LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection*

Ramesh A. Dandekar[1], Josep Domingo-Ferrer[2], and Francesc Sebé[2]

[1] Energy Information Administration, U. S. Department of Energy
1000 Independence Ave, Washington DC 20585, USA
ramesh.dandekar@eia.doe.gov
[2] Universitat Rovira i Virgili,Department of Computer Science and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
{jdomingo,fsebe}@etse.urv.es

Abstract. In previous work by Domingo-Ferrer *et al.*, rank swapping and multivariate microaggregation has been identified as well-performing masking methods for microdata protection. Recently, Dandekar et al. proposed using synthetic microdata, as an option, in place of original data by using Latin hypercube sampling (LHS) technique. The LHS method focuses on mimicking univariate as well as multivariate statistical characteristics of original data. The LHS-based synthetic data does not allow one to one comparison with original data. This prevents estimating the overall information loss by using current measures. In this paper we utilize unique features of LHS method to create hybrid data sets and evaluate their performance relative to rank swapping and multivariate microaggregation using generalized information loss and disclosure risk measures.

Keywords: Microdata masking, synthetic microdata generation, rank swapping, microaggregation.

1 Introduction

Statistical Disclosure Control (SDC) methods are used in official statistics to ensure confidentiality in statistical databases being released for public use[12]. If the database contains microdata (*i.e.* individual respondent records), confidentiality means trying to avoid disclosing the identity of the individual respondent associated with a published record. At the same time, SDC should preserve the informational content as much as possible. SDC methods are somewhere between encryption of the original data set (no disclosure risk but no informational content released) and straightforward release of the original dataset (no confidentiality but maximal informational content released). SDC methods for microdata are also known as masking methods.

* The second and third authors are partly supported by the European Commission under project IST-2000-25069 "CASC".

J. Domingo-Ferrer (Ed.): Inference Control in Statistical Databases, LNCS 2316, pp. 153–162, 2002.

In [4,5], a comparison of masking methods for microdata was conducted. For numerical microdata, two masking methods were determined to be well-performing in that they achieve a good tradeoff between low disclosure risk and information loss. The first method is rank swapping [11] and the second method is multivariate microaggregation [6]. Both methods will be termed "natural" because they obtain a number of masked data records through transformation of the same number of original data records.

Dandekar *et al.* have proposed synthetic microdata generation as an option to natural masking methods [2]. Specifically, Latin hypercube sampling (LHS) has been used to obtain a synthetic masked data set with statistical properties similar to the original data set; masked data being synthetic, the number of masked data records does not need to be the same as the number of original data records.

We generate hybrid data by combining synthetic data with original sampled data. We then compare the performance of natural and hybrid masking by using several information loss and disclosure risk metrics. Section 2 recalls the essentials of rank swapping and multivariate microaggregation. Section 3 summarizes LHS synthetic data generation and describes several hybrid approaches to obtain masked data as a mixture of synthetic and original data. Section 4 discusses the metrics used to compare the three methods. Computational results are reported in Section 5. Finally, Section 6 is a conclusion.

2 Natural Masking Methods

We will summarize here the principles of rank swapping and multivariate microaggregation.

Rank swapping was originally described for ordinal variables, but it can actually be used for any numerical variable [11]. 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). In [4,5], values p from 1 to 20 were considered in experimentation.

The basic idea of microaggregation is to cluster records into small aggregates or groups of size at least k [3,6]. 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 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 18 have been considered. According to the experimental work [4,5], the best microaggregation variant is microaggregation on unprojected multivariate data considering three or four variables at a time.

3 Synthetic and Hybrid Masking Methods

Generation of synthetic data which preserves some of the characteristics of original data has been proposed as an alternative to natural masking methods. Multiple imputation [9] is one sophisticated approach to synthetic data generation which requires specific software not usually available. LHS synthetic data generation discussed below is a simpler approach which also yields satisfactory results.

3.1 LHS Synthetic Masking

Dandekar *et al.* [2] use Latin hypercube sampling (LHS) developed by [10] along with the rank correlation refinement of [7] to generate a synthetic data set which reproduces both the univariate and the multivariate structure of the original data set. The basic LHS algorithm generates a synthetic data set for a group of uncorrelated variables in which the univariate characteristics of the original data are reproduced almost exactly. In case the variables are not uncorrelated, the restricted pairing algorithm of [7], which makes use of LHS, is designed to produce a synthetic data set that reproduces the rank correlation structure of the original data[1].

