Fast Generation of Accurate Synthetic Microdata

Josep Maria Mateo-Sanz[1],
Antoni Martínez-Ballesté[2], and Josep Domingo-Ferrer[2]

[1] Universitat Rovira i Virgili
Statistics Group, Av. Països Catalans 26, E-43007 Tarragona, Catalonia
jmateo@etseq.urv.es
[2] Dept. of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{anmartin,jdomingo}@etse.urv.es

Abstract. Generation of a synthetic microdata set that reproduces the statistical properties of an original microdata set is a promising approach to statistical disclosure control (SDC) of microdata. In this paper, a new method for generating continuous synthetic microdata is proposed. The covariance matrix and the univariate statistics of the original data set are *exactly* preserved. The method is non-iterative and its complexity grows linearly with the number of records to be protected.

1 Introduction

Statistical databases can either contain tabular data or individual data (microdata). Microdata can be continuous, *e.g.* age and weight, or categorical, for instance sex or hair color. When a microdata set is to be released for public use, confidentiality must be ensured. In that sense, the purpose of Statistical Disclosure Control (SDC) techniques is twofold: on one hand, SDC methods must prevent the identity of the individual respondent from being disclosed; on the other hand, the published set of data should preserve as many statistical properties as possible from the original set.

One possibility for protecting a microdata set is to use a *masking method* (*e.g.* additive noise, microaggregation, etc., cf. [1]) to transform original data into protected, publishable data. An alternative to masking the original data is to generate a new data set (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. A third possibility is to build a *hybrid* data set as a mixture of the masked original values and a synthetic data set [2].

1.1 Background on Synthetic Data Generation

Publication of simulated – *i.e.* synthetic – data was proposed long ago as a way to guard against statistical disclosure. In fact, as early as 1993, Rubin [3] suggested

J. Domingo-Ferrer and V. Torra (Eds.): PSD 2004, LNCS 3050, pp. 298–306, 2004.

creating an entirely synthetic data set based on the real survey data and multiple imputation. Specific case studies of synthetic microdata generated by multiple imputation were presented in [4, 5]. Although the results were fairly promising, the multiple imputation approach requires complex models and software, which greatly reduces its appeal in many situations.

In [6, 7] comparisons were presented for measuring the performance of microdata masking methods in terms of information loss and disclosure risk. Based on the proposed measures, it was shown in [8] how to improve the performance of any particular masking method. In particular, post-masking optimization was discussed for preserving as much as possible the moments of first and second order (and thus multivariate statistics) without increasing the disclosure risk. The technique proposed could also be used for synthetic microdata generation and could be extended for preservation of all moments up to m-th order, for any m. The shortcoming of this approach is its computational complexity: the optimization problem is solved using an iterative refinement approach, which may be quite time-consuming when the involved data sets are large.

Latin Hypercube Sampling (LHS) appears in the literature as another method for generating multivariate synthetic data sets. In [9], authors improve the LHS updated technique of [10], but the proposed scheme is still time-intensive even for a moderate number of records. In [11], LHS is used along with a rank correlation refinement to reproduce both the univariate (*i.e.* mean and variance) and multivariate structure (in the sense of rank correlation) of the original data set. This method also permits flexibility in the size of the synthetic data set that is generated. In summary, LHS-based methods rely on iterative refinement, are time-intensive and their running time does not only depend on the number of values to be reproduced, but on the starting values as well.

1.2 Contribution and Plan of This Paper

In this paper, a non-iterative method for generating continuous synthetic microdata is proposed. The implementation of this method results in a fast algorithm which *exactly* reproduces the means and the covariance matrix of the original data set and whose running time grows *linearly with the number of records*. Exact preservation of the original covariance matrix implies that variances and Pearson correlations are also exactly preserved in the synthetic data set. Like in any synthetic data generator, the number of records in the synthetic data set can differ from the number of records in the original data set.

Section 2 describes our proposal for generating synthetic data. Section 3 analyzes the complexity and the data utility properties of the proposed method. Empirical results are reported in Section 4. Finally, Section 5 contains some conclusions.

