lower information loss.

Figure 1. Variable-Sized Groups Versus Fixed-Sized Groups



Exactly solving the microaggregation problem–-that is, finding a grouping where groups have maximal homogeneity and size at least k—was recently shown to be NP-hard[1] (Oganian and Domingo-Ferrer 2001). Methods in the literature are heuristic and can be univariate or multivariate:

- Univariate methods deal with multivariate datasets by microaggregating one variable at a time—that is, variables are sequentially and independently microaggregated. This approach is known as individual ranking or blurring (Defays and Nanopoulos 1993) and, while it causes low information loss, it can lead to rather high disclosure risks (see Chapter 6).

- Multivariate methods either rank multivariate data by projecting them onto a single axis—for example, using the first principal component[2] or the sum of z-scores[3] (Defays and Nanopoulos 1993)—or dealing directly with unprojected data (Domingo-Ferrer and Mateo-Sanz 2002; Mateo-Sanz and Domingo-Ferrer 1999). When working with unprojected data, one can microaggregate all variables of the dataset at a time, or independently microaggregate groups of two variables at a time, three variables at a time, and so on.

Resampling. Originally proposed for protecting tabular data (Domingo-Ferrer and Mateo-Sanz 1999; Heer 1993), resampling can also be used for microdata. Let V be an original variable in a dataset with n records. Take with replacement t independent samples $X_1, \ldots, X_t$ of size n of the values of V. Independently rank each sample (using the same ranking criterion for all samples). Finally, for $j = 1$ to n, compute the j-th value v'_j of the masked variable V' as the average of the j-th ranked values in $X_1, \ldots, X_t$.

Lossy Compression. This method is new and is recommended for continuous data. The idea is to regard a numerical microdata file as an image (with records being rows, variables being columns, and values being pixels[4]). Lossy compression—for example, JPEG (Joint Photographic Experts Group 2001) is then used on the image, and the compressed image is interpreted as a masked microdata file. Depending on the lossy compression algorithm used, appropriate mappings between

---

1 A problem is NP-hard if it cannot be solved in time polynomial in the input size unless P = NP, where P is the class of problems solvable in polynomial time and NP is the class of problems for which correctness of a solution can be verified in polynomial time. It is conjectured that P is strictly included in NP (Garey and Johnson 1979).

2 Principal component analysis aims to transform the observed variables into a new set of variables which are uncorrelated and arranged in decreasing order of importance. The principal aim is to reduce the dimensionality of the dataset to make it easier to understand. In particular, the first principal component can be interpreted as a one-dimensional 'summary' of the dataset.

3 The sum of z-scores is an alternative way to obtain a one-dimensional 'summary' of a dataset. Each variable is standardized, and for each record the standardized values of all variables are added up.

4 Pixel stands for 'picture element' and corresponds to a dot in a digital image. For a black-and-white image, the value of a pixel is a grayscale level. If b bits are used to encode a pixel, then 2^b grayscale levels can taken by the pixel. Similarly, for a color image, if b bits are used, 2^b combinations of the basic colors (red, green, and blue) are possible for the pixel.

variable ranges and color scales will be needed. The example below illustrates the use of lossy compression to obtain a masked dataset from an original dataset.

Example 2

The upper part of Table 2 shows an original dataset with eight variables and eight records; the lower part of the table shows the disclosure protected version of the dataset. The algorithm used is as follows. First, original values have been scaled to integers in the interval *[0,255]*, that is, pixel grayscale values; this gives a grayscale image. Second, JPEG compression with 80 percent quality has been used on the image. Third, the compressed image has been unscaled (using the inverse of the scaling transformation of the first step) to get the masked dataset. The disclosure risk associated to this masking could be estimated in an empirical way, as suggested for the masking of Example 1.