3.2 Hybrid Masking Methods

While pure synthetic data may reproduce univariate and multivariate characteristics of the original data, one-to-one comparison of original and synthetic records poses problems. In other words, given an original record or a subset of original records, it is possible that there is no similar synthetic record or no similar subset of synthetic records. To minimize such a possibility, Dandekar *et al.* [2] recommends using the LHS procedure at subpopulation levels to the extent possible.

Hybrid masking consists of computing masked data as a combination of original and synthetic data. Such a combination allows better control over individual characteristics of masked records. Additive as well as multiplicative combinations could be used. For hybrid masking to be feasible, a rule must be used to pair one original data record with one synthetic data record. A sensible option is to go through all original data records, and pair each original data record with the nearest synthetic record according to some distance.

Example 1 (Euclidean Record Pairing). Assume an original data set consisting of n records, and a synthetic data set consisting of m records. Assume further that both data sets refer to the same d *numerical variables*. Then the d-dimensional Euclidean distance can be used for pairing in the following way:

1. First, variables in both data sets are standardized (subtract to the values of each variable their average value and divide them by their standard deviation).

2. Pair each record in the original standardized data set with the nearest record in the synthetic standardized data set, where "nearest" means at the smallest d-dimensional Euclidean distance.

Once the pairing of original and synthetic records has been done, we need a model to mix variables in paired records in order to get a hybrid masked data set. For numerical variables, at least two different hybrid models are conceivable for combining a variable in the original data set with the corresponding variable in the synthetic data set:

Definition 1 (Additive Hybrid Model). *Let X be a variable in the original data set. Let X_s be the variable corresponding to X in the synthetic data set. Let α be a real number in $[0, 1]$. Then the additive hybrid masked version X_{ah} can be obtained from X and X_s as*

$$X_{ah} = \alpha X + (1 - \alpha) X_s \tag{1}$$

Definition 2 (Multiplicative Hybrid Model). *Let X be a variable in the original data set. Let X_s be the variable corresponding to X in the synthetic data set. Let α be a real number in $[0, 1]$. Then the multiplicative hybrid masked version X_{mh} can be obtained from X and X_s as*

$$X_{mh} = X^{\alpha} \cdot X_s^{(1-\alpha)} \tag{2}$$

Note that the above pairing strategy yields as many pairs as there are original records. This implies that mixing pairs using hybrid models (1) and (2) will result in a masked data set with the same number of records as the original data set. Such a constraint is a flexibility loss in comparison to pure synthetic data sets, whose cardinality can be made independent from the cardinality of the original data sets. A way to remedy the above rigidity is to use *resampled* original data instead of original data. Assume that the original data set consists of n records and that an n'-record masked data set is desired. Then we can obtain the n' masked records by using the following algorithm:

Algorithm 1. *1. Resample with replacement the n-record original data set to obtain an n'-record resampled data set.*
2. For each of the n' resampled records, pair it with the nearest record in the synthetic data set, where nearest means at smallest d-dimensional Euclidean distance.
3. Within each record pair, mix corresponding variables using one of the models (1) or (2) above.

4 Metrics for Method Comparison

In [5] a metric was proposed for method comparison. That metric needs some adaptation to deal with the case where the number of records in the original and the masked data sets are not the same. We first recall the [5] and then discuss some adaptations.