2 A Low-Cost Method for Synthetic Microdata Generation

Let X be an original microdata set, with n records and m variables. Let X' be the synthetic microdata set to be generated, with n' records and m variables. In

fact, X can be viewed as an $n \times m$ matrix and X' can be viewed as an $n' \times m$ matrix. The method presented in this section will ensure that both univariate and multivariate statistical properties of X, such as mean and covariance, are exactly reproduced in the resulting X'.

The algorithm below constructs X' from X:

Algorithm 1 (Basic Procedure)

1. *Generate A, which is a random $n' \times m$ matrix, such that the covariance matrix of A is the identity matrix.*
2. *Compute the covariance matrix C of the original microdata matrix X.*
3. *Use the Cholesky 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. *Obtain the synthetic microdata set X' as a matrix product:*

$$X' = A \cdot U$$

Note that the covariance matrix of X' equals the covariance matrix of X [12].

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

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

We now need to specify how to construct a random $n' \times m$ matrix A, whose covariance matrix is the $m \times m$ identity matrix.

Algorithm 2 (Construction of Matrix A)

1. *Generate A as an $n' \times m$ matrix with random elements $a_{i,j}$. View the m columns of A as samples of variables $A_1, \cdots, A_m$. If $Cov(A_j, A_{j'})$ is the covariance between variables 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:*

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

The mean of the adjusted A_1 is 0.

3. *In order to reach the desired identity covariance matrix, some values of variables* $A_2, \cdots, A_m$ *must change. For* $v = 2$ *to* m *do:*
   (a) *Let* $\bar{a}_v$ *be the mean of variable* A_v.
   (b) *For* $j = 1$ *to* $v - 1$, *the covariance between variables* 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* $Cov(A_j, A_v) = 0, \quad j = 1 \ldots v - 1$, *some elements* $a_{i,v}$ *in the* v*-th column of* A *are assigned a new value. Let* $x_1, \ldots, x_{v-1}$ *be the unknowns for the following linear system of* $v - 1$ *equations:*

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

   that is

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

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

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

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

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

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

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

With the construction proposed in this section, the number of records n' in X' does not depend on the number of records n in X. Thus, disclosure of n is prevented, which may be useful in some situations. On the other hand, Algorithm 2 does not need to be run each time Algorithm 1 is run. In other words, if $X_1, X_2, \ldots X_u$ are original microdata sets, each with n_i records, $i = 1 \ldots u$, and m variables, then u synthetic microdata sets $X'_1, X'_2, \ldots X'_u$ can be generated, each with n' records and m variables, *with a single* $n' \times m$ *matrix* A.

3 Properties of the Proposed Scheme

3.1 Performance and Complexity

To simplify the performance and complexity analysis presented here, we assume that a synthetic data set of size $n \times m$ is generated from an original data set

Table 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 variables			
Number of records	5	10	25	50
1,000	0.00	0.00	0.05	0.31
10,000	0.05	0.19	1.26	5.31
100,000	0.49	1.93	12.41	51.15

of the same size, *i.e.* $n' = n$. The method has been tested with several data set sizes and execution times are shown in Table 1.

The computational complexity for the proposed method will next be estimated. Let n be the number of records, m the number of variables and assume for simplicity $n' = n$. Then the complexities of the various operations are as follows:

- Calculation of the covariance matrix: $\mathcal{O}(n + m^2)$.
- Cholesky decomposition: $\mathcal{O}(m^3/6)$ (see [13]).
- 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 [13].
- 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 variables m, *i.e.* $n >> m$. The strong point of this proposal is that *its complexity is linear in the number of records.* It must also be kept in mind that, as pointed out at the end of Section 2, matrix A can be re-used to generate several synthetic microdata sets, which greatly reduces computation.

3.2 Data Utility

As stated in Section 1.2, the proposed scheme exactly reproduces the statistical properties of the original data set. In particular:

- The means of variables 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 [12]). Thus, in particular:
  - The variance of each variable 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.

