

Chapter 6

A Quantitative Comparison of Disclosure Control Methods for Microdata

Josep Domingo-Ferrer*
Universitat Rovira i Virgili

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial-CSIC[spell out CSIC?]

1. Introduction

As described in Chapter 5, there is a plethora of statistical disclosure control (SDC) methods to protect microdata. This chapter provides guidance in choosing a particular SDC method by comparing some of the methods discussed in Chapter 5 on the basis of both information loss and disclosure risk. Information loss can be readily quantified using analytical measures (either generic or data-use-specific). It is far more difficult to assess disclosure risk in a way that is both analytical and applicable to all methods. For this reason, our approach to disclosure risk evaluation is empirical, based on reidentification experiments carried out using record-linkage algorithms.

Methodology

The methodology we use to compare SDC methods is as follows:

- *Test data collection.* We obtained test data from publicly available microdata files. This guarantees public-domain reproducibility of the experiments reported here. The price paid is that the data we start from are not original but have already undergone some amount of disclosure protection.

* This work 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 and by the European Commission under Project ‘CASC’ IST-2000-25069. We are indebted to William Winkler for providing a U.S. Census Bureau implementation of probabilistic record linkage. Thanks go to Francesc Sebé, Narcís Macià, and Àngel Torres for their help in automating the probabilistic record linkage software and running the experiments. Comments by Josep Maria Mateo-Sanz, the editors, and several reviewers are gratefully acknowledged as well.

- *Information loss metrics.* Information loss actually depends on the data uses to be supported by the masked (*i.e.*, SDC-protected) data. However, potential data uses are so diverse that it is hard even to identify them. An alternative and more pragmatic approach for a general purpose comparison is to use a battery of generic, simple, and easily understandable information loss metrics that try to capture structural differences between the original and the masked data files.
- *Disclosure risk assessment.* This risk is empirically quantified as explained above.
- *Empirical work.* Experiments are conducted to obtain t-uples of the form (*method, parms, risk, loss*), where *parms* are the input parameters to *method*, *risk* is the percentage of reidentified records in the test dataset, and *loss* is the information loss. Given *risk*, it is possible to find *method* and *parms* such that $risk(loss, method(parms))$ is minimal (at least over the set of available t-uples). Given *loss*, it is possible to find *method* and *parms* such that $risk(loss, method(parms))$ is minimal (at least over the set of available t-uples).**[in the last two sentences, is “it is possible” okay in both places to make the construction the same? (before edit, the second one read “it should be possible”)]**

Structure of This Chapter

In Section 2, SDC methods included in the comparison are briefly reviewed, together with their parameterizations. Empirical ways to assess disclosure risk are presented in Section 3. The comparison rationale and results are presented in Section 4 for continuous data and in Section 5 for categorical data. Conclusions are summarized in Section 6.

2. SDC Methods Included in the Comparison

This chapter compares a subset of the methods described in Chapter 5. In this section, we briefly review the methods to be compared and specify the parameterizations we consider. Methods for continuous microdata and categorical microdata are discussed separately.

SDC Methods for Continuous Microdata

Microdata SDC methods can be classified as perturbative and nonperturbative (Willenborg and De Waal 2001). Perturbative methods distort records before release, which allows release of the whole population microdataset. Nonperturbative methods do not alter data but partially suppress or reduce the detail of the original dataset. We consider only perturbative methods for continuous data because reidentification is very easy for unperturbed continuous variables. The reason is that, if a continuous variable V_i is left unperturbed in the masked file and is present in an external administrative public file, unique matches are very likely, because for a

continuous variable (even one truncated due to digital representation) V_i is not likely to take the same value for two different records (*i.e.*, it is unlikely that $V_i(o_1) = V_i(o_2)$ if $o_1 \neq o_2$). Thus, distortion turns out to be the only effective way to protect continuous microdata. The subset of perturbative methods considered is:

- *Additive noise* (Noise p for short, where p is a parameter). Gaussian noise is added to the original data to get the masked data. If the standard deviation of the original variable is s , noise is generated using a $N(0,ps)$. Values of p considered in the experiments below are 0.01, 0.02, 0.04, 0.06, 0.08 up to 0.2 with 0.02 increments.
- *Data distortion by probability distribution* (abbreviated Distr (Liew *et al.* 1985)). For each variable in the original microdataset, the best fitting distribution is found; then the fitted distribution is used to generate the masked dataset. There are no parameters. Crystal Ball software (Crystal Ball 2001) is used to find the best fitting distribution.
- *Resampling*. If n is the number of records in the dataset, take with replacement t independent samples X_1, \dots, X_t of size n of the values of an original variable V_i . Independently rank each sample (using the same ranking criterion for all samples). Build the masked variable V'_i by taking as first value the average of the first values of the samples, as second value the average of the second values, and so on. Resampling is tested for $t = 1$ (Resamp1 for short) and $t = 3$ (Resamp3 for short).
- *Microaggregation*. Records are clustered into small aggregates or groups of size at least k (Defays and Nanopoulos 1993; Domingo-Ferrer and Mateo-Sanz 2002). Rather than publishing a variable for a given individual, the average of the values of the variable over the group to which the individual belongs is published. Variants of microaggregation considered include individual ranking (abbr. MicIR k); microaggregation on projected data using z-scores projection (abbr. MicZ k), and principal components projection (abbr. MicPCP k); and microaggregation on unprojected multivariate data considering two variables at a time (abbr. Mic2mul k), three variables at a time (abbr. Mic3mul k), four variables at a time (abbr. Mic4mul k), or all variables at a time (abbr. Micmul k). Values of k between 3 and 10 have been considered.
- *Lossy compression* (abbr. JPEG q , where q is a parameter). This method is new and we propose it for continuous data. The idea is to regard a numerical microdata file as an image (with rows being records and columns being variables). Lossy compression, and more specifically the JPEG algorithm (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 variable ranges and color scales will be needed. The JPEG quality q is taken as a parameter with values from 5 percent up to 100 percent with 5 percent increments.
- *Rank swapping* (abbr. Rank p , where p is a parameter). Although originally described only for ordinal variables, this method can be used for any numerical

variable (Moore 1996). 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). We consider values of p from 1 to 20.

SDC Methods for Categorical Microdata

For categorical data we consider both perturbative and nonperturbative methods and parameterizations. Each method depends on a single parameter and the set of variables to be masked, as described below.

- *Top-coding* (abbr. Tpv , where p is a parameter and v is a set of variables). This method is applied to ordinal categorical variables. In this case the last p values of the variable are recoded into a new category. We consider values of p from 1 to 9. For example, T5o corresponds to the experiment with a set of variables "o" with the parameter $p = 5$ (i.e., five categories are top-coded).
- *Bottom-coding* (abbr. Bpv , where p and v are as above). This method is also applied to ordinal categorical variables. In this case the first p values of the variable are recoded into a new category. We consider values of p from 1 to 9. For example, B3s corresponds to the experiment with a set of variables "s" with the parameter $p = 3$ (i.e., three categories are bottom-coded).
- *Global recoding* (abbr. Gpv , where p and v are as above). In global recoding, some of the categories of the variable are recoded into new ones. Our experimentation considers the following parameterization: recode the p lowest frequency categories into a single one. As before, we consider values of p from 1 to 9.
- *Post-Randomization Method or PRAM* (abbr. Ppv , where p and v are as above). The scores of some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism (a Markov matrix). We select the approach described in Kooiman *et al.* (1998) to define the PRAM matrix, as follows. Let $TV=(TV(1), \dots, TV(K))_t$ be the vector of frequencies of the K categories of variable V in the original file (assume without loss of generality that $TV(k) \geq TV(K) > 0$ for $k < K$) and let θ be such that $0 < \theta < 1$. Then the PRAM matrix for variable V is defined as:

$$p_{lk} = \begin{cases} 1 - \theta T_V(K) / T_V(k) & \text{if } l = k \\ \theta T_V(K) / ((K-1)T_V(k)) & \text{if } l \neq k \end{cases}$$

Let parameter p be $p: = 10\theta$. For each variable we have built nine matrices generated with p taking integer values between 1 and 9.

