Concepts for the Evaluation of Anonymized Data

Josep Domingo-Ferrer, Josep M. Mateo and Àngel Torres [*)]

Abstract: We present in this paper some criteria for empirical comparison of SDC methods for continuous microdata. Based on re-identification experiments, we try to optimize the tradeoff between information loss and disclosure risk. SDC methods compared include additive noise, distortion by probability distribution, microaggregation, resampling, rank swapping and a novel approach based on lossy compression. Generic information loss measures (not targeted to specific data uses) are defined, and two approaches to empirical re-identification are used: Euclidean record linkage and probabilistic record linkage. Some weighting schemes to aggregate information loss and disclosure risk measures are discussed and empirical results are given for one of them.

Keywords: Statistical disclosure control, Continuous microdata, Record linkage, Re-identification experiments, Information loss measures.

1. Introduction

This paper describes an empirical approach to evaluating the protection of statistical microdata. This work was started in the context of the U.S. Census Bureau contract OTTILIE-R (Optimizing the Tradeoff between Information Loss and dIsclosurE risk for continuous microdata) and continued during the European project CASC (Computational Aspects of Statistical Confidentiality). The idea is to define information loss and disclosure risk measures and then construct a score which combines both types of measures. Specifically, the following steps have been taken in this work:

– *Literature analysis*. Literature on SDC for microdata has been analyzed to identify those methods which are relevant for protecting continuous data. In addition, SDC of continuous microdata based on lossy compression has been introduced.
– *Test data*. Test data have been obtained from publicly available microdata files.
– *Disclosure risk assessment*. Two record linkage algorithms have been used to establish the disclosure risk associated to a particular SDC method. In addition, an interval disclosure measure has been defined.
– *Metrics definition*. Information loss actually depends on the data uses to be supported by masked data. Since data uses did not fall within the scope of OTTILIE-R, we have defined a battery of

[*)] Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics
E-43006 Tarragona, Catalonia, Spain, E-mail jdomingo@etse.urv.es

generic, robust information loss metrics which try to capture structural differences between the original and masked data files.

– *Empirical work*. Experiments carried out are directed to obtaining t-uples of the form *(method, parms, risk, loss)*, where *parms* are the input parameters to *method*, *risk* is the percent of re-identified records in the test data set and *loss* is the information loss. The obtained t-uples can be aggregated to rank methods; depending on the aggregation used, different method rankings are conceivable.

Section 2 reviews relevant SDC methods for the protection of continuous microdata. Section 3 lists information loss measures which have been taken into account in experimentation. Section 4 describes record linkage approaches to assessing disclosure risk. Section 5 reports on actual comparison results. Section 6 discusses alternative score constructions for aggregating information loss and disclosure risk. Section 7 is a conclusion.

2. Relevant SDC methods for continuous microdata

Sampling methods consist of publishing a sample of records from the original microdata set instead of publishing the whole original microdata set. Sampling methods are suitable for categorical microdata, but their adequacy for continuous microdata is less clear in a general disclosure scenario. The reason is that such methods leave a continuous variable *V* unperturbed for all individuals in the sample. Thus, if variable *V* is present in an external administrative public file, unique matches with *V'* (the version of variable *V* restricted to the published sample) are likely, since it is unlikely that a continuous variable (even one truncated due to digital representation) takes exactly the same value for more than one individual. Thus, we will concentrate in what follows on perturbative methods, which have the additional advantage of allowing the entire microdata set to be released.

Perturbative methods distort the microdata set before publication. Perturbative methods considered in our work are a subset of those making sense for continuous microdata:

– *Additive noise* (Noise*p* for short). Gaussian noise is added to the original data to get the masked data (Kim 1986). 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* (Distr for short,(Liew/Choi/Liew 1985)). For each variable in the original variable, the best fitted distribution is found; then the fitted distribution is used to generate the masked data set. There are no parameters.

– *Resampling*. Originally proposed for protecting tabular data (Domingo-Ferrer/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 $X_1, ..., X_t$ 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 $X_1, ..., X_t$. Resampling has been tested for *t=1* (Resamp1) and *t=3* (Resamp3).