Table 2. Values of *Score'* for the synthetic data

Measure	Value
IL_1	544.53
IL_2	3.53565e-05
IL_3	3.25579e-04
IL_4	1.81034e-03
IL_5	1.90171e-04
IL	108.90
DLD	9.10
ID	33.21
$Score'$	65.0338

4 Empirical Work

4.1 The Score

In the experiments conducted to measure the disclosure risk and the information loss in the synthetic data sets produced by our method, we use the $Score'$ defined in [2]. $Score'$ is a modification of the original $Score$ defined in [7] to deal with synthetic data generation in which the number of records of the synthetic data set differs from the number of records in the original data set. We briefly recall the definition of $Score'$:

$$Score' = 0.5 \cdot IL + 0.25 \cdot DLD + 0.25 \cdot ID$$

where IL stands for information loss, DLD refers to distance-based record linkage and ID stands for interval disclosure. DLD and ID are disclosure risk measures. IL measures how different is the synthetic data set from the original one. IL_1 is a component of IL which compares the individual original and synthetic values, whereas the remaining components for IL reflect how different are univariate and multivariate statistics between X and X'. See [2] on how to compute IL, DLD and ID.

4.2 The Data Set

The microdata set for testing was constructed using the Data Extraction System of the U.S. Census Bureau (http://www.census.gov/DES) and contains $n = 1080$ registers for $m = 13$ continuous variables. This data set was also used in [2, 6, 7].

4.3 The Results

As mentioned in Step 1 of Algorithm 2, A is initially composed of random values. It must be noticed that whatever the magnitude of the values in X is, the range in which the initial random values for A are picked – say between 0 and 100 – does not affect the results. The score for a typical execution is shown in Table 2.

Note that, since most of the values in A are random, the result for IL_1 shows that most of values in X are substantially different in X'. The point is that the

Table 3. DLD and ID values for synthetic data sets with n' records

Number of records	DLD	ID
500	13.40	33.27
2,000	12.65	45.65
8,000	15.14	34.14
10,000	12.54	34.77
20,000	11.86	39.60

remaining components IL_2, IL_3, IL_4 IL_5 of the information loss show that the statistical properties listed in Section 3.2 are exactly fulfilled (those measures do not appear as exactly 0 due to rounding errors). On the other hand, the disclosure risk measures DLD and ID are lower than those obtained for the LHS-Based method and reported in [2].

Due to the randomness of matrix A, different runs of the method will result in different synthetic data sets and, consequently, the resulting *Score'* will change. In 10 executions, an average value of 12.14 for DLD was obtained, with a standard deviation of 0.97; the average obtained for ID was 37.99 with a standard deviation of 5.11.

If synthetic data sets are generated whose number n' of records is not the same as the number n of original records, the values for DLD and ID are maintained (see Table 3). Hence, the disclosure risk measures do not depend on the number of records of the synthetic data set.

4.4 Non-random Matrix *A*

In order to reduce the information loss component IL_1 (individual record comparison), one could think of choosing the initial values of matrix A in a "clever" way rather than using initial random values. For example, A could be the result of masking the original data set X using a perturbative masking method (see [1]). This leads to a number of records in X' which equals the the number of records in X.

Table 4 shows the results obtained when different perturbative masking methods are used. The lowest value for DLD and ID is reached when A has been obtained using microaggregation [14] with parameter $k = 20$ The lowest value for IL occurs for additive noise with parameter 2%.

5 Conclusions

In this paper, a new method for generating a synthetic data set X' from a microdata set X has been presented. This method, specified in Algorithms 1 and 2, is suitable for continuous microdata.