3. Disclosure Risk Measures

Chapter 5 discusses ways to measure the information loss caused by SDC methods for microdata. However, the assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk must also be measured. The method that optimizes the trade-off between both magnitudes subject to some user requirements turns out to be the best option. To understand the trade-off, consider the two extreme cases between which SDC methods lie:

- If masking consists of encrypting original data, then no disclosure is possible, but no information at all is released (maximum information loss, minimum disclosure risk).
- If no masking is performed and the original data are released, users can perform fully accurate computations, but disclosure of individual respondent data is very likely, especially for microdata (minimum information loss, maximum disclosure risk).

Literature on disclosure risk basically relates to nonperturbative methods, in which a sample of the original dataset is published. Disclosure risk is measured here as the probability that a sample unique is a population unique (Elliot *et al.* 1998; Skinner *et al.* 1994). Unless the sample size is much smaller than the population size, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample can be almost certain that there is a single individual in the population with that value, which is very likely to lead to that individual's identification.

The uniqueness property as stated above is not relevant for perturbative methods, because, even though the whole microdataset is released, it is released with some distortion. Because there is little in the literature on disclosure risk that can be used for a broad class of perturbative methods—disclosure risk measures tend to be method-specific (Adam and Wortmann 1989 is still up-to-date)—empirical methods, such as record linkage techniques, provide a more unified approach. We briefly describe below two approaches to record linkage that yield empirical disclosure risk measures and one analytical measure based on interval disclosure.

Distance-Based Record Linkage

This approach to record linkage is described in Pagliuca and Seri (1999) for the specific case of microaggregation masking using the Euclidean distance. It can be generalized, however, for any perturbative method provided that a distance between the original and the masked value can be defined. As in any record linkage context, it is assumed that an intruder has an external dataset containing as key variables the same variables present in the released masked dataset. The intruder is assumed to try to link the masked dataset with the external dataset using the key variables.

Linkage then proceeds by computing the distances between records in the original and the masked datasets. The distances used are standardized to avoid scaling problems. For each record in the masked dataset, the distance to every record in the original dataset is computed. Then the ‘nearest’ and ‘second nearest’ records in the original dataset are considered. A record in the masked dataset is labeled as ‘linked’ when the nearest record in the original dataset turns out to be the corresponding original record. A record in the masked dataset is labeled as ‘linked to 2nd nearest’ when the second nearest record in the original dataset turns out to be the corresponding original record. In all other cases, a record in the masked dataset is labeled as ‘not linked’. The percentage of ‘linked’ and ‘linked to 2nd nearest’ is a measure of disclosure risk.

Probabilistic Record Linkage

In Jaro (1989), a record linkage method was described and illustrated on the 1985 census of Tampa, Florida. The matching algorithm uses the linear sum assignment model to ‘pair’ records in the two files to be matched (the original file and the masked file in our case). The percentage of correctly paired records is a measure of disclosure risk.

Although less simple than the Euclidean method, this approach is attractive because it requires the user to provide only two probabilities as input: an upper bound of the probability of a false match and an upper bound of the probability of a false non-match. For that reason, this method of record linkage will be termed probabilistic in what follows. The Euclidean method above requires rescaling variables as well as an assumption on the weight of variables when computing a distance. For instance, in the proposal of Pagliuca and Seri (1999), all variables have the same weight.

We use the U.S. Census Bureau implementation of probabilistic record linkage provided by William Winkler (U.S. Bureau of the Census 2000; Winkler 1998) (with some additions) in our experimentation.

Interval Disclosure

For a record in the masked dataset, compute rank intervals as follows: Each variable is independently ranked and a rank interval is defined around the value the variable takes on each record. The ranks of values within the interval for a variable around record r should differ less than p percent of the total number of records and the rank in the center of the interval should correspond to the value of the variable in record r . If so, the proportion of original values that fall into the interval centered around their corresponding masked value is a measure of disclosure risk. A 100 percent proportion means that an attacker is completely sure that the original value lies in the interval around the masked value (interval disclosure).

4. Comparison for Continuous Microdata

This section details the steps of the methodology outlined in the introduction for the case of continuous data.

Test Data Collection

We constructed a microdataset using the Data Extraction System (DES) of the U.S. Census Bureau (<http://www.census.gov/DES/www/welcome.html>). From the available data sources, we chose the Current Population Survey corresponding to 1995—specifically, the file group ‘March Questionnaire Supplement—Person Data Files’. Variables and records were selected as follows:

- *Variable selection.* Our continuous variable selection was based on the requirement that the values of each span a wide range. Thirteen variables were selected: AFNLWGT (Final weight), AGI (Adjusted gross income), EMCONTRB (Employer contribution for health insurance), ERNVAL (Business or farm net earnings), FEDTAX (Federal income tax liability), FICA (Social Security retirement payroll reduction), INTVAL (Amount of interest income), PEARNVAL (Total personal earnings), POTHVAL (Total other persons income), PTOTVAL (Total personal income), STATETAX (State income tax liability), TAXINC (Taxable income amount), WSALVAL (Amount: Total wage and salary).
- *Record selection.* Our selection of 1,080 records was based on the need to keep the number of repeated values for each variable low (in principle, one would not expect repeated values for a continuous variable, but there were repetitions in the dataset).
- The resulting dataset had three properties that were important to our work:
 - 1) The number of records was fewer than 1,200, which allowed repeated experimentation with the probabilistic record linkage software in reasonable time.
 - 2) Seven variables had no repeated values: FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX. Because absence of repeated values is a distinguishing feature of really continuous variables, these seven were chosen as key variables for record linkage.
 - 3) 1,080 (the number of records) is the largest integer less than 1,200, which is a multiple of 5, 8, and 9. Thus, when the microaggregation SDC method is used, the dataset can be microaggregated with minimal group sizes $k = 3, 4, 5, 6, 8, 9,$ and 10 so that all groups have exactly size k .

We used the resulting data to carry out the empirical work described in the ‘Empirical Work’ section below.

Information Loss Metrics

Let X and X' be the matrices representing the original and the masking datasets, respectively. Let V and R be the covariance matrix and the correlation matrix of X ; let \bar{X} be the vector of variable averages for X and let S be the diagonal of V . Define V' , R' , \bar{X}' , and S' analogously from V' . The Information Loss (IL) is computed by averaging the mean variations of $X - X'$, $\bar{X} - \bar{X}'$, $V - V'$, $S - S'$, and the mean absolute error of $R - R'$ and multiplying the resulting average by 100. According to the formulae given in Chapter 5 we obtain:

$$IL = 100 * \left(\frac{\sum_{i=1}^p \sum_{j=1}^n \frac{|x_{ij} - x'_{ij}|}{|x_{ij}|}}{np} + \frac{\sum_{j=1}^p \frac{|\bar{x}_j - \bar{x}'_j|}{|\bar{x}_j|}}{p} + \frac{\sum_{i=1}^p \sum_{1 \leq i \leq j} \frac{|v_{ij} - v'_{ij}|}{|v_{ij}|}}{\frac{p(p+1)}{2}} + \frac{\sum_{i=1}^p \sum_{1 \leq i \leq j} \frac{|r_{ij} - r'_{ij}|}{|r_{ij}|}}{\frac{p(p-1)}{2}} \right) / 5$$

Term 2 is analogous to term 1 but only with respect to the averages of variables. Term 4 is analogous to term 3 but only with respect to the variances of variables (not covariances).

Disclosure Risk Assessment

The Distance Linkage Disclosure risk (*DLD*) is the average percentage of linked records using distance-based record linkage; the average is computed over the number of key variables that the intruder is assumed to know (we consider knowledge of anywhere from one to seven variables). Similarly, the Probabilistic Linkage Disclosure risk (*PLD*) is the average percentage of correctly paired records using probabilistic linkage. The Interval Disclosure (*ID*) is the average percentage of original values falling in the intervals around their corresponding masked values (averages that have been computed over all parameter values, *i.e.*, 1 percent to 10 percent with 1 percent increments).

