Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 507–524

A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables

Krishnamurty Muralidhar[1] *and Rathindra Sarathy*[2]

Statistical disclosure limitation techniques are designed to provide legitimate users with access to useful data while simultaneously preventing disclosure of sensitive information. Two techniques that can be used to limit disclosure of sensitive numerical data are multiple imputation and data perturbation. While many studies have addressed the effectiveness of perturbation and multiple imputation individually, no studies have directly compared the two techniques. In this study, we compare the effectiveness of multiple imputation and data perturbation for numerical microdata. The results indicate that, in the absence of missing data, data perturbation performs better than multiple imputation. In addition, since only a single perturbed data set is released (unlike the multiply-imputed data sets that are released), data perturbation eases the burden on users of such data.

Key words: Confidentiality; privacy; data dissemination.

1. Introduction

Statistical agencies that gather and release data have dual, conflicting responsibilities. On the one hand, they must preserve the privacy and confidentiality of the individuals from whom they have gathered data. On the other hand, they have to ensure that released data is useful for analysis. The problem is particularly acute if the statistical agency releases microdata. Yet, releasing summary data rather than microdata limits the types of analyses that can be conducted on such data (Duncan and Pearson 1991; Citteur and Willenborg 1993). While there is no question that all types of microdata release are important, in this study we focus on the release of numerical, confidential microdata values.

A variety of techniques have been proposed for releasing numerical, confidential microdata useful for analysis, while preventing disclosure. Generally referred to as "masking" techniques, they include data perturbation, micro-aggregation, multiple imputation, swapping, rounding, etc. For a comprehensive review of these techniques, please refer to Willenborg and de Waal (2001). Not all techniques are equally effective. For example, it has been shown that data swapping can result in reduced data utility and high disclosure risk (Moore 1996; Muralidhar and Sarathy 2003). Micro-aggregation has been shown to result in dangerously high disclosure risk (Winkler 2002) and reduced data utility. Only multiple imputation and data perturbation appear to provide high data utility and low disclosure risk.

[1] University of Kentucky, School of Management, Lexington, KY 40506, U.S.A. Email: krishm@uky.edu
[2] Oklahoma State University, Spears School of Business, Stillwater, OK 74078, U.S.A. Email: sarathy@okstate.edu

Data perturbation for numerical variables has received considerable attention in the literature. Beginning with the simple "noise addition" approach, data perturbation has been enhanced in several ways (Kim 1986; Fuller 1993; Tendick 1991; Tendick and Matloff 1994; Muralidhar et al. 1999, 2001).[3] Originally developed as a methodology for analysis with missing data, multiple imputation was first proposed for statistical disclosure limitation by Rubin (1993) and has been addressed in greater detail recently by Reiter (2002) and Raghunathan et al. (2003). In practice, data perturbation and multiple imputation methodologies can be considered as alternative, competing techniques for masking numerical microdata. When viewed from this perspective, it is desirable to compare the two techniques in terms of data utility and disclosure risk requirements. The objective of this study is to present such a comparison for numerical confidential variables.

The remainder of the article is organized as follows. In the next section, we briefly review the theoretical and philosophical underpinnings of perturbation. In the third section, we numerically illustrate the new perturbation approach. In the fourth section, we empirically compare the new perturbation approach and multiple imputation. In the fifth section, we discuss the limitations of each approach. The last section contains the conclusions.

2. A Sufficiency-Based Approach for Data Perturbation

The basis for data perturbation has been addressed by a variety of authors including Fuller (1993), Fienberg et al. (1998) and, most recently, Muralidhar and Sarathy (2003). In their paper, Muralidhar and Sarathy (2003) suggest generating the perturbed values of the confidential variables from the conditional distribution of the confidential variables given the nonconfidential variables. This provides high data utility and low disclosure risk, under the assumption that the data set itself is the population.

Formally, we can describe the conditional distribution approach proposed by Muralidhar and Sarathy (2003) as follows. Consider the data set *as a finite population* consisting of a set of nonconfidential variables $\boldsymbol{S}$ and a set of confidential variables $\boldsymbol{X}$. For each observation, generate a vector $\mathbf{y_i}$ from the conditional distribution $\mathrm{f}(\boldsymbol{X}|\boldsymbol{S} = s_i)$. Muralidhar and Sarathy show that the collection of values ($\boldsymbol{S}$ and $\boldsymbol{Y}$) has the same distribution as the original values ($\boldsymbol{S}$ and $\boldsymbol{X}$), thereby providing the highest possible level of data utility. Dalenius (1977) and Duncan and Lambert (1986) define disclosure risk in general terms as the improved predictive ability of a data intruder (snooper) when provided access to the masked data (compared with the snooper's predictive ability prior to access to the masked data). The conditional distribution approach results in $\mathrm{f}(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{Y}) = \mathrm{f}(\boldsymbol{X}|\boldsymbol{S})$. Thus, providing access to the masked microdata ($\boldsymbol{Y}$) does not improve the predictive ability of a snooper. Consequently, the conditional distribution approach minimizes disclosure risk.

The conditional distribution approach, while appropriate for perturbing any confidential variable, is difficult to implement for categorical variables (Fienberg et al. 1998). Even for numerical variables, deriving the conditional density $\mathrm{f}(\boldsymbol{X}|\boldsymbol{S})$ can be very difficult in the general case. However, under specific assumptions regarding the underlying population

[3] Please note that we focus on data perturbation for numerical data only. Several studies have addressed data perturbation for categorical data (Fienberg et al. 1998).

and/or depending on the specific characteristics of the data set to be preserved, it is possible to implement the conditional distribution approach. One such case occurs when the underlying population has a multivariate normal distribution and/or it is necessary to preserve the mean vector and covariance matrix of the data set (Muralidhar et al. 1999, 2001).

