

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**

(Skopje, The former Yugoslav Republic of Macedonia,
14-16 March 2001)

Working Paper No. 5
English only

Topic I: Application of statistical disclosure control (SDC) methodology and software in business statistics and social and demographic statistics

**AN EMPIRICAL COMPARISON OF SDC METHODS FOR CONTINUOUS
MICRODATA IN TERMS OF INFORMATION LOSS AND DISCLOSURE RISK**

Contributed paper

Submitted by Universitat Rovira i Virgili, Catalonia Spain¹

Abstract: We present in this paper the first 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 the 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.

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

I. INTRODUCTION

1. This paper describes some preliminary results of OTTILIE-R (Optimizing the Tradeoff between Information Loss and dIsclOsurE risk for continuous microdata), which is a project funded by U.S. Bureau of the Census and conducted at the Universitat Rovira i Virgili.

2. The purpose of the experimentation work carried out under OTTILIE-R is to demonstrate a methodology for optimizing the tradeoff between information loss and disclosure risk. The approach used is as follows:

- *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 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

¹ Prepared by Josep Domingo-Ferrer and Josep M. Mateo-Sanz, Dept. of Computer Engineering and Mathematics (e-mail {jdomingo, jmateo}@etse.urv.es).

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.

3. Section II reviews relevant SDC methods for the protection of continuous microdata. Section III lists information loss measures which have been taken into account in experimentation. Section IV describes record linkage approaches to assessing disclosure risk. Section V reports on actual comparison results. Section VI is a conclusion.

II. RELEVANT SDC METHODS FOR CONTINUOUS MICRODATA

4. 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 unperturbed for all individuals in the sample. Thus, if variable V_i is present in an external administrative public file, unique matches are very likely, since for a continuous variable (even a digitally represented one) it is highly unlikely that $V_i(o_1) = V_i(o_2)$ if $o_1 \neq o_2$. Thus, we will consider only perturbative methods.

5. Perturbative methods considered are a subset of those making sense for continuous microdata:

- *Additive noise* (`NoiseP` for short). Gaussian noise is added to the original data to get the masked data [Kim86]. 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,[Liew85]). 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*. Take t independent samples X_1, \dots, X_t of the values of an original variable V_i . Sort all samples using the same ranking criterion. 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 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 [Defa93,Domi02]. 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 (`MicIRk`); microaggregation on projected data using z-scores projection (`MicZk`) and principal components projection (`MicPCPk`); microaggregation on unprojected multivariate data considering two variables at a time (`Mic2mulk`), three variables at a time (`Mic3mulk`), four variables at a time (`Mic4mulk`) or all variables at a time (`Micmulk`). Values of k between 3 and 10 have been considered.
- *Lossy compression* (`JPEGq`). This method is new and proposed by these authors 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 [JPEG], 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* (`Rankp`). Although originally described only for ordinal variables, this method can be used for any numerical variable [Moor96]. 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.

III. INFORMATION LOSS MEASURES

6. 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:

- (i) 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!).
- (ii) Compare some statistics computed on the original and the masked data sets.

7. 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
$X - X'$	$\frac{\sum_{j=1}^p \sum_{i=1}^n (x_{ij} - x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n x_{ij} - x'_{ij} }{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n \frac{ x_{ij} - x'_{ij} }{ x_{ij} }}{np}$
$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^p (\bar{x}_j - \bar{x}'_j)^2}{p}$	$\frac{\sum_{j=1}^p \bar{x}_j - \bar{x}'_j }{p}$	$\frac{\sum_{j=1}^p \frac{ \bar{x}_j - \bar{x}'_j }{ \bar{x}_j }}{p}$
$V - V'$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} (v_{ij} - v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} v_{ij} - v'_{ij} }{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} \frac{ v_{ij} - v'_{ij} }{ v_{ij} }}{\frac{p(p+1)}{2}}$
$S - S'$	$\frac{\sum_{j=1}^p (v_{jj} - v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^p v_{jj} - v'_{jj} }{p}$	$\frac{\sum_{j=1}^p \frac{ v_{jj} - v'_{jj} }{ v_{jj} }}{p}$
$R - R'$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} (r_{ij} - r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} r_{ij} - r'_{ij} }{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} \frac{ r_{ij} - r'_{ij} }{ r_{ij} }}{\frac{p(p-1)}{2}}$

IV. DISCLOSURE RISK MEASURES

8. 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.