Table 2. Example of SDC Through Lossy Compression

Top, original file; down, protected file with JPEG 80%

4173	4621	4527	1428	27	27	3480	4550
2639	6045	4208	1902	1008	808	3136	4100
3315	4765	5645	1903	485	485	4284	5600
1619	3932	2380	1177	700	700	1750	2288
4604	4349	2151	1219	751	1	1606	2100
3433	2463	3217	830	167	50	2448	3200
824	372	8730	186	1030	22	589	7700
4145	629	2500	693	1	1	1912	2500
4100	4710	4498	1451	1	10	3718	4296
2751	5889	4369	1862	1021	808	3052	4098
3314	4465	5763	1903	448	507	4168	5657
1683	4154	2151	1243	654	706	1791	2385
4604	4154	2305	1222	779	1	1559	2100
3536	2374	3363	812	93	124	2443	3263
824	460	8730	199	1030	1	589	7590
4174	572	2538	670	1	23	1965	2605

Multiple Imputation. This method (Rubin 1993) relies on releasing simulated continuous microdata created by multiple imputation techniques based on the original microdata. A way to perform multiple imputation is on a variable-by-variable basis, using a randomized regression (with normal errors) to impute missing values of each continuous variable (Kennickell 1998).

Camouflage. Vector camouflage (Gopal *et al.* 1998) is a method for giving unlimited, correct numerical responses to *ad hoc* queries to a database while not compromising confidential numerical data. No probabilistic assumptions are made, and optimization techniques are used to camouflage the sensitive record (exact answer) in an infinite set of records, thus providing an interval answer. The information loss is the transformation of a point answer into an interval answer. Because of its nature, this method is suitable only for continuous data.

PRAM. The Post-Randomization Method (PRAM) (Gouweleeuw *et al.* 1997) is a probabilistic, perturbative method for disclosure protection of categorical variables in microdata files. In the masked file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix. The Markov approach makes PRAM very general, because it encompasses noise addition, data suppression, and data recoding.

PRAM information loss and disclosure risk depend largely on the choice of the Markov matrix (De Wolf *et al.* 1999) and are still open research topics.

The PRAM matrix contains a row for each possible value of each variable to be protected. This rules out using the method for continuous variables, unless these are converted into discrete form**[is this what is meant by "are previously discretized"?]** (in the same way discussed below in the case of global recoding).

Rank Swapping. Although originally described only for ordinal variables (Moore 1996), this method can be used for any numerical variable. First, values of variable V_i are ranked in ascending order. Then, each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (*e.g.*, the rank of two swapped values cannot differ by more than p percent of the total number of records). The use of rank swapping is illustrated below.

Example 3

In Table 3, we can see an original microdata file on the left and its rank swapped version on the right. There are four variables and ten records in the original file; the second variable is alphanumeric, and the standard alphabetic order has been used to rank it. A value of $p = 15$ percent has been used for all variables. The rank swapped data file has been sorted by its first variable.

Table 3. Example of Rank Swapping
Left four columns, original file; right four columns, rank swapped file

1	K	3.7	4.4	1	H	3.0	4.8
2	L	3.8	3.4	2	L	4.5	3.2
3	N	3.0	4.8	3	M	3.7	4.4
4	M	4.5	5.0	4	N	5.0	6.0
5	L	5.0	6.0	5	L	4.5	5.0
6	H	6.0	7.5	6	F	6.7	9.5
7	H	4.5	10.0	7	K	3.8	11.0
8	F	6.7	11.0	8	H	6.0	10.0
9	D	8.0	9.5	9	C	10.0	7.5
10	C	10.0	3.2	10	D	8.0	3.4

Rounding. Rounding methods replace original values of variables with rounded values. For a given variable V_i, rounded values are chosen among a set of rounding points defining a *rounding set*. In a multivariate original dataset, rounding is usually performed one variable at a time (*univariate* rounding); however, multivariate rounding is also possible (Willenborg and De Waal 2001). The operating principle of rounding makes it suitable only for continuous data.

Example 4

Assume a continuous variable *V*. Then we have to determine a set of rounding points $\{p_1, \ldots, p_r\}$. One possibility is to take rounding points as multiples of a base value b, that is, $p_i = b \cdot i$ for $i = 1, \ldots, r$. The set of attraction for each rounding point p_i is defined as the interval $[p_i - b/2, p_i + b/2]$, $(i = 2, \ldots, r-1)$; for p_1 and p_r, respectively, the sets of attraction are $[0, p_1 + b/2]$ and $[p_r - b/2, V_{max}]$, where V_{max} is the largest possible value for variable V. Now an original value v of V is replaced with the rounding point corresponding to the set of attraction where v lies.