The perturbation approach suggested by Muralidhar et al. (1999, 2001) generates the values of $\mathbf{y_i}$ using the linear model:

$$\boldsymbol{y}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\boldsymbol{s}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are computed by regressing $\boldsymbol{X}$ on $\boldsymbol{S}$ so that we have $\boldsymbol{\beta_1} = \boldsymbol{\Sigma}_{XS}(\boldsymbol{\Sigma}_{SS})^{-1}$ and $\boldsymbol{\beta_0} = \boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\Sigma}_{XS}(\boldsymbol{\Sigma}_{SS})^{-1}\boldsymbol{\mu}_{\mathbf{S}}$, $\varepsilon \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{X|S})$, $\boldsymbol{\Sigma}_{X|S} = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}(\boldsymbol{\Sigma}_{SS})^{-1}\boldsymbol{\Sigma}_{SX}$, and $\boldsymbol{\Sigma}_{XX}, \boldsymbol{\Sigma}_{SS}$, and $\boldsymbol{\Sigma}_{XS}$ represent the covariance matrix of $\boldsymbol{X}$, $\boldsymbol{S}$, and ($\boldsymbol{X}$ and $\boldsymbol{S}$). We can verify that the random variable $\boldsymbol{Y}$ is distributed with mean vector $\boldsymbol{\mu}_{\mathbf{X}}$, covariance matrix $\boldsymbol{\Sigma}_{XX}$, and $\boldsymbol{\Sigma}_{YS} = \boldsymbol{\Sigma}_{XS}$. We can also verify that the $\boldsymbol{\Sigma}_{X|S,Y} = \boldsymbol{\Sigma}_{X|S}$, and hence providing access to $\boldsymbol{Y}$ does not improve a snooper's ability to predict confidential values using linear models.

In the above approach, the mean and the covariance matrix of the perturbed data will not be **identical** to those of the original data, but will approach the original values asymptotically. This "sampling error" is likely to be a problem in situations where the user treats the data as a sample from an unknown population and attempts inferences regarding a parameter of this unknown population. In other words, the observed values $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_{XX}$ are themselves sample statistics. If $\boldsymbol{Y}$ were used in place of $\boldsymbol{X}$, inferences reached regarding unknown population parameters using $\boldsymbol{Y}$ may be different from inferences using $\boldsymbol{X}$ (Rubin 1987).

Burridge (2003) suggests a simple modification to overcome this problem, relying on the fact that the information contained in a data set about a population parameter can be summarized by its sufficient statistics. If we can reproduce another data set having the same sufficient statistics as the original data set, then "information has been preserved," and all inferences reached about the population parameter using the reproduced data would be the same as inferences using the original data (Burridge 2003). An ideal solution would be to maintain sufficient statistics for *any statistical analysis* in the perturbed data. In practice, this would be very difficult to achieve. Fortunately, the mean vector and covariance matrix of the sample are sufficient statistics for many statistical models that are commonly used in practice. Examples include hypothesis testing of the mean (both univariate and multivariate), ANOVA, MANOVA, multiple regression analysis, multivariate multiple regression, factor analysis, canonical correlation analysis, etc. Hence, maintaining these statistics in the masked data to be exactly the same as those in the original data would ensure that the results of *most* statistical analyses performed using the masked data would yield *exactly* the same results as the original data (Lehmann and Casella 1998).

Burridge (2003) provides an approach for generating $\boldsymbol{\varepsilon}$ such that, regardless of sample size, the mean vector and covariance matrix of the masked data *are exactly the same* as those of the original data. For small sample sizes, this approach could potentially increase disclosure risk since it does not ensure that $\boldsymbol{\Sigma}_{X|S,Y} = \boldsymbol{\Sigma}_{X|S}$. We modify Burridge's procedure to ensure that $\boldsymbol{\Sigma}_{X|S,Y} = \boldsymbol{\Sigma}_{X|S}$ as follows:

- Generate a(ny) matrix of random numbers $\boldsymbol{A}$ of size $(n \times k)$ where n is the number of observations in the data set, and k is the number of confidential variables.

- Regress $\boldsymbol{A}$ on($\boldsymbol{X}$ and $\boldsymbol{S}$) and compute the residuals $\boldsymbol{R}$. Note that $\boldsymbol{R}$ is orthogonal to ($\boldsymbol{X}$ and $\boldsymbol{S}$) and has mean vector $\mathbf{0}$.
- Compute $\boldsymbol{\Sigma}_{\boldsymbol{RR}}$ the covariance matrix of $\boldsymbol{R}$.
- Compute $\boldsymbol{R}^* = (\boldsymbol{\Sigma}_{\boldsymbol{RR}})^{-0.5}\boldsymbol{R}$. $\boldsymbol{R}^*$ has mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$.
- Compute $\boldsymbol{\varepsilon} = (\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{S}})^{0.5}\boldsymbol{R}^*$. $\boldsymbol{\varepsilon}$ is orthogonal to $\boldsymbol{X}$ and $\boldsymbol{S}$, has mean vector $\mathbf{0}$, and covariance matrix $= \boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{S}}$.
- For each observation in the data set, compute $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{s}_i + \boldsymbol{\varepsilon_i}$ where $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are derived by regressing $\boldsymbol{X}$ on $\boldsymbol{S}$.

The difference between the above procedure and the one suggested in Burridge (2003) is in Step (2) above. While Burridge regresses only $\boldsymbol{A}$ on $\boldsymbol{S}$, the above procedure regresses $\boldsymbol{A}$ ***on both*** ($\boldsymbol{X}$ and $\boldsymbol{S}$). This ensures that $\boldsymbol{X}$ and $\boldsymbol{A}$ are orthogonal and $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{S},\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{S}}$. This step has important implications in assessing disclosure risk.

Our assessment of disclosure risk resulting from a masking procedure is based on an incremental approach. Our approach is to isolate the disclosure risk that would result from the masking procedure. Prior to the release of such data, we assume that the agency has already provided the users with aggregate information and access to the nonconfidential variables, $\boldsymbol{S}$. Let us assume that the disclosure risk due to this information is Π_{S}. The only objective of the data masking approach is to provide access to the masked microdata for confidential variables. Hence, we assess the *incremental* disclosure risk resulting from the masking procedure. Assume that the masked values are generated as

$$\boldsymbol{y_i} = g(\boldsymbol{s}_i, \boldsymbol{\varepsilon}_i) \tag{2}$$

where the noise term $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{S}$, and g(.) is any mathematical function (that is an approximation to f($\boldsymbol{X}|\boldsymbol{S}$)). Equation (1) is a special case of Equation (2). Under these conditions, we can show that

$$\mathrm{P}(\boldsymbol{X} \leq \boldsymbol{x}_i|\boldsymbol{S} \leq \boldsymbol{s}_i, \boldsymbol{Y} \leq \boldsymbol{y}_i) = \mathrm{P}(\boldsymbol{X} \leq \boldsymbol{x}_i|\boldsymbol{S} \leq \boldsymbol{s}_i) \tag{3}$$

