WP.5
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

NUMERICAL DATA MASKING TECHNIQUES FOR MAINTAINING SUB-DOMAIN CHARACTERISTICS

Invited Paper

Prepared by Krish Muralidhar, University of Kentucky and
Rathindra Sarathy, Oklahoma State University

Numerical Data Masking Techniques for Maintaining Sub-Domain Characteristics

Krish Muralidhar
Gatton Research Professor
University Of Kentucky
Lexington KY 40513

Rathindra Sarathy
Ardmore Professor
Oklahoma State University
Stillwater OK 74078

1. Introduction

In a recent paper, Muralidhar and Sarathy (2007a) showed that data shuffling and sufficiency-based linear models performed better than other techniques for masking numerical data. This conclusion was based on assessment of both data utility and disclosure risk. Specifically, in terms of data utility, data shuffling (Muralidhar and Sarathy 2006) was shown to: (a) maintain the marginal distribution of all the confidential variables to be the same after data masking, and (b) maintain monotonic relationships between the variables. The sufficiency-based linear models approach (Burridge 2003, Muralidhar and Sarathy 2007b) was shown to maintain the mean vector and covariance matrix of the masked data to be identical to that of the original data. In terms of disclosure risk, both methods were shown to minimize disclosure by ensuring that, given the non-confidential variables, the original and masked data were independent.

In addition to the traditionally used measures for assessing data utility, one of the desirable properties of masking techniques is that they maintain the characteristics of the data not just in the overall data set, but also in sub-domains of data (Winkler 2006). For numerical variables, it is possible to generate an infinite number of possible sub-groups and it becomes difficult to evaluate all possible sub-groups. However, categorical non-confidential variables usually result in a finite number of sub-domains. Furthermore, data consisting of both categorical and numerical variables are very common in practice. Hence, we evaluate the performance of the two selected techniques (sufficiency-based linear models and data shuffling) on sub-domains. For each sub-domain, we evaluate the extent of information loss and disclosure risk resulting from the masked data.

2. A Brief Introduction to Sufficiency-based Linear Models and Data Shuffling

2.1. Sufficiency-based Linear Models

Masking approaches based on linear models generate the perturbed values **Y** using some variation of the following model:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\beta}_2\mathbf{X} + \boldsymbol{\varepsilon}, \qquad (1)$$

where **S** represents a set of (categorical and numerical) non-confidential variables, **X** represents a set of confidential variables, and **e** represents the noise term. Muralidhar et al. (1999) originally proposed a model of the form as shown in (1), but with the requirements that the covariance matrix of the released data (**S** and **Y**) be the same as that of (**S** and **X**). This specification imposes a specific structure on the covariance of **ε**. In order to improve the disclosure risk characteristics they proposed a modified model of the form

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\varepsilon}. \qquad (2)$$

In equation (2), $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}$, $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}\boldsymbol{\mu}_{\mathbf{S}}$, and $\boldsymbol{\Sigma}_{\mathbf{ee}} = \left(\boldsymbol{\Sigma}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}\boldsymbol{\Sigma}_{\mathbf{SX}}\right)$, where $\boldsymbol{\Sigma}_{\mathbf{XX}}$ is the covariance of **X**, $\boldsymbol{\Sigma}_{\mathbf{SS}}$ is the covariance of **S**, $\boldsymbol{\Sigma}_{\mathbf{XS}}$ is the covariance between (**X** and **S**), and $\boldsymbol{\Sigma}_{\mathbf{ee}}$ is the covariance of the noise term **e**. With these specifications, for large data sets, the mean vector and covariance matrix of the released data (**S** and **Y**) will be the same as that of the original data (**S** and **X**). However, there is some information loss in estimates of the covariance matrix, due to sampling error in smaller data sets. Since **Y** is generated as a function of **S** and **ε** this procedure also minimizes the risk of both identity and value disclosure.

An important variant of the linear model was suggested by Burridge (2003). In this approach, by appropriately generating the values of $\boldsymbol{\varepsilon}$, it is possible to ensure that the mean vector and covariance matrix of the released data are *identical* to that of the original data. Hence, for all statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the masked data will yield *identical results* to that using the original data. That is, the (statistical) information loss will be *zero*. Note that for most traditional statistical analyses (including, but not limited to, comparison of means, ANOVA, regression analysis, and even such multivariate procedures such as canonical correlation analysis), the mean vector and covariance matrix serve as sufficient statistics. Hence, if this procedure is employed to mask the data, a user who analyzes the masked data will get exactly the same results as using the original unmasked data. In addition, this procedure also minimizes disclosure risk.