Empirical Work

Table 1 contains a ranking of methods described in Section 2 (we try the parameter values described in that section for each method). The table contains columns specifying *IL*, *DLD*, *PLD*, and *ID* and also an overall score constructed as follows:

$$Score = 0.5(IL) + 0.125(DLD) + 0.125(PLD) + 0.25 (ID)$$

The rationale of the above weighting is to give equal weight to information loss (0.5) and to disclosure risk. The 0.5 weight of disclosure risk is equally divided among *ID* (0.25) and record linkage. The 0.25 weight of record linkage is equally divided among both approaches to record linkage. The correlation between *DLD* and *PLD* is actually 0.962, so both approaches are very similar. The (*IL*, *DLD*), (*IL*, *PLD*), and (*IL*, *ID*) correlations are -0.605 , -0.551 , and -0.807 ; thus, the lower the

information loss, the higher the disclosure risk, as one would expect. The *ILRank*, *DLDRank*, *PLDRank*, and *IDRank* columns contain the ranking of each method with respect to *IL*, *DLD*, *PLD*, and *ID*; the lower the rank, the better a method performs (*i.e.*, lower information loss and disclosure risk).

Table 1: Comparison Results for Continuous Microdata

Method	IL	DLD	PLD	ID	Score	IL- Rank	DLD- Rank	PLD- Rank	ID- Rank
Rank15	19.01	1.19	0.15	35.05	18.44	53	6	7	21
Rank19	22.95	0.93	0.08	28.04	18.61	59	2	2	2
Rank16	20.91	1.39	0.11	32.18	18.69	56	8	5	16
Rank13	16.77	2.17	0.12	40.35	18.76	48	12	6	28
Rank14	19.72	1.92	0.07	37.00	19.36	55	10	1	25
Rank11	14.32	2.43	0.25	47.81	19.45	44	13	14	39
Rank12	16.37	2.50	0.25	43.73	19.46	47	14	11	35
Rank20	25.81	0.69	0.09	26.83	19.71	64	1	3	1
Rank18	25.74	0.95	0.09	29.25	20.31	63	4	4	6
Rank10	13.37	3.90	0.38	53.17	20.51	41	24	17	45
Rank17	25.12	1.52	0.20	30.95	20.51	61	9	9	10
Rank09	11.66	5.01	0.52	57.58	20.91	38	37	29	49
Rank08	11.60	6.07	0.85	63.37	22.51	37	39	39	56
Rank07	9.25	7.51	1.08	68.71	22.87	30	41	43	63
Rank06	7.87	9.02	2.79	73.80	23.86	26	43	56	71
Mic3mul07	11.06	19.34	4.70	72.34	26.62	36	68	65	69
Rank05	6.78	16.80	13.60	78.89	26.91	22	58	70	77
Mic3mul09	13.46	19.22	3.44	69.91	27.04	42	67	60	65
Mic3mul10	14.84	17.99	3.44	68.61	27.25	46	64	59	62
Mic4mul04	12.14	19.76	6.67	71.85	27.33	39	69	68	68
Mic4mul05	14.50	17.43	5.45	69.09	27.39	45	61	66	64
Mic3mul08	13.51	20.81	4.15	70.68	27.54	43	71	63	66
Mic4mul08	18.89	17.78	3.35	62.84	27.80	52	62	58	55
Mic3mul06	10.24	20.41	13.90	74.00	27.91	33	70	71	72
Mic4mul07	19.36	17.10	2.08	64.41	28.18	54	60	53	58
Mic4mul06	17.91	17.82	3.98	66.41	28.28	50	63	62	60
Mic4mul09	21.35	15.93	2.00	61.66	28.33	58	57	52	54
Mic4mul10	22.98	16.85	2.37	60.56	29.03	60	59	55	51
Mic3mul05	9.73	23.78	18.29	76.59	29.27	31	76	73	74
Mic3mul04	7.45	23.49	22.75	79.14	29.29	24	75	75	79
Mic4mul03	10.69	22.88	16.69	76.89	29.51	35	74	72	75
Rank04	5.90	22.77	22.78	84.12	29.67	20	73	76	86
Micmul03	27.67	14.26	1.88	57.23	30.16	65	54	50	47

Table 1: Comparison Results for Continuous Microdata (Continued)

Method	IL	DLD	PLD	ID	Score	IL- Rank	DLD- Rank	PLD- Rank	ID- Rank
Micmul04	31.74	13.72	1.38	52.44	30.86	67	53	48	44
Mic3mul03	6.29	29.70	29.06	82.95	31.23	21	79	80	85
Micmul05	35.12	11.73	1.14	48.43	31.27	70	46	44	41
Micmul07	37.68	13.20	1.20	43.46	31.50	72	52	45	34
Micmul06	38.77	13.00	1.22	45.76	32.60	73	50	46	37
Micmul08	41.53	13.12	0.99	42.66	33.19	75	51	42	32
Rank03	5.07	31.73	36.92	89.53	33.50	18	80	83	93
Mic2mul10	10.68	49.38	27.29	77.43	34.28	34	86	78	76
Micmul10	44.69	14.66	0.50	40.41	34.34	76	55	27	29
Noise0.16	32.56	15.65	4.66	64.39	34.91	68	56	64	57
Micmul09	45.98	12.82	0.85	40.99	34.95	79	49	40	30
Mic2mul09	9.93	51.03	33.04	78.94	35.21	32	87	81	78
Mic2mul08	8.55	54.31	33.70	79.77	35.22	27	88	82	80
Mic2mul07	7.53	54.72	37.41	81.40	35.63	25	89	84	83
Noise0.12	25.24	22.21	22.39	71.58	36.09	62	72	74	67
Noise0.1	21.14	27.70	29.03	75.20	36.46	57	78	79	73
Mic2mul06	7.03	56.38	42.00	82.89	36.54	23	90	86	84
JPEG080	33.97	19.13	6.93	66.35	36.83	69	65	69	59
Noise0.14	35.13	19.21	6.24	67.62	37.65	71	66	67	61
Noise0.18	41.12	11.96	3.52	60.95	37.73	74	47	61	52
Noise0.08	17.43	36.06	39.76	79.84	38.15	49	82	85	81
Rank02	2.90	47.26	57.47	94.56	38.18	11	85	90	96
JPEG070	44.92	9.66	2.34	57.28	38.28	77	44	54	48
Noise0.2	45.97	10.01	0.97	57.63	38.77	78	45	41	50
Mic2mul05	5.88	58.97	56.84	85.40	38.77	19	92	89	88
JPEG085	29.47	23.85	24.48	72.80	38.98	66	77	77	70
Mic2mul04	4.90	61.53	60.69	87.26	39.54	17	94	91	89
JPEG090	18.17	35.37	46.98	80.87	39.60	51	81	87	82
Noise0.06	13.03	45.54	56.22	84.16	40.28	40	84	88	87
Mic2mul03	3.28	66.97	64.79	90.51	40.74	15	95	92	94
Noise0.04	8.93	58.51	65.28	88.95	42.18	28	91	94	90
JPEG075	50.45	12.67	2.90	61.27	42.49	80	48	57	53
JPEG095	9.06	60.11	66.56	89.23	42.67	29	93	96	92
Resamp3	3.15	67.90	67.63	96.81	42.72	14	96	97	97
Rank01	2.34	69.19	66.35	99.54	43.00	9	97	95	106
JPEG065	57.77	7.02	1.90	53.87	43.47	81	40	51	46
Noise0.02	4.24	77.34	71.32	94.42	44.31	16	99	98	95
Resamp1	3.11	75.42	71.85	98.36	44.56	13	98	99	99
MicPCP03	69.62	3.16	0.77	38.41	44.90	84	17	38	26

Table 1: Comparison Results for Continuous Microdata (Continued)

