Distance-based and probabilistic record linkage for re-identification of records with categorical variables *

Josep Domingo-Ferrer† and Vicenç Torra‡

† Dept. of Computer Science and Mathematics, Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona
‡ Institut d'Investigació en Intel·ligència Artificial,
Campus de Bellaterra, E-08193 Bellaterra
jdomingo@etse.urv.es; vtorra@iiia.csic.es

Abstract

Record linkage methods are methods for identifying the presence of the same individual in different data files (re-identification). This paper studies and compares the two main existing approaches for record linkage: probabilistic and distance-based. The performance of both approaches is compared when data are categorical. To that end, a distance over ordinal and nominal scales is defined. The paper shows that, for categorical data, distance-based and probabilistic-based record linkage lead to similar results. This is parallel to comparisons in the literature for numerical data, which also showed a similar behaviour between both record-linkage approaches. As a consequence, the distance proposed for ordinal and nominal scales is implicitly validated.

Keywords: Record linkage, re-identification, distances over categorical scales.

1 Introduction

Large amounts of data are nowadays at the disposal of public and private institutions, due to the fact that data are easy to capture and store. While in the past only a few records about individuals or households were stored in each circumstance, at present times, large amounts of information are recorded. [1] estimates that the disk storage per person (DSP) expressed in megabytes (MB) is about 472 in 2000 while it was only 28 in 1996 and 0.02 in 1983.

*Work partly supported by the European Commission under project IST-2000-25069 "CASC"

Nonetheless, available data are not centralized but highly distributed. Partially overlapping data can be found under different forms in different databases. Moreover, data are usually non-uniformly standardized (different standardizations are applied pursuant to recoding by institutions) and subject to all kind of errors.

Re-identification procedures are tools developed to detect the presence of the same individual in different data files. Record linkage is a strategy for re-identification which links records in separate files that correspond to the same individual or household. For the sake of simplicity, we consider here only the case of linking two different data files.

The usual assumption in re-identification (*e.g.* [2], [3], [4], [5] - the latter work describes and compares some of the existing approaches) is to consider the presence of a set of common variables in both files. However, re-identification in this case is far from trivial and it is usually not enough to have a matching procedure among pairs of records to establish links between them. This is so due to the presence of errors in the files. As [3] points out, "the normal situation in record linkage is that identifiers in pairs of records that are true matches disagree by small or large amounts and that different combinations of the non-unique, error-filled identifiers need to be used in correctly matching different pairs of records".

An alternative re-identification scenario (see [6], [7]) is to consider files which do not share a set of common variables even if they refer to the same individuals. This is relevant when comparing data files with similar information (*e.g.* financial variables) corresponding to consecutive periods of time (*e.g.* two different years) and referring to almost the

same individuals (*e.g.* companies in a certain region). In this case, re-identification is based on the structural similarities found in the different files. In other words, re-identification is based on the structures that are kept constant through files. For example, [6] use partitions to denote these common structures.

In the case of re-identification assuming common variables, the two most successful re-identification methods are probabilistic record linkage and distance-based record linkage. [9] describes both approaches and includes a comparison for numerical data files. [8] describes a promising method for re-identification based on clustering techniques; the rationale of the latter proposal is similar to the one of [6].

In the case of re-identification without common variables, methods operate in two stages. First, underlying structures are identified in both files (*e.g.* partitions or centroids) and, second, re-identification procedures are applied to these structures.

Two main users of re-identification procedures can be distinguished: 1) companies that want to exploit their distributed information; 2) National Statistical Offices that use re-identification procedures to evaluate the disclosure risk associated to the data they intend to release.

In this work, we study re-identification procedures when files share a set of common variables. Two re-identification procedures for categorical data are considered: probabilistic record linkage and distance-based record linkage. First, distances for categorical data are defined which allow distance-based record linkage to be carried out. Then probabilistic and distance-based record linkage are compared (no comparison for categorical data existed so far in the literature). The comparison is based on extensive experimentation. Experiments consist of applying both record linkage procedures to a set of file pairs. Given a file A, a file B is generated through the application of some particular masking method to the original file A; then record linkage between A and B is performed.