Muralidhar and Sarathy (2007) recently proposed a further modification of this linear model in equation (1) with the following restrictions:

$$\boldsymbol{\beta_0} = (\mathbf{I} - \boldsymbol{\beta_2})\boldsymbol{\mu_X} - \boldsymbol{\beta_1}\boldsymbol{\mu_S}, \tag{3}$$

$$\boldsymbol{\beta_1} = (\mathbf{I} - \boldsymbol{\beta_2})\boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}, \text{ and} \tag{4}$$

$$\boldsymbol{\Sigma_{\varepsilon\varepsilon}} = \left(\boldsymbol{\Sigma_{XX}} - \boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}\boldsymbol{\Sigma_{SX}}\right) - \boldsymbol{\beta_2}\left(\boldsymbol{\Sigma_{XX}} - \boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}\boldsymbol{\Sigma_{SX}}\right)\boldsymbol{\beta_2^T}. \tag{5}$$

With the above specifications and appropriately selecting the value of $\boldsymbol{\beta_2}$ it is possible to ensure that the masked data (**Y** and **S**) has exactly the same mean vector and covariance matrix as the original data (**X** and **S**). That is, this approach preserves sufficient statistics underlying linear relationships. Consequently, there is zero information loss in estimating any of the linear relationships among the different variables. When $\boldsymbol{\beta_2}$ is zero, this model reduces to that shown in equation (2). For non-zero $\boldsymbol{\beta_2}$ the resulting masked variables *do not* provide the lowest possible level of disclosure risk. However, these masked values may have greater acceptance among users who may have reservations using the more "synthetic" data generated from the model in equation (2).

While the sufficiency-based linear models (SBLM) procedure provides significant advantages over other perturbation procedures, it is not without problems. First and foremost, this procedure results in information loss in the marginal distribution of the masked variable (**Y**). The exact form of the distribution of **Y** would depend on the selection of the distribution for the error term $\boldsymbol{\varepsilon}$. However, unless **X** was normally distributed, the marginal distribution of **Y** will be different from that of **X**. One common problem that arises as a consequence is that the masked values may consist of negative values whereas the original data may be all positive. In addition, while this procedure maintains linear relationships among all variables, non-linear relationships are not preserved in the masked data. Thus, while this procedure is a *complete solution* to the masking problem when the joint distribution of (**X** and **S**) is multivariate normal, resulting in zero information loss and zero incremental disclosure risk, in other cases, it has some shortcomings.

2.2. Data Shuffling

Data shuffling is a new patented procedure (US Patent # 7200757) developed by Muralidhar and Sarathy (2006). It is a hybrid procedure where the original variables are first perturbed using the copula based perturbation approach (Sarathy et al. 2002). The resulting perturbed values are then reverse-mapped on to the original values, resulting in the shuffled data set. Superficially, data shuffling can be considered to be a multivariate version of data swapping since it is performed on the entire data set rather than on a variable by variable basis.

Data shuffling has the following desirable properties. First and foremost, the perturbed values are generated independent of **X** (given **S**) and hence have no incremental disclosure risk. Second, like data swapping, the shuffled values are actually the original values of the confidential variables assigned to a different observation. Hence, the marginal distribution of the masked data is identical to the marginal distribution of the original data. Third, the use of the copula-based perturbation approach enables data shuffling to maintain the rank order correlation of the masked data to be the same as that of the original data. This implies that data shuffling results in minimal information loss in linear and monotonic non-linear relationships among variables. It does not maintain non-monotonic non-linear relationships.

3. Empirical Assessment

We performed an empirical assessment of the two masking techniques using two data sets. The first masking technique used was data shuffling that does not require any parameter specifications. The second masking technique was the SBLM procedure with the requirement that $\boldsymbol{\beta}_2$ be a diagonal matrix with the value d ($0 \leq d \leq 1$) in the diagonal and 0 in the off-diagonal terms. This simple specification implies that when d = 0, the resulting model is the one shown in equation (6) and when d = 1, the entire data set is released unmodified. Thus, the selection of d directly influences the extent to which the original values are used in the masking. Note that when d > 0, this method does not provide minimum security.

3.1. Experimental Assessment Using Simulated Data Set

The first data set was simulated and consisted of 50000 observations. The data consisted of 3 categorical non-confidential variables Gender (male or female), Marital Status (married or other), and Age group (1 to 6). The 3 confidential numerical variables (Home value, Mortgage balance, Total net value of assets) were generated using the NORTA approach for generating related multivariate non-normal variables. Of the three confidential variables, two (Home value and Mortgage balance) had non-normal marginal distributions, while the third had a normal distribution. The relationship between the last two variables was linear while the other relationships were non-linear. Twenty four sub-groups were formed as a combination of the Gender × Martial status × Age group. Data shuffling was applied to the entire data set. In addition, 3 different levels of masking were applied for linear model approach (d = 0.00, 0.50, 0.90). As indicated earlier, when d = 0.00, given the non-confidential variables, the perturbed variables are independent of the original variables and are sometimes considered synthetic data.

3.1.1. Assessment of Disclosure Risk. As indicated earlier, the first step in the assessment of the masking techniques was to compute the risk of identity disclosure for each sub-domain. Table 1 provides the results of the *identity disclosure* (or *re-identification risk*) assessment performed using the procedure suggested by Fuller (1993). There are many approaches for assessing identity disclosure and we could use any one of these procedures. However, the primary objective of this assessment is to compare the different methods rather than assess the extent of disclosure. While the specific results of using another procedure for assessing identity disclosure may be different, the relative performance of the different methods will be the same. Table 2 provides, for each sub-group defined by the categorical variables (a total of 24 sub-groups), the number of observations in each sub-group and the number of observations that were re-identified. As indicated earlier, when shuffling and perturbation with d = 0.00 are used to mask the variables, within a given sub-group, the original and masked variables are independent. Hence, the probability of re-identification within a sub-group is $(1/n_k)$ where n_k is the size of the sub-group. The results in Table 2 clearly show that this is indeed the case. The probability of re-identification is much higher for the other perturbed values, with the higher re-identification occurring when d = 0.90. Thus, in terms of disclosure risk, it is easy to see that the data shuffling and perturbation with d = 0.00 provide the best results, with re-identification occurring by chance alone.