Nonperturbative Methods

Nonperturbative methods do not rely on distortion of the original data but on partial suppressions or reductions of detail. Some of the methods are usable on both categorical and continuous data, but others are not suitable for continuous data. Table 4 lists the nonperturbative methods described below. For each method, the table indicates whether it is suitable for continuous and/or categorical data.

Table 4. Nonperturbative Methods Versus Data Types

Method	Continuous data	Categorical data
Sampling		X
Global recoding	X	X
Top and bottom coding	X	X
Local suppression		X

Sampling. Instead of publishing the original microdata file

$$\boldsymbol{V}:\boldsymbol{O} \rightarrow D(V_1) \times D(V_2) \times \ldots \times D(V_m)$$

what is published is

$$\boldsymbol{V}':\boldsymbol{S} \rightarrow D(V_1) \times D(V_2) \times \ldots \times D(V_m)$$

where $\boldsymbol{S} \subset \boldsymbol{O}$ is a sample of the original set of records and $\boldsymbol{V}'$ stands for the original function $\boldsymbol{V}$ restricted to $\boldsymbol{S}$.

Sampling is suitable for categorical microdata, but its adequacy for continuous microdata is less clear in a general disclosure scenario. The reason is that the method leaves a continuous variable $V_i(\bullet)$ unperturbed for all records in S. Thus, if variable V_i is present in an external administrative public file, unique matches with V' are very likely, because for a continuous variable (even one truncated due to digital representation) it is unlikely that V_i takes the same value for two different records (*i.e.,* $V_i(o_1) = V_i(o_2)$ if $(o_i \neq o_2)$.

If, for a continuous identifying variable, the score of a respondent is only approximately known by an attacker (as assumed in Willenborg and De Waal 1996), it might still make sense to use sampling to protect that variable. However, assumptions about attacker resources are perilous and may prove too optimistic if good quality external administrative files are at hand. For the purpose of illustration, Example 5 gives the technical specifications of a real-world application of sampling.

Example 5

In 1995, Statistics Catalonia released a sample from the 1991 population census of Catalonia (IDESCAT-Statistics Catalonia 1995). The information released corresponded to 36 categorical variables (including the recoded versions of initially continuous variables); some of the variables were related to the individual person and some to the household. The technical specifications of the sample were as follows:

- *Sampling algorithm:* Simple random sampling.
- *Sampling unit:* Individuals in the population whose residence was in Catalonia as of March 1, 1991.
- *Population size:* 6,059,494 inhabitants.
- *Sample size:* 245,944 individual records.
- *Sampling factor:* 0.0406.

With the above sampling fraction, the maximum absolute error for estimating a maximum-variance proportion is 0.2 percent.

Global Recoding. For a categorical variable V_i, several categories are combined to form new (less specific) categories, thus resulting in a new V'_i with $|D(V'_i)| < |D(V_i)|$ where $|\bullet|$ is the cardinality operator. For a continuous variable V_i, global recoding means replacing V_i by another variable V'_i which is a discretized version of V_i. In other words, a potentially infinite range $D(V_i)$ is mapped onto a finite range $D(V'_i)$. This is a technique used in the μ-Argus SDC package (Hundepool *et al.* 1998).

This technique is more appropriate for categorical microdata, where it helps disguise records with strange combinations of categorical variables. Global recoding is used heavily by statistical offices.

Example 6

If there is a record with 'Marital status = Widow/er' and 'Age = 17', global recoding could be applied to 'Marital status' to create a broader category 'Widow/er or divorced', so that the probability of the above record being unique would diminish. Global recoding can also be used on a continuous variable, but the inherent discretization very often leads to an unaffordable loss of information. Also, arithmetical operations that were straightforward on the original V_i are no longer easy or intuitive on the discretized V'_i.