4.1 Same Number of Original and Masked Records

Let X and X' be the original and the masked data sets. 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 information loss measures proposed. In this table, d 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 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
$X - X'$	$\frac{\sum_{j=1}^{d}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{nd}$	$\frac{\sum_{j=1}^{d}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{nd}$	$\frac{\sum_{j=1}^{d}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{nd}$
$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^{d}(\bar{x}_j-\bar{x'}_j)^2}{d}$	$\frac{\sum_{j=1}^{d}\lvert\bar{x}_j-\bar{x'}_j\rvert}{d}$	$\frac{\sum_{j=1}^{d}\frac{\lvert\bar{x}_j-\bar{x'}_j\rvert}{\lvert\bar{x}_j\rvert}}{d}$
$V - V'$	$\frac{\sum_{j=1}^{d}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{d(d+1)}{2}}$	$\frac{\sum_{j=1}^{d}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{d(d+1)}{2}}$	$\frac{\sum_{j=1}^{d}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{d(d+1)}{2}}$
$S - S'$	$\frac{\sum_{j=1}^{d}(v_{jj}-v'_{jj})^2}{d}$	$\frac{\sum_{j=1}^{d}\lvert v_{jj}-v'_{jj}\rvert}{d}$	$\frac{\sum_{j=1}^{d}\frac{\lvert v_{jj}-v'_{jj}\rvert}{\lvert v_{jj}\rvert}}{d}$
$R - R'$	$\frac{\sum_{j=1}^{d}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{d(d-1)}{2}}$	$\frac{\sum_{j=1}^{d}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{d(d-1)}{2}}$	$\frac{\sum_{j=1}^{d}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{d(d-1)}{2}}$

Disclosure risk can be measured using record linkage. Two record linkage methods were combined in [5]:

Distance-based record linkage. Let the original and masked data sets consist both of d variables (it is assumed that both data sets contain the same variables). We define that a record in the masked data set corresponds to the nearest record in the original data set, where "nearest" means at shortest d-dimensional Euclidean distance. Assume further that the intruder can only access i key variables of the original data set and tries to link original and masked record based on these i variables. Linkage then proceeds by computing i-dimensional distances between records in the original and the masked data sets (distances are computed using only the i key variables) The variables used are standardized to avoid scaling problems. A record in the masked data set is labeled as "correctly linked" when the nearest record using i-dimensional distance is the corresponding one (*i.e.* the nearest record using d-dimensional distance).

Probabilistic record linkage. Defined in [8], uses a matching algorithm to pair records in the masked and original data sets. The matching algorithm is based on the linear sum assignment model. The definition of "correctly linked" records is the same as in distance-based record linkage. This method 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 a false non-match. Unlike distance-based record linkage, probabilistic record linkage does not require rescaling variables nor makes any assumption on their relative weight (by default, distance-based record linkage assumes that all variables have the same weight). Furthermore, probabilistic record linkage can be used on both numerical and categorical data.

In [5], a score was constructed to rate methods which combined some of the above information loss and disclosure risk measures. The components of the score were as follows:

IL Information Loss: 100 times the average of the mean variation of $X - X'$ (called IL_1), the mean variation of $\bar{X} - \bar{X}'$ (called IL_2), the mean variation of $V - V'$ (called IL_3), the mean variation of $S - S'$ (called IL_4) and the mean absolute error of $R - R'$ (called IL_5).

DLD Average of DLD-1, $\cdots$, DLD-7. DLD-i is the percent of records correctly linked using distance-based record linkage with Euclidean distance when the intruder knows i key variables of the original file.

PLD Same as DLD, but for probabilistic record linkage.

ID Average percent of original values falling in the intervals around their corresponding masked values. The average is over interval widths $p = 1\%$ to $p = 10\%$.

Overall score

$$Score = 0.5 \cdot IL + 0.125 \cdot DLD + 0.125 \cdot PLD + 0.25 \cdot ID \qquad (3)$$

The lower $Score$, the better is a method.

4.2 Different Number of Original and Masked Records

Computation of the IL_1 component in the score described in Subsection 4.1 implicitly assumes that there exists a one-to-one mapping between original and masked records. For natural masking, the number of masked and original records is the same and it is (in theory) possible to track from which original record a masked record originates (for example using metadata such as record identifier); so a one-to-one mapping can be established.

When the number of masked records is not the same as the number of original records (as it can happen with synthetic or hybrid data), then there is no one-to-one mapping any more. In this case, a new way to compute IL_1 must be defined. A natural way is to map each published masked record to the nearest original record, using the d-dimensional Euclidean distance between records (where d is

the number of variables in the data sets). Then we compute a new IL_1' as the sum of differences between masked records and the original records to which they are mapped. Replacing IL_1 by IL_1' leads to a modified information loss measure IL'.

