1

Information Loss in Continuous Hybrid Microdata: Subdomain-Level Probabilistic Measures

Josep Domingo-Ferrer, Josep Maria Mateo-Sanz and Francesc Sebé

Rovira i Virgili University of Tarragona,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
e-mail {josep.domingo,josepmaria.mateo,francesc.sebe}@urv.net

Summary. The goal of privacy protection in statistical databases is to balance the social right to know and the individual right to privacy. When microdata (*i.e.* data on individual respondents) are released, they should stay analytically useful but should be protected so that it cannot be decided whether a published record matches a specific individual. However, there is some uncertainty in the assessment of data utility, since the specific data uses of the released data cannot always be anticipated by the data protector. Also, there is uncertainty in assessing disclosure risk, because the data protector cannot foresee what will be the information context of potential intruders. Generating synthetic microdata is an alternative to the usual approach based on distorting the original data. The main advantage is that no original data are released, so no disclosure can happen. However, subdomains (*i.e.* subsets of records) of synthetic datasets do not resemble the corresponding subdomains of the original dataset. Hybrid microdata mixing original and synthetic microdata overcome this lack of analytical validity. We present a fast method for generating numerical hybrid microdata in a way that preserves attribute means, variances and covariances, as well as (to some extent) record similarity and subdomain analyses. We also overcome the uncertainty in assessing data utility by using newly defined probabilistic information loss measures.
Keywords: Privacy in statistical databases, Hybrid microdata generation, Probabilistic information loss measures.

1.1 Introduction

Statistical databases come in two flavors: tabular data (aggregated data) and microdata (records on individual persons or entities). Microdata can be continuous, *e.g.* salary or weight, or categorical, for instance sex, hair color or instruction level. Releasing microdata, or any statistical data for that matter, must face the tension between respondent privacy and data utility. In the case of microdata, providing respondent privacy means that an intruder should be unable to make a decision whether a published record corresponds to a specific

respondent. On the other hand, providing data utility means that the published set of data should preserve as many statistical properties as possible from the original set.

However, assessing respondent privacy and data utility is inherently uncertain:

- The privacy of respondents is inversely proportional to the risk of disclosure of their responses, *i.e.* the risk that an intruder can link specific released data to specific respondents. But the ability of an intruder to do so depends on how many and how good are the external identified information sources he can gather (*e.g.* censuses, yearbooks, phonebooks, etc.). Unfortunately, the data protector cannot exactly anticipate how much external information intruders will be able to link to the data being released in anonymized form.
- The utility of the released data depends on the specific data uses it has to serve. But the data protector is very often unable to foresee how the released data will be used by legitimate users.

One possibility for protecting a microdata set is to use a *masking method* (*e.g.* additive noise, microaggregation, etc., cf. [3]) to transform original data into protected, publishable data. An alternative to masking the original data is to generate a a *synthetic* data set not from the original data, but from a set of random values that are adjusted in order to fulfill certain statistical requirements [14, 11, 13, 12]. A third possibility is to build a *hybrid* data set as a mixture of the masked original values and a synthetic data set [6, 1, 2].

The advantage of synthetic data over masking is that deciding whether a synthetic data record corresponds to a particular respondent does not make sense any more. So the problem of assessing the privacy of respondents is circumvented. The drawback of synthetic data is that they have limited data utility: they only preserve the statistics that were taken into account by the data protector when generating the synthetic dataset. This raises an interesting question: why not directly publish the statistics that should be preserved rather than a synthetic dataset preserving them?

For synthetic microdata to be really useful, they ought to preserve record-level similarity to some extent, so that subdomains of the protected dataset still yield acceptably accurate analyses. Hybrid microdata are better than purely synthetic microdata at preserving record-level similarity.