Hence, it follows that the disclosure risk with access to both the nonconfidential variables $\boldsymbol{S}$ and the masked microdata $\boldsymbol{Y}$ is also Π_S. In other words, providing access to the masked microdata $\boldsymbol{Y}$ does not provide *any* additional information regarding the confidential variable $\boldsymbol{X}$, thereby minimizing disclosure risk as defined by Dalenius (1977). It is important to note that we *do not* contend that release of the entire data set prevents disclosure; we only contend that release of the masked microdata $\boldsymbol{Y}$ prevents *additional* disclosure (over and above that resulting from the release of aggregate information and the nonconfidential variables $\boldsymbol{S}$). When the value of Π_{S} itself is considered too high, the agency releasing the data may want to reconsider releasing information from this data set.

Thus, the sufficiency-based perturbation approach provides two distinct advantages. First, many traditional types of statistical analyses using the masked microdata will yield the same results as analyses using the original data, resulting in high data utility. Second, access to the masked microdata will not provide additional information regarding the confidential variables; thereby, disclosure risk from releasing the masked microdata is minimized. When such data is disseminated, however, it is important that the agency releasing the data inform the user that the data is best suited for statistical analyses where the requisite sufficient statistics are the mean vector and the covariance matrix.

Finally, note that the sufficiency-based perturbation approach can be implemented for any data set regardless of its underlying characteristics. If the underlying distribution of the data set is approximately multivariate normal, this procedure provides possibly the best solution to the masking problem, since the results of any analysis using the perturbed data will be identical to results using the original data. However, when the underlying data set is not multivariate normal, the above approach will maintain the mean vector and covariance matrix to be the same, but will not maintain the marginal distribution of nonnormal confidential variables, nor will it maintain nonlinear relationships.

3. A Numerical Illustration

In this section, we provide a simple illustration using an empirical data set consisting of 50 observations with two categorical nonconfidential variables and two numerical confidential variables. The data was masked using the sufficiency-based perturbation approach described in the previous section. The entire data set (both the original and masked values) is provided in Table 1.

In terms of usefulness of the released data, consider a user attempting to estimate the relationship between the confidential variables. Table 2 provides the results of the regression analysis to predict X_2 using S_1, S_2, and X_1 and the results of the same analysis using the masked data. Table 2 shows that the results are identical for both data sets. Hence, for this analysis, using $\boldsymbol{Y}$ in place of $\boldsymbol{X}$ results in no information loss. Similar results will be observed for all analyses for which the mean vector and the covariance matrix are sufficient statistics and can be verified using the data in Table 1.

We can assess disclosure risk by considering the information regarding the confidential variables that is contained in the released data. Table 3 provides the results of the regression analysis to predict the value of variable X_1 using the nonconfidential variables $\boldsymbol{S}$. Even without access to the masked microdata $\boldsymbol{Y}$, an intruder would be able to predict the values of the confidential variables, and the results in Table 3 provide this baseline. Table 3 also provides the results of a regression analysis to predict X_1 using *both* $\boldsymbol{S}$ and $\boldsymbol{Y}$. The table shows that the regression coefficients for the masked variables Y_1 and Y_2 are 0.0000. In other words, the masking procedure does not provide the intruder with any additional information regarding the confidential variables. At the same time all other measures of effectiveness are maintained before and after the availability of the masked data $\boldsymbol{Y}$. We can easily show that these results hold for other variables and identity disclosure as well.

In conclusion, the results in this section show that the sufficiency-based data perturbation approach provides a high level of data utility (or low information loss) and simultaneously minimizes disclosure risk. It is easy to show that this approach performs better than all other perturbation approaches for most types of statistical analysis. In the following section, we compare the performance of the sufficiency-based perturbation approach with that of multiple imputation. As discussed in the introduction, multiple imputation has been proposed as a technique for limiting disclosure of confidential, numerical microdata. One strength claimed for multiple imputation is the ability to make valid inferences about population parameters based on multiply-imputed sample data. Based on its ability to maintain sufficient statistics in the perturbed data while minimizing disclosure risk, we show that the sufficiency-based perturbation approach has advantages over multiple imputation.

Table 1. Original and Perturbed Data

S_1	S_2	X_1	X_2	Y_1	Y_2
1	0	500.4600	1155.8300	482.6735	1424.7832
1	1	493.9200	1377.0200	359.0649	1071.8574
1	1	634.3200	1299.1300	624.4686	1301.0551
0	0	508.8400	1354.3300	434.0697	1335.2988
1	0	525.2200	1107.1000	665.4042	1929.2955
1	1	526.7600	1235.1200	371.3863	1102.3653
0	0	594.3600	1181.6300	491.9644	1294.8621
0	0	443.6900	1300.6100	497.0892	1271.2183
0	0	597.2000	1465.0700	555.7086	1341.8809
1	0	546.0100	1497.0500	476.3101	878.0547
1	0	480.4500	1316.3900	397.8660	1099.8688
0	1	378.6100	1219.5600	412.8643	950.9305
1	1	518.3300	1521.5600	561.0805	1121.9023
0	1	471.6900	1028.4900	587.6252	1093.2226
1	0	503.7900	1330.5200	509.7884	1093.1565
1	1	485.5200	872.8400	475.6626	1119.4988
0	1	368.5500	1252.7200	539.0616	1351.5073
1	0	676.1800	1652.2600	611.7734	1299.2261
1	0	528.1300	1318.2800	528.3772	1267.4939
0	1	454.8200	1223.0200	361.0217	765.5149
0	1	415.2200	1166.7200	340.2120	1128.0408
1	1	495.2000	1253.5000	507.4785	1427.8009
1	0	631.0800	1212.7100	648.9842	1353.4810
0	1	450.6300	883.6500	302.5341	1036.0298
1	1	515.6400	1038.3600	529.3986	1187.7625
0	0	339.3100	918.3800	580.4005	861.9066
1	0	467.4300	989.3000	537.7450	1193.0941
1	1	555.3500	1347.9100	460.9777	1277.1995
1	0	604.2600	1541.7800	468.5067	1419.7107
1	0	519.2400	1429.8500	561.6572	1266.0748
1	1	436.3900	733.3200	499.6173	1291.7136
1	1	408.3600	1211.0600	555.4182	914.2747
1	0	475.3800	921.5900	474.6856	1179.2390
1	0	686.9900	1629.1500	496.7289	1275.3090
1	0	431.8600	955.9400	527.3035	1261.7240
1	0	405.8100	1283.3900	500.5734	1274.2354
1	0	637.7700	1070.3400	418.3473	1114.4093
0	0	554.5700	1189.4900	634.0938	1273.2714
0	1	466.7700	862.2300	462.9712	1311.1441
1	0	419.4800	1349.0500	607.6823	1718.6883
1	0	430.4400	1147.7900	603.7399	1420.8918
1	0	478.8200	1519.4800	520.6576	1601.5956
1	0	463.5500	1424.8100	533.0257	1270.0153
1	0	643.9300	1512.4100	550.5419	892.2651
1	1	481.9900	1112.7500	607.2267	1187.1399
1	0	529.6700	1242.8000	461.4771	1526.3195
1	0	609.9600	1326.4100	478.3422	1326.5562
0	0	546.9800	1093.4500	422.6095	1070.7836
1	0	370.1500	1409.4600	503.8587	1258.2023
0	0	520.4600	949.6600	489.4743	1003.3982