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
JPEG055	63.70	5.57	1.26	49.70	45.13	83	38	47	42
Noise0.01	2.57	85.19	74.13	97.03	45.46	10	100	103	98
JPEG100	3.06	87.14	73.03	99.14	46.34	12	101	100	101
MicIR10	1.19	97.37	74.07	99.12	46.81	8	102	102	100
MicIR08	1.03	97.84	74.07	99.29	46.83	6	108	101	103
MicIR09	1.14	97.96	74.40	99.24	46.93	7	109	104	102
MicIR06	0.87	97.66	75.28	99.51	46.93	5	106	105	105
MicIR05	0.69	97.58	75.99	99.58	46.94	3	104	106	107
MicIR03	0.45	97.39	78.96	99.79	47.22	1	103	107	109
MicIR04	0.64	97.63	79.78	99.67	47.41	2	105	108	108
MicIR07	0.81	97.79	88.06	99.42	48.49	4	107	109	104
MicPCP04	78.84	3.43	0.62	36.00	48.92	87	19	32	23
JPEG050	73.20	4.26	0.67	47.96	49.21	86	31	36	40
JPEG060	71.24	7.66	1.52	51.71	49.69	85	42	49	43
MicPCP05	82.55	3.94	0.69	34.10	50.38	88	25	37	20
MicPCP07	89.28	4.02	0.62	32.56	53.36	91	27	33	17
MicPCP09	90.78	4.54	0.25	31.40	53.84	94	34	12	13
MicPCP06	90.26	3.37	0.50	33.42	53.97	93	18	26	19
MicZ03	90.25	3.16	0.61	35.71	54.52	92	16	31	22
JPEG035	88.80	3.65	0.44	43.20	55.71	90	20	23	33
JPEG045	87.55	4.15	0.67	46.78	56.07	89	30	35	38
MicZ04	94.94	3.70	0.53	33.04	56.26	96	21	30	18
MicPCP08	96.93	3.97	0.34	32.04	57.02	97	26	16	14
MicPCP10	97.82	4.13	0.46	31.19	57.28	98	29	24	11
JPEG040	90.99	3.72	0.66	44.98	57.29	95	22	34	36
MicZ07	102.87	4.27	0.38	30.53	59.65	99	32	20	9
MicZ06	103.92	3.88	0.41	30.43	60.10	100	23	21	8
MicZ05	104.06	4.03	0.42	31.30	60.41	101	28	22	12
MicZ08	107.92	4.55	0.52	29.60	61.99	102	35	28	7
MicZ10	109.79	4.83	0.38	28.20	62.59	103	36	18	3
MicZ09	110.91	4.35	0.38	28.36	63.14	105	33	19	4
Distr	58.62	43.05	64.88	88.98	65.04	82	83	93	91
JPEG030	110.48	3.02	0.48	41.79	66.12	104	15	25	31
JPEG025	155.15	2.13	0.25	38.76	87.56	106	11	13	27
JPEG020	164.91	1.36	0.29	36.11	91.69	107	7	15	24
JPEG015	202.66	1.10	0.15	32.06	109.50	108	5	8	15
JPEG010	269.38	0.93	0.22	28.44	141.94	109	3	10	5

Because publishing the original data without masking yields a score of 50 ($IL = 0$ and $DLD = PLD = ID = 100$), methods scoring above 50 in Table 1 are of no use.

Figure 1. Comparison of SDC Methods for Continuous Microdata With Best Parameter Choice

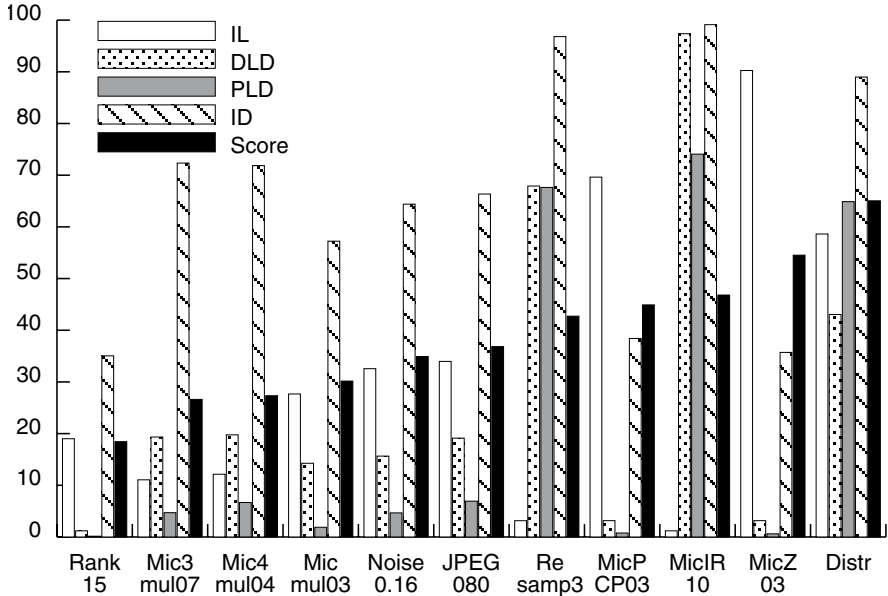


Figure 1 gives a comparison of SDC methods for continuous microdata when the best parameter choice is made. Rank swapping with parameter 15 (Rank15) stands out as the best performer. For multivariate microaggregation, taking three variables at a time and groups of size at least 7 (Mic3mul07) turns out to be the best parameter choice. See Section 6 for expanded conclusions.

5. Comparison for Categorical Microdata

This section details steps of the methodology outlined in the introduction for the case of categorical data.

Test Data Collection

In order to compare masking methods for categorical microdata, we used data from the *American Housing Survey 1993* (also obtained from the U.S. Census Bureau using the DES). We selected the variables BUILT (year structure was built), DEGREE (long-term average degree days), GRADE1 (highest school grade), METRO (metropolitan areas), SCH (schools adequate), SHP (shopping facilities adequate),

TRAN1 (principal means of transportation to work), WHYMOVE (primary reason for moving), WHYTOH (main reason for choice of house), WHYTON (main reason for choosing this neighborhood). We took the first 1,000 records from the corresponding data file (we chose a small dataset for the same reasons mentioned above for continuous microdata).

Five subsets of variables were defined from the set of selected variables, and the same analysis was performed for each of them in the testing process. Three groups were defined by grouping variables with similar number of categories; ‘s’, ‘m’, and ‘l’ correspond to groups of variables with small, medium, and large number of categories, respectively. Additionally, ‘u’ corresponds to the group of variables with medium or large number of categories (union of ‘m’ and ‘l’); ‘o’ corresponds to the subset of variables defined as ordered. The variables used and the groups of variables are given in Table 2. This table also includes the number of categories for each variable.

**Table 2: Variables Used in the Masking Process
and in
the Reidentification Process**

Variables	u	l	s	m	o	N. of Cat.
BUILT	X	X			X	25
DEGREE			X		X	8
GRADE1	X	X			X	21
METRO			X			9
SCH			X			6
SHP			X			6
TRAN1	X			X		12
WHYMOVE	X	X				18
WHYTOH	X			X		13
WHYTON	X			X		13

Information Loss Metrics

For categorical data, we considered three kinds of information loss measures: direct comparison of categorical values, comparison of contingency tables, and entropy-based measures (see Chapter 5 for details).

Direct Comparison of Categorical Values. A distance is defined over the range of categorical variables. When the range of a variable is an ordinal scale, the distance between category a and b is proportional to the number of categories be-

tween a and b . When the range of a variable is not ordinal, the distance is 1 if the values are different and 0 if they are not. We denote this information loss measure by *Dist*.

Comparison of Contingency Tables. For a given subset of variables, contingency tables are computed for a file before and after applying the masking process. The number of differences between the two contingency tables is denoted by *CT-BIL*. Because the number of cells in a contingency table depends on the number of categories in the variable, we also consider the normalizing of *CTBIL* by dividing it by the number of cells in all tables. We denote the resulting information loss measure by *ACTBIL*.

Entropy-Based Measures. In Kooiman *et al.* (1998), the use of Shannon’s entropy to measure information loss is discussed. The idea is that this information-theoretic measure 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. Because this measure depends only on the masked dataset and does not account for its relation with the original data, we define a new information loss measure.

Let V be a variable in the original dataset and V' be the corresponding variable in the PRAM-masked dataset (we take PRAM because it is a very general method encompassing the rest of the masking methods considered for categorical data). Then the entropy-based information loss measure *EBIL* is defined as:

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

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

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

$P_{V, V'} = \{p(V' = j|(V = i))\}$ being the PRAM Markov matrix.