It is also easy to assess the risk of *value disclosure*. As indicated earlier, for a given sub-domain, the shuffled data and perturbed data with d = 0.00 are independent of the original data. This implies that the covariance between the original and masked data are close to zero for shuffled data and exactly 0.00 for the perturbed data with d = 0.00. Hence, the correlation between the original and masked data for these two methods will be 0.00, resulting in no predictive ability. By contrast, for the other two approaches, the correlation between the original and masked variables will be d and the intruder would be able to explain d^2 proportion of the variability in the values of the original variables using the masked variables.

As an illustration, consider the sub-group Gender = 0, Marital = 0, and Age = 1. The mean and standard deviation of the Home value variable in this sub-group are 2.872 and 8.643, respectively. With only this information, for any observation in this sub-set, the best prediction of a 99% interval estimate of the true value of the Home value variable would have an interval of approximately (3 × 8.643). Now assume that the shuffled data is released. The correlation between the original and the shuffled home values is 0.03. Hence, if we perform regression analysis to predict the original value of the confidential variable using the shuffled values, the resulting R^2 would be 0.0009 resulting in a standard error of 8.642. Using this information, a simple 99% confidence interval would have an interval of approximately (3 × 8.642), which for all practical purposes is almost exactly the same as the interval constructed without access to the masked data. In other words, releasing the shuffled data does not allow the intruder to estimate the value of the confidential variable with any greater level of security. Similar results will be observed for the perturbed data when d = 0.00.

Gender	Marital	Age	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	1	1220	3	1	5	40
		2	1181	0	1	13	47
		3	1193	1	1	8	42
		4	1162	3	1	4	39
		5	1159	2	1	5	29
		6	1181	0	1	4	42
	1	1	4672	2	1	7	56
		2	4723	0	1	12	73
		3	4671	1	1	9	54
		4	4719	2	1	5	48
		5	4635	1	1	6	61
		6	4650	2	1	7	58
1	0	1	515	2	1	5	25
		2	468	2	1	6	33
		3	502	3	1	1	30
		4	511	0	1	3	24
		5	503	2	1	2	34
		6	464	0	1	2	21
	1	1	2019	2	1	7	59
		2	1968	0	1	6	43
		3	2044	0	1	3	49
		4	1940	0	1	4	35
		5	1960	1	1	5	52
		6	1940	0	1	4	50

Table 1. Risk of Identity Disclosure (Simulated Data)

The above result does not hold for the other two perturbation parameters (d = 0.50, 0.90). When d = 0.50, if we perform a regression analysis to predict the original Home value variable using the perturbed values, the resulting standard error is 7.485. A 99% confidence interval estimate would have an interval of approximately (3 × 7.485). This implies that the intruder is able to gain a more accurate estimate compared to not having the perturbed values. When d = 0.90, the resulting standard error from the regression analysis is 3.767. If we construct a 99% confidence interval using this information, it results in an interval of approximately (3 × 3.767). Compared to the original interval, the width of this interval is less than 50% of the original width. This allows the intruder to gain a far more accurate estimate of the value of the confidential variable.

Thus, an intruder would have a much better estimate of the original values when the data is masked using the perturbation approach with d = 0.50 and 0.90. In conclusion, when considering disclosure risk, because of their inherent property of conditional independence, data shuffling and perturbation with d = 0.00 perform better than perturbation with d = 0.50 and 0.90. If disclosure risk were the only criterion, data shuffling and perturbation with d = 0.00 would be the preferred methods.

5.1.2. Assessment of Information Loss. In assessing information loss, we focus our attention on sub-domain performance, rather than on the entire data set. We know that, for the entire data set, *data shuffling maintains the marginal distribution of the masked variables to be exactly the same as that of the original variables*. By contrast, the SBLM approach is capable of maintaining the marginal distribution of the variables only when the variable has a normal distribution.

One of the attractive features of the data shuffling procedures is that *the marginal distribution of the shuffled data within any sub-group defined by the non-confidential categorical variables is exactly the same as that of the original variable*. The marginal distribution of the perturbed data are different from that of the original data for sub-groups. To illustrate this, consider the case for the sub-group where Gender = 0, Marital Status = 0, and Age = 1. Figure 1 provides the marginal distribution of the original and the 3 perturbed data sets for the Home value variable. We do not provide the shuffled data since it will coincide exactly with the original data.