– *Microaggregation*. Records are clustered into small aggregates or groups of size at least *k* (Defays/Nanopoulos 1993; Domingo-Ferrer/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 (MicIR*k*); microaggregation on projected data using z-scores projection (MicZ*k*) and principal components projection (MicPCP*k*); microaggregation on unprojected multivariate data considering two variables at a time (Mic2mul*k*), three variables at a time (Mic3mul*k*), four variables at a time (Mic4mul*k*) or all variables at a time (Micmul*k*). Values of *k* between 3 and 10 have been considered.

– *Lossy compression* (JPEG*q*). This method is new and proposed by these authors for continous 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, Standard IS 10918-1 (ITU-T T.81), http://www.jpeg.org), 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* has been taken as a parameter with values from 5% up to 100% with 5% increments.

– *Rank swapping* (Rank*p*). Although originally described only for ordinal variables, this method can be used for any numerical variable (Moor 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%* of the total number of records). The following values of *p* have been considered in experimentation: 1, 2, 3, 4, 5, 6, 7 and 10.

3. Information loss measures

To evaluate the information loss caused by an SDC method on a continuous microdata set, we want to assess how different the masked data set is from the original data set. We will say there is little information loss if the structure of the masked data set is very similar to the structure of the original data set. In fact, the motivation for preserving the structure of the data set is to ensure that the masked data set will be analytically valid and interesting. We can actually try several complementary ways to assess the preservation of the structure of the original data set:

1. Compare the data in the original and the masked data sets. The more similar the SDC method to the identity function, the less impact (but the higher the disclosure risk!).

2. Compare some statistics computed on the original and the masked data sets.

Let X and X' be the original and the masked data set. Let V and V' be the covariance matrices of X and X', respectively; similarly, let R and R' be the correlation matrices. Table 1 summarizes the measures proposed. In this table, p is the number of variables, n the number of records, and components of matrices are represented by the corresponding lowercase letters (e.g. x_{ij} is a component of matrix X). Regarding X –X' measures, it also makes sense to compute those on the averages of variables rather than on all data (see the $\bar{X} - \bar{X}'$ row in Table 1). Similarly, for V – V' measures, it is also sensible to compare only the variances of the variables, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (see the S – S' row in Table 1).

Table 1. Information loss measures

	Mean square error	Mean abs. Error	Mean variation
X-X'	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$\bar{X}-\bar{X}'$	$\frac{\sum_{j=1}^{p}(\bar{x}_j-\bar{x'}_j)^2}{p}$	$\frac{\sum_{j=1}^{p} ¿\lvert\bar{x}_j-\bar{x'}_j\rvert^2}{p}$ ¿	$\frac{\sum_{j=1}^{p}\frac{\lvert\bar{x}_j-\bar{x'}_j\rvert}{\lvert\bar{x}_j\rvert}}{p}$
V-V'	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
S-S'	$\frac{\sum_{j=1}^{p}(v_{jj}-v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^{p}\lvert v_{jj}-v'_{jj}\rvert}{p}$	$\frac{\sum_{j=1}^{p}\frac{\lvert v_{jj}-v'_{jj}\rvert}{\lvert v_{jj}\rvert}}{p}$
R-R'	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$

4. Disclosure risk measures

The assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk is another magnitude that should be measured. The method that optimizes the tradeoff between both magnitudes subject to some user requirements turns out to be the best option.

Literature on disclosure risk is basically related to sampling methods, in which a sample of the original data set is published. Disclosure risk here is measured as the probability that a sample unique is a population unique (Skinner/Marsh/Openshaw/Wymer 1994). If the size of the sample is similar to the size of the whole population, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample could be almost sure that there is a single individual in the population with that value. This could lead to identification of that individual.

The uniqueness property as stated above is no longer relevant for perturbative methods, since in this case the whole microdata set is published, but with some distortion. There is not much literature on disclosure risk that can be used for a broad class of perturbative methods; disclosure risk measures tend to be method-specific (measures described in (Adam/Wortmann 1989) are still up-to-date). Empirical methods, like record linkage techniques, provide a more unified approach to disclosure risk assessment for perturbative methods. We briefly describe below two approaches to record linkage and one measure of interval disclosure.