Table 2. Results of Regression Analysis to Predict X_2 (Y_2)

	Original data	Perturbed data
Regression statistics		
R-Square	0.2370	0.2370
Standard error of model	192.6845	192.6845
Overall significance of model	0.0057	0.0057
Coefficient estimates		
Estimate of intercept	767.8866	767.8866
(Standard error of estimate)	(184.9393)	(184.9393)
Estimate of coefficient of S_1	78.3935	78.3935
(Standard error of estimate)	(61.5696)	(61.5696)
Estimate of coefficient of S_2	-78.2139	-78.2139
(Standard error of estimate)	(59.1628)	(59.1628)
Estimate of coefficient of X_1 (Y_1)	0.8603	0.8603
(Standard error of estimate)	(0.3572)	(0.3572)

4. An Experimental Comparison of Multiple Imputation and Sufficient Perturbation

In this section, we describe a simulation experiment conducted to evaluate the relative performance of multiple imputation and the sufficiency-based perturbation approach. First, we briefly describe multiple imputation. Assume that the data set is a finite population of size N, consisting of two sets of variables $\boldsymbol{S}$ and $\boldsymbol{X}$, where $\boldsymbol{S}$ represents the design variables and is observed for the entire data set. $\boldsymbol{X}$ represents the survey variables of interest. Let $\boldsymbol{X_{obs}}$ represent the observed portion of the n sampled units ($n << N$). Using this information, the agency releasing the data imputes $\boldsymbol{X_{nobs}}$, the missing values for the $N - n$ units using the posterior predictive distribution of $(X|S,\ Y_{obs})$, so that a complete data set is created. A random sample (of size, say, n_1) is then selected from the $(N - n)$ imputed values. This process is repeated M times to generate M synthetic data sets. The agency then releases the M data sets of size n_1. The user analyzes each of the M data sets using traditional complete data techniques and estimates the population parameter Q with some estimator q_i, and the variance of q_i with some estimator v_i. These values are then aggregated as illustrated in Reiter (2002) and can be used for inferences. In addition, since the masked values are generated in the manner described in Equation (2) and since the imputed values are only generated for those observations for which the survey variables are not observed, the imputed values are independent of the original values.

Table 3. Assessment of Disclosure Risk

	Predict X_1 using S_1 and S_2	Predict X_1 using S_1, S_2 and masked data (Y_1 and Y_2)
Estimate of intercept	513.17625	513.17625
Estimate of coefficient of S_1	-8.46898	-8.46898
Estimate of coefficient of S_2	10.40958	10.40958
Estimate of coefficient of $S_1 \times S_2$	-83.70625	-83.70625
Estimate of coefficient of Y_1	N/A	0.00000
Estimate of coefficient of Y_2	N/A	0.00000

In practice, the imputed values are often generated based on a linear model as in Equation (1) (Reiter 2002; Raghunathan et al. 2003). In these cases, the imputed data have similar characteristics to those using the sufficiency-based perturbation approach. Hence, as in the case of perturbation approach, the imputed data set(s) allow the user to draw valid inferences for those analyses for which the mean vector and the covariance matrix are sufficient statistics. Just as with the perturbation approach, when the underlying data set is not normal, multiple imputation will also not maintain the marginal distribution of the variables, nor will it maintain nonlinear relationships among variables.

Note that the values of $\boldsymbol{S}$ must be observed for the entire set of N values (in order to impute the values of $\boldsymbol{Y}$ for the $N - n$ observations). This makes it difficult to use multiple imputation for data sets such as those shown in Table 1. Hence, in this study we use two experiments that lend themselves to multiple imputation. The first example involves simple random sampling of a single variable and focuses on estimating the mean of this sample. The second example involves multiple variables and focuses on the estimate of the regression coefficient from this data. Reiter (2002) used the first example, and Raghunathan et al. (2003) the second example, to illustrate the effectiveness of multiple imputation in these situations. We show that the perturbation procedure suggested in this study performs better than multiple imputation for the same examples.

4.1. Experiment 1

The first experiment involves the release of a single variable X and no nonconfidential variables. The parameter of interest is the population mean (μ) estimated by constructing a 95 percent confidence interval. The objective of the experiment is to evaluate the effectiveness of the masking procedure in estimating the population mean. It is assumed that the population is of size N and is of normal distribution with mean 0 and variance 100. A simple random sample of size $n = 100$ is drawn from this population and represents the original confidential data (X). The mean ($\bar{x}$) and variance ($\hat{\sigma}^2$) of the sample were computed. The multiple imputation procedure was implemented as follows.

- Draw a random variable g from χ^2_{n-1} and compute $\sigma^2_* = \hat{\sigma}^2 \frac{(n-1)}{g}$
- Generate $\mu*$ from a multivariate normal distribution with mean $\bar{x}$ and variance σ^2_*/n
- Generate $N = 1,000$ observations from a normal distribution with mean μ^* and variance σ^2_*/n
- Obtain a simple random sample of size $n (= 100)$ from N
- Repeat steps 1-4 M times to obtain the M imputed samples

The parameter of interest in this case is the population mean μ estimated through a 95 percent confidence interval. For each imputed sample l, the sample mean ($q^{(l)}$) and the variance of the sample mean ($v^{(l)} =$ imputed sample variance/n) are computed. Using this information, the following quantities are computed:

$$\bar{q}_n = \sum_l \frac{q^{(l)}}{M}; \; b_M = \sum_l \frac{(q^{(l)} - \bar{q}_M)^2}{M-1}; \; \bar{v}_M = \sum_l \frac{v^{(l)}}{M}; \; T_S = (1 + M^{-1})b_M - \bar{v}_M$$

Reiter (2002) indicates that T_S could be negative, and in these cases, $T_S = \bar{v}_M$. The 95 percent confidence interval for the mean is then computed as:

$$\bar{q}_M \pm t_{v_S}(0.025)\sqrt{T_S}$$

where $v_s = (1 + M^{-1})(1 - r_M^{-1})^2$, $r_M = (1 + M^{-1})b_M/\bar{v}_M$, and t_k represents the t-distribution with k degrees of freedom.[4]

The implementation of the perturbation approach in this case is simplified since the data consists of a single confidential variable and no nonconfidential variables. The perturbation is implemented as follows:

- Generate n observations from a normal distribution with mean 0 and variance 1. Let this be A.
- Regress A on X and compute the residuals A^*.
- Normalize A^* to have mean 0 and variance 1.
- For $i = 1$ to n, compute $y_i = a_i^* \hat{\sigma} + \bar{x}$. $Y(= y_1, y_2, \ldots, y_n)$ represents the masked values.

Once the masked values Y were generated, the computation of the confidence intervals was performed exactly as it would have been performed on the original data set.

The entire simulation was then repeated 1,000 times. The results of the simulation, provided in Table 4, are self-explanatory. The average parameter estimate, the average variance of the estimator, and the 95 percent confidence interval coverage for *the original data and the perturbed data are identical*. The results for the perturbation approach are not the results of "averaging" over the 1,000 replications as is the case with multiple imputation. For each replication, the results obtained from perturbation are *exactly the same as those obtained from the original data*. This is not true for multiple imputation. The average standard error for the multiple imputation is slightly higher and the 95 percent confidence interval coverage is also slightly higher.

Figure 1 shows the original sample estimates on the X-axis and the perturbed sample estimates as well as the multiply-imputed sample estimates on the Y-axis. It is easy to see that the original estimates and the perturbed estimates fall exactly on a 45-degree line. By contrast, the estimates from the multiply-imputed sample are different from that of the original mean, even with a large number of imputed samples (100).

In order to illustrate the effect of decreasing the number of imputed samples, Figure 2 provides the results of the same experiment above, but with $M = 10$, instead of 100. Comparing Figures 1 and 2, obviously the sample estimates show considerable increase in variability when $M = 10$, compared with when $M = 100$. In practice however, analyzing all 100 imputed samples would require more effort than analyzing only 10 imputed samples. Figure 3 provides the results of a similar analysis performed for $n = 30$ for comparison purposes.

Table 5 provides the results of imputation and perturbation for several sample sizes and several specifications for the number of imputed samples. In each case, we provide the

[4] Reiter (2002) uses two different estimation procedures and concludes that the procedure described here provides more efficient results that the other procedure. Hence, we use only this procedure.

Table 4. Results for Experiment 1 ($n = 100$ and $m = 100$)

Method	Average parameter estimate of mean	Average variance of estimate	Confidence interval coverage
Original data	−0.00161	0.09976	94.30%
Perturbed data	−0.00161	0.09976	94.30%
Multiply imputed data	0.00094	0.10059	94.70%

estimate and its standard error for both the original sample and the multiply-imputed samples averaged over 1,000 replications. The results in Table 5 indicate that it is necessary for the number of imputed samples to be rather large in order for the estimates to be accurate. If the number of imputed samples is small (say 3 or 5), the standard errors of the estimate from using the multiply-imputed data are much higher than those of the original sample.

We also repeated this experiment by generating the confidential values from several nonnormal populations. The results of the experiment for nonnormal data are similar to those observed for normal data. This is not surprising: the confidence interval estimate for the mean is unaffected by the characteristics of the underlying population. Further, both imputation and perturbation use the same basic approach (linear models) to generate the masked data. Hence, the results observed in the experiments above can be generalized to all data sets.

4.2. Experiment 2

In this experiment, we consider a 5-variate data set having a multivariate normal distribution with means equal to 0, variances equal to 1, and a common correlation equal to 0.5 (Raghunathan et al. 2003 (p. 5, Section 3.1). In this experiment, a 95 percent confidence interval was constructed for the regression coefficient of X_2 when regressing X_1 on