The structure of the rest of this paper is as follows. In Section 2, both record linkage methods for re-identification are described. In Section 3, both methods are analyzed and compared; in particular, this section proposes distances for categorical data and describes the structure of the experiments that were performed. Section 4 contains some conclusions and mentions future work.

2 Re-identification methods

In this section, we review the two main approaches for re-identification between files sharing a set of variables. Let A and B be two files sharing a set of common variables. Both files are defined over the same set of individuals. As it is commonly the case, we cannot assure that the values in both files are the same for the same individuals. In other words, even though variables are the same, values for a particular individual may differ due to errors.

The section begins with the description of probabilistic record linkage. Then, distance-based record linkage is considered. The last part of the section is a discussion of both approaches.

2.1 Probabilistic record linkage

Probabilistic record linkage is described in [10], [11] and [3]. In this section, we outline only some of its elements. See the above mentioned references for details.

Let us consider two files A and B with a single variable V each. Let r_A and r_B be records belonging to files A and B, respectively. Probabilistic record linkage applied to files A and B is based on the computation of an index for each pair (r_A, r_B). Some index thresholds are then used to label the pair as a linked pair (LP), a clerical pair (CP) or a non-linked pair (NP). Equivalently, when the index is larger than, say, *linkThreshold*, the pair is linked; when the index is lower than, say, *nonLinkThreshold*, the pair is non-linked; when the index is between both thresholds the pair is classified as a clerical pair. A clerical pair is one that cannot be automatically classified as linked or non-linked; human inspection is needed to classify it.

Let the values of records r_A and r_B for variable V be a and b, respectively. That is, $r_A = a$ and $r_B = b$. Then, the index $R(a, b)$ is computed as follows:

$$R(a,b) = log(\frac{P(a = b|(a,b) \in \mathbf{M})}{P(a = b|(a,b) \in \mathbf{U})}) \qquad (1)$$

where $\mathbf{M}$ corresponds to the set of *matched pairs* and $\mathbf{U}$ corresponds to the set of *unmatched pairs*. Pairs in $\mathbf{M}$ are those that can be proven to be true matches (the ones that a perfect re-identification method would detect as corresponding to the same individual) and pairs in $\mathbf{U}$ are those that can be proven to be non-related (the ones that a perfect re-identification procedure would not relate).

When a set of variables are considered in both files rather than a single variable, an expression equivalent to Expression (1) is used. In this case, a and b correspond to vectors of values rather than values for a single variable V. It is usually assumed for computing $R(a, b)$ that different variables are statistically independent and thus products of conditional probabilities can be used. Alternative approaches not assuming statistical independence have also been considered in the literature (see [5]).

To use probabilistic record linkage in an effective way, we need to set the thresholds (*e.g.* the values *linkThreshold* and *nonLinkThreshold*) and the conditional probabilities in Expression (1).

The thresholds are usually determined from the probabilities:

$$P(LP|\mathbf{U})$$

$$P(NP|\mathbf{M})$$

In plain words, thresholds are computed from: (i) the probability of linking a pair that is an unmatched pair (a *false positive* or *false linkage*) and (ii) the probability of not linking a pair that is a match pair (a *false negative* or *false unlinkage*).

Conditional probabilities in Expression (1) are usually estimated using the EM algorithm [12].

2.2 Distance-based record linkage

This approach, described in [13] in a very restricted formulation, consists of computating distances between records in the two data files being considered. The method was applied in [13] for disclosure risk assessment. An original data file A was considered together with a distorted version B of the same file. Record linkage was used to find out to what extent distorted records could be re-identified.