1.1.1 Contribution and plan of this paper

In this paper, a method for generating continuous hybrid microdata is proposed such that:

- It is non-iterative and fast, because its running time grows *linearly* with the number of records in the original dataset;

- It *exactly* reproduces the means and the covariance matrix of the original dataset, which implies that variances and Pearson correlations are also exactly preserved;
- It allows preservation of record-level similarity to some extent, so that subdomains of the protected dataset can still be analytically valid (this is especially true if the relevant subdomains can be decided before applying the method).

When assessing the utility of the data output by the proposed method, we propose to deal with the uncertainty about data uses using probabilistic information measures.

Section 1.2 describes our proposal for generating hybrid data. Section 1.3 deals with the complexity and the data utility properties of the proposed method. Probabilistic information loss measures are described in Section 1.4. Empirical results are reported in Section 1.5. Finally, Section 1.6 contains some conclusions and suggestions for future research.

1.2 A low-cost method for hybrid microdata generation

Let X be an original microdata set, with n records and m attributes. Let X'' be a hybrid microdata set to be generated, also with n records and m attributes. In fact, both X and X'' can be viewed as $n \times m$ matrices. The method presented is a hybrid evolution of the synthetic data generation method in [8]. Like [8], it exactly preserves both univariate and multivariate statistical properties of X, such as means, covariances and correlations. As we show below, the improvement with respect to [8] is that, since X'' is hybrid rather than synthetic, a fair amount of record-level similarity between X and X'' is preserved, which allows for subdomain analysis. The algorithm below constructs X'' from X:

Algorithm 1 (Basic procedure)

1. *Use a masking method to transform the original dataset X into a masked dataset X'.*
2. *Compute the covariance matrix C of the original microdata matrix X.*
3. *Use Cholesky's decomposition on C to obtain*

$$C = U^t \times U$$

where U is an upper triangular matrix and U^t is the transposed version of U.

4. *Compute an $n \times m$ matrix A as*

$$A := X' \cdot U^{-1}$$

5. *Use Algorithm 2 to modify matrix A so that its covariance matrix is the $m \times m$ identity matrix.*
6. *Obtain the hybrid microdata set as*

$$X'' = A \cdot U$$

By construction, the covariance matrix of X'' equals the covariance matrix of X (see [15]).

7. *Due to the construction of matrix A, the mean of each attribute in X'' is 0. In order to preserve the mean of attributes in X, a last adjustment is performed. If $\bar{x}_j$ is the mean of the j-th attribute in X, then $\bar{x}_j$ is added to the j-th column (attribute) of X'':*

$$x''_{ij} := x''_{ij} + \bar{x}_j \ \text{ for } i = 1, \cdots, n \text{ and } j = 1, \cdots, m \tag{1.1}$$

Note that modification of matrix A is needed for two reasons:

- To preserve covariances;
- To prevent the hybrid data X'' from being exactly the masked data X'.

We now need to specify how to modify an $n \times m$ matrix A so that its covariance matrix is the $m \times m$ identity matrix.

Algorithm 2 (Modification of matrix A)

1. *Given an $n \times m$ matrix A with elements $a_{i,j}$, view the m columns of A as samples of attributes $A_1, \cdots, A_m$. If $Cov(A_j, A_{j'})$ is the covariance between attributes A_j and $A_{j'}$, the objective of the algorithm is that*

$$Cov(A_j, A_{j'}) = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise} \end{cases}$$

for $j, j' \in \{1, \cdots, m\}$.

2. *Let $\bar{a}_1$ be the mean of A_1. Let us adjust A_1 as follows:*

$$a_{i,1} := a_{i,1} - \bar{a}_1 \ \ i = 1, \ldots, n$$

The mean of the adjusted A_1 is 0.