4.1 Distance-based record linkage

This approach to record linkage is described in (Pagliuca/Seri 1998) for the specific case of microaggregation masking and using the Euclidean distance. However, it can be generalized 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 data set containing as key variables some of the same variables present in the released masked data set. The intruder is assumed to try to link the masked data set with the external data set.

Linkage then proceeds by computing the distances between records in the original and the masked data sets. The distances used are standardized to avoid scaling problems. For each record in the masked data set, the distance to every record in the original data set is computed. Then the "nearest" and "second nearest" records in the original data set are considered. A record in the masked data set is labelled as "linked" when the nearest record in the original data set has the same record number is the corresponding original record). A record in the masked data set is labelled as "linked to 2nd nearest" when the second nearest record in the original data set has the same record number. In all other cases, a record in the masked data set is labelled as "not linked". The percent of "linked" and "linked to 2nd nearest" is a measure of disclosure risk.

4.2 Probabilistic record linkage

In (Jaro 1989), a probabilistic 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 percent of correctly paired records is a measure of disclosure risk.

Although less simple than the Euclidean method described in the previous section, this approach is attractive because it only requires the user to provide two probabilities as input: one is an upper bound of the probability of a false match, and the other an upper bound of the probability of false non-match. 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/Seri 1998), all variables have the same weight.

The U.S. Census Bureau implementation of probabilistic record linkage provided by W. Winkler (U. S. Bureau of Census 2000; Winkler 1998) has been used (with some additions) in the experimentation.

4.3 Interval disclosure

For a record in the masked data set, take a rank interval centered on the values of that record 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%* 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*. Then the measure is the proportion of original values which fall into the interval centered around their corresponding masked value. A 100% proportion means that an intruder is completely sure that the original value lies in the interval around the masked value (interval disclosure). Values of *p* ranging between 1% and 10% have been considered for experimentation.

5. Comparison results

A microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (http://www.census.gov/DES). 13 continuous variables were chosen and 1080 records were selected so that there were not many repeated values for any of the variables (in principle, one would not expect repeated values for a continuous variable, but there were repetitions in the data set). Table 2 contains a ranking of methods described in Section 2 (the parameter values described in that section were tried for each method). The Information Loss column (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'; the resulting average has been multiplied by 100. The Distance Linkage Disclosure risk column (DLD) contains the average percent 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 have considered knowledge of 1 up to 7 variables). Similarly, the Probabilistic Linkage Disclosure risk column (PLD) is the average percent of correctly paired records using probabilistic linkage. The Interval Disclosure (ID) column contains the average percent of original values falling in the intervals around their corresponding masked values (averages have been computed over all parameter values, i.e. 1% to 10% with 1% increments). Finally, the column Score has been used to rank Table 2 and has been computed as

$$\text{Score} = 0.5 \cdot \text{IL} + 0.125\ \text{DLD} + 0.125 \cdot \text{PLD} + 0.25 \cdot \text{ID}.$$

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 IL Rank, DLD Rank, PLD Rank and ID Rank 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 2. Comparison results

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**
iRank20	**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**
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**

Table 2. Comparison results

Method	**IL**	DLD	**PLD**	ID	**Score**	IL Rank	**DLD Rank**	PLD Rank	**ID Rank**
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**
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**

Table 2. Comparison results

Method	**IL**	DLD	**PLD**	ID	**Score**	IL Rank	**DLD Rank**	PLD Rank	**ID Rank**
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**

6. Alternative score constructions

In (Torres 2003), an alternative score construction is proposed which can be used instead of the one proposed in the previous section. Like in the previous score, information loss and disclosure risk are both assigned 0.5 weight. The differences are in the way of information loss and disclosure risk are computed.

IL is computed as the 100 times the average of IL1, IL2 and IL3, where IL1 is the mean variation of X-X'; IL2 is the average of the mean variation of $\bar{X}$- $\bar{X}$' and the mean variation of S-S'; and IL3 is the average of the mean variation of V-V' and the mean absolute error of R-R'. Thus, IL1 represents the discrepancy between raw original and raw protected data, IL2 represents the discrepancy between univariate statistics for original and protected data, while IL3 represents the discrepancy between bivariate statistics for original and protected data. The philosophy here is to give equal weight to those three discrepancies.