Also, the lack of a one-to-one mapping between original and masked records forces a redefinition of disclosure risk measures DLD and PLD. If the masked and the original data sets have d variables, we will now say that a masked record is correctly linked to an original record if they are at the shortest possible d-dimensional Euclidean distance. Finally, ID can be redefined so that "corresponding values" mean values in records at shortest d-dimensional Euclidean distance. Call DLD', PLD' and ID' the resulting redefined disclosure risk measures.

Call $Score'$ the new score arising from replacing IL and DLD with IL' and DLD' in Equation 3 as well as dropping PLD for computational reasons:

$$Score' = 0.5 \cdot IL' + 0.25 \cdot DLD' + 0.25 \cdot ID' \tag{4}$$

5 Computational Results

Two original data sets have been tried:

Data set 1. This microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau[1]. $d = 13$ continuous variables were chosen and 1080 records were selected so that there were not many repeated values for any of the attributes (in principle, one would not expect repeated values for a continuous attribute, but there were repetitions in the data set).

Data set 2. $d = 13$ variables were drawn from the Commercial Building Energy Consumption Survey carried out by the Energy Information Administration of the U.S. Department of Energy. There are two categorical variables and 11 continuous variables. The number of records taken was 1080 (like for data set 1). Variables in this data set were more skewed than in data set 1.

A first round of experiments involved using rank swapping to mask the original data sets. Parameter values from $p = 1$ to $p = 20$ were considered. For each parameter choice, $Score'$ was computed using Equation (4). For each data set, Table 2 gives the parameter choices yielding the best $Score'$, the best IL', the best DLD' and the best ID'. For all measures, "best" means "lowest". Table 2 also gives the best values reached for each measure.

A second round of experiments used multivariate microaggregation taking three variables at a time (`Mic3mul`k) and four variables at a time (`Mic4mul`k), which were the best forms of microaggregation according to the score defined in [5]. Values for the parameter k between 3 and 18 were tried (k is the minimal size of microaggregates). For each data set, Table 3 gives the parameter choices

[1] `http://www.census.gov/DES`.

Table 2. Rank-swapping. Best parameter choices to minimize $Score'$, IL', DLD', and ID'

Data set		$Score'$	IL'	DLD'	ID'
1	Best parameter p	14	1	18	20
	Best measure value	25.663	1.95	12.355	29.541
2	Best parameter p	12	1	20	20
	Best measure value	23.687	2.037	11.839	29.022

yielding the best $Score'$, the best IL', the best DLD' and the best ID'. The best values reached for each measure are given as well.

Table 3. Multivariate microaggregation. Best parameter choices to minimize $Score'$, IL', DLD', and ID'

Data set	Method		$Score'$	IL'	DLD'	ID'
1	`Mic3mul`k	Best parameter k	14	3	18	17
		Best measure value	30.595	6.359	23.704	59.599
1	`Mic4mul`k	Best parameter k	14	3	17	18
		Best measure value	30.252	1.1	16.786	52.543
2	`Mic3mul`k	Best parameter k	14	3	18	17
		Best measure value	30.576	6.365	23.823	59.465
2	`Mic4mul`k	Best parameter k	17	3	16	18
		Best measure value	30.590	1.1	16.812	52.471

A third experimental round involved hybrid data computed using Algorithm 1 and an additive model (Equation (1)). The parameter here was α, which ranged between $\alpha = 0$ (pure LHS synthetic data) and $\alpha = 1$ (pure resampled original data). The size of the hybrid masked data set took values $n' = 500, 2000, 4000, 8000$. For each data set, Table 4 gives the parameter choices yielding the best $Score'$, the best IL', the best DLD' and the best ID'. The best values reached for each measure are given as well.

Note 1. An alternative to using simple random resampling with replacement in the first step of Algorithm 1 is to generate $n' - n$ LHS synthetic data records and replace each synthetic record with the nearest original record; the result is also a n'-record resampled original data set. Interesting as it may seem, this modification of the resampling procedure in Algorithm 1 does not lead to significant variation of the results shown in Table 4.

Note 2. For the data sets discussed in this section, the additive model of Equation (1) yields better results than the multiplicative model of Equation (2). However, this may change for other data sets.