9. 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 [Skin94]. 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.

10. 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 [Adam89] 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.

IV.1 Distance-based record linkage

11. This approach to record linkage is described in [Pagl98] 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 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.

12. 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.

IV.2 Probabilistic record linkage

13. In [Jaro89], 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.

14. 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 [Pagl98], all variables have the same weight.

15. The U.S. Census Bureau implementation of probabilistic record linkage provided by W. Winkler [USBC,Wink98] has been used (with some additions) in the experimentation.

IV.3 Interval disclosure

16. 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.

V. COMPARISON RESULTS

17. 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 (I.L.) 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 (D.L.D.) contains the percent of linked records using distance-based record linkage. Similarly, the Probabilistic Linkage Disclosure risk column (P.L.D.) is the percent of correctly paired records using probabilistic linkage. The Interval Disclosure (I.D.) 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

$$\text{Score} = 0.5(\text{I.L.}) + 0.125(\text{D.L.D.}) + 0.125(\text{P.L.D.}) + 0.25(\text{I.D.}).$$

18. 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 I.D. (0.25) and record linkage. The 0.25 weight of record linkage is equally divided among both approaches to record linkage. The correlation between D.L.D. and P.L.D. is actually 0.962, so both approaches are very similar. The (I.L.,D.L.D.), (I.L.,P.L.D.) and (I.L.,I.D.) 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 I.L. Rank, D.L.D. Rank, P.L.D. Rank and I.D. Rank columns contain the ranking of each method with respect to I.L., D.L.D., P.L.D. and I.D.; the lower the rank, the better a method performs (i.e. lower information loss and disclosure risk).

Table 2. Comparison results

Method	I.L.	D.L.D.	P.L.D.	I.D.	Score	I.L. Rank	D.L.D. Rank	P.L.D. Rank	I.D. Rank
Rank10	13.4	3.9	0.4	53.2	20.5	39	14	7	35
Rank7	9.2	7.5	1.1	68.7	22.9	30	29	31	51
Rank6	7.9	9.0	2.8	73.8	23.9	26	31	44	59
Mic3mul7	11.1	19.3	4.7	72.3	26.6	36	56	53	57
Rank5	6.8	16.8	13.6	78.9	26.9	22	46	58	65
Mic3mul9	13.5	19.2	3.4	69.9	27.0	40	55	48	53
Mic3mul10	14.8	18.0	3.4	68.6	27.3	43	52	47	50
Mic4mul4	12.1	19.8	6.7	71.8	27.3	37	57	56	56
Mic4mul5	14.5	17.4	5.4	69.1	27.4	42	49	54	52
Mic3mul8	13.5	20.8	4.2	70.7	27.5	41	59	51	54
Mic4mul8	18.9	17.8	3.3	62.8	27.8	47	50	46	44
Mic3mul6	10.2	20.4	13.9	74.0	27.9	33	58	59	60
Mic4mul7	19.4	17.1	2.1	64.4	28.2	48	48	41	46
Mic4mul6	17.9	17.8	4.0	66.4	28.3	45	51	50	48
Mic4mul9	21.4	15.9	2.0	61.7	28.3	50	45	40	43
Mic4mul10	23.0	16.9	2.4	60.6	29.0	51	47	43	40
Mic3mul5	9.7	23.8	18.3	76.6	29.3	31	64	61	62
Mic3mul4	7.5	23.5	22.8	79.1	29.3	24	63	63	67
Mic4mul3	10.7	22.9	16.7	76.9	29.5	35	62	60	63
Rank4	5.9	22.8	22.8	84.1	29.7	20	61	64	74
Micmul3	27.7	14.3	1.9	57.2	30.2	53	42	38	37