Top- and Bottom-Coding. Top- and bottom-coding is a special case of global recoding that can be used on variables that can be ranked, that is, continuous or categorical ordinal. The idea is that top values (those above a certain threshold) are lumped together to form a new category. The same is done for bottom values (those below a certain threshold) (see Hundepool *et al.* 1999). **[Hundepool et al. in References is 1998; might this refer to Hundepool and Willenborg 1999?]**

Local Suppression. Certain values of individual variables are suppressed with the aim of increasing the set of records agreeing on a combination of values. Ways to combine local suppression and global recoding are discussed in De Waal and Willenborg (1995) and implemented in the m-Argus SDC package (Hundepool et al. 1998).

If a continuous variable V_i is part of a set of variables, then each combination of values is probably unique. Because it does not make sense to systematically suppress all the values of V_i, we conclude that local suppression is more oriented to categorical variables.

4. Information Loss Measures

Strictly speaking, information loss depends on the data uses to be supported by the masked (*i.e.,* SDC-protected) data. However, potential data uses are very diverse and it may be hard even to identify them all at the moment of data release by a statistical office. It is thus desirable for the data protector to be able to measure information loss in a generic way that reflects how much harm is being inflicted to the data by a given masking method; the amount of information loss measured in this generic way should roughly correspond to the amount of information loss for a reasonable range of data uses. The approach described here to derive generic information loss measures is based on assessing how different the masked dataset is from the original dataset. We will say there is little information loss if the analytic structure of the masked dataset is very similar to the structure of the original dataset. In fact, the motivation for preserving the structure of the dataset is to ensure that the masked dataset will be analytically valid and interesting. Winkler (1998) has determined that

- A microdata set is *analytically valid* if the following are approximately preserved (some conditions apply only to continuous variables):
  1) Means and covariances on a small set of subdomains.
  2) Marginal values for a few tabulations of the data.
  3) At least one distributional characteristic.
- A microdata file is *analytically interesting* if six variables 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 it may be, the above definition of analytical validity does shed some light on what preserving the dataset structure means. We can actually try several complementary ways to assess the preservation of the original dataset's structure:

- Compare the data in the original and the masked datasets. The more similar the SDC method to the identity function, the less the impact (but the higher the disclosure risk!).
- Compare some statistics computed on the original and the masked datasets. Little information loss should translate to little differences between the statistics.
- Analyze the behavior of the particular SDC method used to measure its impact on the structure of the original dataset.

Information Loss Measures for Continuous Data

Assume a microdata set with n individuals (records) $I_1, I_2, \ldots, I_n$ and p continuous variables $Z_1, Z_2, \ldots, Z_p$. Let X be the matrix representing the original microdata set (rows are records and columns are variables). Let X' be the matrix representing the

masked microdata set. The following tools are useful to characterize the information contained in the dataset:

- Covariance matrices V (on X) and V' (on X').
- Correlation matrices R and R'.
- Correlation matrices RF and RF' between the p variables and the p factors $PC_1, \ldots, PC_p$ obtained through principal components analysis.
- Commonality between each of the p variables and the first principal component PC_1 (or other PC_i's). Commonality is the percentage of each variable 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 variable 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 that completely reflects those structural differences. Therefore, we propose 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 masked dataset. In particular, discrepancy between correlations is related to the information loss for data uses such as regressions and cross tabulations.

Matrix discrepancy can be measured in at least three ways:

- *Mean square error:* Sum of squared componentwise differences between pairs of matrices, divided by the number of cells in either matrix.
- *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 percentage variation of components in the matrix computed on masked 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 variables.

Table 5 summarizes the measures proposed. In this table, p is the number of variables and n the number of records. 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 sense to compute them on the averages of variables rather than on all data (call this variant $\overline{X} - \overline{X}'$). It would also be sensible to use $V - V'$ measures to compare only the variances of the variables, that is, to compare the diagonals of the covariance matrices rather than the whole matrices (call this variant $S - S'$).

Table 5. Information Loss Measures for Continuous Microdata

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

Information Loss Measures for Categorical Data

Straightforward computation of measures in Table 5 on categorical data is not possible. The following alternatives have been considered in the literature:

- Direct comparison of categorical values.
- Comparison of contingency tables.
- Entropy-based measures.

Direct Comparison of Categorical Values. Comparison of matrices X and X' for categorical data requires the definition of a distance for categorical variables. Definitions consider only the distances between pairs of categories that can appear

when comparing a record and its masked version (possible pairs depend on the particular SDC method being used).