3. *In order to reach the desired identity covariance matrix, some values of attributes $A_2, \cdots, A_m$ must change. For $v = 2$ to m do:*
   a) *Let $\bar{a}_v$ be the mean of attribute A_v.*
   b) *For $j = 1$ to $v - 1$, the covariance between attributes A_j and A_v is*

$$Cov(A_j, A_v) = \frac{\sum_{i=1}^{n} a_{i,j} \cdot a_{i,v}}{n} - 0 \cdot \bar{a}_v = \frac{\sum_{i=1}^{n} a_{i,j} \cdot a_{i,v}}{n}$$

c) In order to obtain $Cov(A_j, A_v) = 0, \quad j = 1 \dots v-1$ *, some elements* $a_{i,v}$ *in the v-th column of A are assigned a new value. Let* $x_1, \dots, x_{v-1}$ *the unknowns for the following linear system of* $v-1$ *equations:*

$$\frac{\sum_{i=1}^{n-v+1} a_{i,j} \cdot a_{i,v} + \sum_{i=1}^{v-1} a_{n-v+1+i,j} \cdot x_i}{n} = 0 \ \textit{for } j = 1 \dots v-1$$

that is

$$\sum_{i=1}^{n-v+1} a_{i,j} \cdot a_{i,v} + \sum_{i=1}^{v-1} a_{n-v+1+i,j} \cdot x_i = 0 \ \textit{for } j = 1 \dots v-1$$

Once the aforementioned linear system is solved, the new values are assigned:

$$a_{n-v+1+i,v} := x_i \ \textit{for } i = 1 \dots v-1$$

d) Let $\bar{a}_v$ *be the mean of attribute* A_v*. A final adjustment on* A_v *is performed to make its mean 0:*

$$a_{i,v} = a_{i,v} - \bar{a}_v \ \textit{for } i = 1 \dots n$$

4. In the last step, values in A are adjusted in order to reach $Cov(A_j, A_j) = 1$ *for* $j = 1 \dots m$*. If* σ_j *is the standard deviation of attribute* A_j*, the adjustment is computed as:*

$$a_{i,j} := \frac{a_{i,j}}{\sigma_j}, \quad i = 1 \dots n, j = 1 \dots m$$

1.3 Properties of the proposed scheme

1.3.1 Performance and complexity

The computational complexity for the proposed method will next be estimated; we exclude the masking step, because it is not part of the method itself. Let n be the number of records and m the number of attributes. Then the complexities of the various operations are as follows:

- Calculation of the covariance matrix: $\mathcal{O}(n + m^2)$;
- Cholesky's decomposition: $\mathcal{O}(m^3/6)$ (see [10]);
- Inversion of the triangular matrix U: $\mathcal{O}(m^2/2)$;
- Calculation of A: $\mathcal{O}(2nm + 2m^3 + 2m^4/3)$ where the term $2m^4/3$ is the cost of solving a Gauss system m times [10];
- Matrix product: $\mathcal{O}(nm^2)$;
- Mean adjustment: $\mathcal{O}(nm)$.

In summary, the overall complexity is $\mathcal{O}(nm + 2m^4/3) = \mathcal{O}(n + m^4)$. To understand this complexity, one should realize that, in general, the number of records n is much larger than the number of attributes m, *i.e.* $n >> m$. Thus this proposal maintains the strong point of [8] that *its complexity is linear in the number of records.*

The method has been tested with several data set sizes and execution times to generate X'' from X' (that is, excluding the running time for masking X into X') are shown in Table 1.1.

Table 1.1. Running time (in seconds) on a 1.7 GHz desktop Intel PC under a Linux OS. Note that time for random matrix generation is included

	Number of attributes m			
Number of records n	5	10	25	50
1,000	0.00	0.01	0.06	0.32
10,000	0.06	0.20	1.28	5.33
100,000	0.50	1.95	12.43	51.19

1.3.2 Data utility

As stated above, the proposed scheme exactly reproduces the statistical properties of the original data set.

- The means of attributes in the original data set X are exactly preserved in the synthetic data set X''.
- The covariance matrix of X is exactly preserved in X'' (see [15]). Thus, in particular:
  - The variance of each attribute in X is preserved in X'';
  - The Pearson correlation coefficient matrix of X is also exactly preserved in X'', because correlations are obtained from the covariance matrix.

