Abstract

There is a significant need among government agencies and other organizations for tools and techniques that facilitate analysis, dissemination, and sharing of data without compromising privacy and/or confidentiality. Data perturbation techniques are a set of statistical disclosure limitation techniques that try to preserve confidentiality of sensitive numeric data, while retaining the validity of the protected data for analysis. This paper details the evolution of data perturbation techniques, discusses the statistical basis for each technique and compares their effectiveness in terms or protecting the data from statistical disclosure (data security) while retaining analytical validity (data utility).

1. Introduction

Recent advances in information technology have enabled organizations to gather, store, analyze, disseminate, and share large quantities of data. The reduced cost of hardware, the advent of the internet and e-commerce, and other information technology advances have given organizations a wealth of data about individuals and other entities—information collected from a variety of sources. Most organizations consider this data a rich source of potentially valuable information. In many cases, they attempt to glean information from this data for competitive advantage, often using data mining techniques to identify new and interesting relationships among variables.

Organizations are also realizing that sharing data with their partners has potential benefits. The “merged” data from multiple organizations could provide insights that were not apparent when data from a single source was analyzed. Consider, for example, a large organization (like the now defunct CitiGroup) that offers banking (CitiBank), investment (Solomon Smith Barney), and insurance (Travelers) services. It is easy to see that by creating a large customer database, the organization may analyze the combined data. Similar examples can be drawn from the healthcare industry where a hospital and an insurance firm share data in order to provide better services. Some provisions of the Patriot Act explicitly allow agencies to share information in order to improve their analytical capabilities. Recently, the National Institutes of Health adopted a requirement that “…investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing.” The recent Health Insurance Portability and Accountability (HIPAA) of 1996 allows “…the flow of health information needed to provide and promote high quality health care and to protect the public’s health and well being.”
Last, but not least, is the issue of data dissemination by government agencies. With the ability to store and organize large quantities of data, government agencies such as the Census Bureau can disseminate data in a variety of forms. Improvements in computer technology have increased government agencies’ abilities to gather and store data exponentially. Simultaneously, users of such data have increased their demands, since it is now possible to disseminate the data in various ways.

At the same time, privacy advocates have reacted to this increase in the ability to analyze, disseminate, and share data by demanding means for preserving the privacy and confidentiality of sensitive data, particularly individual health and financial data. Civil liberties advocates have also raised cautions regarding the unrestricted analysis, dissemination, and sharing of sensitive data. Thus, while organizations and individuals could benefit from analysis, dissemination, and sharing of data, the process could also adversely affect privacy and confidentiality. Hence, the need is dire to develop tools and techniques that facilitate analysis, dissemination, and data sharing without compromising privacy and/or confidentiality.

This dilemma is not new. For several decades, government agencies such as the Census Bureau have faced this issue in disseminating data that they gathered under conditions of anonymity. Considerable research exists on the topic of disseminating useful data while preventing disclosure of sensitive information. For this purpose, a variety of tools and techniques have been developed. The objective of this chapter is to trace the evolution of one particular technique, data perturbation, and evaluate it comparatively. Although data perturbation techniques have been proposed for protecting both numerical and categorical data, we focus on data perturbation for protecting confidential numerical data.

2. Definition of Data Utility and Disclosure Risk for Perturbation Methods
Prior to describing the evolution of the perturbation methods, we first define the problem. Consider an organization that has a data set of size N consisting of a set of K confidential variables \( X \), a set of L nonconfidential variables \( S \), and a set of M identifier or key variables (such as name, social security number, etc.). In the general case, \( X \) and \( S \) can be any type of variables. In this study, we address the particular case where the \( X \) are assumed to be continuous numerical variables, while \( S \) can be continuous, discrete, or categorical. The organization wishes to provide microdata access to users so that they can perform unrestricted analysis on the data. The specific type of access depends on the situation. In some cases (such as with release of data by most government agencies), all identifiers are removed prior to providing access to the data since anonymity is to be preserved. In other organizational contexts, anonymity may not be an issue and it may not be necessary to remove identifiers prior to release. However, in both situations, it is necessary to prevent the disclosure of the true values of \( X \). Hence, the values of the confidential variables \( X \) are perturbed (or masked or modified). Let \( Y \) represent the perturbed variables. Users are provided access to \( S \) and \( Y \), but no access to \( X \). Ideally, the released microdata (\( S \) and \( Y \)) should satisfy two major criteria, namely, maximum data utility and minimum disclosure risk. In this section, we define data utility and disclosure risk.

2.1. Data Utility
In this study, we assume that the objective of perturbation is to provide users with a dataset whose statistical characteristics closely resemble the original dataset; that is, we assume
that the available data set represents a finite population. This assumption is commonly used in
prior literature (Fienberg 1997; Fuller 1993; Muralidhar et al. 1999, 2001; Sarathy et al. 2002;
Willenborg and de Waal 2001). Data utility refers to the extent to which the characteristics of
the released data resemble the original dataset. Ideally, the characteristics of the released data
should be identical to the original data so that results of any analyses performed on the perturbed
data are exactly the same as the results of the analyses performed on the original data. However,
given the nature of perturbation, this may not be possible since modifying individual values will
result in some changes for some arbitrary analysis. More often, a statistically based definition of
data utility is used whereby the results of analyses performed on the perturbed data should be the
same as that using the original dataset, except for sampling error. The larger the dataset, the less
the difference between analyses performed on the original and perturbed datasets.

Mathematically, we can state the ideal data utility requirements in terms of the marginal and
joint distributions as follows:

\[ f(Y, S) = f(X, S). \]  

(1)

where, \( f(.) \) represents the joint pdf. We will base our evaluation of the different methods of
perturbation on this general definition of data utility.

A secondary consideration regarding data utility is the extent to which statistical analyses using
the masked data will provide the same results as those using the original data. This is important
to consider in situations where the primary purpose of the data release is for statistical inference.
Since most perturbation techniques rely on random noise, sampling error plays an important role
in these techniques. Hence, in many cases, evaluating the standard error of at least some known
statistics may be necessary in order to evaluate the inferential effectiveness of the masking
procedure.