For a nominal variable V (a categorical variable that takes values over an unordered set), the only permitted operation is comparison for equality. This leads to the following distance definition:

$$d_v(c, c') = \begin{cases} 0 \ (\textit{if } c = c') \\ 1 \ (\textit{if } c \neq c') \end{cases}$$

where c is a category in the original dataset and c' is the category corresponding to c in the masked dataset. Correspondence between pairs of categories is determined by the masking method being used.

For an ordinal variable V (a categorical variable that takes values over a totally ordered set), let $\leq_V$ be the total order operator over the range $D(V)$ of variable V. Define the distance between categories c and c' as the number of categories between the minimum and the maximum of c and c' divided by the cardinality of the range:

$$d_v(c, c') = \frac{|c'' : (c, c') \leq c'' \leq \max(c, c')|}{|D(V)|}$$

Comparison of Contingency Tables. An alternative to directly comparing the values of categorical variables is to compare their contingency tables. Given two datasets F and G (the original and the masked set, respectively) and their corresponding t-dimensional contingency tables for $t \leq K$, we can define a contingency table-based information loss measure *CTBIL* for a subset W of variables as follows:

$$CTBIL(F, G; W, K) = \sum_{\substack{\{V_{j1} \cdots V_{jt}\} \subseteq W \\ |\{V_{j1} \cdots V_{jt}\}| \leq K}} \sum_{i_1 \ldots i_t} \left| x^F_{i_1 \ldots i_t} - x^G_{i_1 \ldots i_t} \right|$$

where $x^{file}_{subscripts}$ is the entry of the contingency table of *file* at position given by *subscripts*.

Because the number of contingency tables to be considered depends on the number of variables $|W|$, the number of categories for each variable, and the dimension K, a normalized version of expression (1) may be desirable. This can be obtained by dividing expression (1) by the total number of cells in all considered tables.

Distance between contingency tables generalizes some of the information loss measures used in the literature. For example, μ-Argus (Hundepool *et al.* 1998) measures information loss for local suppression by counting the number of suppressions. The distance between two contingency tables of dimension one returns twice the number of suppressions. This is because, when category A is suppressed for one record, two entries of the contingency table are changed: The count of

records with category *A* decreases and the count of records with the 'missing' category increases.

Entropy-Based Measures. In De Waal and Willenborg (1999) and Kooiman *et al.* (1998), the use of Shannon's entropy to measure information loss is discussed for local suppression, global recoding, and PRAM. Entropy is an information-theoretic measure, but it can be used in SDC if the masking process is modeled as the noise that would be added to the original dataset in the event of its being transmitted over a noisy channel.

As noted earlier, PRAM is a method that generalizes noise addition, suppression, and recoding methods. Therefore, our description of the use of entropy will be limited to PRAM.

Let V be a variable in the original dataset and V' be the corresponding variable in the PRAM-masked dataset. Let $P_{V,V'} = \{p(V' = j|V = i)\}$ be the PRAM Markov matrix. Then the conditional uncertainty of V given that $V' = j$ is:

$$H(V|V' = j) = -\sum_{i=1}^{n} p(V = i|V' = j)\log p(V = i|V' = j)$$

The probabilities in expression (2) can be derived from $P_{V,V'}$ using Bayes's formula. Finally, the entropy-based information loss measure (*EBIL*) is obtained by accumulating expression (2) for all individuals r in the masked dataset G

$$EBIL(P_{V,V'},\mathrm{G}) = \Sigma_{r \in G}\ H(V|V'=j_r)$$

where j_r is the value taken by V' in record r.

An Alternative Information Loss Measure. From our point of view, information loss when measured using the formalism reviewed in the previous subsection presents a drawback: The measure is a function of the masked dataset G but does not depend on the original dataset F. We begin with an example to illustrate this point; we then present an alternative approach that can also be applied to any PRAM-like SDC method.