Micmul4	31.7	13.7	1.4	52.4	30.9	55	41	36	34
Mic3mul3	6.3	29.7	29.1	83.0	31.2	21	67	68	73
Micmul5	35.1	11.7	1.1	48.4	31.3	58	34	32	31
Micmul7	37.7	13.2	1.2	43.5	31.5	60	40	33	26
Micmul6	38.8	13.0	1.2	45.8	32.6	61	38	34	28
Micmul8	41.5	13.1	1.0	42.7	33.2	63	39	30	24
Rank3	5.1	31.7	36.9	89.5	33.5	18	68	71	81
Mic2mul10	10.7	49.4	27.3	77.4	34.3	34	74	66	64
Micmul10	44.7	14.7	0.5	40.4	34.3	64	43	16	21
Noise0.16	32.6	15.6	4.7	64.4	34.9	56	44	52	45
Micmul9	46.0	12.8	0.8	41.0	34.9	67	37	28	22
Mic2mul9	9.9	51.0	33.0	78.9	35.2	32	75	69	66
Mic2mul8	8.6	54.3	33.7	79.8	35.2	27	76	70	68
Mic2mul7	7.5	54.7	37.4	81.4	35.6	25	77	72	71
Noise0.12	25.2	22.2	22.4	71.6	36.1	52	60	62	55
Noise0.1	21.1	27.7	29.0	75.2	36.5	49	66	67	61
Mic2mul6	7.0	56.4	42.0	82.9	36.5	23	78	74	72
JPEG80	34.0	19.1	6.9	66.3	36.8	57	53	57	47
Noise0.14	35.1	19.2	6.2	67.6	37.7	59	54	55	49
Noise0.18	41.1	12.0	3.5	61.0	37.7	62	35	49	41
Noise0.08	17.4	36.1	39.8	79.8	38.2	44	70	73	69
Rank2	2.9	47.3	57.5	94.6	38.2	11	73	78	84
JPEG70	44.9	9.7	2.3	57.3	38.3	65	32	42	38
Noise0.2	46.0	10.0	1.0	57.6	38.8	66	33	29	39
Mic2mul5	5.9	59.0	56.8	85.4	38.8	19	80	77	76
JPEG85	29.5	23.8	24.5	72.8	39.0	54	65	65	58
Mic2mul4	4.9	61.5	60.7	87.3	39.5	17	82	79	77
JPEG90	18.2	35.4	47.0	80.9	39.6	46	69	75	70
Noise0.06	13.0	45.5	56.2	84.2	40.3	38	72	76	75
Mic2mul3	3.3	67.0	64.8	90.5	40.7	15	83	80	82
Noise0.04	8.9	58.5	65.3	89.0	42.2	28	79	82	78
JPEG75	50.4	12.7	2.9	61.3	42.5	68	36	45	42
JPEG95	9.1	60.1	66.6	89.2	42.7	29	81	84	80
Resamp1	3.1	67.9	67.6	96.8	42.7	14	84	85	85
Rank1	2.3	69.2	66.3	99.5	43.0	9	85	83	94
JPEG65	57.8	7.0	1.9	53.9	43.5	69	28	39	36
Noise0.02	4.2	77.3	71.3	94.4	44.3	16	87	86	83
Resamp3	3.1	75.4	71.9	98.4	44.6	13	86	87	87
MicPCP3	69.6	3.2	0.8	38.4	44.9	72	7	27	19
JPEG55	63.7	5.6	1.3	49.7	45.1	71	27	35	32
Noise0.01	2.6	85.2	74.1	97.0	45.5	10	88	91	86
JPEG100	3.1	87.1	73.0	99.1	46.3	12	89	88	89
MicI10	1.2	97.4	74.1	99.1	46.8	8	90	90	88
MicI8	1.0	97.8	74.1	99.3	46.8	6	96	89	91
MicI9	1.1	98.0	74.4	99.2	46.9	7	97	92	90
MicI6	0.9	97.7	75.3	99.5	46.9	5	94	93	93
MicI5	0.7	97.6	76.0	99.6	46.9	3	92	94	95
MicI3	0.5	97.4	79.0	99.8	47.2	1	91	95	97
MicI4	0.6	97.6	79.8	99.7	47.4	2	93	96	96
MicI7	0.8	97.8	88.1	99.4	48.5	4	95	97	92
MicPCP4	78.8	3.4	0.6	36.0	48.9	75	9	21	17
JPEG50	73.2	4.3	0.7	48.0	49.2	74	21	24	30
JPEG60	71.2	7.7	1.5	51.7	49.7	73	30	37	33
MicPCP5	82.5	3.9	0.7	34.1	50.4	76	15	26	15
MicPCP7	89.3	4.0	0.6	32.6	53.4	79	17	22	12
MicPCP9	90.8	4.5	0.3	31.4	53.8	82	24	3	9