In general, for each record in file A, the distance to every record in file B is computed. Then the *nearest* and *second nearest* records in file B are considered. A record in file B is labeled as *linked* when the nearest record in file A turns out to be its corresponding original record (the one that generated the distorted record). A record in file B is labeled as *linked to 2nd nearest* when the second nearest record in file A turns out to be the corresponding original record. In all other cases, records are not linked.

The distance-based approach requires that distances be standardized to avoid scaling problems. Also, an assumption on the weights of variables for computing the distance between a pair of records (equal weight for all variables according to [13]) is required.

2.3 Discussion

Both record linkage methods are focused to find the records in files A and B that correspond to the same individuals. As shown above, both approaches are radically different. The following aspects can be underlined:

- Distance-based record linkage methods are simple to implement and to operate. The main difficulty consists of establishing appropriate distances for the variables under consideration. In particular, distances for categorical variables (in ordinal and nominal scales) are required. On the other hand, distance-based record linkage allows the inclusion of subjective information (about individuals or variables) in the re-identification process.

- Probabilistic record linkage methods are less simple. However, they do not assume rescaling or weighting of variables and require the user to provide only two probabilities as input: the values $P(LP|\mathbf{U})$ and $P(NP|\mathbf{M})$.

For numerical data, it has been proven (see [9]) that both approaches lead to similar re-identification results. For categorical data, no comparison is available in the literature, probably because distances over categorical data are less straightforward than distances over numerical data.

3 Analysis

To compare the re-identification methods described in Section 2, and due to the lack of benchmarks for this purpose, we have performed a set of experiments based on the ones used by National Statistical Offices to evaluate masking procedures and to determine the re-identification risk for a particular data file prior to its publication.

Thus, we have applied re-identification procedures between an original data file and some masked data files obtained through application of several masking methods on the original file. This is consistent with the methodology proposed in [9] for the case of continuous variables.

Several re-identification experiments were performed in order to mitigate the dependency of results on a single dataset. Thus, different sets of

Table 1: Variables used in the analysis.

Variable	Meaning
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	means of transportation to work
WHYMOVE	primary reason for moving
WHYTOH	main reason for choice of house
WHYTON	id. for choosing this neighborhood

Table 2: Groups of variables

Variable	u	l	s	m	o	N. Categ.
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

variables, different masking methods and different method parameterizations were considered. In this section, we detail the experiments and the results obtained so far. We first describing the original data file used (a publicly available data file). Then, we describe how masking methods were applied to obtain different masked data files. We then propose some distances for categorical and explain the re-identification experiments. The last part of this section reports the results obtained.

3.1 Test data collection

Data from the *American Housing Survey 1993* were used (these data can be obtained from the U.S. Census Bureau using the Data Extraction System at http://www.census.gov/DES/www/welcome.html). A set of 10 categorical variables were selected. These variables are displayed in Table 1. To allow a substantial amount of experiments in reasonable time, we took only the first 1000 records from the corresponding data file.

Five groups of variables were defined over the set of selected variables, and the same analysis was performed for each of them. First, three groups were defined by grouping variables with a similar number of categories. Let 's', 'm' and 'l' denote the groups of variables with small, medium and large number of categories, respectively. A fourth group denoted by 'u' was defined that corresponds to the union of the groups 'm' and 'l'; thus, 'u' corresponds to the group of variables with medium or large number of categories. Finally, a fifth group 'o' was defined as the subset of ordered variables (variables that range in an ordinal scale). This latter group was defined after analyzing the meaning of each category in the range of variables. Table 2 presents the variables and the main characteristics of each group; it includes the variables, the number of categories for each variable and the groups to which each variable belongs.

3.2 Generation of file pairs for re-identification

The generation of pairs of files to perform re-identification was achieved by masking the original data file. Each pair was formed by the original data set and one masked version of it. To generate masked versions of the original data, several masking methods were applied to the original 1000-record file containing the 10 variables in Table 1.