Much has been discussed regarding the development of performance measures for assessing data
utility (see, for example, Doming-Ferrer and Torra 2001a, 2001b). Unfortunately, many of these
performance measures are rather ad hoc. Conclusions based on these measures may be subject
to subjective interpretation and, perhaps more importantly, could lead to incorrect assessments of
the techniques. One interesting aspect of this paper is that we show that the evolution of these
techniques clearly demonstrates the superiority of one technique over another without the need
for empirical performance measures of data utility. In fact, with a few minor exceptions, we
show that more recent techniques dominate prior techniques in terms of data utility. As we show
later in this paper, the hierarchy of the techniques can be established theoretically.

2.2. Disclosure Risk

In terms of disclosure risk, the perturbation method should guarantee that a snooper (a devious
user or data spy) would not be able to infer the identity of an individual (in a dataset where the
identity of the individual is protected) or obtain an accurate estimate of the value of a sensitive
variable (Willenborg and de Waal 2001). Dalenius (1977) and Duncan and Lambert (1986)
provided some early and general definitions of disclosure risk. The definition of disclosure risk
by Duncan and Lambert (1986) is based on the gain in knowledge and the subsequent reduction
in uncertainty that results from having access to specific data. Dalenius defines disclosure as
having occurred when released data allows the user to better estimate an unknown confidential
value. Other authors have proposed specific measures of disclosure risk (see, for example, Fuller 1993; Fienberg et al. 1997; Sarathy and Muralidhar 2002) that are practical measurements of the general definitions of Dalenius (1977) and Duncan and Lambert (1986).

Note that even before microdata is released, a snooper can predict confidential variables using relationships between nonconfidential variables and aggregate data (Palley and Simonoff 1987). This disclosure risk can be measured by the amount of information regarding the confidential variables that is available from the nonconfidential variables. When this risk of disclosure is considered high, even aggregate data regarding relationships between variables may not be released. When access to microdata is provided, the snooper may use the masked microdata and nonconfidential variables to predict the original confidential variables (using the distribution of the confidential variables conditioned on both the nonconfidential and masked variables). In other words, the release of masked microdata increases the amount of information available to users and could potentially increase disclosure risk. Hence, in evaluating disclosure risk, we focus on the incremental information that results from the microdata release.

We assume that users have maximum prior information in the form of aggregate data regarding all confidential variables, relationships between confidential and nonconfidential variables, and microdata access to nonconfidential variables. In this case, the maximum information available to the snooper is the distribution of the confidential variables conditioned on the nonconfidential variables. This definition of a snooper is consistent with definitions of a snooper (or intruder) who has “verified information” prior to microdata release (Fienberg 1997; Fuller 1993; Willenborg and de Waal 1996, 2001). Then, our definition of disclosure risk can be described as an increase in identity and value disclosure resulting from the incremental information provided by access to masked microdata, given knowledge of the distribution of the real microdata. This definition of disclosure is consistent with the explicit purpose of perturbation procedures, namely, to provide users with access to masked microdata.

We can formalize this definition of disclosure risk as follows. We assume that users already have the aggregate information regarding \( X \) and \( S \), as well as microdata access to the nonconfidential variables \( S \). Hence, disclosure risk prior to access to perturbed microdata is defined by the ability of a snooper to predict \( X \) using the conditional density \( f(X|S) \). When users are provided access to the perturbed microdata, they have additional information, and would use the conditional density \( f(X|S,Y) \) to predict \( X \). However, if \( f(X|S,Y) = f(X|S) \), then providing access to the masked data does not provide any additional information to the intruder. In other words, given \( S \), the original \( (X) \) and masked \( (Y) \) variables are independent of one another.

One simple approach to achieve this ideal level of data security is to generate the masked values as a function purely of \( S \) and independent of \( X \). Consider the situation where the perturbed values are generated using an arbitrary function of the nonconfidential variables and an independent noise term. In other words,

\[
y_i = u(S = s, \varepsilon = \varepsilon_i), \quad (2)
\]

where \( u(.) \) is some function and \( \varepsilon \) is a noise term independent of all other variables. We can show that for any function \( u(.) \), and therefore for any data set, the perturbed values generated using (2) will satisfy the minimum disclosure risk requirements. When \( y_i \) is generated using (2),

\[
P(X \leq x, Y \leq y | S = s) = P(X \leq x, \varepsilon \leq v(s, y) | S = s) \quad (3)
\]
where \( v(s, y) \) represents the solution of \( y = u(s, e) \) for \( e = e \) (Mood and Graybill, 1963, page 221). Since \( e \) is independent of all other variables, we can write the above equation as:

\[
P(X \leq x, e \leq v(s, y) | S = s) = P(X \leq x | S = s) P(e \leq v(s, y) | S = s) =
\]

\[
P(X \leq x | S = s) P(Y \leq y | S = s).
\]

The above shows that \( f(X, Y | S) = f(X|S) f(Y|S) \), and it follows that \( f(X|S, Y) = f(X|S) \). Hence, any masking procedure that generates the masked values as a function of the nonconfidential variables and random noise (and independent of \( X \)) would provide the lowest level of disclosure risk.

As with the evaluation of data utility, tracing the evolution of masking techniques will clearly show that it is not necessary to evaluate disclosure risk empirically. We can simply check whether a particular masking technique satisfies the conditional independence requirement. If it does, then the corresponding disclosure risk of that technique will be lower than one that does not.

In cases where it is necessary to compare techniques that do not satisfy the conditional independent criterion, we can assess their relative performance using the analytical derivations provided by Fuller (1993) for predicting the value of a confidential variable of a known individual without microdata access (equation 2.5) and with microdata access (equation 2.6). Fuller evaluates the probability that a given released record corresponds to the target record (equation 2.11). Fuller’s expressions show that both identity and value disclosure are a function of the relationship between the original and masked values. Hence, we will use the equations (2.5) and (2.6) in Fuller (1993) to assess disclosure risk. Other measures have been proposed in the literature for both identity disclosure (Fienberg 1997; Winkler 1991, 1993a, 1993b, 2004; Yancey et al. 2002) and value disclosure (Sarathy and Muralidhar 2002). However, since our objective is only to evaluate the relative performance of the perturbation methods, Fuller’s expressions are adequate.