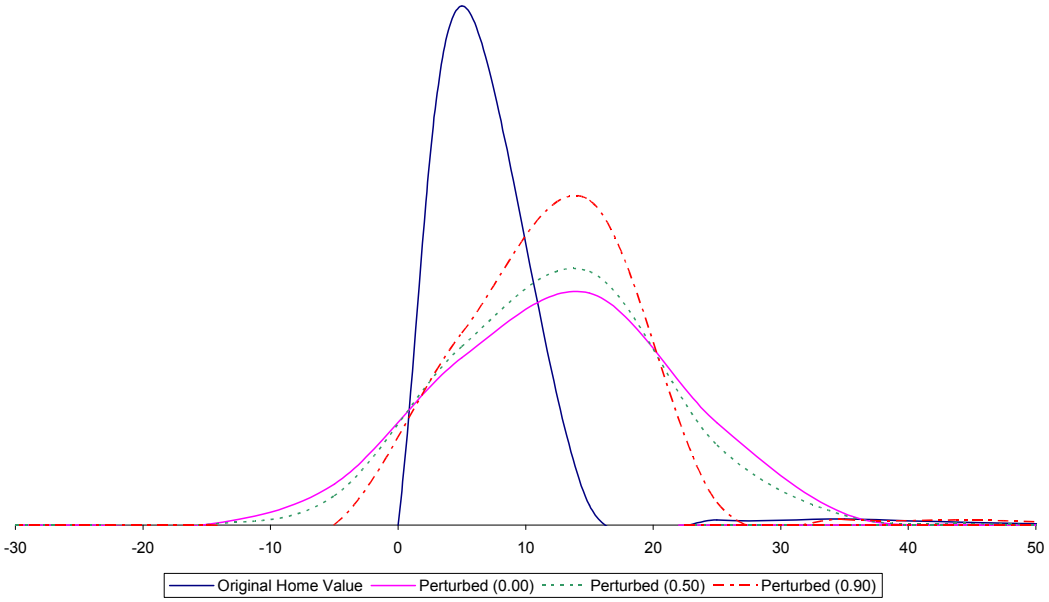

Figure 1. Sub-domain Marginal Distribution of Home Value (Simulated Data)

As can be seen from this example, the marginal distribution of the perturbed data differs considerably from the original data even when d = 0.90. Thus, the "addition of noise" results in a marginal distribution that is closer to normality than the original data. Note that, for the Net Assets variable, the marginal distribution of all the masked variables for all the sub-groups will be similar since the original variable was normally distributed. As discussed earlier, one other attractive feature of all the methods considered in this study is that the mean and variance of all the variables in every sub-group defined by the non-confidential categorical variables will be exactly the same as that of the original data. Hence, we have not provided this data. However, in addition to maintaining the mean and variance, the *shuffled data maintains all the univariate marginal characteristics of the masked data to be the same as that of the original data*.

To assess the extent to which the methods maintain relationships among variables, we computed the product moment correlation between the variables in each sub-group. The results of this analysis are provided in Table 2. As expected, *the product moment correlations of the original data and those of the perturbed data are exactly the same for the data set as a whole and for every sub-group*. The shuffled data does not provide exactly the same results, but *the product moment correlations of the shuffled data and those of the original data are very similar for the data set as a whole and for every sub-group*. Thus, in terms of maintaining product moment correlation, the SBLM approach seems to perform better than the shuffling approach. This is expected since the SBLM approach is intended to maintain first and second order moments (and consequently correlation) among the variables. However, this does not necessarily mean that it is superior to data shuffling as the following discussion shows.

Consider the relationship between the variables Mortgage balance and Net asset value. Figure 2.a provides a scatter plot of the original values with Net asset values on the X-axis and Mortgage balance on Y-axis. It is clear from this figure that the relationship between the two variables is non-linear. In cases where the relationship is non-linear, product moment correlation which measures only the linear relationship is not an appropriate measure. The product moment correlation for these two variables in the data set is 0.719 and all three perturbed values maintain this correlation. By contrast, the correlation between the corresponding shuffled variables slightly different (0.718). Now consider a plot of the perturbed values of Mortgage balance and Net asset values (with d = 0.00) overlaid on top of the original scatter plot (Figure 2.b). Figure 2.b clearly indicates that the perturbation approach has considerably modified the relationship between the variables; the original relationship was non-linear while the perturbed data is almost linear. A plot of the shuffled values of Mortgage balance and Net asset values overlaid on the original data is shown in Figure 2.c. This figure clearly indicates that the shuffled data maintains the (monotonic) non-linear relationship between the two variables better than the perturbed data. Thus, although the SBLM approach maintains the product moment correlation exactly, it does not necessarily maintain non-linear relationships between the variables. By contrast, while the shuffled data does not maintain the product moment correlation exactly, it is capable of maintaining monotonic non-linear correlations much better than the perturbed data.