The difference between this method and the one in [8] is that record-level similarity is preserved to some extent, as shown in Section 1.5 below on empirical results. Unlike [8], which used a random matrix A, the method in our paper uses a matrix A which is derived from a masked version X' of the original data X.

1.4 A generic information loss measure

To measure information loss, we assume that the original dataset X is a population and the hybrid dataset X'' is a sample from the population. Given a population parameter θ on X, we can compute the corresponding sample statistic $\hat{\Theta}$ on X''. Let us assume that $\hat{\theta}$ is the value taken by $\hat{\Theta}$ in a specific

instance of sample X''. The more different is $\hat{\theta}$ from θ, the more information is lost when publishing the sample X'' instead of the population X. We show next how to express that loss of information through probability.

If the sample size n is large, the distribution of $\hat{\Theta}$ tends to normality with mean θ and variance $Var(\hat{\Theta})$. According to [5], values of n greater than 100 are often large enough for normality of all sample statistics to be acceptable. Fortunately, most protected datasets released in official statistics consist of $n > 100$ records, so that assuming normality is safe. Thus, the standardized sample discrepancy

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{Var(\hat{\Theta})}}$$

can be assumed to follow a $N(0, 1)$ distribution.

Therefore, in [7] we defined a generic probabilistic information loss measure $pil(\theta)$ referred to parameter θ as the probability that the absolute value of the discrepancy Z is less than or equal to the actual discrepancy we have got in our specific sample X', that is

$$pil(\hat{\Theta}) = 2 \cdot P(0 \leq Z \leq \frac{|\hat{\theta} - \theta|}{\sqrt{Var(\hat{\Theta})}}) \tag{1.2}$$

Being a probability, the above measure is bounded in the interval $[0, 1]$, which facilitates comparison and tradeoff against disclosure risk (which is also bounded). This is a major advantage over previous non-probabilistic information loss measures [4, 16], which are unbounded.

1.5 Empirical work

1.5.1 Information loss and disclosure risk measures

To assess data utility, we will compute the above generic probabilistic information loss measure for specific population parameters θ and sample statistics $\hat{\Theta}$: quantiles, means, variances, covariances and Pearson correlations. We write below PIL rather than pil when the probabilistic measure has been averaged over all attributes or pairs of attributes, rather than being computed for a single attribute or attribute pair. Also, the measures below have been multiplied by 100 to scale them within $[0, 100]$ rather than within $[0, 1]$.

1. $PIL(Q)$ is 100 times the average $pil(Q_q)$ for all attributes and quantiles Q_q from $q = 5\%$ to $q = 95\%$ in 5% increments over all attributes; this is the average impact on quantiles;
2. $PIL(m_1^0)$ is the average impact on means over all attributes;
3. $PIL(m_2)$ is the average impact on variances over all attributes;
4. $PIL(m_{11})$ is the average impact on covariances over all attribute pairs;

5. $PIL(r)$ is the average impact on Pearson's correlation coefficients over all attribute pairs.

Regarding disclosure risk, it is measured using the following three measures defined in [3, 9], which also take values in $[0, 100]$ as follows:

- DLD (Distance-based Linkage Disclosure) is the average of $DLD-1$ to $DLD-10$, where $DLD-i$ is the percentage of correctly linked pairs of original-hybrid records using distance-based record linkage through i attributes;
- RID (Rank Interval Disclosure) is the average of $RID-1$ to $RID-10$, where $RID-i$ is the percentage of original values of the i-th attribute that fall within a rank interval centered around their corresponding hybrid value with width $p\%$ of the total number of records;
- $SDID$ (Standard Deviation Interval Disclosure) is analogous to RID but using intervals whose width is a percentage of each attribute's standard deviation.