The new information loss measure taking original data into account is:

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

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

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

We compute *EBIL* and *IL* using two different data files to estimate probabilities: the same file masked (*EBILMF* and *ILMF*) and a reference file (*EBILRF* and *ILMF*).

Disclosure Risk Assessment

The Probabilistic Linkage Disclosure risk (*PLD*) is the number of correctly paired records using probabilistic linkage. *PLDRank* is the ranking of the method with respect to *PLD* (normalized with maximal value 100). The inability to use the Euclidean distance for categorical data makes computation of *DLD* and *ID* more cumbersome than for continuous data. Although distance definitions exist for categorical variables (as the ones used for the information loss measures), difficulties arise in ordering pairs of records.

Empirical Work

Table 3 contains a ranking of methods described in the section ‘SDC Methods for Categorical Microdata’ (the parameter values described in that section were tried for each method). The table contains columns with information loss measures (*Dist*, *CTBIL*, *ACTBIL*, *EBILRF*, *ILRF*, *EBILMF*, and *ILMF*), a reidentification measure (*PLD*), and a score of the methods (*Score* and *Ave. Score*) defined as an average between the ranks of disclosure risk and information loss measures. The (*Dist*, *PLD*), (*CTBIL*, *PLD*), (*ACTBIL*, *PLD*), (*EBILRF*, *PLD*), (*ILRF*, *PLD*), (*EBILMF*, *PLD*), and (*ILMF*, *PLD*) correlations are, respectively, -0.4898, -0.4156, -0.5520, -0.345, -0.288, -0.3368 and -0.408. Here again, as expected, the lower the information loss, the higher the disclosure risk.

Table 3: Comparison Results for Categorical Microdata

Method	PLD	Dist	CTBIL	ACTBIL	EBILRF	ILRF	EBILMF	ILMF	PLD Rank	AIL Rank	Score	Ave. Score
T5u	235	2716	31830	11.10	2892.10	3030.30	2952.60	2952.60	5.60	72.82	39.19	41.60
T5s	428	2372	14788	27.00	2453.40	2564.10	2506.90	2506.90	23.90	70.19	47.04	41.60
T5o	347	1093	6148	5.20	1661.00	1727.40	1687.50	1687.50	15.00	56.76	35.88	41.60
T5l	853	319	1820	1.00	486.80	481.40	474.90	474.90	58.30	32.36	45.35	41.60
T5m	769	344	1874	2.70	438.70	466.20	445.70	445.70	45.00	36.11	40.56	41.60
T6u	200	3152	35786	12.50	3764.10	3900.70	3819.50	3819.50	5.00	80.23	42.62	42.19
T6s	457	2789	16400	30.00	3256.00	3357.30	3299.80	3299.80	26.70	74.81	50.74	42.19
T6o	287	1242	6886	5.80	2106.70	2171.90	2130.50	2130.50	10.00	59.77	34.88	42.19
T6l	789	391	2220	1.30	671.20	663.60	655.50	655.50	48.90	37.18	43.03	42.19
T6m	751	363	1966	2.90	508.10	543.40	519.70	519.70	41.70	37.64	39.65	42.19
T3u	288	1081	14154	4.90	942.10	963.80	946.40	946.40	10.60	55.65	33.10	42.90
T3s	748	819	6150	11.20	691.20	717.40	706.70	706.70	41.10	51.62	46.37	42.90
T3o	758	559	3258	2.80	576.40	588.40	579.10	579.10	42.80	41.90	42.34	42.90
T3l	943	177	1028	0.60	179.40	178.90	173.30	173.30	78.90	23.38	51.13	42.90
T3m	825	262	1442	2.10	250.90	246.50	239.70	239.70	52.80	30.37	41.57	42.90
T4u	254	2475	29456	10.30	2289.30	2357.70	2311.40	2311.40	7.20	69.07	38.15	43.15

Table 3: Comparison Results for Categorical Microdata (Continued)

Method	PLD	Dist	CTBIL	ACTBIL	EBILRF	ILRF	EBILMF	ILMF	PLD Rank	AIL Rank	Score	Ave. Score
T4s	492	2144	13756	25.10	1896.40	1943.90	1917.60	1917.60	28.30	67.27	47.80	43.15
T4o	529	854	4896	4.10	1117.70	1135.10	1124.90	1124.90	30.00	50.65	40.32	43.15
T4l	908	233	1348	0.80	304.90	303.10	297.10	297.10	71.10	27.31	49.21	43.15
T4m	775	331	1812	2.60	392.90	413.80	393.80	393.80	46.10	34.40	40.25	43.15
G1u	423	55	770	0.30	0.00	0.00	0.00	0.00	23.30	12.41	17.87	43.25
G1s	1000	51	408	0.90	0.00	0.00	0.00	0.00	98.30	13.80	56.06	43.25
G1o	998	35	210	0.20	0.00	0.00	0.00	0.00	95.60	6.48	51.02	43.25
G1l	998	4	24	0.00	0.00	0.00	0.00	0.00	96.10	0.65	48.38	43.25
G1m	962	4	24	0.00	0.00	0.00	0.00	0.00	84.40	1.39	42.92	43.25
T7u	265	3453	38344	13.40	4389.90	4564.00	4475.60	4475.60	7.80	82.64	45.21	44.15
T7s	687	3067	17418	31.80	3801.90	3943.70	3879.90	3879.90	35.00	80.09	57.55	44.15
T7o	307	1314	7172	6.10	2281.90	2345.20	2303.70	2303.70	12.80	61.76	37.27	44.15
T7l	719	483	2736	1.60	903.20	893.80	885.50	885.50	38.30	41.71	40.02	44.15
T7m	751	386	2072	3.00	588.00	620.30	595.70	595.70	42.20	39.17	40.69	44.15
T9u	429	4163	43976	15.40	4921.40	5230.90	5133.50	5133.50	24.40	88.15	56.30	44.18
T9s	187	3695	19294	35.30	4097.70	4368.10	4305.50	4305.60	4.40	83.56	44.00	44.18
T9o	297	1604	8162	6.90	2925.90	2917.30	2834.90	2834.90	12.20	66.99	39.61	44.18
T9l	534	812	4392	2.50	1665.60	1585.90	1536.00	1536.00	30.60	49.35	39.95	44.18
T9m	725	468	2442	3.50	823.70	862.70	828.00	828.00	38.90	43.19	41.04	44.18
T8u	510	4167	43940	15.30	4957.20	5244.20	5140.20	5140.20	28.90	88.33	58.61	44.75
T8s	44	3774	19386	35.40	4337.20	4583.70	4511.20	4511.20	0.60	84.44	42.50	44.75
T8o	329	1494	7772	6.60	2635.50	2622.50	2542.50	2542.60	14.40	64.81	39.63	44.75
T8l	643	679	3726	2.10	1308.40	1222.40	1175.50	1175.50	33.30	47.08	40.21	44.75
T8m	774	393	2108	3.10	620.00	660.50	629.00	629.00	45.60	40.00	42.78	44.75
G4u	388	1269	16606	6.90	1246.10	1272.40	1241.80	1241.80	16.70	60.88	38.77	45.35
G4s	729	1215	8690	19.80	1183.50	1203.20	1175.60	1175.60	39.40	61.48	50.46	45.35
G4o	810	424	2508	2.40	560.10	597.20	517.50	517.50	51.10	39.21	45.16	45.35
G4l	983	71	420	0.30	81.90	144.90	76.10	76.10	90.00	14.63	52.31	45.35
G4m	893	54	316	0.50	62.60	69.30	66.10	66.10	66.10	13.94	40.02	45.35
T2u	324	655	8802	3.10	406.50	400.40	388.90	388.90	13.90	44.35	29.12	45.36
T2s	890	462	3598	6.60	301.40	299.50	293.30	293.30	65.60	39.72	52.64	45.36
T2o	903	330	1960	1.70	218.00	210.50	205.90	205.90	68.30	30.79	49.56	45.36
T2l	965	132	780	0.40	83.20	79.40	76.10	76.10	85.00	19.49	52.25	45.36
T2m	873	193	1058	1.50	105.00	101.00	95.60	95.60	61.10	25.37	43.24	45.36
T1u	397	222	3058	1.10	0.00	0.00	0.00	0.00	19.40	25.00	22.22	46.88
T1s	1000	172	1362	2.50	0.00	0.00	0.00	0.00	98.90	23.70	61.30	46.88
T1o	1000	109	654	0.60	0.00	0.00	0.00	0.00	99.40	15.56	57.50	46.88
T1l	996	40	240	0.10	0.00	0.00	0.00	0.00	93.90	6.30	50.09	46.88
T1m	933	50	272	0.40	0.00	0.00	0.00	0.00	76.70	9.91	43.29	46.88
G3u	389	546	7384	3.10	440.40	445.50	432.80	432.80	17.20	43.56	30.39	47.03
G3s	916	513	3872	8.80	410.20	412.90	402.20	402.20	73.30	43.80	58.56	47.03