3. The Theoretical Basis for Perturbation Methods

A procedure for generating perturbed microdata values in the general case (where both \( X \) and \( S \) could be any type of variables) that satisfies all the data utility and disclosure risk requirements was provided by Muralidhar and Sarathy (2003):

Generate an observation \( y_i \) from the conditional distribution \( f_{X|S}(X|S = s_i) \) such that, given \( S = s_i \), \( Y \) is independent of \( X \). Thus,

\[
y_i \sim f_{X|S}(X|S = s_i), \text{ and}
\]

\[
f_{X|S|Y}(X,Y|S = s_i) = f_{X|S}(X|S = s_i) f_{Y|S}(Y|S = s_i).
\]

Repeat the process for every observation \( i \) in the dataset.

This approach of using the conditional distribution of \( X|S \) to generate perturbed microdata has been previously investigated, in the context of categorical data by Fienberg et al. (1998), in the context of multiple imputation by Little (1993) and Rubin (1993), and in the context of
evaluating disclosure risk by Willenborg and de Waal (2001). When data is generated in this manner, we can easily show that:

1. The statistical characteristics of $Y$ are the same as that of $X$ (i.e., $f(Y) = f(X)$), and the relationship between $Y$ and $S$ is the same as that between $X$ and $S$ (i.e., $f_{YS}(Y, S) = f_{XS}(X, S)$)
2. Providing access to the perturbed microdata values ($Y$) does not provide the intruder with improved predictive ability (i.e., $f_{XYS}(X, Y) = f_{XS}(X, S)$).

Thus, this approach provides the highest possible level of data utility and lowest possible level of disclosure risk.

The problem with implementing this approach is that, in practice, it will be very difficult (if not impossible) to derive the conditional distribution of a data set and generate the perturbed values from this conditional distribution, except in the special case where the data set has a multivariate normal distribution. Thus, while the description in this section provides the ideal approach for generating perturbed values for any type of data, it is not a practical approach for generating perturbed values. In the following section, we describe the evolution of different methods of perturbation for numerical data.

4. Evolution of Perturbation Methods for Numerical Data

In this section, we trace the evolution of perturbation methods for numerical data. It is our objective to identify and describe the major innovations that occurred in perturbing numerical data. We do not intend to provide a comprehensive literature review but will cite only those references that, in our opinion, provide the best description of the respective technique. The evolution is divided into two major sections, namely, linear perturbation models and nonlinear perturbation models.

4.1. Perturbation Approaches Based on Linear Models

4.1.1. Linear perturbation model 1

The simplest of all perturbation methods is the additive noise method:

$$Y = X + \varepsilon,$$  \hfill (7)

where $\varepsilon$ is random noise. Typically, $\varepsilon$ is assumed to be independent of $X$, with zero mean and covariance matrix $\Sigma_{\varepsilon\varepsilon}$ that has non-zero diagonal terms and zeroes in the off-diagonal terms. This implies that the noise terms are independent of each other as well. The exact source of this particular model is not clear from the literature, although Fuller (1993) provides early references to Steinberg and Pritsker (1967), Bach and Baron (1969) and Clark (1978). The first detailed analysis of this approach was provided by Beck (1980), Traub et al. (1984), and Matloff (1986), and later by Fuller (1993).

4.1.2. Linear perturbation model 2

An important improvement to the basic additive model was proposed by Kim (1986). Kim proposed

$$Y = \beta_0 + \beta_1X + \varepsilon.$$  \hfill (8)

Kim then derived the necessary conditions such that any regression analyses performed on $Y$ will yield the same results as that using $X$. When simplified, Kim’s model is the same as the original additive noise model with the exception that the covariance matrix of the noise terms has a structure of the form:
\[ \Sigma_{ee} = d \Sigma_{XX}, \]  
where \( d \) is a scalar quantity and \( \Sigma_{XX} \) is the covariance matrix of \( X \). We will refer to this model as the Correlated Noise Method. Tendick and Matloff (1994) later showed that this model results in lower value disclosure risk than the original additive noise model. Tendick and Matloff (1994) also proposed a modification to the procedure suggested by Kim (1986) as follows:

\[ Y = (1+d)^{-0.5}(X + \varepsilon), \]

where \( \varepsilon \) had a covariance structure \( \Sigma_{ee} = d \Sigma_{XX} \). The advantage of this approach is that the modifying \( Y \) as shown in (10) results in \( \Sigma_{YY} = \Sigma_{XX} \).

### 4.1.3. Linear perturbation model 3

Franconi and Stander (2002) suggest a “model-based” approach for perturbing numerical data. Note that while this approach is the logical extension of the perturbation procedures thematically, a more general procedure was suggested by Muralidhar et al. (1999) prior to Franconi and Stander (2002). In their study, Franconi and Stander (2002) suggest that the perturbed values for a given confidential variable \( X_i \) be generated as a function of other confidential variables \( X_j (j = 1, 2, \ldots, k; i \neq j) \) and the nonconfidential variables \( S \).

\[ Y_i = \beta_0 + \beta_1 X_j + \beta_2 S + \varepsilon, \quad (i = 1, 2, \ldots, k; j = 1, 2, \ldots, k-1; i \neq j) \]  

Thus, the perturbed values of a given confidential attribute are a linear function of other confidential variables and the nonconfidential variables. The authors also consider alternative specifications for \( \varepsilon \) to increase data utility.

### 4.1.4. Linear perturbation model 4

As discussed earlier, a more general form of the model-based approach was proposed by Muralidhar et al. (1999). This study attempted to provide a general model for additive perturbation methods of the form:

\[ Y = \beta_0 + \beta_1 X + \beta_2 S + \varepsilon. \]  

The authors investigated different forms of the above expression to derive the conditions under which the characteristics of \{Y, S\} are the same as that of \{X, S\}. A further modification of this approach was subsequently presented by the same authors (Muralidhar et al. 2001). In this study, the authors propose the general additive perturbation method (GADP)

\[ Y = \beta_0 + \beta_1 S + \varepsilon, \]

where \( \beta_0 = \mu_X - \Sigma_{xs} \Sigma_{ss}^{-1} \mu_S, \beta_1 = \Sigma_{xs} \Sigma_{ss}^{-1}, \) and \( \varepsilon = (\Sigma_{xx} - \Sigma_{xs} \Sigma_{ss}^{-1} \Sigma_{sx}) \).