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.338	0.363	0.338	0.338	0.338	0.373	0.328	0.373	0.373	0.373	0.693	0.701	0.693	0.693	0.693
2	0.216	0.245	0.216	0.216	0.216	0.214	0.274	0.214	0.214	0.214	0.700	0.707	0.700	0.700	0.700
3	0.402	0.306	0.402	0.402	0.402	0.375	0.319	0.375	0.375	0.375	0.707	0.702	0.707	0.707	0.707
4	0.294	0.338	0.294	0.294	0.294	0.282	0.277	0.282	0.282	0.282	0.705	0.698	0.705	0.705	0.705
5	0.201	0.251	0.201	0.201	0.201	0.250	0.267	0.250	0.250	0.250	0.707	0.716	0.707	0.707	0.707
6	0.320	0.355	0.320	0.320	0.320	0.337	0.344	0.337	0.337	0.337	0.746	0.755	0.746	0.746	0.746
7	0.266	0.229	0.266	0.266	0.266	0.230	0.222	0.230	0.230	0.230	0.698	0.695	0.698	0.698	0.698
8	0.313	0.318	0.313	0.313	0.313	0.281	0.292	0.281	0.281	0.281	0.695	0.705	0.695	0.695	0.695
9	0.276	0.253	0.276	0.276	0.276	0.264	0.264	0.264	0.264	0.264	0.694	0.697	0.694	0.694	0.694
10	0.195	0.210	0.195	0.195	0.195	0.179	0.196	0.179	0.179	0.179	0.708	0.708	0.708	0.708	0.708
11	0.284	0.285	0.284	0.284	0.284	0.274	0.261	0.274	0.274	0.274	0.710	0.700	0.710	0.710	0.710
12	0.259	0.250	0.259	0.259	0.259	0.262	0.243	0.262	0.262	0.262	0.728	0.723	0.728	0.728	0.728
13	0.288	0.243	0.288	0.288	0.288	0.244	0.256	0.244	0.244	0.244	0.698	0.712	0.698	0.698	0.698
14	0.321	0.294	0.321	0.321	0.321	0.310	0.351	0.310	0.310	0.310	0.718	0.730	0.718	0.718	0.718
15	0.356	0.371	0.356	0.356	0.356	0.364	0.376	0.364	0.364	0.364	0.705	0.751	0.705	0.705	0.705
16	0.386	0.354	0.386	0.386	0.386	0.329	0.325	0.329	0.329	0.329	0.694	0.664	0.694	0.694	0.694
17	0.387	0.352	0.387	0.387	0.387	0.393	0.418	0.393	0.393	0.393	0.707	0.714	0.707	0.707	0.707
18	0.195	0.208	0.195	0.195	0.195	0.193	0.228	0.193	0.193	0.193	0.737	0.692	0.737	0.737	0.737
19	0.320	0.389	0.320	0.320	0.320	0.338	0.356	0.338	0.338	0.338	0.676	0.662	0.676	0.676	0.676
20	0.349	0.264	0.349	0.349	0.349	0.288	0.259	0.288	0.288	0.288	0.692	0.663	0.692	0.692	0.692
21	0.357	0.303	0.357	0.357	0.357	0.348	0.318	0.348	0.348	0.348	0.703	0.714	0.703	0.703	0.703
22	0.239	0.236	0.239	0.239	0.239	0.216	0.227	0.216	0.216	0.216	0.701	0.710	0.701	0.701	0.701
23	0.289	0.328	0.289	0.289	0.289	0.272	0.292	0.272	0.272	0.272	0.707	0.717	0.707	0.707	0.707
24	0.307	0.263	0.307	0.307	0.307	0.275	0.253	0.275	0.275	0.275	0.721	0.727	0.721	0.721	0.721
Entire Data	0.223	0.220	0.223	0.223	0.223	0.201	0.197	0.201	0.201	0.201	0.719	0.718	0.719	0.719	0.719
Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)															

Table 2. Original and Masked Product Moment Correlation (Simulated Data)

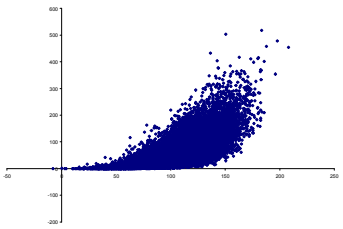

Figure 2.a. Scatter Plot of Net Asset Value and Mortgage Balance (Original Data)

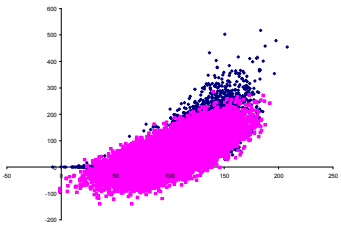

Figure 2.b. Scatter Plot of Net Asset Value and Mortgage Balance (Original and Perturbed)

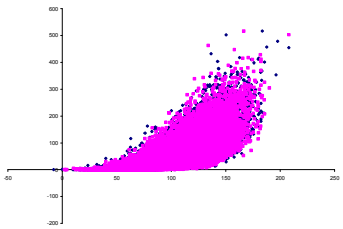

Figure 2.c. Scatter Plot of Net Asset Value and Mortgage Balance (Original and Shuffled)

In situations where the relationship is non-linear, in place of product moment correlation, rank order correlation is used to measure the strength of the relationship. The rank order correlation results, provided in Table 3, indicate that the shuffled data maintain rank order correlation better than the perturbed data. This is to be expected since the shuffling procedure attempts to maintain all monotonic relationships, while the SBLM approach only deals with linear relationships. Note that the shuffling procedure is able to maintain the rank order correlation of the masked data to be very close to that of the original data both at the

overall and sub-group level. This is a significant advantage of the shuffling approach over the SBLM approach. It is also important to note that *any approach based on linear models* (simple additive noise, Kim's method, multiple imputation, among others) are susceptible to the same "linearization" of non-linear relationships. Currently, only data shuffling offers the ability to maintain monotonic relationships among variables.