Table 3: Comparison Results for Categorical Microdata (Continued)

Method	PLD	Dist	CTBIL	ACTBIL	EBILRF	ILRF	EBILMF	ILMF	PLD Rank	AIL Rank	Score	Ave. Score
G3o	943	234	1384	1.30	249.60	246.70	225.90	225.90	79.40	28.24	53.84	47.03
G3l	992	40	236	0.10	39.30	45.40	27.60	27.60	91.70	9.35	50.51	47.03
G3m	918	33	194	0.30	30.20	32.60	30.60	30.60	73.90	9.81	41.85	47.03
G2u	411	211	2858	1.20	118.70	123.70	118.10	118.10	21.10	28.19	24.65	47.47
G2s	1000	194	1466	3.30	110.30	115.40	110.10	110.10	100.00	29.07	64.54	47.47
G2o	992	117	702	0.70	79.30	74.40	71.00	71.00	92.80	19.40	56.09	47.47
G2l	996	11	66	0.00	5.50	7.60	7.00	7.00	94.40	4.63	49.54	47.47
G2m	940	17	100	0.20	8.50	8.30	8.00	8.00	78.30	6.76	42.55	47.47
B9u	591	5061	49270	17.20	6282.70	6359.20	6228.20	6228.20	32.80	95.79	64.28	47.92
B9s	50	4000	20000	36.60	5199.20	5375.00	5291.70	5291.80	1.70	89.40	45.53	47.92
B9o	284	1498	7934	6.70	2553.10	2661.80	2468.10	2468.10	8.90	65.32	37.11	47.92
B9l	733	655	3676	2.10	1046.90	1103.30	931.80	931.80	40.00	45.46	42.73	47.92
B9m	760	1061	5270	7.60	1083.50	984.20	936.40	936.40	43.90	55.97	49.93	47.92
B8u	571	4990	49068	17.10	6018.90	6139.00	6020.20	6020.30	31.70	94.91	63.29	48.66
B8s	50	4000	20000	36.60	5199.20	5375.00	5291.70	5291.80	1.10	89.95	45.53	48.66
B8o	296	1465	7810	6.60	2438.60	2535.20	2351.40	2351.40	11.70	63.70	37.69	48.66
B8l	777	599	3380	1.90	869.70	908.70	749.70	749.70	46.70	42.87	44.77	48.66
B8m	813	990	5068	7.40	819.70	763.90	728.50	728.50	51.70	52.41	52.04	48.66
G9u	778	4746	48140	20.10	6035.90	6223.70	6120.40	6120.50	47.20	94.91	71.06	48.98
G9s	50	4000	20000	45.70	5199.20	5375.00	5291.70	5291.80	2.80	90.79	46.78	48.98
G9o	285	1313	7214	6.80	2336.50	2818.20	2404.20	2404.20	9.40	63.47	36.46	48.98
G9l	764	411	2356	1.50	735.20	1172.60	795.10	795.10	44.40	39.72	42.08	48.98
G9m	779	746	4140	6.90	836.60	848.70	828.70	828.70	47.80	49.21	48.50	48.98
G5u	395	2899	34516	14.40	2933.20	3140.60	3073.50	3073.50	18.90	74.58	46.74	49.22
G5s	834	2809	16918	38.60	2820.80	3007.50	2947.30	2947.30	55.00	74.81	64.91	49.22
G5o	648	631	3698	3.50	886.80	1080.00	914.10	914.10	33.90	46.34	40.12	49.22
G5l	972	116	680	0.40	152.80	285.00	152.00	152.00	87.20	19.77	53.50	49.22
G5m	876	90	502	0.80	112.40	133.10	126.20	126.20	62.20	19.44	40.83	49.22
G8u	858	4283	45544	19.00	5680.40	5891.10	5792.00	5792.00	59.40	93.06	76.25	49.33
G8s	50	4000	20000	45.70	5199.20	5375.00	5291.70	5291.80	2.20	91.34	46.78	49.33
G8o	253	1244	6952	6.60	2195.50	2588.70	2239.80	2239.80	6.70	61.06	33.87	49.33
G8l	825	322	1856	1.20	543.60	893.30	581.10	581.10	53.30	34.40	43.87	49.33
G8m	849	283	1544	2.60	481.20	516.10	500.20	500.20	57.20	34.58	45.90	49.33
G6u	464	3288	37942	15.90	3734.20	3940.40	3854.40	3854.40	27.20	81.34	54.28	49.49
G6s	694	3152	17984	41.10	3547.30	3716.70	3641.70	3641.70	36.70	81.25	58.96	49.49
G6o	471	875	5064	4.80	1394.00	1674.80	1427.80	1427.80	27.80	52.87	40.32	49.49
G6l	925	176	1026	0.60	249.40	475.20	262.50	262.50	75.60	25.23	50.39	49.49
G6m	876	136	766	1.30	186.90	223.80	212.60	212.60	62.80	24.21	43.50	49.49
B7u	667	4886	48578	17.00	5678.90	5814.10	5707.10	5707.10	34.40	93.52	63.98	49.58
B7s	78	3967	19934	36.40	5028.60	5226.80	5146.70	5146.80	3.30	88.06	45.69	49.58
B7o	289	1368	7430	6.30	2220.10	2300.80	2129.80	2129.90	11.10	61.94	36.53	49.58

Table 3: Comparison Results for Categorical Microdata (Continued)