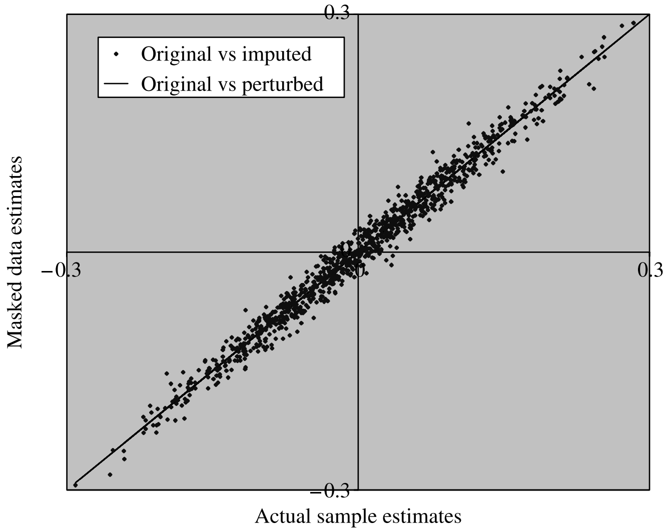

Fig. 1. Results for Example 1 for m = 100

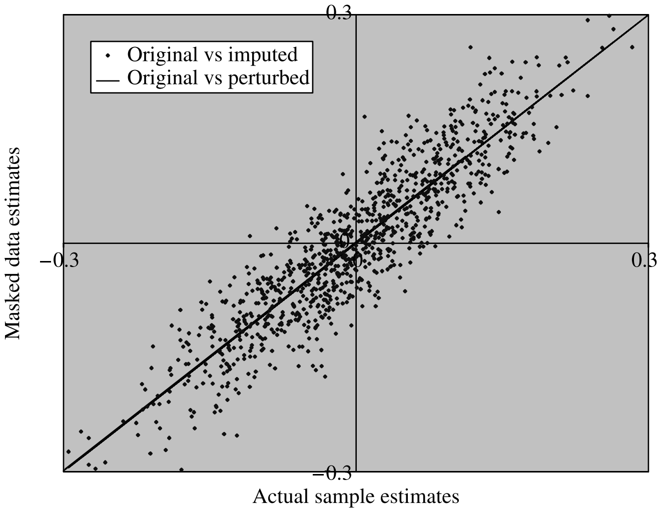

Fig. 2. Results for Example 1 for m = 10

(X_2, X_3, X_4, and X_5). First, a data set of size n was generated to represent the original data. The mean vector ($\bar{\boldsymbol{x}}$) and covariance matrix ($\hat{\boldsymbol{\Sigma}}$) of this data set were computed.

In generating the masked values for imputation, it was assumed that the data was to follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The multiple imputation procedure was implemented as described in Raghunathan et al. (2003), as follows.

- Generate a random variate, W, from a Wishart distribution with $n - 1 = 99$ degrees of freedom and associated matrix $\hat{\boldsymbol{\Sigma}}^{-1}(n-1)$ and let $\boldsymbol{\Sigma}^* = \boldsymbol{W}^{-1}$.
- Generate $\boldsymbol{\mu}^*$ from a multivariate normal distribution with mean $\bar{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{\Sigma}^*/n$.

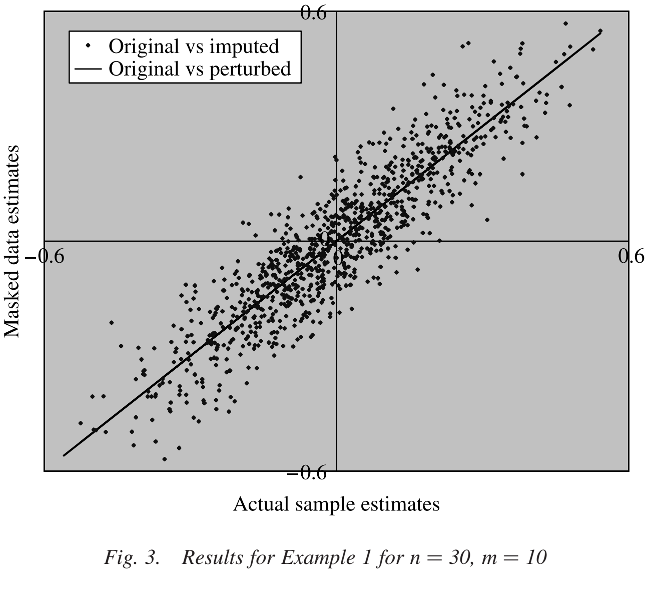

Fig. 3. Results for Example 1 for n = 30, m = 10

Table 5. Results of Experiment 1 for Different Sample Sizes and Different Numbers of Imputed Samples

	Perturbed			Original			Imputed			
n	Estimate	Standard error	Confidence interval coverage	Estimate	Standard error	Confidence interval coverage	Number of imputed samples (*M*)	Estimate	Standard error	Confidence interval coverage
30	− 0.00410	0.18100	94.7%	− 0.00410	0.18100	94.7%	3	− 0.00242	0.24243	97.4%
							5	− 0.01089	0.23008	94.7%
							10	− 0.01010	0.20118	92.4%
							50	0.00045	0.18817	93.7%
							100	0.00803	0.18879	93.4%
50	0.00044	0.14013	93.9%	0.00044	0.14013	93.9%	3	− 0.00306	0.18856	98.5%
							5	− 0.01107	0.17297	94.0%
							10	− 0.00465	0.15326	90.2%
							50	0.00262	0.14508	92.7%
							100	–0.00724	0.14325	93.3%
100	− 0.00161	0.09976	94.3%	− 0.00161	0.09976	94.3%	3	0.00053	0.13415	98.5%
							5	− 0.00098	0.11841	94.1%
							10	0.00392	0.11122	92.6%
							50	0.00014	0.10169	92.5%
							100	0.00094	0.10059	94.7%
500	0.00153	0.04472	94.1%	0.00153	0.04472	94.1%	3	0.00243	0.05861	98.3%
							5	0.00103	0.05268	93.6%
							10	− 0.00183	0.04742	93.9%
							50	0.00129	0.04495	92.7%
							100	0.00093	0.04499	95.6%

- Generate $N = 1,000$ independent multivariate normal random vectors with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$.
- Obtain a simple random sample of size $n(= 100)$ from N.
- Repeat steps 1-4 M times to obtain the imputed samples.

For each imputed sample, the regression coefficient ($q^{(l)}$) and the variance of the regression coefficient ($v^{(l)}$) were computed. From these computations, the average of the estimates

$$\bar{q}_M = \sum_l \frac{q^{(l)}}{M}$$

was computed as the posterior mean of the parameter Q, and

$$T_M = (1 + M^{-1})b_M - \bar{v}_M$$

where $\bar{v}_M = \sum_l \frac{v^{(l)}}{M}$ and $b_M = \sum_l \frac{(q^{(l)} - \bar{q}_M)^2}{M-1}$ as the approximate posterior variance. The 95 percent confidence interval for the regression coefficient of X_2 was then computed as $\bar{q}_M \pm 1.96\sqrt{T_M}$.

For the perturbation approach, the masked values were generated as follows:

- Generate n vectors each of size 5 from a(ny) multivariate normal distribution. Let this be $\boldsymbol{A}$ (of dimension $n \times k$).
- Regress $\boldsymbol{A}$ on $\boldsymbol{X}$ and compute the residuals $\boldsymbol{R}$. Note that $\boldsymbol{R}$ is orthogonal to $\boldsymbol{X}$ and has mean vector $\mathbf{0}$.
- Compute $\boldsymbol{\Sigma}_{\boldsymbol{RR}}$ the covariance matrix of $\boldsymbol{R}$.
- Compute $\boldsymbol{R}^* = (\Sigma_{\boldsymbol{RR}})^{-0.5}\boldsymbol{R}$. $\boldsymbol{R}^*$ has mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$.
- Compute $\mathbf{Y} = \hat{\Sigma}^{0.5}\boldsymbol{R}^* + \bar{\boldsymbol{x}}^T\mathbf{1}$ where $\mathbf{1}$ is an $(n \times 1)$ matrix of 1's.