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.553	0.575	0.329	0.340	0.365	0.624	0.648	0.368	0.368	0.376	0.762	0.786	0.680	0.679	0.708
2	0.527	0.535	0.211	0.216	0.234	0.608	0.625	0.214	0.224	0.259	0.772	0.768	0.680	0.682	0.722
3	0.536	0.539	0.371	0.379	0.388	0.617	0.605	0.357	0.360	0.398	0.764	0.743	0.692	0.696	0.731
4	0.516	0.535	0.275	0.282	0.290	0.603	0.615	0.255	0.255	0.291	0.741	0.744	0.688	0.689	0.718
5	0.540	0.554	0.201	0.205	0.255	0.622	0.647	0.237	0.240	0.285	0.763	0.757	0.692	0.691	0.723
6	0.566	0.566	0.301	0.305	0.315	0.646	0.663	0.322	0.316	0.336	0.792	0.792	0.735	0.737	0.770
7	0.529	0.531	0.252	0.258	0.266	0.623	0.613	0.216	0.224	0.247	0.761	0.767	0.674	0.682	0.707
8	0.523	0.503	0.296	0.300	0.307	0.604	0.591	0.266	0.272	0.297	0.768	0.770	0.677	0.684	0.718
9	0.535	0.532	0.264	0.268	0.282	0.608	0.605	0.253	0.257	0.292	0.759	0.761	0.676	0.679	0.714
10	0.538	0.525	0.183	0.195	0.220	0.613	0.612	0.172	0.183	0.224	0.761	0.759	0.693	0.693	0.724
11	0.539	0.540	0.279	0.285	0.301	0.619	0.621	0.263	0.267	0.297	0.752	0.756	0.692	0.699	0.723
12	0.538	0.532	0.242	0.247	0.273	0.623	0.632	0.245	0.256	0.291	0.768	0.764	0.719	0.718	0.741
13	0.590	0.606	0.287	0.302	0.330	0.669	0.672	0.257	0.273	0.315	0.798	0.816	0.685	0.683	0.705
14	0.510	0.550	0.304	0.310	0.310	0.640	0.652	0.295	0.305	0.321	0.764	0.788	0.704	0.698	0.723
15	0.539	0.552	0.351	0.339	0.342	0.633	0.617	0.349	0.336	0.357	0.752	0.783	0.706	0.689	0.712
16	0.560	0.564	0.378	0.370	0.366	0.598	0.631	0.337	0.317	0.322	0.732	0.742	0.670	0.682	0.692
17	0.523	0.530	0.374	0.365	0.370	0.597	0.603	0.410	0.399	0.412	0.754	0.782	0.678	0.677	0.707
18	0.569	0.542	0.161	0.168	0.217	0.650	0.663	0.170	0.175	0.244	0.763	0.732	0.710	0.713	0.747
19	0.536	0.562	0.314	0.328	0.353	0.638	0.648	0.316	0.332	0.376	0.762	0.758	0.656	0.665	0.697
20	0.538	0.530	0.332	0.330	0.323	0.622	0.607	0.273	0.273	0.284	0.755	0.751	0.676	0.680	0.698
21	0.521	0.525	0.351	0.358	0.364	0.601	0.609	0.328	0.341	0.365	0.759	0.761	0.680	0.685	0.719
22	0.553	0.573	0.228	0.233	0.251	0.638	0.641	0.223	0.226	0.261	0.760	0.766	0.688	0.696	0.727
23	0.521	0.541	0.279	0.291	0.314	0.598	0.610	0.262	0.274	0.316	0.762	0.773	0.696	0.697	0.715
24	0.554	0.566	0.297	0.300	0.294	0.628	0.637	0.261	0.272	0.292	0.767	0.769	0.708	0.715	0.740
Entire Data	0.582	0.583	0.262	0.265	0.283	0.681	0.682	0.255	0.258	0.292	0.782	0.783	0.707	0.711	0.740

Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)

Table 3. Original and Masked Rank Order Correlation (Simulated Data)

5.2. Experimental Assessment Using Census Data

In the previous example, we used a simulated data to highlight the strengths and weaknesses of the two procedures. In this section, we illustrate the applicability of the two procedures to any data set by considering the often used "Census Data". The original Census Data consists of 13 variables and 1080 observations. Of the 13 variables, the variable called PEARNVAL (Total personal earnings) equals PTOTVAL (Personal total income) – POTHVAL (Total other person's income). Hence, rather than using all 3 variables, we only used PEARNVAL in the analysis. Since all 13 of the variables were numerical, in order to illustrate the performance of these procedures for sub-groups, we converted 3 variables (AFLNWGT – Final weight, EMCOMTRB – Employer contribution, and PEARNVAL – Total personal earnings) to categorical variables. For each observation, if the value of each of these variables was less than the average for the entire data set, the value of the corresponding categorical variable was specified as 0 otherwise as 1. This resulted in a total of 8 possible combinations (sub-groups). We used shuffling and perturbation (d = 0.00, 0.50, and 0.90) to mask the data.