1.5.2 The data set

The microdata set X used for testing was obtained from the U.S. Energy Information Authority and contains 4092 records [1]. Initially, the data file contained 15 attributes from which the first 5 were removed as they corresponded to identifiers. We have worked with the attributes: RESREVENUE, RESSALES, COMREVENUE, COMSALES, INDREVENUE, INDSALES, OTHREVENUE, OTHRSALES, TOTREVENUE, TOTSALES. This dataset was also used in [9].

1.5.3 The results

Results on the overall dataset

Three different masking methods have been used in Step 1 of Algorithm 1: microaggregation with parameter 3 to obtain $X'(mic)$, rank swapping with parameter 7 to obtain $X'(sw)$ and random noise with parameter 0.16 to obtain $X'(rn)$ (see [3] for details on those methods). Call $X''(mic)$, $X''(sw)$ and $X''(rn)$ the hybrid datasets obtained with Algorithm 1 using $X'(mic)$, $X'(sw)$ and $X'(rn)$, respectively. By construction, the hybrid datasets preserve means, variances, covariances and Pearson correlations of the original dataset X. Table 1.2 gives the above information loss measures for each masked and hybrid dataset.

It can be seen from Table 1.2 that X'' have higher $PIL(Q)$ than X' for all three masking methods (only slightly higher for random noise); however, X'' have substantially lower values than X' for the remaining information loss

[1] `http://www.eia.doe.gov/cneaf/electricity/page/eia826.html`

Table 1.2. Information loss and disclosure risk measures for the overall masked and hybrid datasets

	$PIL(Q)$	$PIL(m_1^0)$	$PIL(m^2)$	$PIL(m_{11})$	$PIL(r)$	DLD	RID	$SDID$
$X'(mic)$	5.3	0	6.6	2.0	27.0	19.3	93.0	84.5
$X''(mic)$	49.4	0	0	0	0	2.0	41.1	41.4
$X'(sw)$	0	0	0	99.9	100.0	3.8	71.3	62.7
$X''(sw)$	53.3	0	0	0	0	0.1	31.7	32.7
$X'(rn)$	52.6	5.2	29.2	3.4	66.5	7.7	39.1	26.6
$X''(rn)$	62.7	0	0	0	0	0.3	28.9	20.5

Table 1.3. Top-down information loss and disclosure risk measures (induced in subdomains)

	d	$PIL(Q)$	$PIL(m_1^0)$	$PIL(m^2)$	$PIL(m_{11})$	$PIL(r)$	DLD	RID	$SDID$
$X''(mic)$	2	54.2	78.0	72.1	71.5	96.5	2.2	32.2	28.8
$X''(mic)$	3	56.3	84.1	80.5	86.4	97.5	2.5	28.2	23.1
$X''(mic)$	4	57.9	81.6	84.7	90.3	96.9	2.6	25.7	20.5
$X''(mic)$	5	58.8	87.2	86.2	93.5	96.1	2.8	24.4	19.2
$X''(sw)$	2	57.9	85.9	80.5	75.3	95.5	0.2	20.5	20.6
$X''(sw)$	3	59.5	88.1	86.9	89.9	96.8	0.3	16.7	16.9
$X''(sw)$	4	60.6	89.0	92.1	94.1	97.0	0.3	14.6	15.5
$X''(sw)$	5	62.5	92.4	94.7	96.2	96.0	0.4	13.9	14.5
$X''(rn)$	2	63.9	88.1	79.3	77.0	95.2	0.4	21.4	15.8
$X''(rn)$	3	67.3	92.0	86.9	88.3	96.0	0.5	18.2	14.0
$X''(rn)$	4	66.9	85.9	92.8	93.3	95.5	0.5	16.6	12.8
$X''(rn)$	5	70.2	91.1	94.3	95.5	96.2	0.5	15.9	12.3

measures. Disclosure risk measures are also much lower for X'' than for X' for all masking methods. Thus, all in all, the hybrid datasets X'' are much better than the masked X.