The mean vector and covariance matrix of the masked data set ($\boldsymbol{Y}$) will be $\bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}$. Y_1 was regressed on the other four variables and the 95 percent confidence interval for the regression coefficient of Y_2 was computed using traditional regression analysis procedures.

Table 6 provides, for several sample sizes, the estimate of the regression coefficient averaged over all (1,000) replications, the standard error of the estimate also averaged over all replications, and the confidence interval coverage. In the case of multiple imputation, the procedure was also performed for several specifications of M (the number of imputed samples generated). These results for all values of M are provided. The results in Table 6 are very similar to those in Tables 4 and 5. It is clear that in all cases the confidence interval estimates using the original and perturbed data are *identical*. For multiple imputation, the estimate of the regression coefficient as well as that of the standard error is very close to, but not the same as, the original data set. Not surprisingly, the performance of the multiple imputation procedure improves (and approaches the estimate of the original sample) as the sample size and/or the number of imputed samples increases.

We repeated the entire experiment using nonnormal populations to generate the original data. We used the procedure suggested in Clemen and Reilly (2002) to generate data from related multivariate nonnormal populations. The results observed for the nonnormal data are very similar to those observed for the normal data. Since the parameter being estimated is not affected by the characteristics of the underlying population, and since the imputed

Table 6. Results of Experiment 2 for Different Sample Sizes and Different Numbers of Imputed Samples

	Perturbed			Original			Imputed			
n	Estimate	Standard error	Confidence interval coverage	Estimate	Standard error	Confidence interval coverage	Number of imputed samples (*M*)	Estimate	Standard error	Confidence interval coverage
30	0.17292	0.19842	95.2%	0.17292	0.19842	95.2%	3	0.18374	0.26846	98.8%
							5	0.15432	0.23281	92.8%
							10	0.18679	0.20731	90.0%
							50	0.18681	0.18591	93.6%
							100	0.18408	0.18630	94.9%
50	0.18897	0.14664	94.2%	0.18897	0.14664	94.2%	3	0.19115	0.19318	98.1%
							5	0.17928	0.17408	95.1%
							10	0.18383	0.15427	90.0%
							50	0.18721	0.14330	92.3%
							100	0.18901	0.14381	92.4%
100	0.18196	0.10098	95.3%	0.18196	0.10098	95.3%	3	0.18621	0.13382	97.5%
							5	0.18456	0.11954	93.2%
							10	0.17956	0.10548	90.2%
							50	0.18415	0.09990	92.4%
							100	0.18181	0.10052	94.7%
500	0.17965	0.04411	94.5%	0.17965	0.04411	94.5%	3	0.18432	0.05829	97.9%
							5	0.18294	0.05410	94.9%
							10	0.18416	0.04737	92.0%
							50	0.18173	0.04433	93.2%
							100	0.17972	0.04350	93.5%

and perturbed values are generated using the same model, it is not surprising that the results observed for nonnormal data are the same as those for normal data. Hence, the results observed in this experiment can be generalized to all data sets.

4.3. Summary of Experimental Results

The results of the experiments indicate that in all cases the sufficiency-based perturbation approach provides estimates with standard errors exactly the same as those using the original data. Multiple imputation always results in estimates with standard error larger than standard error using the original data. In addition, while the experiments above were limited to estimating the mean and a regression coefficient, we can see from the numerical illustration that for all inferential analyses for which the mean vector and covariance matrix are sufficient statistics, the standard error of the perturbed data will be the same as that using the original data, regardless of the characteristics of the underlying data set. Multiple imputation does not offer this guarantee, even when the imputed samples are very large.

5. Why Multiple Imputation?

Rubin (1993, p. 463) provided two important reasons for using multiple imputation in place of perturbation; namely, “that released microdata (1) should look like actual individual microdata in the sense that they must be analyzable using the full range of standard complete data statistical tools, and (2) valid inferences for legitimate estimands should be easily obtainable.” He also argued that perturbation methods (at that time) did not satisfy both of these requirements and hence suggested the use of multiple imputation.

We contend that the criticisms of perturbation are no longer valid. The perturbation method described in this study provides users with microdata that can be analyzed using standard complete-data statistical tools. The method also guarantees valid inferences concerning any statistical model for which the mean vector and covariance matrix are sufficient statistics. In addition, when this perturbation method is employed, the user must analyze only a single data set, unlike the multiple data sets that must be analyzed for multiple imputation. Furthermore, as shown above, the results from multiple imputation only *approach* the results observed using the original data set. By contrast, the perturbed data set provides *exactly the same* results as the original data set.

In his paper, Rubin (1993) states that users should be told:

> Although valid inferences will be obtained, the standard errors will be larger than those from actual microdata because there is a reduction in information relative to the actual microdata, and this is reflected by the between imputation variability (page 463).

Note that for statistical models for which the mean vector and covariance matrix are sufficient statistics, the standard errors of the perturbed data set will be exactly the same as those of the original data set (see Tables 2 and 3) when sufficiency-based perturbation is used. In other words, there is no reduction in information relative to the actual microdata.

The perturbation approach also provides the same disclosure risk characteristics as multiple imputation. In both cases, the masked data set is generated independently of the original data set. Consequently, an intruder would not be able to predict either the identity or the original value of a variable for a given individual using the masked data. In other words, both procedures minimize disclosure risk.

It is also important to note that perturbation and multiple imputation share many assumptions. Multiple imputation assumes that users will employ certain statistical models and analyses. In practice, the assumption for generating the imputed values is the linear model shown in (1). This same is true for the perturbation procedure. We assume that users will employ linear models for analyzing the data and reproduce the sufficient statistics for such analysis.