5.2.1. Assessment of Disclosure Risk. As before, we assessed identity disclosure risk using the procedure described in Fuller (1993). The results of this assessment are provided in Table 4. As with the simulated data set, it is easy to see that the shuffled data and perturbed data (d = 0.00) provide the lowest risk of identity disclosure, with just one or two records being

identified in each sub-group. The other two perturbed data sets do not fare quite as well. Using the perturbed data with d = 0.50, an intruder could identify a greater proportion of individuals in each sub-group. With the perturbed data with d = 0.90, the level of identity disclosure is extremely high.

Categorical Variable 1	Categorical Variable 2	Categorical Variable 3	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	0	156	4	1	8	83
		1	89	2	1	9	57
	1	0	57	4	1	5	35
		1	156	2	1	7	68
1	0	0	203	4	1	8	90
		1	103	4	1	13	47
	1	0	96	2	1	10	52
		1	220	3	1	10	82

Table 4. Risk of Identity Disclosure (Census Data)

In terms of value disclosure, the width of the confidence interval estimate for the perturbed data with d = 0.00 is exactly 100% of the original width. For the shuffled data, the width of the confidence interval is very close to 100% of the original data. For perturbed data with d = 0.50, the width of the confidence interval is 86.6% $[(1 - 0.5^2)^{0.5}]$ of the width of the original interval. The width of the confidence interval for the perturbed data with d = 0.90 is only 43.6% $[(1 - 0.9^2)^{0.5}]$ of the original width. Thus, the shuffled data and perturbed data with d = 0.00 minimize the risk of value disclosure. The perturbed data with d = 0.50 results in value disclosure which may be considered acceptable. The value disclosure risk resulting from the perturbed data with d = 0.90 is very high and allows the intruder to estimate the values of the confidential variables with much greater accuracy than without access to the data.

5.2.2. Assessment of Information Loss. Figure 3 shows the marginal distribution of the original and perturbed data for the INTVAL variable for the first sub-group (when the value of all the categorical variables is zero). The marginal distribution of the perturbed values are very different from the original values. As with the previous example, there are many negative values while the original variable does not consist of any negative values. While we have limited our discussion to this particular variable for one sub-group, this behavior is observed for practically all variables in all sub-groups. In our opinion, this is a significant problem with the perturbation approach. We also experimented with using alternative distributions for the noise term. The results however are similar to those observed in these cases.

One major advantage of the shuffling approach is that for all variables and all sub-groups, the shuffled data have exactly the same marginal distribution as the original variables. When the data is shuffled, users will be able to analyze individual variables within sub-groups without any information loss. SBLM at least maintains the mean and variance of the variables within the sub-groups. The other procedures (simple additive noise, Kim's method, multiple imputation, micro-aggregation, and swapping) do not typically maintain even mean and variance. Thus, from the perspective of univariate analysis of the masked data for the complete data set and sub-groups, data shuffling provides the best alternative among existing procedures.

As in the previous example, we analyzed both product moment and rank order correlation among the variables. In this case, with as many as 8 variables, there are a total of 21 different correlations to be considered for each of the 8 sub-groups and 4 methods. For the sake of brevity, we did not reproduce the entire set of results. Instead, Table 5 provides the product moment correlation of FICA (Social security deduction) and WSALVAL (Annual total wage and salary). We selected this particular example because of the fact that in one of the sub-groups, the correlation among the two variables is exactly 1.0. The results in Table 5 are similar to those observed for the simulated data. The perturbed data maintains the product moment correlation to be exactly the same for the overall data set and for each sub-group. The product moment correlation of the shuffled data, while very close to the original data, is not exactly the same. As observed earlier however, we do not believe that the product moment correlation is the best method for assessing the relationship among these variables. Hence, we computed the rank order correlation among the variables as an additional measure of information loss. Hence we computed the rank order correlation which is provided in Table 5.