With these specifications, the resulting perturbed variable \( Y \) has the same mean vector and covariance matrix as \( X \). The covariance matrix of \{Y, S\} is also the same as that of \{X, S\}. Note that when \( X \) and \( S \) together have a multivariate normal distribution, the values of \( Y \) are generated from the conditional distribution of \( X|S \). Hence, when the entire data set has a normal distribution, this approach provides maximum data utility and minimum disclosure risk.

### 4.1.5. Linear perturbation model 5

One problem with model 4 is that, while it performs extremely well for large data sets, the sampling error for small data sets can particularly impact both utility (for statistical inference) and security. Recently, Burridge (2003) proposed a modification to the GADP approach that overcomes this problem. Specifically, Burridge proposes that the perturbed values be generated from the conditional distribution \( f(Y|S,T) \) where \( T \) represents the sufficient statistics. Although it may be theoretically possible to reproduce the complete sufficient statistics, it is unlikely that we can achieve this in practice. In this particular case, the sufficient statistics that are reproduced are the mean vector and covariance matrix. A further enhancement was proposed by Muralidhar and Sarathy (2005) along the lines suggested by Burridge to ensure that the procedure also minimizes disclosure risk.
4.2. Perturbation Approaches Based on Nonlinear Models

4.2.1. Multiplicative Model
The basic nonlinear perturbation model was the multiplicative model proposed by Hwang (1985) as:
\[ Y = X^* \varepsilon \]  
This model can be expressed as the additive model (see equation (7)) in logarithmic form. Subsequent investigations of this model can be found in Muralidhar et al. (1995) and Kim and Winkler (2003). This model was originally proposed in cases where the original data are skewed and the relationship between the variables is nonlinear. An obvious extension that has not yet been investigated is to use the conditional distribution approach for the log-transformed variables.

4.2.2. Sullivan’s Model
Another type of nonlinear perturbation model was proposed by Sullivan (1989) in cases where the marginal distribution of the variables is not normal. This approach attempts to preserve the marginal distribution of the masked variables to be the same as that of the original variables. While this approach is also applicable to categorical variables, we limit our discussion to its implementation for numerical variables. Sullivan’s approach transforms each observation of every confidential variable into a uniform random variable, using its empirical CDF. The uniform random variable is then retransformed to a standard normal random variable. Let \( x_i^* \) represent the transformed variable. Then
\[ y_i^* = \Phi^{-1}(F_i(x_i)), \quad i = 1, \ldots, n \]  
where \( F_i(.) \) is the empirical CDF of confidential variable \( x_i \), and \( \Phi^{-1} \) is the standard normal inverse. An appropriate level of noise is then added to the standard normal variable to result in \( y_i^* \) as follows:
\[ y_i^* = x_i^* + \varepsilon_i^* \]  
where \( \varepsilon_i^* \) represents the independent noise term. Once noise addition is completed, the entire process is reversed to yield perturbed values that have the same empirical distribution as the original confidential values as follows:
\[ y_i = F_i^{-1}(F_i(y_i^*)) \]  
The empirical nature of this approach makes it difficult to derive the distributional characteristics of the perturbed variables, in the general case.

4.2.3. Copula Model
Sarathy et al. (2002) recently proposed an alternative approach based on approximating the joint distribution of the entire set of variables using the multivariate normal copula. The procedure can be described as follows:

1. Identify the marginal distribution of attributes \( X_1, \ldots, X_n, S_1, \ldots, S_m \).
2. Compute rank order correlation matrix \( (R) \) of the original database.
3. Compute product moment correlation matrix \( \rho \) using \( R \).
4. Construct the multivariate normal copula and compute the new variables \( X^* \) and \( S^* \) from the copula distribution (see equation (19)).
5. Apply GADP (see equation (13)) to variables \( X^* \) and \( S^* \) to generate \( Y^* \).
6. Compute \( Y \) from \( Y^* \) using \( y_j = F_i^{-1}(\Phi(y^*_j)) \). (18)

where
\[
\begin{align*}
x^*_i &= \Phi^{-1}(F_i(x_i)), \ i = 1,\ldots,n \\
s^*_j &= \Phi^{-1}(F_j(s_j)), \ j = 1,\ldots,m, \text{ and} \\
y^*_k &= \Phi^{-1}(F_k(y_k)), \ k = 1,\ldots,n.
\end{align*}
\]

and \( F(.) \) represents the cumulative density function. Sarathy et al. (2002) showed that when this approach is implemented, the marginal distribution of \( Y \) is the same as that of \( X \), and the pairwise rank order correlation matrix of \{\( Y, S \)\} is the same as that of \{\( X, S \)\}. Thus, in addition to preserving the marginal distribution of the variables, this approach preserves all monotonic relationships between variables.

### 4.2.4 Data Shuffling

Muralidhar and Sarathy (2005b) recently proposed a new approach called “data shuffling” for protecting numerical confidential variables that combines the perturbation/swapping approaches. In data swapping, for each confidential variable, pairs of observations are selected and the values of these observations are exchanged (or swapped). The advantage of swapping is that, unlike perturbation approaches where the actual values are modified, the swapped values are not modified. Data shuffling is implemented as follows:

1. Compute rank order correlation matrix \( \mathbf{R} \) of the original database.
2. Compute product moment correlation matrix \( \mathbf{\rho} \) using \( \mathbf{R} \).
3. Compute the rank of individual observations in \( S \).
4. Compute the new variables \( S^* \) using equations (19).
5. Generate \( Y^* \) using equation (13).
6. Replace the values of \( y^*_j \) by \( x_{(i),j}, j = 1, \ldots, M; i = 1,\ldots, N \) to result in \( Y \), where \( x_{(i),j} \) is the observation corresponding to the \( j \)th confidential variable with rank \( i \) (similarly for \( y^*_j \)).

The following section provides a theoretical basis, showing that it is possible to derive the specific data utility and disclosure risk characteristics of data shuffling.

### 5. Evaluation of Perturbation Methods

In this section, we compare the performance measures of the different perturbation methods with respect to data utility and disclosure risk. As discussed earlier, the process of evaluation is performed using the theoretical characteristics of the masking procedure.