Four masking methods were considered and for each one nine different parameterizations were applied. Masking methods were selected among those commonly used by National Statistical Offices to protect data files (see [14] for a detailed survey of masking methods used in different countries). Different parameterizations were taken so that different levels of data protection were experimented with. The consideration of both aspects led to $4 * 9 = 36$ different masked data files.

Masking methods tried in the experiments are listed below (see [15] for a detailed analysis of masking methods and their application by National Statistical Offices):

Top-coding: Categories over a certain threshold are recorded into a given value.

Bottom-coding: Categories below a certain threshold are recoded into a given value.

Global recoding: Several categories are recoded into a new category. This corresponds to defining a new variable with a number of categories smaller than the number of categories in the

original variable. For example, if a variable corresponds to *marital status* and consists of categories *widow*, *divorced*, *married*, etc., we can recode this variable so that *widow* and *divorced* are merged into a single category *widow-or-divorced*.

PRAM: The Post-Randomization Method (PRAM) is based on a Markov matrix P. The matrix contains the probabilities of replacing categories in the original file by other categories. Thus, a probability p_{kl} in P means that a category c_k in the original file is replaced by a category c_l in the masked file with probability p_{kl}.

Note that, from the point of view of re-identification procedures, top-coding reduces the probability of re-identifying an individual with a large value while bottom-coding reduces the probability of re-identifying an individual with a small value. Global recoding would make it more difficult to re-identify individuals because categories become broader. PRAM makes re-identification difficult as categories for an individual can change as a result of masking.

The following parameterizations were considered for the four masking methods described above:

Top-coding: This method was applied to ordinal categorical variables. Given a parameter p, the last p values of the variable were recoded into a new category. We considered p from 1 to 9.

Bottom-coding: This method was also applied to ordinal categorical variables. In this case, given a parameter p, the first p values of the variable were recoded into a new category. Values of p from 1 to 9 were considered.

Global recoding: A recoding scheme was defined based on frequencies of categories. Given a parameter p, the p lowest frequency categories were recoded into a single one. As before, values of p between 1 and 9 were considered.

PRAM: Parameterization of this method requires a Markov Matrix to be specified. The approach described in [16] was chosen for selecting the matrix. Let $T_V = (T_V(1), \cdots, T_V(K))^t$ be the vector of frequencies of the K categories of variable V in the original file (assume without loss of generality that $T_V(k) \geq T_V(K) > 0$ for $k < K$) and let θ be such that $0 < \theta < 1$. Then, the PRAM matrix for variable V is defined as:

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

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

Table 3: Re-identification results for the 's' group of variables using probabilistic record linkage

parameter	Bottom	Global	PRAM	Top
1	1000	1000	966	1000
2	699	1000	921	891
3	577	917	897	749
4	447	730	881	493
5	279	835	843	429
6	161	695	803	458
7	79	355	789	688
8	51	51	759	45
9	51	51	734	188

Application of the four masking methods above with nine different parameterizations per method led to 36 different masked data files.

3.3 Re-identification experiments

For each record linkage method (probabilistic record linkage and distance-based record linkage) and for each pair of *(original-file, masked-file)*, five re-identification experiments were performed. More specifically, each of the five experiments corresponded to one of the five groups of variables 'u', 'l', 's', 'm' and 'o' defined in Table 2. Since there were 36 different file pairs, 36*5=180 re-identification experiments were performed for each record linkage method.

The implementation of probabilistic record linkage used in the experimentation was the U.S. Census Bureau software provided by W. Winkler [17], [5] with some additions. The EM algorithm was used for the estimation of the probabilities.