Method	PLD	Dist	CTBIL	ACTBIL	EBILRF	ILRF	EBILMF	ILMF	PLD Rank	AIL Rank	Score	Ave. Score
B7l	854	478	2720	1.60	593.70	607.40	465.20	465.20	58.90	38.52	48.70	49.58
B7m	842	919	4842	7.00	650.20	587.30	560.30	560.30	55.60	50.46	53.01	49.58
B6u	688	4761	47670	16.60	5139.10	5323.30	5230.60	5230.70	35.60	90.42	62.99	49.76
B6s	160	3855	19644	35.90	4552.60	4795.10	4728.20	4728.20	3.90	85.93	44.91	49.76
B6o	320	1287	7044	6.00	1922.40	2011.50	1852.40	1852.40	13.30	59.49	36.41	49.76
B6l	899	420	2406	1.40	428.60	432.00	299.30	299.30	67.80	35.05	51.41	49.76
B6m	847	906	4796	7.00	586.50	528.20	502.50	502.50	56.70	49.49	53.08	49.76
G7u	880	3867	42628	17.80	4839.90	5078.10	4985.60	4985.70	63.90	87.08	75.49	50.25
G7s	354	3681	19362	44.20	4546.80	4745.60	4665.60	4665.70	15.60	85.19	50.37	50.25
G7o	246	1181	6710	6.30	2062.70	2430.60	2101.80	2101.80	6.10	59.54	32.82	50.25
G7l	881	243	1408	0.90	371.50	695.60	403.50	403.50	64.40	30.23	47.34	50.25
G7m	875	186	1044	1.70	293.10	332.40	320.00	320.00	61.70	28.80	45.23	50.25
B5u	691	4581	46568	16.30	4524.60	4697.50	4608.60	4608.60	36.10	87.08	61.60	50.36
B5s	278	3694	19288	35.30	4007.60	4247.40	4181.10	4181.10	8.30	83.01	45.67	50.36
B5o	393	1207	6634	5.60	1633.60	1724.00	1576.50	1576.40	17.80	57.31	37.55	50.36
B5l	915	394	2280	1.30	351.40	363.60	241.90	241.90	72.80	33.43	53.10	50.36
B5m	861	887	4728	6.90	517.00	450.20	427.50	427.50	60.00	47.78	53.89	50.36
B4u	824	3516	37970	13.30	2937.10	2998.90	2937.50	2937.50	52.20	78.94	65.58	52.55
B4s	446	2650	15784	28.90	2502.60	2625.10	2585.60	2585.60	25.60	71.20	48.38	52.55
B4o	522	818	4780	4.00	1065.10	1149.10	1032.30	1032.30	29.40	49.77	39.61	52.55
B4l	974	119	704	0.40	132.80	188.10	90.10	90.10	87.80	19.44	53.61	52.55
B4m	882	866	4640	6.70	434.50	373.80	351.90	351.90	65.00	46.11	55.56	52.55
B3u	825	3150	34974	12.20	2053.00	2066.90	2017.70	2017.70	53.90	73.61	63.75	52.71
B3s	576	2294	14336	26.20	1661.50	1742.00	1713.70	1713.70	32.20	67.13	49.68	52.71
B3o	710	606	3550	3.00	604.40	668.50	587.20	587.20	37.80	43.29	40.53	52.71
B3l	985	102	604	0.30	86.50	128.50	64.60	64.60	90.60	16.85	53.70	52.71
B3m	896	856	4598	6.70	391.50	324.80	304.00	304.00	66.70	45.09	55.88	52.71
P9u	394	6086.3	1858	0.90	5678.30	5499.00	6146.30	5676.60	18.30	78.70	48.52	54.83
P9s	733	3086.3	886	2.60	3125.20	3839.40	3243.10	3738.60	40.60	65.28	52.92	54.83
P9o	800	98.462	662	0.70	5318.10	6890.70	4916.00	4718.10	50.00	44.72	47.36	54.83
P9l	947	1012	248	0.20	5305.10	6112.10	5005.10	4164.70	81.10	51.67	66.39	54.83
P9m	907	3000	142	0.30	2553.10	1659.70	2903.20	1938.00	70.60	47.36	58.96	54.83
B2u	795	2738	31508	11.00	1197.40	1205.40	1185.20	1185.20	49.40	66.20	57.82	55.20
B2s	698	1949	12786	23.40	997.90	1042.30	1028.10	1028.10	37.20	63.47	50.35	55.20
B2o	921	336	2006	1.70	193.80	248.90	214.10	214.10	75.00	31.81	53.40	55.20
B2l	992	81	484	0.30	48.60	74.80	46.70	46.70	92.20	14.72	53.47	55.20
B2m	945	789	4346	6.30	199.50	163.10	157.10	157.10	80.60	41.34	60.95	55.20
P8u	399	6077	1748	0.90	5624.80	5464.30	6091.80	5641.80	20.00	77.73	48.87	55.42
P8s	758	3077	804	2.40	3077.30	3808.70	3193.40	3707.00	43.30	64.17	53.75	55.42
P8o	828	87.916	604	0.60	5276.60	6879.40	4871.70	4696.90	54.40	43.33	48.89	55.42
P8l	960	1010.9	230	0.20	5290.60	6110.10	4990.30	4155.40	83.90	50.74	67.31	55.42

Table 3: Comparison Results for Categorical Microdata (Continued)

Method	PLD	Dist	CTBIL	ACTBIL	EBILRF	ILRF	EBILMF	ILMF	PLD Rank	AIL Rank	Score	Ave. Score
P8l	904	3000	140	0.30	2547.50	1655.60	2898.40	1934.80	69.40	47.08	58.26	55.42
P7u	387	6068.8	1544	0.80	5564.20	5426.50	6031.00	5605.60	16.10	76.99	46.55	55.44
P7s	788	3068.8	706	2.10	3024.80	3779.00	3139.80	3678.10	48.30	62.96	55.65	55.44
P7o	850	78.787	522	0.60	5236.10	6861.30	4829.40	4671.00	57.80	42.08	49.93	55.44
P7l	958	1009.9	224	0.20	5272.20	6102.00	4974.00	4139.50	83.30	49.95	66.64	55.44
P7m	905	3000	134	0.30	2539.30	1647.50	2891.30	1927.40	70.00	46.81	58.40	55.44
P6u	410	6059.3	1426	0.70	5503.80	5367.20	5971.60	5544.80	20.60	75.74	48.15	56.37
P6s	802	3059.3	668	2.00	2971.80	3736.80	3086.70	3633.40	50.60	61.90	56.23	56.37
P6o	862	68.418	480	0.50	5188.40	6845.20	4782.90	4645.10	60.60	40.60	50.58	56.37
P6l	966	1009	186	0.10	5255.80	6107.80	4962.60	4137.00	86.10	49.03	67.57	56.37
P6m	910	3000	118	0.20	2532.00	1630.50	2884.90	1911.40	72.20	46.39	59.31	56.37
B1u	537	1950	23888	8.30	0.00	0.00	0.00	0.00	31.10	51.76	41.44	57.54
B1s	1000	1302	9198	16.80	0.00	0.00	0.00	0.00	96.70	49.72	73.19	57.54
B1o	1000	146	876	0.70	0.00	0.00	0.00	0.00	97.20	18.80	58.01	57.54
B1l	1000	59	354	0.20	0.00	0.00	0.00	0.00	97.80	9.63	53.70	57.54
B1m	970	648	3770	5.50	0.00	0.00	0.00	0.00	86.70	36.02	61.34	57.54
P5u	421	6050.2	1228	0.60	5445.30	5288.20	5912.90	5466.10	22.20	74.40	48.31	58.09
P5s	842	3050.2	578	1.70	2922.40	3677.50	3035.10	3575.20	56.10	60.46	58.29	58.09
P5o	903	58.17	432	0.50	5143.50	6815.80	4739.30	4610.10	68.90	39.58	54.24	58.09
P5l	975	1007.9	182	0.10	5234.60	6106.00	4946.10	4130.80	88.90	48.38	68.63	58.09
P5m	925	3000	98	0.20	2522.90	1610.70	2877.90	1891.00	76.10	45.83	60.97	58.09
P4u	429	6041.6	954	0.50	5379.50	5225.40	5848.40	5402.80	25.00	72.78	48.89	58.78
P4s	880	3041.6	468	1.40	2866.00	3629.60	2978.10	3526.80	63.30	58.94	61.13	58.78
P4o	908	48.183	376	0.40	5092.30	6785.80	4690.50	4571.30	71.70	37.45	54.56	58.78
P4l	974	1006.5	164	0.10	5214.80	6105.30	4932.20	4122.90	88.30	47.64	67.99	58.78
P4m	939	3000	66	0.10	2513.50	1595.80	2870.30	1876.00	77.80	44.91	61.34	58.78
P3u	446	6034.1	842	0.40	5316.40	5186.40	5786.00	5363.30	26.10	71.76	48.94	59.63
P3s	896	3034.1	406	1.20	2812.70	3592.20	2923.50	3488.70	67.20	57.45	62.34	59.63
P3o	933	39.313	322	0.30	5041.30	6753.90	4641.70	4531.20	77.20	35.93	56.57	59.63
P3l	979	1005.1	142	0.10	5191.00	6091.00	4915.40	4101.60	89.40	46.71	68.08	59.63
P3m	944	3000	60	0.10	2503.70	1594.20	2862.50	1874.60	80.00	44.40	62.20	59.63
P2u	419	6026.2	706	0.40	5248.30	5142.60	5718.40	5320.70	21.70	70.23	45.95	60.30
P2s	920	3026.2	340	1.00	2754.30	3557.10	2863.80	3454.40	74.40	56.06	65.25	60.30
P2o	951	29.846	242	0.30	4986.80	6724.90	4588.80	4494.50	81.70	33.94	57.80	60.30
P2l	992	1003.5	106	0.10	5165.30	6082.30	4896.60	4084.50	93.30	45.60	69.47	60.30
P2m	953	3000	42	0.10	2494.00	1585.50	2854.60	1866.20	82.20	43.84	63.03	60.30
P1u	422	6013.3	360	0.20	5172.10	5048.10	5642.10	5226.80	22.80	65.79	44.28	61.36
P1s	965	3013.3	178	0.50	2687.40	3463.20	2794.80	3360.60	85.60	52.45	69.00	61.36
P1o	987	15.666	126	0.10	4920.50	6659.70	4526.30	4422.00	91.10	30.74	60.93	61.36
P1l	996	1002.2	50	0.00	5133.50	6071.40	4874.00	4067.70	95.00	44.31	69.65	61.36
P1m	956	3000	16	0.00	2484.70	1584.90	2847.40	1866.20	82.80	43.10	62.94	61.36