In addition, multiple imputation also assumes that the data set consists of a population of size N, and that the administrative variables for all N observations are known, while the survey variables for the n sampled observations are known. This is a critical assumption for multiple imputation that allows for the generation of the "missing values" of the survey variables for the $N - n$ observations. The assumption involves both to advantages and disadvantages. When satisfied, the assumption allows for the release of only the synthetic data (i.e., none of the values of the original n survey variables are ever released). Furthermore, since $N >> n$, it is possible to release a data set whose size is actually larger than that of the original sample. However, this assumption also implies that when administrative variables are not available for the nonsurveyed observations (such as for the data presented in Table 1), it will not be possible to release imputed data. One of the key aspects of multiple imputation is that, in order to minimize disclosure risk, only synthetic data will be released, as evidenced by the statement, "Your data will be used to create only synthetic data for public-use; none of your data values will ever be released" (Rubin 1993, p. 463). Hence, if the administrative data for the nonsurveyed observations are not available, it is not possible to generate the "synthetic" data set.

In summary:

- The perturbation approach guarantees that, for inferential analysis for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the perturbed data would be identical to the results using the original data (resulting in increased data utility). Multiple imputation does not offer this guarantee.
- With the perturbation approach, the user needs only to analyze a single data set. With multiple imputation, this analysis must be repeated M times.
- The user who analyzes the perturbed data in exactly the same manner as the original data reaches exactly the same inference using the masked data as the original data. This is not true for multiple imputation.
- In terms of disclosure risk, perturbation performs as well as multiple imputation.
- In order to implement multiple imputation, it is necessary that administrative variables be observed for a larger set N and that the survey variables be observed for a smaller set n. This may not always be possible. By contrast, the perturbed data can be generated using only the smaller data set n.

6. Limitations

The sufficiency-based perturbation approach represents the first step towards developing a perturbation procedure capable of providing a complete solution to the masking problem where the response to any arbitrary query using the masked data is identical to that using the original data. This approach cannot maintain the marginal distribution of nonnormal confidential variables or relationships that are nonlinear. (For that matter, neither can multiple imputation.) There are other procedures that are capable of maintaining nonnormal marginal distributions and monotonic relationships (Sarathy et al. 2002), but they do not maintain sufficient statistics. In addition, in practice, the use of the perturbation procedure or multiple imputation may result in values that are unacceptably small, negative, or large. In these cases, it may be necessary to modify the procedure to eliminate the occurrence of such values. Further, for any model-based approach, complex sample designs could pose a problem in generating perturbed values. Finally, this study focuses on the use of the perturbation approach for linear models with continuous confidential variables, but this procedure cannot be used for categorical data. The development of a perturbation procedure that overcomes all these problems and is applicable to all types of data remains a future research challenge.

7. Conclusions

Rubin (1993) proposed multiple imputation as an alternative to perturbation approaches since, at that time, perturbed data did not meet the requirements that "they must be analyzable using the full range of standard complete-data statistical tools, and valid inferences for legitimate estimands should be easily obtainable" (p. 463). In this study, we provide a modified perturbation procedure that meets these requirements. This procedure also minimizes disclosure risk.

Using empirical experiments, we show that the perturbation method performs better than the multiple imputation method. We can further generalize these results to those cases where multiple imputation uses linear models. In these cases, unlike multiple imputation, the perturbation approach will provide results identical to those using the original. In addition, the perturbed data requires the user to analyze a single data set, whereas multiple imputation requires the user to analyze a large number of data sets (in some cases as many as 100). The performance of the multiple imputation procedure is also shown to be related to both the size of the original sample and to the number of imputed samples. When the original sample is small and/or the imputed samples are small, multiple imputation performs unfavorably compared with the perturbation method. In conclusion, for the specific purpose of masking numerical confidential variables, the results of this study provide strong evidence that the sufficiency-based perturbation approach is to be preferred to multiple imputation.

8. References

Burridge, J. (2003). Information Preserving Statistical Obfuscation. Statistics and Computing, 13, 321–327.

Clemen, R.T. and Reilly, T. (1999). Correlations and Copulas for Decision and Risk Analysis. Management Science, 45, 208–224.

Citteur, C.A.W. and Willenborg, L.C.R.J. (1993). Public Use Microdata Files: Current Practices at National Statistical Bureaus. Journal of Official Statistics, 9, 783–794.

Dalenius, T. (1997). Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 5, 429–444.

Duncan, G.T. and Lambert, D. (1986). Disclosure Limited Data Dissemination. Journal of the American Statistical Association, 81, 10–18.

Duncan, G.T. and Pearson, R.W. (1991). Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. Statistical Science, 6, 219–239.

Fienberg, S.E., Makov, U.E., and Steele, R.J. (1998). Disclosure Limitation Using Perturbation and Related Methods for Categorical Data. Journal of Official Statistics, 14, 485–502.

Fuller, W.A. (1993). Masking Procedures for Microdata Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. Proceedings of the American Statistical Association, Section on Survey Research Methods, Washington, D.C., 370–374.

Lehmann, E.L. and Casella, G. (1998). Theory of Point Estimation. New York: Springer Verlag.

Moore, R.A. (1996). Controlled Data Swapping for Masking Public Use Microdata Sets. U.S. Census Bureau Research Report RR96/04.

Muralidhar, K., Parsa R., and Sarathy R. (1999). A General Additive Data Perturbation Method for Database Security. Management Science, 45, 1399–1415.

Muralidhar, K. and Sarathy, R. (2003). A Theoretical Basis for Perturbation Methods. Statistics and Computing, 13, 329–335.

Muralidhar, K., Sarathy, R., and Parsa, R. (2001). An Improved Security Requirement for Data Perturbation with Implications for E-Commerce. Decision Sciences, 32, 683–698.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.

Reiter, J.P. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. Journal of Official Statistics, 18, 531–543.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Sarathy, R., Muralidhar, K., and Parsa, R. (2002). Perturbing Non-Normal Confidential Variables: The Copula Approach. Management Science, 48, 1613–1627.

Tendick, P. (1991). Optimal Noise Addition for Preserving Confidentiality in Multivariate Data. Journal of Statistical Planning and Inference, 27, 341–353.

Tendick, P. and Matloff, N. (1994). A Modified Random Perturbation Method for Database Security. ACM Transactions on Database Systems, 19, 47–63.

Willenborg, L. and de Waal, T. 2001. Elements of Statistical Disclosure Control. New York: Springer Verlag.

Winkler, W.E. (2002). Single Ranking Micro-aggregation and Re-identification. U.S. Census Bureau Research Report RRS2002/08.

Received February 2004
Revised May 2006