The implementation of distance-based record linkage was especially written in C for the experimental work reported in this paper. An essential point was to define a distance for categorical variables, which was done as follows:

Table 4: Re-identification results for the 's' group of variables using distance-based record linkage

parameter	Bottom	Global	PRAM	Top
1	502	986	978	853
2	263	861	938	617
3	176	651	916	447
4	101	326	905	200
5	66	103	882	95
6	43	83	828	78
7	36	9	792	42
8	3	3	780	42
9	3	3	753	42

Definition 1 *1. For a nominal variable V, the only permitted operation is comparison for equality. This leads to the following distance definition:*

$$d_V(c, c') = \begin{cases} 0 & \text{if } c = c' \\ 1 & \text{if } c \neq c' \end{cases}$$

where c and c' correspond to categories for variable V.

2. For an ordinal variable V, let $\leq_V$ be the total order operator over the range of V. Then, the distance between categories c and c' is defined as the number of categories between the minimum and the maximum of c and c' divided by the cardinality of the range (denoted by $D(V)$):

$$d_V(c, c') = \frac{|c'' : min(c, c') \leq_V c'' \leq_V max(c, c')|}{|D(V)|}$$

The distance for pairs of records was computed assuming equal weight for all variables.

3.4 Results

Results of both re-identification procedures turn out to be similar. In Tables 3-4, the number of correctly re-identified records is displayed for the 's' group of variables (PRL is the probabilistic record linkage and DBRL corresponds to distance-based one. Similar results are obtained for the other groups of variables (they are not displayed for the sake of space).

The average number of re-identified records per experiment was computed as a measure of similarity between both record linkage methods:

$$\frac{\sum \texttt{number of correct re} - \texttt{identifications}}{\texttt{number of experiments}}$$

For distance-based record linkage, an average number of 593.06 re-identified records (over 1000) was obtained. For probabilistic record linkage, the average was 579.32 re-identified records. Thus, the performance of both methods is similar.

Additional analysis have been carried to assess the performance of both record linkage methods. In particular, correlation statistics have been computed between the number of correctly re-identified records and some standard information loss measures. Information loss measures are used by National Statistical Offices to reflects how much harm is being inflicted to the data by a given masking method; the amount of information loss measured in this generic way should roughly correspond to the amount of information loss for a reasonable range of data uses.

In particular, we have considered the following information loss measures (see [18] and [15] for details):

Direct comparison of categorical values: Using the distance in Definition 1, we compare the original records and the masked ones. We will denote this information loss measure by Dist.

Comparison of contingency tables: Given a subset of variables, contingency tables are computed for both the original and the masked data. The number of differences between both contingency tables is denoted by CTBIL (Contingency Table Based Information Loss measure). An alternative measure ACTBIL (Average CTBIL) is defined (dividing CTBIL by the number of cells) so that the number of cells does not affect the information loss.

Entropy-based measures: After [16] Shannon's entropy is used to measure information loss. 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 data set in the event of it being transmitted over a noisy channel. We use EBIL to denote this information loss measure. As this measure only depends on the masked data set and it does not account for its relation with the

Table 5: Correlations between two alternative approaches for record linkage and the information loss measures

Inf. Loss	PRL	DBRL
Dist	-0.846	-0.911
CTBIL	-0.822	-0.924
ACTBIL	-0.706	-0.823
EBILRF	-0.833	-0.842
ILRF	-0.908	-0.878
EBILMF	-0.910	-0.892
ILMF	-0.909	-0.889

original data, a new information loss measure was defined in [18]. We use IL to denote this alternative measure.

We have used two different approaches to estimate the probabilities of EBIL and IL: the same file being masked (in this case the measures are denoted by EBILMF and ILMF) and a reference file (in this case, we denote the measures by EBILRF and ILRF).

Correlation between the two approaches for record linkage (distance-based and probabilistic) and the information loss measures are given in Table 5. This table shows that correlations are all quite similar and around -0.9 (note that the value is negative because the larger the information loss, the more difficult the re-identification). The results also show that for some information loss larger correlations are obtained with probabilistic record linkage (ILRF, EBILMF, ILMF) while for others larger correlations are obtained with distance-based record linkage (Dist, CTBIL, ACTBIL and EBILRF). Therefore, in relation to information loss both re-identification methods are also similar.