In addition to allowing a different number of records in the original and the synthetic data sets (a property shared by all synthetic data generation methods), the main properties of the method are:

Table 4. Results when using a masked data set as A

Masking method	$Score'$	IL'	DLD'	ID'
noise.02	36.42	36.41	20.85	52.00
noise.08	38.04	40.69	19.48	51.32
noise.14	35.61	36.50	18.20	51.24
microag.k=5	37.74	39.70	20.19	51.37
microag.k=10	40.96	45.21	23.28	50.15
microag.k=20	37.42	42.85	16.30	47.68
rankswap.5	38.76	43.89	16.77	50.50
rankswap.15	42.28	50.77	16.93	50.64

- Its computational complexity is linear in the number of records. The implementation of Algorithms 1 and 2 is simple and executions are efficient. One of the computations is solving $m - 1$ linear systems, where m is the number of variables of X and X' (note that the number of variables is usually much smaller than the number of records). The remaining computations, *i.e.* Cholesky decomposition, matrix product, etc., are also of low complexity.
- Algorithms 1 and 2 are non-iterative, so that the number of steps for obtaining X' can be known *a priori*. Other methods for obtaining synthetic data are based on iterative algorithms whose running time cannot be predicted before execution.
- The following statistical properties of the original data set X are preserved by the synthetic data set X': mean, variances, the covariance matrix and the Pearson correlation matrix. Note that a good deal of methods for synthetic data generation in the literature do not preserve the variance-covariance matrix nor the Pearson correlation matrix.
- Empirical work shows that the disclosure risk is lower than for masking methods.

As shown in Section 4.3, IL is higher than for usual masking methods when the initial values used for A are random. To keep IL_1 low while preserving the covariance matrix and the univariate statistics, initial values for A can be obtained using a perturbative masking method.

Acknowledgments

This work has been partly supported by the European Commission under project IST-2000-25069 "CASC".

References

1. J. Domingo-Ferrer and V. Torra, "Disclosure protection methods and information loss for microdata", in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, Amsterdam: North-Holland, 2001, pp. 91-110.

2. R. A. Dandekar, J. Domingo-Ferrer and F. Sebé, "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection", in *Inference Control in Statistical Databases*, ed. J. Domingo-Ferrer, vol. LNCS 2316, pp. 153-162, Springer, 2002.
3. D. B. Rubin, "Discussion on statistical disclosure limitation", *Journal of Official Statistics*, vol. 9, no. 2, pp. 461-468.
4. A. B. Kennickell, "Multiple imputation and disclosure control: the case of the 1995 Survey of Consumer Finances", in *Record Linkage Techniques*, Washington DC: National Academy Press, 1999, pp. 248-267.
5. A. B. Kennickell, "Multiple imputation and disclosure protection: the case of 1995 Survey of Consumer Finances", in *Statistical Data Protection*, Luxemburg: Office for Official Publication of the European Communities, 1999, pp. 177-206.
6. 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. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, Amsterdam: North-Holland, 2001, pp. 111-134.
7. J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Proceedings of ETK-NTTS 2001*, Luxemburg: Eurostat, pp. 807-825, 2001.
8. F. Sebé, J. Domingo-Ferrer, J. Mateo-Sanz and V. Torra, "Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets", in *Inference Control in Statistical Databases*, ed. J. Domingo-Ferrer, vol. LNCS 2316, pp. 163-171, Springer, 2002.
9. D. E. Huntington and C. S. Lyrintzis, "Improvements to and limitations of Latin hypercube sampling", *Probabilistic Engineering Mechanics*, vol. 13, no. 4, pp. 245-253, 1998.
10. A. Florian, "An efficient sampling scheme: updated latin hypercube sampling", *Probabilistic Engineering Mechanics*, no. 7, pp. 123-130, 1992.
11. R. A. Dandekar, M. Cohen and N. Kirkendall, "Sensitive micro data protection using latin hypercube sampling technique", in *Inference Control in Statistical Databases*, vol. LNCS 2316, pp. 245-253, Springer, 2002.
12. E. M. Scheuer and D. S. Stoller, "On the generation of normal random vectors", *Technometrics*, no. 4, pp. 278-281, 1962.
13. W. Press, W. T. Teukolsky, S. A. Vetterling and B. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1993.
14. 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, no. 1, pp. 189-201, Feb. 2002.