Example 7

Assume that, in a household survey file, variable V contains the town where the household is located. Now consider that V is masked into a new variable V' where the town has been replaced by the state. Locations like 'New York City' and 'Albany' will be recoded into 'NY'. Living in Albany is more specific and identifying (in the sense of being less anonymous) than living in New York City. The information loss measure should somehow reflect that there is more information loss when a household in 'Albany' becomes a household in 'New York State' than when a household in 'New York City' becomes a household in 'New York State'.

Note that in Example 7

$$P(V = \text{'Albany'} \mid V' = NY) < P(V = \text{'New York City'} \mid V' = NY)$$

According to the U.S. Census Bureau's *American FactFinder* (U.S. Census Bureau 2001), the population of New York State in 2000 was 17,990,455, the population of New York City was 7,322,564, and the population of Albany was 101,082. Thus, the above probabilities are $P(V = \text{'Albany'} \mid V' = NY) = 101{,}082/17{,}990{,}455 = 0.05$ and $P(V = \text{'New York City'} \mid V' = NY) = 7{,}322{,}564/17{,}990{,}455 = 0.407$.

More generally, the smaller the conditional probability $P(V = i \mid V' = j)$, the larger the information loss. Based on this, we can define the information loss for a variable V as a function of three elements: the conditional probability, the original category i, and the masked category j. If we use minus the logarithm of $P(V = i \mid V' = j)$, the resulting information loss measure satisfies the monotonicity requirement of increasing as the conditional probability decreases. Thus the per-record information loss when $V = i$ is masked as $V' = j$ can be defined as

$$PRIL(P_{V, V'}, i, j) = -\log P(V' = j \mid V = i)$$

Note that it does not make sense to compute *PRIL* for categories i, j such that $P(V = i \mid V' = j) = 0$, because category i will never be masked as j. So *PRIL* is well defined. The information loss for the entire datasets F,G is

$$IL(P_{V, V'}, F, G) = \Sigma_{r \in G} PRIL(P_{V, V'}, i_r, j_r)$$

where i_r is the value taken by V in record r of F and j_r is the value taken by V' in record r of G.

5. Conclusions

The literature on statistical disclosure control for microdata is becoming increasingly vast. This chapter has presented an overview of current proposals as well as a set of measures to assess the extent to which a method damages the informational content of the data being protected. However, the reader should not forget that there is a trade-off between information loss and disclosure risk. In Chapter 6 we compare SDC methods in terms of this inevitable trade-off.

References [need first names for all authors]

Adam, N.R., and J.C. Wortmann (1989) 'Security-Control Methods for Statistical Databases: A Comparative Study', *ACM Computing Surveys,* 21(4), pp.515-56.

Crystal Ball (2001), http://www.cbpro.com/.

Defays, D., and P. Nanopoulos (1993) 'Panels of Enterprises and Confidentiality: The Small Aggregates Method', in *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada, pp.195-204.

De Waal, A.G., and L.C.R.J. Willenborg (1995) 'Global Recodings and Local Suppressions in Microdata Sets', in *Proceedings of Statistics Canada Symposium 95*, Ottawa: Statistics Canada, pp.121-32.

——— (1999) 'Information Loss Through Global Recoding and Local Suppression', *Netherlands Official Statistics* (special issue on SDC), 14, pp.17-20.

De Wolf, P.-P., J.M. Gouweleeuw, P. Kooiman, and L.C.R.J. Willenborg (1999) 'Reflections on PRAM', in *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, pp.337-49.

Domingo-Ferrer, J., and J.M. Mateo-Sanz (1999) 'On Resampling for Statistical Confidentiality in Contingency Tables', *Computers & Mathematics with Applications*, 38, pp.13-32.

——— **[is this indeed the same two authors as the previous?]**(2002) 'Practical Data-Oriented Microaggregation for Statistical Disclosure Control, *IEEE Transactions on Knowledge and Data Engineering* (forthcoming March 2002).

Duncan, G.T., and R.W. Pearson (1991) 'Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future, *Statistical Science*, 6, pp.219-39.

Garey, M.R., and D.S. Johnson (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, New York: Freeman.

Gopal, R., P. Goes, and R. Garfinkel (1998) 'Confidentiality via Camouflage: The CVC Approach to Database Query', in J. Domingo-Ferrer (ed) *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, pp.19-28.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg, and P.-P. De Wolf (1997) 'Post Randomisation for Statistical Disclosure Control: Theory and Implementation', Research Paper No. 9731, Voorburg: Statistics Netherlands.