In this table, *Dist*, *CTBIL*, *ACTBIL*, *EBILRF*, *ILRF*, *EBILMF*, and *ILMF* are the information loss measures described in Subsection 5.2. *AILRank* is an average rank defined from the ranks of these measures (*DistRank*, *CTBILRank*, *ACTBILRank*, *EBILRFrank*, *ILRFrank*, *EBILMFrank*, *ILMFrank*—not displayed) that gives the same weight to the three classes of information loss measures (distance, contingency table, and entropy-based measures). Within a class, all measures also have the same importance. Therefore, *AILRank* is defined as:

$$AILRank = \frac{DistRank + \frac{CTBILRank + ACTBILRank}{2} + \frac{EBILRFrank + ILRFrank + EBILMFrank + ILMFMRank}{4}}{3}$$

PLD is the number of correctly paired records using probabilistic linkage, and *PLDRank* is the ranking of the method with respect to *PLD* (normalized with maximal value 100).

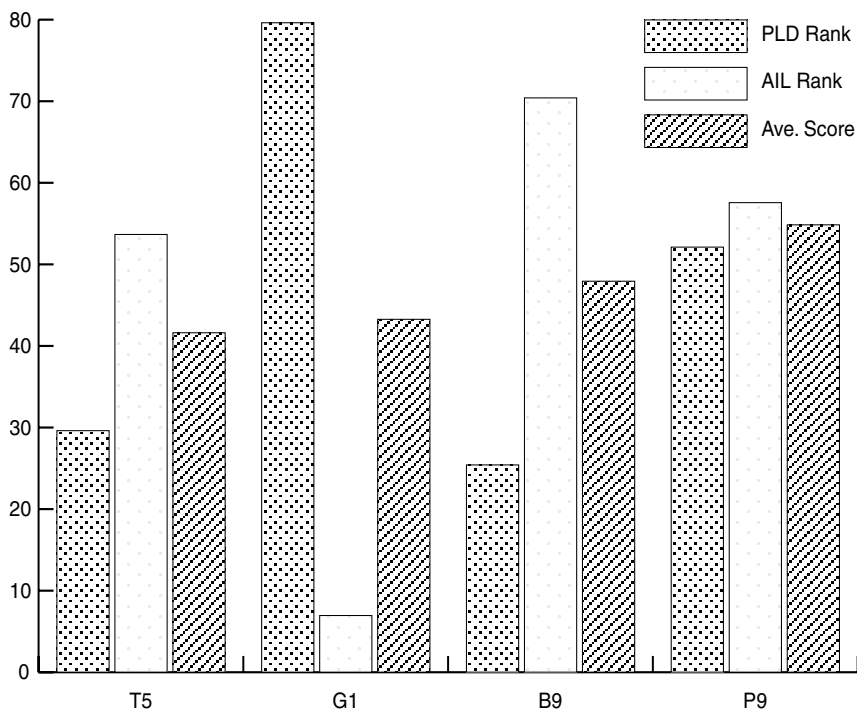
The score is defined as

$$Score = (PLDRank + AILRank)/2$$

to give the same importance to disclosure risk and information loss, and an average score is computed for each pair (*masking method, parameter*). The average is defined over the different choices of variables considered. This is the column labeled *Ave. Score*. Figure 2 is a graphical comparison of the selected SDC methods when the best parameter choice is made (average values over the considered groups of variables—‘m’, ‘u’, ‘s’, ‘o’, and ‘l’—are displayed). Top-coding with several parameterizations turns out to be the best performer (the best parameter choice being 5). See Section 6 for more details.

The results presented here are based on ranks instead of on the values of the information loss measures, because the ranges of the latter are different and difficult to compare with one another.

Figure 2. Comparison of SDC Methods for Categorical Microdata With Best Parameter Choice



6. Conclusions

There is a rich array of methods for microdata disclosure control. A set of proposals for continuous and categorical [“**and categorical**” okay? if not, why not?] microdata has been identified and described in this chapter. Measures for assessing information loss have also been described.

Regarding methods for masking continuous microdata (Table 1), *Distr*, *MicZk*, *MicIRk*, *MicPCPk*, and resampling score around or above 50 for all tried k , so their use is not recommended. Multivariate microaggregation on unprojected data, proposed in Domingo-Ferrer and Mateo-Sanz (2002), is the only form of microaggregation scoring well. Taking three variables at a time seems the best strategy, but even in this case, microaggregation is second to rank swapping. Rank swapping outperforms the best microaggregation and is the best performer (for p around 15 percent). However, being a stochastic method, it is not reproducible and this may lead to disclosure in on-line databases allowing repeated queries. Microaggregation on unprojected data can still be a good option in such cases. Lossy compression

sion (JPEG) in its current form is not excellent, but the general approach is promising; it is a new proposal, so there is room for improvement (such as using other compression algorithms). It is worth noting that random noise masking as implemented to carry out this experimentation is a simple algorithm based on univariate Gaussian noise generation; refinements using multivariate Gaussian noise (Kim 1986) should perform much better.

Regarding methods for categorical microdata, parameterizations of top-coding are the best rated, while the PRAM methods (with the parameters described above) are poorly rated. In general, for the former methods the reidentification risk is low while the information loss is moderate. For the PRAM methods, both reidentification risk and information loss are high. However, the results show that the information loss and the number of reidentifications are highly dependent on the set of variables and the number of categories in each variable. In this sense, the selected parameterizations for PRAM seem particularly ill-suited for variables with a large number of categories.

References [need full names of all authors]

Adam, N.R., and J.C. Wortmann (1989) 'Security-Control Methods for Statistical Databases: A Comparative Study', *ACM [spell out 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.

Domingo-Ferrer, J., and J.M. Mateo-Sanz (2002) 'Practical Data-Oriented Microaggregation for Statistical Disclosure Control', *IEEE [spell out IEEE?] Transactions on Knowledge and Data Engineering* (forthcoming). [still forthcoming?]

Elliot, M.J., C.J. Skinner, and A. Dale (1998) 'Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk', *Research in Official Statistics*, 1(2), pp. 53-67.

Jaro, M.A. (1989) 'Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida', *Journal of the American Statistical Association*, 84, pp.414-20.

Joint Photographic Experts Group (2001) Standard IS 10918-1 (ITU-T T.81) <http://www.jpeg.org/>.

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

Moore, R. (1996) 'Controlled Data Swapping Techniques for Masking Public Use Microdata Sets', U. S. Bureau of the Census (unpublished manuscript).

Pagliuca, D., and G. Seri (1999) *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*, Esprit SDC Project, Deliverable MI-3/D2.[**any more info needed?**]

Skinner, C., C. Marsh, S. Openshaw, and C. Wymer (1994) 'Disclosure Control for Census Microdata', *Journal of Official Statistics*, 10, pp.31-51.

U.S. Bureau of the Census (2000) 'Record Linkage Software: User Documentation'. Available from U. S. Bureau of the Census.

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 *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, 1999.[**date at beginning of entry is 1998, presumably reflecting first publication in journal. Does this "1999" mean the EC published it a year later or what? might readers be confused? (same query in Chapter 5 listing)**] Journal version in *Research in Official Statistics*, 1(2), pp.50-69.