Top-down generation: posterior subdomains

The next question is how does the proposed hybrid data generation method behave for subdomains that were not predictable at the moment of generating the hybrid data (posterior subdomains). This is relevant if the user is interested in a subset of records selected according to the values of a (maybe external) categorical attribute (*e.g.* the records corresponding to women).

To answer that question, X and X'' have been partitioned into d corresponding subdomains, for $d = 2, 3, 4, 5$. The various information loss measures have been computed between each subdomain in X and its corresponding subdomain in X''. Table 1.3 gives the average results, *i.e.* for each d the measures averaged over the d pairs of $X - X''$ subdomains.

At least two things can be observed from Table 1.3, regardless of the masking method used:

Table 1.4. Bottom-up information loss and disclosure risk measures

	d	$PIL(Q)$	$PIL(m_1^0)$	$PIL(m^2)$	$PIL(m_{11})$	$PIL(r)$	DLD	RID	$SDID$
$X''(mic)$	2	45.0	0	0	0	0	2.5	36.6	22.1
$X_t''(mic)$	2	37.8	0	0	0	0	2.3	50.0	54.6
$X''(mic)$	3	45.5	0	0	0	0	5.6	38.8	21.7
$X_t''(mic)$	3	37.3	0	0	0	0	5.1	57.3	56.9
$X''(mic)$	4	44.5	0	0	0	0	6.4	37.3	18.7
$X_t''(mic)$	4	37.5	0	0	0	0	5.3	58.7	58.5
$X''(mic)$	5	47.0	0	0	0	0	5.4	34.6	15.1
$X_t''(mic)$	5	37.2	0	0	0	0	4.2	61.1	59.4
$X''(sw)$	2	50.9	0	0	0	0	0.3	23.1	12.7
$X_t''(sw)$	2	46.5	0	0	0	0	0.2	37.7	46.8
$X''(sw)$	3	54.0	0	0	0	0	0.4	20.6	9.6
$X_t''(sw)$	3	45.0	0	0	0	0	0.3	41.5	48.1
$X''(sw)$	4	54.4	0	0	0	0	0.5	17.9	7.3
$X_t''(sw)$	4	41.5	0	0	0	0	0.3	42.8	49.7
$X''(sw)$	5	53.5	0	0	0	0	0.5	17.7	6.8
$X_t''(sw)$	5	40.5	0	0	0	0	0.4	45.2	51.0
$X''(rn)$	2	53.8	0	0	0	0	0.4	21.1	11.1
$X_t''(rn)$	2	47.3	0	0	0	0	0.4	32.1	40.6
$X''(rn)$	3	57.0	0	0	0	0	0.5	18.4	8.5
$X_t''(rn)$	3	45.2	0	0	0	0	0.6	34.6	42.0
$X''(rn)$	4	55.6	0	0	0	0	0.7	16.5	7.4
$X_t''(rn)$	4	42.6	0	0	0	0	0.6	37.0	46.1
$X''(rn)$	5	55.4	0	0	0	0	0.6	15.4	6.3
$X_t''(rn)$	5	42.8	0	0	0	0	0.5	38.7	47.0

- Means, variances, covariances and Pearson correlations of the original dataset are *not* preserved in the subdomains; further, the discrepancy of these statistics between the subdomains in X and X'' is remarkable;
- The more subdomains are made, the larger the information loss and the smaller the interval disclosure risks RID and $SDID$.

Bottom-up generation: prior subdomains

We now turn to the performance of the proposed hybrid data generation when directly applied to subdomains that can be decided *a priori*. By construction, means, variances, covariances and Pearson correlations of subdomains of X are preserved in subdomains of X''. The interesting point is that the overall X_t'' obtained as union of d hybrid subdomains also preserves means, variances, covariances and Pearson correlations of X. Table 1.4 shows average information loss and disclosure risk measures for subdomains in a way analogous to Table 1.3; in addition, for each partition into d subdomains, it gives the measures between the overall X and X_t''.