#### 5.1 Evaluation of Perturbation Approaches Based on Linear Models

##### 5.1.1 Linear perturbation model 1

We will assume that \( \epsilon \) has mean 0, (non-zero) covariance matrix \( \Sigma_{\epsilon \epsilon} \), and is independent of \( X \). We will not assume any specific distributional characteristics for \( \epsilon \). For large data sets, regardless of the distributional characteristics of \( \epsilon \), since noise is being added to the original variables, the distribution of the \( i \)th variable \( Y_i \) will be different from that of \( X_i \). The mean vector of \( Y \) will be same as that of \( X \) and the variance of \( Y_i \) will be different from the variance of \( X_i \). We can also show that:
(1) The covariance of $Y = \Sigma_{XX} + \Sigma_{\varepsilon\varepsilon}$ is different from the covariance of $X$ ($\Sigma_{XX}$). Hence correlations among the masked variables will be different from those of the original variables.

(2) The covariance between $Y$ and $S$ ($\Sigma_{YS}$) is the same as that covariance between $X$ and $S$ ($\Sigma_{XS}$). However, since $\Sigma_{YY}$ is different from $\Sigma_{XX}$, the correlation between the masked and nonconfidential variables will be different from the correlation between the original confidential and nonconfidential variables.

In addition, since non-zero variance random noise is being added to the original values, the rank order correlation of the masked data will be different from that of the original data. In summary, with the exception of equality in the mean vectors, the simple additive noise model results in information loss for all criteria that we have selected.

In terms of disclosure risk, since this model does not satisfy the conditional independence requirement, we can conclude that it results in higher-than-desirable level of disclosure risk. For the sake of completeness, we now show that is indeed the case. Consider the case where the data set consists of a single confidential variable $X$ and a single nonconfidential variable $S$. Let $\rho$ be the correlation between $X$ and $S$. Then, the proportion of variability in $X$ that is explained using $S$ ($R^2_{X:S}$) is $\rho^2$. Let the proportion of variability in $X$ that is explained by using both $S$ and $Y$ be ($R^2_{X:SY}$). For this perturbation model ($Y = X + \varepsilon$), we can derive:

$$R^2_{X:SY} = \rho^2 + \frac{(1-\rho^2)^2}{(1-\rho^2)+\sigma^2}\quad(20)$$

where $\sigma^2$ represents the variance of $\varepsilon$. Further,

$$R^2_{X:SY} - R^2_{X:S} = \frac{(1-\rho^2)^2}{(1-\rho^2)+\sigma^2}\quad(21)$$

Obviously, $R^2_{X:SY} \geq R^2_{X:S}$ for $\sigma > 0$, resulting in an increased disclosure risk for this model, when microdata is released. This result can easily be generalized for the case where there are multiple nonconfidential and/or confidential variables.

5.1.2 Linear perturbation model 2:

We consider this model superior to the model of the form $Y = (X + \varepsilon)$ with $\Sigma_{\varepsilon\varepsilon} = d\Sigma_{XX}$ since it provides the additional advantage that $\Sigma_{YY} = \Sigma_{XX}$ without compromising any other aspect of data utility or disclosure risk. From a univariate perspective, this model does not satisfy the requirement that the marginal distribution of $Y$ be identical to that of $X$. In addition, since noise is being added to the original values, the empirical distribution of $Y$ is different from that of $X$. While the mean and variance of each individual variable $Y_i$ is the same as that of $X_i$, higher order moments will be different.

Evaluating this model from a multivariate perspective, we can verify that $\Sigma_{YY} = \Sigma_{XX}$ and, hence, linear relationships between the masked variables is the same as the linear relationship between the original variables. Since $\text{Cov}(Y, S) = \text{Cov}((1+d)^{0.5}(X + \varepsilon), S) = (1+d)^{0.5}\text{Cov}(X, S)$, linear relationships between $\{Y, S\}$ are different from that of $\{X, S\}$. In addition, because perturbation is based on a linear model, this procedure will maintain only linear relationships within $Y$ to be the same as that of $S$, but will not maintain any nonlinear relationships.

In terms of disclosure risk, just as with the previous model, we can show that when we consider $X$, $S$, and $Y$, the resulting $R^2_{X:SY}$ is the same as shown in equation (20) and is greater than $R^2_{X:S}$, resulting in increased disclosure risk.
5.1.3 Linear perturbation model 3:
In this model proposed by Franconi and Stander (2002), the level of data utility and disclosure risk would depend on the characteristics of the data set. As indicated earlier, this model is a special case of the model proposed by Muralidhar et al. (1999, 2001). However, for the sake of completeness, we show that this model results in disclosure risk that is higher than that provided by Muralidhar et al. (2001).
Consider a data set consisting of one nonconfidential variable S and two confidential variables X1 and X2. For simplicity and without loss of generality, assume that all three variables have mean 0 and variance 1. Using this model, for the first confidential variable X1, the perturbed values for Y1 would be generated as:
\[ Y_1 = \beta_0 + \beta_1 S + \beta_2 X_2 + \epsilon. \]  
(22)
We now show that that this model increases disclosure risk since the masked values are not generated independent of X. From the above model, the covariance matrix of X1, S, and Y1 can be derived as:

\[
\begin{pmatrix}
X_1 & S & Y_1 \\
X_1 & 1.00 & \beta_1 \rho_{S1} + \beta_2 \rho_{12} \\
S & 1.00 & \beta_1 + \beta_2 \rho_{S2} \\
Y_1 & & \beta_1^2 + \beta_2^2 + \sigma^2 + 2 \beta_1 \beta_2 \rho_{S2}
\end{pmatrix}
\]

where \( \rho_{S1}, \rho_{S2}, \) and \( \rho_{12} \) represent the covariance between (S and X1), (S and X2), and between (X1 and X2), respectively. Now consider the ability of an intruder to predict the confidential variable X1 using only the nonconfidential variable S. The proportion of variability explained in X1 using S can be written as:
\[ R^2_{X_1|S} = (\rho_{S1})^2. \]  
(23)
When the perturbed data Y1 is released, the intruder would use this information to predict the value of the confidential variable (in addition to S). To minimize disclosure risk, it is necessary that \( R^2_{X_1|S,Y} = R^2_{X_1|S} \). In the above example, we can derive
\[ R^2_{X_1|S,Y} - R^2_{X_1|S} = \frac{\beta_2^2 (\rho_{12} - \rho_{S1} \rho_{S2})^2}{|D|} \]  
(24)
where \( |D| \) is the determinant of the covariance matrix of (S,Y1). It is easy to see that the above expression is always positive (both numerator terms are squared terms and \( |D| \) is the determinant of a positive definite matrix). Thus, providing access to the masked microdata increases disclosure regarding the confidential variables. In the above expression, the values of \( \rho_{S1}, \rho_{S2}, \) and \( \rho_{12} \) are data specific and there is no reason to expect the value \( (\rho_{12} - \rho_{S1} \rho_{S2}) \) to equal 0.
Hence, in order for \( R^2_{X_1|S,Y} = R^2_{X_1|S} \), it is necessary that \( \beta_2 = 0 \). We can easily extend this result to the multivariate case involving multiple confidential and nonconfidential variables to show that, to minimize disclosure risk, the perturbed values Y must be generated as a function only of the nonconfidential variables S and independent of the confidential variables X. In addition, in order to maximize data utility, the values of Y must be generated from the conditional distribution of X|S. When both conditions are applied, the model-based approach of Franconi and Stander (2002) reduces to the conditional distribution model proposed by Muralidhar et al. (1999; 2002).