MicPCP6	90.3	3.4	0.5	33.4	54.0	81	8	17	14
MicZ3	90.2	3.2	0.6	35.7	54.5	80	6	20	16
JPEG35	88.8	3.7	0.4	43.2	55.7	78	10	13	25
JPEG45	87.5	4.2	0.7	46.8	56.1	77	20	25	29
MicZ4	94.9	3.7	0.5	33.0	56.3	84	11	19	13
MicPCP8	96.9	4.0	0.3	32.0	57.0	85	16	6	10
MicPCP10	97.8	4.1	0.5	31.2	57.3	86	19	14	7
JPEG40	91.0	3.7	0.7	45.0	57.3	83	12	23	27
MicZ7	102.9	4.3	0.4	30.5	59.7	87	22	10	6
MicZ6	103.9	3.9	0.4	30.4	60.1	88	13	11	5
MicZ5	104.1	4.0	0.4	31.3	60.4	89	18	12	8
MicZ8	107.9	4.6	0.5	29.6	62.0	90	25	18	4
MicZ10	109.8	4.8	0.4	28.2	62.6	91	26	8	1
MicZ9	110.9	4.4	0.4	28.4	63.1	93	23	9	2
Distr	58.6	43.1	64.9	89.0	65.0	70	71	81	79
JPEG30	110.5	3.0	0.5	41.8	66.1	92	5	15	23
JPEG25	155.2	2.1	0.3	38.8	87.6	94	4	4	20
JPEG20	164.9	1.4	0.3	36.1	91.7	95	3	5	18
JPEG15	202.7	1.1	0.1	32.1	109.5	96	2	1	11
JPEG10	269.4	0.9	0.2	28.4	141.9	97	1	2	3

VI. CONCLUSIONS

19. 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.

Acknowledgments

This work was partly funded by the U.S. Bureau of the Census under contract no OBLIG-2000-29158-0-0. Thanks go to Francesc Sebé for his help in automating the probabilistic record linkage software and running the experiments.

References

- [Adam89] Adam, N. R., Wortmann, J. C., (1989), Security-control methods for statistical databases: a comparative study, *ACM Computing Surveys*, vol. 21(4):515-556.
- [Defa93] 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.
- [Domi02] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002), Practical Data-Oriented Microaggregation for Statistical Disclosure Control, *IEEE Transactions on Knowledge and Data Engineering*, (to appear, March 2002).
- [Jaro89] 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.
- [JPEG] Joint Photographic Experts Group, Standard IS 10918-1 (ITU-T T.81) <http://www.jpeg.org>.

- [Kim86] 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.
- [Liew85] 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.
- [Moor96] Moore, R., (1996), Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (unpublished manuscript).
- [Pagl98] 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.
- [Skin94] Skinner, C., Marsh, C., Openshaw, S., Wymer, C., (1994), Disclosure Control for Census Microdata, *Journal of Official Statistics*, vol. 10:31-51.
- [USBC] U. S. Bureau of the Census, (2000), Record Linkage Software: User Documentation. Available from U. S. Bureau of the Census.
- [Wink98] 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.