We can see from Table 1.4 that, regardless of the masking method used:

- When assembling the hybrid subdomains into X_t'', $PIL(Q)$ and DLD decrease (even if the latter is already very low). At the same time, the interval disclosure measures RID and $SDID$ increase substantially;
- As the number d of subdomains increases, interval disclosure measures decrease for each subdomain, but they increase for the overall dataset X_t''.

1.6 Conclusions and future research

In this paper, we have presented a new method for generating numerical hybrid microdata, based on combining a masking method with Cholesky's decomposition. Excluding masking computation (dependent on the particular masking method used), the method is very fast, because its running time is linear in the number of records. The method preserves a fair amount of record-level similarity, which allows for subdomain analysis. The best results are obtained when the relevant subdomains can be decided in advance by the data protector: in that case, the method can be applied independently for each subdomain and the information loss and disclosure risk measures for the overall dataset are still very good (bottom-up approach). If the relevant subdomains cannot be anticipated, then the only option is to apply the method to the overall dataset and hope for the best when subdomains are analyzed (top-down approach); in this case, the method does not guarantee intra-subdomain preservation of any statistics, but it distorts those statistics less than sheer synthetic data generation.

A first line for future research is to extend the notion of subdomain from a subset of records to a subset of records *and* attributes. The idea is to evaluate the bottom-up and the top-down approaches for subdomains comprising only a subset of attributes for each record.

A second line for future research is to give some guidance to help the data protector anticipate the relevant subdomains in the bottom-up approach (*e.g.* use categorical variables to partition records in subdomains, etc.).

Acknowledgments

The authors are partly supported by the Spanish Ministry of Science and Education through project SEG2004-04352-C04-01 "PROPRIETAS", by the Government of Catalonia under grant 2002 SGR 00170 and by Cornell University under contract no. 47632-10043.

References

1. J. M. Abowd and S. D. Woodcock (2004) Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and

V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 290–297, Berlin Heidelberg: Springer.

2. R. Dandekar, J. Domingo-Ferrer, and F. Sebé (2002) LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *LNCS*, pages 153–162, Berlin Heidelberg: Springer.
3. J. Domingo-Ferrer and V. Torra (2001) Disclosure protection methods and information loss for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 91–110, Amsterdam: North-Holland.
4. J. Domingo-Ferrer and V. Torra (2001) A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134, Amsterdam: North-Holland.
5. M. G. Kendall, A. Stuart, S. F. Arnold J. K. Ord, and A. O'Hagan (1994) *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory (6th Edition)*. London: Arnold.
6. A. B. Kennickell (1999) Multiple imputation and disclosure protection: the case of the 1995 survey of consumer finances. In J. Domingo-Ferrer, editor, *Statistical Data Protection*, pages 248–267, Luxemburg: Office for Official Publications of the European Communities.
7. J. M. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé (2005) Probabilistic information loss measures for continuous microdata. *Data Mining and Knowledge Discovery*, to appear.
8. J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer (2004) Fast generation of accurate synthetic microdata. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 298–306, Berlin Heidelberg: Springer.
9. J. M. Mateo-Sanz, F. Sebé, and J. Domingo-Ferrer (2004) Outlier protection in continuous microdata masking. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 201–215, Berlin Heidelberg: Springer.
10. W. Press, W. T. Teukolsky, S. A. Vetterling, and B. Flannery (1993) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.
11. T. J. Raghunathan, J. P. Reiter, and D. Rubin (2003) Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16.
12. J. P. Reiter (2005) Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 131(2):365–377.
13. J. P. Reiter (2005) Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 168:185–205.
14. D. B. Rubin (1993) Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468.
15. E. M. Scheuer and D. S. Stoller (1962) On the generation of normal random vectors. *Technometrics*, 4:278–281.

16. W. E. Yancey, W. E. Winkler, and R. H. Creecy (2002) Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *LNCS*, pages 135–152, Berlin Heidelberg: Springer.