Table 4. Additive LHS hybrid data. Best parameter choices to minimize $Score'$, IL', DLD', and ID'

Data set			$Score'$	IL'	DLD'	ID'
1	$n' = 500$	Best parameter α	0	1	0	0
		Best measure value	33.466	11.27	16.6	48.037
1	$n' = 2000$	Best parameter α	0	1	0	0
		Best measure value	32.069	5.363	19.336	50.775
1	$n' = 4000$	Best parameter α	0.1	1	0	0
		Best measure value	34.546	2.357	21.368	52.911
1	$n' = 8000$	Best parameter α	0	1	0	0
		Best measure value	38.854	1.982	23.964	55.139
2	$n' = 500$	Best parameter α	0.3	1	0	0
		Best measure value	34.225	7.504	18	46
2	$n' = 2000$	Best parameter α	0.1	1	0	0
		Best measure value	32.245	3.33	20	50.712
2	$n' = 4000$	Best parameter α	0.1	1	0	0
		Best measure value	31.940	6.929	21.614	52.827
2	$n' = 8000$	Best parameter α	0.4	1	0	0
		Best measure value	38.3	1.8	24.29	54.9

6 Conclusions

Looking at the performance in terms of $Score'$, it can be seen that rank swapping appears as the best performer, followed by multivariate microaggregation and then by LHS hybrid data (the difference between the last two approaches is small). The best scores of all methods are similar on both data sets, but the best parameter choice for hybrid data shows some dependency on the particular data set:

- For the first data set, LHS hybrid data score best when the data are purely synthetic ($\alpha = 0$) or with a weak resampled original component ($\alpha = 0.1$ for $n' = 2000$).
- For the second (more skewed) data set, a stronger resampled original component (α up to 0.4) may be needed to attain the best scores.

Regarding the global information loss IL', there is no clear winner. However, note that hybrid data get the lowest information loss when $\alpha = 1$, *i.e.* when the masked data are just resampled original data.

Note 3. If the IL'_1 component is suppressed from IL', then hybrid data tend to be the best performer (regardless of the parameter α): LHS nearly preserves averages, covariances and correlations, which brings IL_2, IL_3, IL_4 and IL_5 close to 0. In situations where IL'_1 is of critical importance, the LHS procedure should be performed at the subpopulation levels.

From the disclosure risk standpoint (DLD' and ID' measures), rank swapping is best, while hybrid data and multivariate microaggregation perform similarly.

Two additional lessons that can be learned from the tables in Section 5 are that:

- `Mic3mul`k behaves similarly to `Mic4mul`k.
- Although there is no big influence of n' on the performance of additive hybrid data, experiments on both data sets show that taking $n' \approx 2n$ seems to be a wise option.

To summarize, best parameter choices for LHS synthetic microdata and multivariate microaggregation yield similar results, a bit behind those obtained with rank swapping.

References

1. R.A. Dandekar, "Performance improvement of restricted pairing algorithm for Latin hypercube sampling", in *ASA Summer Conference* (unpublished).
2. R.A. Dandekar, M. Cohen and N. Kirkendall, "Applicability of Latin hypercube sampling technique to create multivariate synthetic microdata", in *Proc. of ETK-NTTS 2001.* Luxembourg: Eurostat, pp. 839-847, 2001.
3. D. Defays and P. Nanopoulos, "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, 1993.
4. J. Domingo-Ferrer, J.M. Mateo-Sanz, and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Proc. of ETK-NTTS 2001.* Luxemburg: Eurostat, pp. 807-825, 2001.
5. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, pp. 111-133, 2001.
6. J. Domingo-Ferrer and J.M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 189-201, 2002.
7. R.L. Iman and W.J. Conover, "A distribution-free approach to inducing rank correlation among input variables", *Communications in Statistics*, vol. B11, no. 3, pp. 311-334, 1982.
8. 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, pp. 414-420, 1989.
9. A.B. Kennickell, "Multiple imputation and disclosure protection: the case of the 1995 survey of consumer finances", in *Statistical Data Protection.* Luxembourg: Office for Official Publications of the European Communities, pp. 381-400, 1999.
10. M.D. McKay, W.J. Conover, and R.J. Beckman, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code", *Technometrics*, vol. 21, no. 2, pp. 239-245, 1979.
11. R. Moore, "Controlled data swapping techniques for masking public use microdata sets", U. S. Bureau of the Census, 1996 (unpublished manuscript).
12. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control.* New York: Springer-Verlag, 2001.