5.1.4 Linear perturbation model 4:
This model generates the perturbed values from the linear model as shown in equation (13). When generated in this manner, it is easy to derive the following (Muralidhar and Sarathy 2003; Muralidhar et al. 1999, 2001):

\[ \mu_Y = \mu_X, \quad \Sigma_{YY} = \Sigma_{XX}, \quad \text{and} \quad \Sigma_{YS} = \Sigma_{XS}. \]

Further,

\[ R_{X|S,Y}^2 = R_{X|S}^2. \]  

(25)

Thus, when the underlying data set has a multivariate normal distribution, the resulting joint distribution of \((Y, S)\) is the same as that of \((X, S)\). However, when the underlying distribution is not normal, this procedure results in modifying the marginal distribution of the confidential variables, causing information loss. In terms of disclosure risk, since the masked variables are generated in a manner satisfying equation (2), the conditional independence requirement is satisfied, and hence this procedure provides the lowest possible level of disclosure risk.

5.1.5 Linear perturbation model 5:

One problem in the procedure suggested by Muralidhar et al. (1999, 2001) is that, for small samples, the noise term that is generated may not have the exact desired mean vector and covariance matrix, resulting in “sampling error.” Burridge’s Information Preserving Statistical Obfuscation (IPSO) (Burridge 2003) overcomes this problem by generating the noise terms such that the mean vector and covariance matrix of the released data \((Y,S)\) are identical to those of \((X,S)\). Burridge’s procedure in essence maintains the necessary sufficient statistics for a multivariate normal distribution. Burridge acknowledges that, for small samples, this procedure may be still vulnerable to disclosure risk. Muralidhar and Sarathy (2005a) suggest a modification to the Burridge procedure that ensures that, regardless of sample size, \(R_{X|S,Y}^2 = R_{X|S}^2\). This is achieved by ensuring that the noise terms that are generated are orthogonal to the original confidential variables \(X\). We refer to this model as the Exact General Additive Data Perturbation (EGADP) model. The procedure for implementing this model would be as follows:

(i) Regress \(X\) on \(S\) and compute \(\hat{\beta}_1 = \hat{\Sigma}_{XS} \hat{\Sigma}_{SS}^{-1}\) and \(\hat{\beta}_0 = \bar{X} - \hat{\Sigma}_{XS} \hat{\Sigma}_{SS}^{-1} \bar{S}\). Compute the covariance of the residuals \(\hat{\Sigma}_{ee}\).

(ii) Generate a \((n \times M)\) matrix of random numbers \(A\) from a standard multivariate normal distribution.

(iii) Regress \(A\) on \((S \text{ and } X)\). Compute the residuals from this regression. Let these residuals be \(B\). The new noise terms \(B\) is orthogonal to both \(X\) and \(S\). In addition, the mean vector of \(B\) is 0.

(iv) Compute the covariance matrix \((\hat{\Sigma}_{BB})\) of \(B\).

(v) Compute a new variable \(C\) where \(c_i = \hat{\Sigma}_{ee}^{0.5} \hat{\Sigma}_{BB}^{-0.5} b_i, \ i = 1, 2, \ldots, n\). Compute

\[ y_i = \hat{\beta}_0 + \hat{\beta}_1 s_i + c_i , \ i = 1, 2, \ldots, n. \]  

(26)

When this model is implemented, for a data set of any size, the following results hold true:

\[ \mu_Y = \mu_X, \quad \Sigma_{YY} = \Sigma_{XX}, \quad \Sigma_{YS} = \Sigma_{XS}, \quad R_{X|S,Y}^2 = R_{X|S}^2. \]  

(27)
where ≡ represents identical to. Such a data set can be released with the explicit assurance that if used for statistical analyses for which the mean vector and covariance matrix are sufficient statistics users are assured of results that are identical to the original data. Finally, given $S$, since the values of $Y$ are generated conditionally independent of $X$, this procedure also provides the lowest possible level of disclosure risk.

5.1.6 Summary of Comparison of Linear perturbation models

It is easy to see from the development of the perturbation approaches using linear models that EGADP dominates the performance of all other models, making it the ideal, and perhaps the only, choice for among linear perturbation models. However, linear perturbation models do cause information loss in other respects. For instance, they modify the marginal distributions of the individual confidential variables. Where it is important to maintain the marginal distributions of variables with non-normal marginal distribution and/or maintain nonlinear relationships, it is necessary to consider alternative perturbation approaches.

5.2 Comparison of Nonlinear perturbation models

5.2.1 Multiplicative Model

A comprehensive discussion of the multiplicative model is found in Muralidhar et al. (1995) and Winkler and Kim (2003). The simple multiplicative model is similar to the original additive model and results in information loss in all aspects of the model. Thus, this model results in:

1. Individual masked variables having a different marginal distribution than the original confidential variables,
2. The covariance matrix of the masked variables is different from the original variables, and
3. The covariance matrix of the masked and nonconfidential variables is different from the original and nonconfidential variables.

In terms of disclosure risk, as in other methods, since the masked values of $Y$ are not generated independent of $X$, we conclude that this method does not result in the lowest possible level of disclosure risk. However, as in the other cases, for completeness we show that this procedure increases disclosure risk. Consider the case with one confidential variable $X$ and a nonconfidential variable $S$. For simplicity and without loss of generality, assume that both $X$ and $S$ have mean 0 and variance 1. Assume that the perturbed values of $Y$ have been generated as $Y = X \times \epsilon$, with $\epsilon$ having mean 1 and a specified variance $\sigma^2$. Let $\rho$ represent the correlation between $X$ and $S$. The covariance matrix of $X$, $S$, and $Y$ can be derived as:

\[
\begin{array}{ccc}
X & S & Y \\
X & 1 & \rho & 1 \\
S & 1 & \rho \\
Y & 1 + \sigma^2 \\
\end{array}
\]

The resulting $R^2_{X|S,Y}$ can be derived as:

\[
R^2_{X|S,Y} = \rho^2 + \frac{(1-\rho^2)^2}{(1-\rho^2) + \sigma^2} \tag{28}
\]

and

\[
R^2_{X|S,Y} - R^2_{X|S} = \frac{(1-\rho^2)^2}{(1-\rho^2) + \sigma^2} \tag{29}
\]
This expression is the same as equation (21) and results in increasing disclosure risk. Thus, in terms of performance, the simple multiplicative approach performs very similarly to the simple additive approach.

5.2.2 Sullivan’s Model
As indicated earlier, given the nature of Sullivan’s model, it is difficult to derive the exact characteristics of the resulting perturbed variable. It is, however, possible to derive the characteristics of the transformed variables \( x_i^* \) and \( y_i^* \). Since \( y_i^* = x_i^* + \epsilon_i^* \), the results derived for the simple linear perturbation model (section 5.1.1) hold true for the transformed variable. In other words, in terms of information loss, when this procedure is used:

1. All relationships among the masked variables are different from those of the original confidential variables, and
2. All relationships among the masked and nonconfidential variables are different from those of the confidential and nonconfidential variables.

The major advantage of this procedure is that it results in the same marginal distribution as the original variables. In terms of disclosure risk, since the masked values of \( Y \) are not independent of \( X \), this procedure does not provide the lowest possible level of disclosure risk. In fact, we can show that when we consider the transformed variables \( x_i^* \) and \( y_i^* \), this procedure increases disclosure risk as shown in equation (21).

5.2.3 Copula Model
The copula model attempts to generate the perturbed values from an assumed conditional distribution \( f_{X|S}(X|S) \) using the multivariate normal copula. In terms of data utility, the copula model maintains the marginal distribution of the masked variables to be the same as that of the original variables. In terms of relationships, this approach maintains both linear and monotonic nonlinear relationships between all variables. However, it does not maintain non-monotonic relationships. In terms of disclosure risk, since the perturbed values are generated independent of the confidential values, we can show that this procedure provides the lowest possible disclosure risk (see Sarathy et al. 2002, section 6, pages 1623-1625).

5.2.4 Data Shuffling
The data shuffling procedure is similar to the copula model since the perturbed values are generated using the copula model. The important aspect of data shuffling is the reverse mapping of the perturbed values with those of the original values. This results in the case where the actual original values are used as the masked values. Thus, in addition to providing the same benefits as the copula model, data shuffling provides the additional advantage that the original confidential values are unmodified. So, for each individual variable, the masked marginal distribution is identical to the original confidential distribution. Like the copula model, this model also maintains linear and monotonic nonlinear (but not non-monotonic) relationships among all variables. In terms of disclosure risk, this procedure also ensures that it provides the lowest possible level of disclosure risk.

5.2.5 Summary of Comparison of Nonlinear perturbation models
As with linear perturbation models, it is easy to see that in terms of information loss, the performance of the simple multiplicative approach is the worst, followed by Sullivan’s model, then by the copula model, and finally, data shuffling. It is also evident that, in terms of disclosure risk, the multiplicative model and Sullivan’s model result in increased disclosure risk, while the copula model and data shuffling do not. When compared to the copula model, data shuffling offers the additional advantage that the original values of the confidential variables are unmodified. In this sense, data shuffling provides greater data utility (and lower information
loss) than the copula model. Thus, data shuffling represents the best alternative among nonlinear perturbation models.

5.3 Comparison of Linear and Nonlinear perturbation models
Among linear perturbation models, EGADP provides data utility that is superior to all other linear perturbation models and disclosure risk that is the lowest among all linear perturbation models. Hence, this model should be preferred over all other linear perturbation models. Similarly, among nonlinear perturbation models, the performance of data shuffling dominates the other approaches and should be preferred. Comparing across models is somewhat more difficult. In terms of disclosure risk, both EGADP and data shuffling provides the lowest possible level of security since they both ensure that given $S$, $X$ and $Y$ are independent. In terms of data utility, data shuffling provides some advantages that the linear perturbation model does not and vice versa. Specifically, data shuffling has the following advantages over EGADP:

1. The original values of the confidential variables are unmodified, and
2. Monotonic nonlinear relationships are preserved.

This would imply that data shuffling would be preferred. However, EGADP provides one significant advantage that data shuffling does not; namely, that for all statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the EGADP perturbed data will be identical to those using the original data. By contrast, it is reasonable to expect that, for large sample sizes, the results of such analyses using data shuffling would be very close (but not identical) to those using the original data.

The choice of the specific procedure will depend on the specific application for which the procedure is being considered. EGADP would be the preferred method if it:

1. is relatively small,
2. will be used mainly for traditional, parametric statistical analyses, and
3. will be used for inferential analyses.

The above characteristics are commonly seen in situations where the data has been collected for experiments and the data will be used for statistical inference of population parameters. By contrast, data shuffling would be the preferred method if the data set:

1. is relatively large,
2. will be used mainly for nontraditional statistical analyses (such as data mining), and
3. consists of important nonlinear relationships.

The above characteristics are commonly seen in data that business organizations have collected and stored in data warehouses. Such data is often used to identify potentially new relationships using nontraditional data mining techniques. In addition, the primary objective of this analysis is not inferential but descriptive, since the available data is often considered the entire population and/or the data sets are so large that practically any difference would be statistically significant. In these cases, data shuffling offers a viable alternative to EGADP.

6. Comparison of Perturbation with other Masking Methods
In this section, we compare perturbation approaches with other masking approaches.

6.1 Comparison of Micro-aggregation and Perturbation
In micro-aggregation, a set of k-sorted values of the confidential variable are “aggregated” or “averaged” to generate the perturbed values. Let $x_{(i,j)}$ represent the values of $X$ sorted by variable $j$. Then, the masked values of $Y$ are generated as:

$$y_{(m),j} = \sum_{i=1}^{k} x_{(i,j),m = i, i + 1, i + k - 1; j = 1, 2, \ldots, L} \quad (30)$$
The value of the masking parameter $k$ could vary from $2$ to $n$. Hansen and Mukherjee (2003) and Domingo-Ferrer and Mateo-Sanz (2002) have shown that micro-aggregation modifies the marginal distributions, the relationship between confidential variables, and relationship between confidential and nonconfidential variables. Winkler’s (2002) experiments with micro-aggregation indicate that its disclosure risk characteristics are extremely poor, with unacceptably high levels of identity disclosure. When the value of $k$ is small, it results in very high disclosure risk but lower information loss compared to larger $k$. When $k$ is large, there is considerable information loss, but the disclosure risk is lower compared to smaller $k$. However, some information loss occurs even when $k$ is small, and disclosure risk is not minimized even when $k$ is large.

Compared to both EGADP and data shuffling, micro-aggregation fares very poorly. First and foremost, both EGADP and data shuffling minimize disclosure risk, but micro-aggregation increases disclosure risk. In terms of data utility, by maintaining sufficiency, EGADP maintains inferential validity. Micro-aggregation does not. Data shuffling preserves all monotonic relationships for large data sets. Micro-aggregation does not. Thus, when comparing EGADP, data shuffling, and micro-aggregation, the choice is still between EGADP and data shuffling.

### 6.2 Comparison of Data Swapping and Perturbation

Data swapping was originally proposed by Dalenius and Reiss (1982) for categorical variables. Moore (1996) proposed the rank-based proximity swap (hereafter referred to simply as data swapping) for numerical variables. Swapping can be described as follows. Sort the data by confidential variable $j$. Exchange the values of $x_{(i),j}$ with $x_{(k),j}$. Repeat the process of every $i$ and every $j$. In performing data swapping, Moore (1996) suggests a masking parameter called the “swapping distance” parameter. Moore (1996, page 6) defines this parameter as follows:

Determine a value $P(a)$, with $0 < P(a) < 100$. The intent of the procedure is to swap the value of $a_i$ with that of $a_j$, so that the percentage difference of the indices, $i$ and $j$, is less than $P(a)$ of $N$. That is $|i – j| < P(a)N/100$.

The larger the value of $P(a)$, the larger the value of $|i – k|$, and greater the distance between the swapped values and vice versa. When $P(a) = 1$, then the swapping is essentially performed randomly. This destroys all relationships among variables but minimizes disclosure risk. When $P(a)$ is very small, the resulting disclosure risk is very high, but the information loss is small.

Muralidhar and Sarathy (2005) have directly compared data shuffling and data swapping. Their results indicate that disclosure risk of data shuffling is the same as that of data swapping with $P(a) = 1$. However, unlike data swapping, data shuffling provides the same level of data utility comparable to very small values of $P(a)$. Hence, data shuffling always performs better than data swapping (either in data utility or disclosure risk or both) and is the preferred method. Thus, when comparing EGADP, data shuffling, and data swapping, the choice is still between EGADP and data shuffling.

### 6.3 Comparison of Multiple Imputation and Perturbation

The multiple imputation process (Rubin 1993, Raghunathan et al. 2003) for generating masked microdata can be briefly described as follows. Assume that the data set is a finite population of size $N$, consisting of two sets of variables $S$ and $X$, where $S$ represents the $K$ design variables and
is observed for the entire data set. \( \mathbf{X} \) represents the \( L \) survey variables of interest. Let \( \mathbf{X}_{\text{obs}} \) represent the observed portion of the \( n \) sampled units. In other words, \( \mathbf{S} \) is a matrix of size \( (N \times K) \), and \( \mathbf{X}_{\text{obs}} \) is a matrix of size \( (n \times L) \), \( n << N \). Using this information, the agency releasing the data imputes \( \mathbf{X}_{\text{nobs}} \), the missing values for the \( N - n \) units so that a complete data set is created. The missing values are generated from the posterior predictive distribution of \( (\mathbf{X} | \mathbf{S}, \mathbf{Y}_{\text{obs}}) \). A random sample (of size say \( n_1 \)) is then selected from the \( (N - n) \) imputed values. This process is repeated by \( m \) 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. The most important aspect of multiple imputation is that it provides valid inferences regarding population parameters.

When comparing multiple imputation and perturbation approaches, remember that multiple imputation also involves the use of the appropriate model for generating the imputed values. Although alternative models have been proposed (Raghunathan et al. 2003), the most common implementation of multiple imputation uses the general linear perturbation model approach used in GADP. In this case, it can be shown that multiple imputation leads to valid inferences regarding parameters such as the mean, variance, regression coefficients, etc., or, in general, for those parameters for which the mean vector and covariance matrix are sufficient parameters. However, EGADP provides precisely the same advantage as multiple imputation in this context. In fact, it provides a greater degree of confidence in the sense that the results of analyses performed using the masked data will be identical to those using the original data. Furthermore, EGADP does not require that multiple data sets be analyzed and aggregated. Thus, when comparing EGADP, data shuffling, and multiple imputation, the choice is still between EGADP and data shuffling.

7. Conclusions

The objective of this paper is to trace the evolution of perturbation methods and evaluate their effectiveness for protecting numerical data. This paper clearly reveals that as perturbation techniques have evolved over time, their effectiveness has improved correspondingly. The results indicate that among all perturbation approaches, EGADP and data shuffling dominate the performance of other approaches in terms of both disclosure risk and information loss. We use the term “dominate” to imply that the performance of the technique is superior in both dimensions or superior in one dimension and equal in the other. The same conclusion is true when we consider other techniques that are commonly used for masking numerical data, namely, micro-aggregation, data swapping, and multiple imputation. This conclusion, however, is not true when we compare EGADP and data shuffling. While both procedures minimize disclosure risk, they provide different advantages in terms of data utility. Finally, we recommend that: EGADP should be used for perturbing data sets when the primary purpose is to conduct parametric, inferential statistical data analyses, and data shuffling should be used for perturbing relatively large data sets where we seek to preserve non-linear relationships.

---

1 The model based approach is the only exception to this general statement. This approach was proposed after the GADP method but provides higher disclosure risk and higher information loss than GADP.
References