Disclosure risk assessment does not use probabilistic record linkage (only distance-based record linkage DLD is used), which allows larger experiments to be carried out in less time. Two kinds of interval disclosure are distinguished: ID1, which is based on ranks and exactly corresponds to ID as defined in Sections 4.3 and used in Section 5, and ID2, which is based on standard deviations. ID2 is defined as the average percent of original values falling in the intervals around their corresponding masked values, where intervals are centered on the actual masked values (not on their rank) and interval widths are *p%* of the standard deviation of the variable.

Thus, this score can be formalized as:

$$\text{Score2} = 0.5\cdot((\text{IL1}+\text{IL2}+\text{IL3})/3)+0.25\cdot\text{DLD} +0.125\cdot\text{ID1}+0.125\cdot\text{ID2}$$

Results obtained with this new score are similar to those reported in Section 5 and Table 2 regarding the relative ranking of families of methods. Rank swapping scores best, closely followed by microaggregation. The differences show up in the best scoring parameterizations: while in Table 2, Rank15 was best, the best method with Score2 is Rank07. Also, with Score2 microaggregation Micmul16 and Micmul10 appear as the fourth and eighth best scoring methods, respectively.

7. Conclusions

There is a rich array of methods for microdata disclosure limitation. A set of proposals for continuous microdata have been identified and described in this paper. Measures for assessing information loss have also been described. Experimental results presented in Table 2 are self-explanatory. One thing that stands out is that rankswapping with parameter around 10% is a very good option; next follows multivariate microaggregation taking groups of three or four variables at a time; for microaggregation, the group size has no significant effect. Data distortion by probability distribution turns out to perform very poorly. For most methods, performance depends on parameter choice, even if some methods are more parameter-dependent than other.

Changing the way the score is constructed can change the experimental results to a limited extent: the relative ranking between families of methods (rank swapping, microaggregation, etc.) is likely to stay the same as long as the score strikes a balance between information loss and disclosure risk.

Acknowledgments

This work was partly funded by the U.S. Bureau of the Census under contract OTTILIE-R (ref. no OBLIG-2000-29158-0-0) and by the European Commission under project CASC (ref. no. IST-2000-25069). Thanks go to Francesc Sebé for his help in automating the probabilistic record linkage software and running the experiments. The contribution of Sarah Giessing as Ko-Referent is also gratefully acknowledged.

References

Adam, N. R., Wortmann, J. C., (1989), Security-control methods for statistical databases: a comparative study, ACM Computing Surveys, vol. 21(4):515-556.

Defays, D., Nanopoulos, P., (1993), Panels of enterprises and confidentiality: the small aggregates method, in Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada, 195-204.

Domingo-Ferrer, J., Mateo-Sanz, J. M., (1999), On resampling for statistical confidentiality in contingency tables, Computers & Mathematics with Applications, vol 38: 13-32.

Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002), Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering, vol 14(1): 189-201.

Heer, G. R., (1993), A bootstrap procedure to preserve statistical confidentiality in contingency tables, in Proceedings of the International Seminar on Statistical Confidentiality (ed. D. Lievesley), Luxemburg: Office for Official Publications of the European Communities, 261-271.

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, vol. 84:414-420.

Joint Photographic Experts Group, 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 Proc. of the ASA Sect. on Survey Res. Meth., pp. 303-308.

Liew, C. K., Choi, U. J., Liew, C. J., (1985), A data distortion by probability distribution, ACM Transactions on Database Systems, vol. 10: 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., Seri, G., (1998), Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.

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

Torres, A. (2003), Contributions to Microaggregation for Statistical Data Protection, Ph.D. Dissertation (in Catalan), Polytechnical University of Catalonia, Barcelona, 2003. Advisors: J. Domingo-Ferrer and J. M. Mateo-Sanz.

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

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. Journal version in Research in Official Statistics, vol. 1(2): 50-69, 1998.