Heer, G.R. (1993) 'A Bootstrap Procedure to Preserve Statistical Confidentiality in Contingency Tables', in D. Lievesley (ed) *Proceedings of the International Seminar on Statistical Confidentiality*, Luxembourg: Office for Official Publications of the European Communities, pp.261-71.

Hundepool, A., L. Willenborg, A. Wessels, L. Van Gemerden, S. Tiourine, and C. Hurkens (1998) μ-*Argus Users Manual Version 2.5*, Voorburg: Statistics Netherlands, March.

[okay if this is reference cited in text (p. 18) as Hundepool et al. 1999; if not, this work is not cited in text—delete?]Hundepool, A., and L. Willenborg (1999) 'ARGUS: Software From the SDC Project', in *Proceedings of Joint UNECE-Eurostat Work Session on Statistical Data Confidentiality,* Luxembourg: UNECE-Eurostat, pp.87-98.

IDESCAT-Statistics Catalonia (1995) *Sample of 1991 Population Census of Catalonia*.**[more info needed]**

Joint Photographic Experts Group [JPEG] (2001) Standard IS 10918-1 (ITU-T T.81), http://www.jpeg.org.

Kennickell, A. (1998) 'Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances', in J. Domingo-Ferrer (ed) *Proceedings of Statistical Data Protection '98*, Luxembourg: Office for Official Publications of the European Communities, pp.381-400.

Kim, J.J. (1986) 'A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation', in *Proceedings of the ASA Section on Survey Research Methodology*,**[please spell out ASA]** pp.303-8.

Kooiman, P., L. Willenborg, and J. Gouweleeuw (1998) *PRAM: A Method for Disclosure Limitation of Microdata*, Research Report, Voorburg: Statistics Netherlands.

Liew, C.K., U.J. Choi, and C.J. Liew (1985) 'A Data Distortion by Probability Distribution', *ACM Transactions on Database Systems*, 10, pp.395-411.

Little, R.J.A. (1993) 'Statistical Analysis of Masked Data', *Journal of Official Statistics*, 9, pp.407-26.

Mateo-Sanz, J.M., and J. Domingo-Ferrer (1999) 'A Method for Data-Oriented Multivariate Microaggregation', in **[any editor?]***Statistical Data Protection,* Luxembourg: Office for Official Publications of the European Communities, pp.89-99.

Matloff, N.E. (1986) 'Another Look at the Use of Noise Addition for Database Security', in *Proceedings of IEEE Symposium on Security and Privacy*, pp.173-80.

Moore, R. (1996) 'Controlled Data Swapping Techniques for Masking Public Use Microdata Sets', Washington, D.C.: U.S. Bureau of the Census, Statistical Research Division RR96/04.

Oganian, A., and J. Domingo-Ferrer (2001) 'On the Complexity of Microaggregation', in *Second Joint UNECE-Eurostat Work Session on Statistical Data Confidentiality*, Skopje, March.

Rubin, D.B. (1993) 'Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata, *Journal of Official Statistics*, 9, pp.461-8.

Sullivan, G., and W.A. Fuller (1989) 'The Use of Measurement Error to Avoid Disclosure', in *Proceedings of the ASA* **[spell out ASA]***Section on Survey Research Methodology*, pp.802-7.

——— (1990) 'Construction of Masking Error for Categorical Variables', in *Proceedings of the ASA* **[spell out ASA]***Section on Survey Research Methodology*, pp.435-9.

U.S. Bureau of the Census (2001) *American FactFinder,* http://factfinder.census.gov/.

Willenborg, L., and T. De Waal (1996) *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics 111, New York: Springer-Verlag.

Willenborg, L., and T. De Waal (2001) *Elements of Statistical Disclosure Control*, New York: Springer-Verlag.

Winkler W. (1998) 'Re-Identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata', in**[any editor?]** *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, 1999.**[which year is right—it says 1998 after author? is the journal version 1998 and the other 1999, or vice versa, or what?]** Journal version in *Research in Official Statistics*, 1(2), pp.50-69.