In the experiments reported in this paper, no information is fed to record linkage procedures about the masking method applied to protect the original file. In fact, this is not the usual case in disclosure risk assessment. It can be proven that, if a distance is used which takes into account the masking method applied, the distance-based record linkage largely improves its results. A simple way for a distance to take masking into account is as follows: assign a distance *infinity* when category c cannot be recoded as c' using the current masking method. In this case, the average number of correctly re-identified records increases to 663.49, which should be compared to 593.06 for the original distance-based record linkage and to 579.32 for probabilistic record linkage.

4 Conclusions and future work

Two record linkage methods for re-identification of categorical microdata have been studied in this paper: probabilistic and distance-based. Since distance-based record linkage only existed in the literature for numerical data, a distance for categorical data has been defined to extend this kind of linkage to categorical data. We have shown that the number of re-identifications is similar for both record linkage procedures, but that the re-identified individuals/records are not the same. This is consistent with existing comparisons of both record linkage methods for numerical data. Beyond implicit validation of the proposed distance for categorical data, results in this paper show that both methods are complementary rather than antagonistic and best results are obtained if they are combined.

It has also been pointed out that distance-based record linkage can substantially improve (and thus outperform probabilistic record linkage) if information about the masking method is embedded into the distance function. It is important to note that the knowledge about the masking method does not uniquely determine the links between original and masked categories but only permit to modify conditional probabilities (e.g. some combinations are known to be impossible) and thus increase the performance of the system. This approach is relevant as NSOs usually describe the used masking methods.

Future refinements of distance-based record linkage is to give a different weight to each variable when computing the distance. This would be problem-dependent and would require a learning mechanism to adjust weights beforehand.

References

[1] L. Sweeney, "Information explosion", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz, Elsevier, 43–74, 2001.

[2] H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, "Automatic linkage of vital records", *Science*, vol. 130, 954–959, 1959.

[3] W. E. Winkler, "Matching and record linkage", in *Business Survey Methods*,ed. B. G. Cox, Wiley, 355–384, 1995.

[4] J. F. Robinson-Cox,"A record-linkage approach to imputation of missing data: analyzing tag retention in a tag-recapture experiment", *Journal of Agricultural, Biological and Environmental Statistics*, vol. 3, 48–61, 1998.

[5] W. E. Winkler, "Advanced methods for record linkage", *Proc. of the American Statistical Assoc. Section on Survey Research Methods*, 467–472, 1995.

[6] V. Torra, "Towards the re-identification of individuals in data files with non-common variables", in *Proceedings of ECAI'2000*, 326–330.

[7] V. Torra, "Re-identifying individuals using OWA operators", *Proceedings of the 6th Intl. Conference on Soft Computing*, Iizuka, Fukuoka, Japan, 2000.

[8] J. Bacher, S. Bender and R. Brand, "Empirical re-identification – Evaluation of a simple clustering technique", *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, forthcomming.

[9] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes, L. M. Zayatz, Elsevier, 111–133, 2001.

[10] I. P. Fellegi and A. B. Sunter, "A theory of record linkage", *Journal of the American Statistical Association*, vol. 64, 1183–1210, 1969.

[11] M. A. Jaro, "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, 1989.

[12] A. P. Dempster, N. N. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, vol. 39, 1–38, 1977.

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

[14] F. Felsö, J. Theeuwes and G. G. Wagner, "Disclosure limitation methods in use: results of a survey", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz, Elsevier, 17–42, 2001.

[15] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, Springer-Verlag, 2001.

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

[17] U. S. Bureau of the Census, *Record Linkage Software: User Documentation*. Available from U. S. Bureau of the Census, 2000.

[18] Domingo-Ferrer, J., Torra, V., (2001), Disclosure Control Methods and Information Loss for Microdata, 91-110, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.