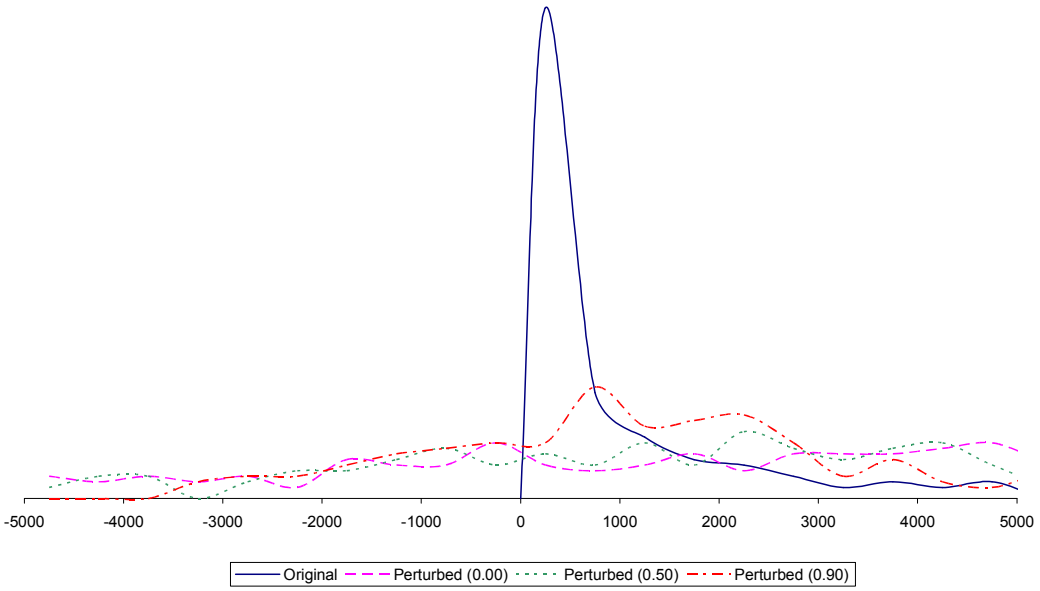

Figure 3. Marginal Distribution of INTVAL Variable for Sub-Domain 1 (Census Data)

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.642	0.800	0.642	0.642	0.642
2	1.000	1.000	1.000	1.000	1.000
3	0.817	0.899	0.817	0.817	0.817
4	0.863	0.915	0.863	0.863	0.863
5	0.529	0.734	0.529	0.529	0.529
6	0.988	0.971	0.988	0.988	0.988
7	0.766	0.885	0.766	0.766	0.766
8	0.929	0.943	0.929	0.929	0.929
All Observations	0.910	0.946	0.910	0.910	0.910

Table 5. Product Moment Correlation between FICA and WSALVAL (Census Data)

The results in Table 7 clearly indicate that the shuffled data maintains the rank order correlation among these two variables better than the perturbed data for the overall data set as well as for practically every sub-group. Note that the data shuffling procedure is able to maintain the perfect correlation among the variables in sub-group 2 as does the perturbed data with d = 0.00. Thus, in addition to maintaining linear correlation, data shuffling performs better in maintaining non-linear relationships among variables while the perturbed data do not.

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.857	0.858	0.597	0.642	0.770
2	1.000	1.000	1.000	0.955	1.000
3	0.876	0.930	0.821	0.779	0.857
4	0.938	0.930	0.834	0.914	0.913
5	0.807	0.787	0.502	0.472	0.653
6	0.975	0.977	0.986	0.879	0.985
7	0.923	0.945	0.737	0.654	0.846
8	0.965	0.948	0.932	0.922	0.954
All Observations	0.953	0.968	0.930	0.919	0.943

Table 6. Rank Order Correlation between FICA and WSALVAL (Census Data)

4. Conclusions

In summary, data shuffling offers the following advantages: (1) Disclosure risk is minimized for every sub-domain, (2) The marginal distribution of the shuffled data is exactly the same as that of the original data for the complete data set as well as for every sub-domain, and (3) The rank order correlation of the shuffled data is very similar to that of the original data for the complete data set as well as for every sub-domain. The SBLM approaches offers the following advantages: (1) Disclosure risk is minimized for every sub-domain for the perturbed data set when d = 0, but not in the other cases, (2) The mean vector and covariance matrix of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-domain. However, the marginal distribution of the perturbed data is different from that of the original data, and (3) The product moment correlation of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-domain. However, the rank order correlation of the perturbed data is very different from the original rank order correlation.

The selection of the specific approach would depend on the characteristics of the data. If the numerical data does not deviate significantly from normality and/or we are only interested in estimating linear relationships among variables, then the SBLM perturbation approach may be preferred since it offers the advantage that the results of traditional statistical analyses conducted on the masked data would yield *exactly* the same results as those using the original data. However, if the data is known to be non-normal and/or we are interested in estimating non-linear monotonic relationships, then shuffling would be preferred since it maintains the marginal distribution *exactly* and is also capable of maintaining monotonic non-linear relationships among variables. In practice, since data sets that exhibit multivariate normality are not very common, data shuffling would generally be the preferred approach.

References

1. Burridge, J. 2003. Information preserving statistical obfuscation. *Statistics and Computing*, 13 321-327.
2. Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.
3. Muralidhar K., R. Parsa, R. Sarathy. 1999. A general additive data perturbation method for database security. *Management Science*, 45 1399-1415.
4. Muralidhar, K. and R. Sarathy. 2006. Data shuffling - A new masking approach for numerical data. *Management Science*, 52(5), 658-670.
5. Muralidhar, K. and R. Sarathy. 2007a. 'Easy to implement' is putting the cart before the horse – Effective techniques for masking numerical data. 2007 Federal Committee On Statistical Methodology Research Conference, Arlington VA, November 5-7.
6. Muralidhar, K. and R. Sarathy. 2007b. Generating Sufficiency Based Non-Synthetic Perturbed Data. Working paper.
7. Winkler, W. 2002. Single-ranking micro-aggregation and re-identification. *Research Report Series (Statistics 2002-08)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2002-08.pdf.
8. Winkler, W. 2006. "Modeling and quality of masked microdata," *Research Report Series (Statistics 2006-01)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf.