Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 527–544

Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection

Tapan K. Nayak[1], *Bimal Sinha*[2], *and Laura Zayatz*[3]

This article investigates statistical properties of random noise multiplication as a data masking procedure, especially for tabular magnitude data. It is shown that (i) the original data moments and correlations can be unbiasedly recovered from noise multiplied data (ii) for both finite and infinite population sampling, all polynomial estimators for the original data can be adopted easily for the masked data and (iii) for tabular magnitude data, multiplicative noises affect the quality of a cell total more for sensitive cells than for nonsensitive cells. Disclosure risk assessment and the choice of the noise distribution are discussed using the prediction error variance in a conservative scenario, where an intruder knows the perturbed cell total and all values within the cell, except the target unit's value. We also derive some interesting properties of a balanced noise method, and ascertain the reduction in the variance of a cell total by using the balancing mechanism.

Key words: Data quality; disclosure risk; noise variance; tabular data; unbiasedness; variance inflation.

1. Introduction

The main goal of most statistical agencies is to collect and publish data relevant to important national and regional public policy issues, but they also need to protect the privacy of survey respondents for legal reasons and to maintain public trust. Typically, a microdata set contains records of n sampling units on k variables, some of which are key variables (Bethlehem et al. 1990), and some which are confidential or sensitive that need protection against disclosure. To reduce disclosure risk, statistical agencies often release a perturbed or masked version of the original data, sacrificing some statistical information. Various masking procedures, such as grouping, cell suppression, data swapping, multiple imputation and random noise inoculation have been developed for practical use. The

[1] Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, and Department of Statistics, George Washington University, Washington, DC 20052, U.S.A. Email: tapan@gwu.edu
[2] Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, and Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, U.S.A. Email: sinha@math.umbc.edu
[3] Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: laura.zayatz@census.gov

Acknowledgments: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. We thank Paul Massell, Richard Moore, John Slanta and Tommy Wright for some discussions and constructive comments on an earlier draft of the article. We also thank an anonymous referee for some critical and useful comments that helped to improve the article.

books by Doyle et al. (2001) and Willenborg and De Waal (2001) discuss many issues germane to disclosure avoidance and various disclosure control techniques.

Disclosure is a difficult topic (cf., Lambert 1993) and it can occur in different forms depending on the disclosure scenario (see Willenborg and De Waal 2001). Broadly speaking, disclosure occurs when the released data enable an intruder to predict the values of some confidential variables for a specific unit *too accurately*. Identity disclosure, which happens when an intruder correctly identifies the record of a survey unit using externally available values of some key variables and thus learns the values of all confidential variables of the identified unit, is most serious. Measures of identity disclosure risk have been discussed by Bethlehem et al. (1990), Greenberg and Zayatz (1992), Willenborg and De Waal (2001), Skinner and Elliot (2002), Reiter (2005) and others.

Another type of disclosure that has received much attention is predictive disclosure, which occurs when the released data enable one to infer about a confidential variable value of a respondent with high accuracy. Obviously, predictive disclosure depends not only on the released data set but also on the intruder's prior knowledge, and should be assessed by comparing the intruder's knowledge before and after data release (see Duncan and Lambert 1986, 1989; Lambert 1993; Keller-McNulty et al. 2005).

Commonly used masking procedures dilute, suppress, and in some cases distort the information in the original data. So, in practice, one should attempt to strike a balance between disclosure risk and information loss when selecting disclosure control methods. We refer to Duncan and Fienberg (1999), Duncan and Stokes (2004), Karr et al. (2006), and Keller-McNulty et al. (2005) for excellent discussions of data utility and disclosure risk issues.

Methods and formulas for analyzing a data set may not be appropriate for analyzing a masked version of it; masking may destroy known properties, such as unbiasedness, of standard estimators. Obviously, the sampling distribution of an estimator and hence its statistical properties depend not only on the sampling design but also on the masking method. So, a full knowledge of the masking process is necessary for investigating properties of any statistical procedure and for deriving suitable inferential methods. Little (1993) presents a likelihood theory that is applicable to a wide variety of masked data. In general, likelihood theories require information about the masking procedure, which can be viewed as a process for selecting the values that are to be masked and a mechanism for masking the selected values. Thus, to allow data users to derive valid inferences, data providers need to release full information about the masking procedure along with the masked data.

There are multiple paradigms for addressing privacy protection and different procedures are suitable in different paradigms. One paradigm advocates that inferential methods for the original data should remain valid, at least approximately, for the perturbed data, so that users will not need to develop new methods for data analysis (see Rubin 1993). This goal seems to be the main motivation for creating synthetic data. However, analytical validity is retained fully if and only if the sampling distributions of original and masked data are the same, which does not hold for most masking methods available in the literature.

Another paradigm, which we subscribe to in this article, has the following features: (i) data providers disseminate masked data and full information about the masking procedure,

(ii) data users derive proper inference procedures for the released data, taking their sampling distribution (and established statistical principles and theory) appropriately into account, and (iii) data providers use masking procedures for which (a) adjustments to standard analyses, additional theoretical derivations and programming are not too complex or burdensome and (b) protection of private information can be assessed and communicated reasonably well. With this perspective, we investigate statistical properties of random noise multiplication as a disclosure avoidance technique, especially in the context of magnitude tabular data.

Most papers on noise perturbation deal with additive noise and assume that the data are generated by random sampling from an infinite population; see Brand (2002) for a nice review and further references. Some distinguishing features of our article are that it (i) focuses on multiplicative noise, which is well-suited for uniform privacy protection, as the noise CV is held constant, (ii) includes estimation in finite population sampling, (iii) covers magnitude tabular data, in addition to standard microdata, and (iv) appraises confidentiality protection rendered by multiplicative noise masking.

In Section 2, we discuss statistical properties of multiplicative noise masking at microdata level. Multiplicative noise provides uniform protection, in terms of noise CV, to all values in the data set. Population moments are easy to estimate unbiasedly, along with their standard errors, for both finite and infinite populations. Also, in finite population sampling, all polynomial estimators for the original data can be adopted easily for applying to noise multiplied data. In Section 3, we discuss certain properties of a procedure, proposed by Evans et al. (1998), for protecting confidentiality in magnitude tabular data. We theoretically prove that the cell level noise CV decreases as the contributing values to the cell become more homogeneous. This indicates that the total of a nonsensitive cell is likely to be less affected than that of a sensitive cell. We address confidentiality protection and the choice of the noise distribution by considering the variance of the prediction error under a fairly conservative scenario. In Section 4, we consider a variation of Evans et al.'s (1998) procedure, viz., a balanced noise masking method introduced by Massell and Funk (2007a, b). We show that the procedure is unbiased in the sense that the noisy total of *any* set of units is an unbiased estimator of the corresponding total based on the original data. We also ascertain the reduction in cell level noise variance from using the balancing mechanism. Section 5 contains some concluding remarks.

2. Random Noise Perturbation

Several forms of data masking using random noise have been discussed by Kim (1986), Tendick (1991), Fuller (1993), Evans et al. (1998), Brand (2002), Yancey et al. (2002), Kim and Winkler (2003) and others. Typical data sets contain values of several variables for n units, usually sampled from a population. First, let us consider a single quantitative sensitive variable Y with values $y_1, \ldots, y_n$ for the n units. The basic mechanism for random noise perturbation is to independently generate n numbers $r_1, \ldots, r_n$ from a known noise distribution, and then apply them to the y-values, either additively or multiplicatively, to create a masked data set $z_1, \ldots, z_n$, where $z_i = y_i + r_i$ or $z_i = y_i r_i$, $i = 1, \ldots, n$. The data agency selects the noise distribution, usually with mean zero for additive noise, and mean 1 for multiplicative noise, so that $E[Z_i|y_i] = y_i$. In this article, we shall focus mainly on multiplicative noise, which may be described by

$$Z = YR \tag{2.1}$$

where R denotes a noise variable. Let $\nu_j = E(R^j)$, $j = 1, 2, \ldots$, denote the raw moments of the noise distribution and σ_R^2 the noise variance, and assume that $\nu_1 = 1$.

How much protection does noise multiplication provide to individual data values? Specifically, what can an intruder infer about the original value (y) of a specific unit, whose identity he has ascertained correctly, from its perturbed value z? From (2.1) it follows that

$$E[Z|y] = y, \quad \text{and} \quad \sigma_{Z|y}^2 = V[Z|y] = y^2\sigma_R^2$$

So, z is an unbiased estimate of y and the standard deviation $\sigma_{Z|y} = |y|\sigma_R$ is a measure of an intruder's uncertainty about y. An intruder may estimate $\sigma_{Z|y}$ by $|z|\sigma_R$. As $\sigma_{Z|y}$ is proportional to $|y|$, the relative size of perturbation is the same for all y, viz., $\frac{1}{|y|}\sigma_{Z|y} = \sigma_R$ is a constant, which also provides a practical interpretation of σ_R. A constant noise CV is desirable in some applications, where one feels that small $|y|$ should be perturbed little to avoid excessive distortion and large $|y|$ should be perturbed more to protect y reasonably well. In contrast, for additive noise, $V[Z|y] = \sigma_R^2$ is the same for all y, which is too much for small $|y|$ and too little for large $|y|$. Thus, with additive noise, the level of masking may vary widely depending on the range of y-values.

One can easily calculate confidence intervals for the original data values and use them to assess disclosure risk. Suppose R is positive and has a continuous unimodal distribution, which usually hold in practice. For given α, let $[a,b]$ be the shortest interval satisfying $P(a \le R \le b) = 1 - \alpha$. Note that the interval $[a,b]$ can be computed from the known distribution of R. For any $y > 0$, it follows from (2.1) that $P(ay \le Z \le by|y) = 1 - \alpha$ or $P(Z/b \le y \le Z/a|y) = 1 - \alpha$. Similarly, for $y < 0$, $P(Z/a \le y \le Z/b|y) = 1 - \alpha$. Since z and y have the same sign, a $100(1 - \alpha)\%$ confidence interval for y, based on z, is $(z/b, z/a)$ if $z > 0$ and $(z/a, z/b)$ if $z < 0$.

2.1. *Estimation of Infinite Population Moments*

Certain inferences, e.g., estimates of the mean, variance and moments of Y, can be derived easily from noise multiplied data. First, consider random sampling from an infinite population or simple random sampling with replacement (SRSWR) from a finite population. Letting μ_Y and σ_Y^2 denote the mean and variance of Y, it can be easily seen that $E[Z] = \mu_Y$ and

$$\begin{aligned} V[Z] &= V[E(Z|Y)] + E[V(Z|Y)] \\ &= \sigma_Y^2 + \sigma_R^2\left[\sigma_Y^2 + \mu_Y^2\right] \qquad (2.2) \\ &= \left(1 + \sigma_R^2\right)\sigma_Y^2 + \mu_Y^2\sigma_R^2 \end{aligned}$$

From these, it follows that the mean ($\bar{Z}$) of the masked data is an unbiased estimator of μ_Y, but the variance S_Z^2 over-estimates σ_Y^2. However, unbiased estimation of higher order moments of Y is fairly easy. Note that for all $j \ge 1$, $E[Z^j|y] = \nu_j y^j$ and hence $E[Z^j] =$

$E[Y^j]E[R^j] = \nu_j E[Y^j]$. So, z_i^j/ν_j is an unbiased estimate of y_i^j and $(1/\nu_j)\left(\sum_i Z_i^j/n\right)$ is an unbiased estimator of $E[Y^j]$. Thus, all sample moments of noise multiplied masked data can be modified easily to make them unbiased estimators of the corresponding moments of Y. In particular, if T is any unbiased estimator of μ_Y^2, e.g., $T = 1/n(n-1)\sum_{i \neq j} Z_i Z_j$, then from (2.2),

$$\delta = \frac{S_Z^2 - \sigma_R^2 T}{1 + \sigma_R^2} = \frac{1}{n(n-1)}\left[\left(\frac{n + \sigma_R^2}{1 + \sigma_R^2}\right)\sum_{i=1}^{n} Z_i^2 - \left(\sum_{i=1}^{n} Z_i\right)^2\right] \tag{2.3}$$

is an unbiased estimator of σ_Y^2. Unbiased estimators of μ_Y based on the original and the masked data are $\bar{Y}$ and $\bar{Z}$, respectively, and their variances are σ_Y^2/n and σ_Z^2/n, which can be estimated unbiasedly, from the masked data, by δ/n and S_Z^2/n, respectively. So, $(S_Z^2 - \delta)/n$ is an unbiased estimator of the variance inflation due to multiplicative noise, in the context of estimating μ_Y. Kim and Winkler (2003) have discussed estimation of μ_Y and σ_Y^2 when the noise distribution is truncated normal.

Practical datasets contain values of many variables, several of which may be sensitive. Noise multiplication may be applied easily to more than one variable. It is convenient to generate the noise factors independently, but possibly from different distributions for different variables. Noise multiplication (or addition) distorts correlations among the variables. For simplicity, suppose Y and W are two variables in the original file and the masked file contains W (unchanged) and Z, which is noise multiplied Y, as described before. Then, it can be seen, using (2.2), that

$$\rho(Z, W) = \left[\frac{\sigma_Y^2}{\left(1 + \sigma_R^2\right)\sigma_Y^2 + \mu_Y^2\sigma_R^2}\right]^{1/2} \rho(Y, W) \tag{2.4}$$

where $\rho(.,.)$ denotes the correlation between the two variables within the parentheses. Generally, as can be seen from (2.4), noise multiplication (or addition) deflates correlations. Note that

$$cov(Z_1, Z_2) = E(Z_1 Z_2) - E(Z_1)E(Z_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = cov(Y_1, Y_2)$$

and hence independent noise multiplication (with mean 1) does not bias the sample means and covariances, but inflates the variances as seen in (2.2). Thus, for valid estimates of correlations, only the variances need to be estimated appropriately, perhaps using (2.3).

Unbiased estimation of correlations and joint moments from noise multiplied data is also quite straightforward. Suppose Y_1 and Y_2 are two original variables and the corresponding masked variables are $Z_i = Y_i R_i$, $i = 1, 2$, where R_1 and R_2 are independently (but possibly not identically) distributed. Then, for all $k_1, k_2 \in R$,

$$E\left[Z_1^{k_1} Z_2^{k_2} | y_1, y_2\right] = y_1^{k_1} y_2^{k_2} E\left[R_1^{k_1}\right] E\left[R_2^{k_2}\right]$$

which shows that $\left[Z_1^{k_1} Z_2^{k_2}\right] / \left\{E\left[R_1^{k_1}\right] E\left[R_2^{k_2}\right]\right\}$ is an unbiased estimator of $y_1^{k_1} y_2^{k_2}$. A data user would simply need to divide the masked sample joint raw moment of order (k_1, k_2) by $E\left[R_1^{k_1}\right] E\left[R_2^{k_2}\right]$ to get an unbiased estimate of the corresponding original sample moment. A similar approach can be used to obtain consistent estimators of regression coefficients and their standard errors (see Hwang 1986). Analogous adjustments for additive noise (with mean 0) are fairly simple for estimating means, variances and covariances (see Kim 1986; Kim and Winkler 1995), but can be tedious for higher order moments.

2.2. Finite Population Estimation

Commonly used finite population estimators, viz., all polynomial estimators, can be modified easily to account for the effects of multiplicative noise masking. Suppose the original data came from a subset s of a finite population, selected using a sampling design $p(s)$, and N denotes the population size. First, consider one survey variable Y. Since z_i^j/ν_j is an unbiased estimate of y_i^j it follows that if $w_0 + \sum_{i\in s}\sum_{j=1}^{k} w_{ij} Y_i^j$ is an unbiased estimator of a population parameter based on the original data, then $w_0 + \sum_{i\in s}\sum_{j=1}^{k} w_{ij}\left[Z_i^j/\nu_j\right]$ is an unbiased estimator of the same parameter but based on the masked data.

As linear estimators are most commonly used in practice, we now discuss them in more detail. Suppose $T = \sum_{i\in s} w_{si}Y_i$ is a homogeneous linear unbiased estimator of a population parameter θ based on the original data and $V_p(T)$ is its design based variance. Then $T^* = \sum_{i\in s} w_{si}Z_i$ is an unbiased estimator of θ based on noise multiplied data and

$$\begin{aligned} V[T^*] &= E_p[V_R(T^*|s)] + V_p[E_R(T^*|s)] \\ &= E_p\left[\sum_{i\in s} w_{si}^2 \sigma_R^2 Y_i^2\right] + V_p(T) \\ &= \sigma_R^2 \sum_{i=1}^{N} Y_i^2 \sum_{s\ni i} w_{si}^2 p(s) + V_p(T) \end{aligned} \tag{2.5}$$

where E_p and E_R denote expectations with respect to the sampling design and the noise distribution, respectively. The first term of (2.5) is the variance inflation due to noise multiplication, for which an unbiased estimator, based on the original data, is $\sigma_R^2 \sum_{i\in s} w_{si}^2 Y_i^2$. So, an unbiased estimator of it based on the masked data is

$$\sigma_R^2 \sum_{i\in s} w_{si}^2 \left(\frac{Z_i^2}{\nu_2}\right) = \left(\frac{\sigma_R^2}{1+\sigma_R^2}\right) \sum_{i\in s} w_{si}^2 Z_i^2 \tag{2.6}$$

It can be seen that (e.g., Hedayat and Sinha 1991, Sec. 3.1)

$$V_p(T) = \sum_{i=1}^{N} b_i Y_i^2 + \sum\sum_{i\neq j} b_{ij} Y_i Y_j$$

where

$$b_i = \sum_{s\ni i} w_{si}^2 p(s) - 1 \quad \text{and} \quad b_{ij} = \sum_{s\ni i,j} w_{si} w_{sj} p(s) - 1$$

Hence an unbiased estimator of $V_p(T)$, based on the original data, is

$$\hat{V}_p(T) = \sum_{i\in s} b_i \frac{Y_i^2}{\pi_i} + \sum\sum_{i,j\in s,\, i\neq j} b_{ij} \frac{Y_i Y_j}{\pi_{ij}}$$

where $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$. Clearly, $\hat{V}_p(T)$ is a quadratic estimator in Y and it can be easily adopted for the masked data. Specifically, an unbiased estimator of $V_p(T)$, based on the masked data, is

$$\tilde{V}_p(T) = \left(\frac{1}{1+\sigma_R^2}\right) \sum_{i \in s} b_i \frac{Z_i^2}{\pi_i} + \sum_{i,j \in s,} \sum_{i \neq j} b_{ij} \frac{Z_i Z_j}{\pi_{ij}} \tag{2.7}$$

Thus, from noise multiplied masked data, we can easily obtain an unbiased estimator (T^*) of θ and also its variance, which is the sum of (2.6) and (2.7). Note that (2.7) gives a data user an estimate of the variance of an estimator of θ based on the original data. Thus, the estimates of the two components of $V[T^*]$, given by (2.6) and (2.7), are useful for ascertaining information loss (for estimating θ) due to noise multiplication. Many agencies grant researchers access to original data, but it often involves a lengthy application and review process and conducting research at agencies locations. The numerical values of T^* along with (2.6) and (2.7) are directly useful to researchers for (i) ascertaining the worth of the original data as compared to the masked data, (ii) suggesting a suitable level of data masking to the data agency, and (iii) making a case for gaining access to the original data, subject to appropriate pledge of maintaining confidentiality. We may note that $\hat{V}_p(T)$ and hence $\tilde{V}_p(T)$ can be negative. However, alternative estimators of $V_p(T)$ based on the original data, that are available in the literature, can easily be adopted for noise multiplied data. We conclude this subsection by noting that the usual finite sample estimators of joint moments can also be adjusted easily, through simple divisions by appropriate raw moments of relevant noise distributions.

2.3. *Comments on Other Noise Methods*

Mathematically speaking, noise masking can always be treated as additive if the noise distribution is allowed to depend on y (see Fuller 1993). However, there are two disadvantages of this approach. First, generating the noise values is not as simple as it is for the iid case. Second, and more importantly, proper analysis of masked data is generally more difficult for dependent noise.

Some researchers, e.g., Kim (1986) and Fuller (1993), have suggested that one should preserve the means and the covariance matrix of the survey variables, which are important summary statistics. One approach is to use a data dependent linear transformation after noise inoculation (e.g., Kim 1986). Suppose k variables are to be masked and the vector $\vec{y}_i$ represents the values of the k variables for unit i. Then, $\vec{y}_i (i = 1, \ldots, n)$ are first changed to $\vec{z}_i = \vec{y}_i + \vec{\varepsilon}_i$, where $\vec{\varepsilon}_1, \ldots, \vec{\varepsilon}_n$ are random noise vectors, independently generated from a common k-dimensional distribution with zero mean and covariance matrix Λ (diagonal when the noise values are independent). Next, $\vec{z}_i (i = 1, \ldots, n)$ are changed to $\vec{z}_{i*} = A\vec{z}_i + \vec{b}$, via a linear transformation, where the matrix A (of order $k \times k$) and the vector $\vec{b}$ are so chosen that the mean vector and the covariance matrix of $\vec{z}_{i*} (i = 1, \ldots, n)$ are the same as those of $\vec{y}_i (i = 1, \ldots, n)$. Clearly, A and $\vec{b}$ depend on $(\vec{y}_i, \vec{z}_{i*})$, $i = 1, \ldots, n$, which makes calculation of the probability distribution of masked data and assessment of the masking effect on various inferences very difficult. Because of the second step, viz., the data dependent linear transformation, known properties of additive noise do not continue to hold for the overall masking process.

We also note that the mean vector and the covariance matrix can be preserved, using a data dependent linear transformation, from arbitrarily generated $\vec{z}_1, \ldots, z_n$, not necessarily through additive noise. Let $\mathcal{Y}$ and $\mathcal{Z}$ be two data matrices (of same order) with mean vectors $\bar{y}$ and $\bar{z}$ and nonsingular covariance matrices S_y and S_z. Let $\vec{z}_{i*} = A\vec{z}_i + \vec{b}$, $i = 1, \ldots, n$, where $A = S_y^{1/2} S_z^{-1/2}$ and $\vec{b} = \bar{y} - A\bar{z}$. Then it can be seen easily that $\{\vec{y}_i, i = 1, \ldots, n\}$ and $\{\vec{z}_i^*, i = 1, \ldots, n\}$ have the same mean vector and the same covariance matrix. Thus, the task of modifying the original data while preserving the means and the covariance matrix can be accomplished easily (and fairly arbitrarily). Interestingly, Kim and Winkler (1995) proved that if the covariance matrix of the original data is nonsingular, then it is possible to change the values in one record arbitrarily and yet preserve the means and the covariance matrix, by modifying other records suitably.

Another approach is to generate the noise vectors from a distribution with mean 0 and covariance matrix $\delta\Sigma$, where δ is a constant chosen by the data provider and Σ is the covariance matrix of the survey variables. Then, the noise added variables would have mean 0 and covariance matrix $(1 + \delta^2)\Sigma$. Since δ is known, Σ can be estimated unbiasedly (and consistently) from the masked data (see Brand 2002). Mathematical treatment of additive noise is most convenient when both the survey variables and noise vectors are normally (multivariate) distributed. Thus, Fuller (1993) suggested transforming observed variables into pseudo normal variables, adding independent normal noise vectors to the transformed records, and finally back-transforming the noise added values to the original scale.

The main reason for publishing microdata is to facilitate the performance of different types of analyses by data users. Preserving the overall mean vector and the covariance matrix is of limited help if the analysis involves other features of the data that are perturbed by the masking procedure. For example, deriving unbiased estimators of subdomain means can be very difficult, even if the masking procedure is revealed fully. Generally, the dependency of the transformations on both the original and noise added data makes proper analysis of the masked data and assessment of disclosure risk very difficult.

3. Tabular Magnitude Data

Often the mean or the total of a quantitative variable for various subgroups is of interest. The estimates are presented conveniently in the form of a table, whose cells represent the subgroups and are defined by cross classification of some geographic and demographic variables. A published table may report for each cell its frequency, an estimate of the quantity of interest and its standard error. Tables of magnitudes are commonly used for disseminating information in data generated by economic surveys of establishments.

Usually, the variables that define the cells of a magnitude table are key variables, and based on external information it may be possible to identify the cell in which a target unit falls or even all units falling in a cell. If a cell contains only a few units, an estimate for that cell, based on the original data, may induce high disclosure risk for all units in that cell. Obviously, cells with only one or two contributors are highly sensitive. Let n denote the number of contributors to a cell. For $n \geq 3$, one common rule for defining sensitive cells is the *p*% rule (see Federal Committee on Statistical Methodology 2005), by which a cell is sensitive if

$$y_1 \geq \frac{100}{p} \sum_{i=c+2}^{n} y_i \tag{3.1}$$

where $y_1 \geq \cdots \geq y_n$ are the ordered values of the units in the cell, and $0 < p < 100$ and $1 \leq c \leq (n-2)$ are two prespecified numbers.

One widely used technique for dealing with sensitive cells is cell suppression, which begins by suppressing the values of all sensitive cells. In addition, the values of some other cells are also suppressed, called secondary suppressions, so that primary suppression values cannot be recovered from nonsuppressed cell totals and the marginal totals. Cell suppression has certain disadvantages (see Evans et al. 1998), including withholding too much information in many cases and the possibility of disclosure based on information from multiple tables.

As an alternative to cell suppression, Evans et al. (1998) suggested creating magnitude tables after noise multiplying the original microdata values. Often many tables are published from the same microdata, and for maintaining consistency among different tables, it is desirable to first create a masked microdata set and then generate all tables for public release from it. We shall examine effects of iid noise multiplication on both confidentiality and data quality for tabular magnitude data, assuming that the survey variable is nonnegative, as is the case in most applications. For establishment survey data, Evans et al. (1998) perturb all establishment values within a company in the same direction (up or down), which makes some noise factors dependent. We do not consider that case here and for simplicity assume that all noise factors are generated independently from a common noise distribution.

3.1. *Effects of Multiplicative Noise on Data Quality*

We shall now consider the effect of multiplicative noise on a cell total. Suppose a cell contains n units with values $y_1, \ldots, y_n$ and $T = y_1 + \cdots + y_n$ is the cell total. Let T_* denote the perturbed total, i.e.,

$$T_* = \sum_{i=1}^{n} y_i R_i$$

where R_i are iid random noise multipliers. It follows easily that $E[T_* | y_1, \ldots, y_n] = T$, i.e., T_* is an unbiased estimator of T, and the cell level noise variance is

$$\sigma_C^2 = V(T_* | y_1, \ldots, y_n) = \sigma_R^2 \sum_{i=1}^{n} y_i^2 \tag{3.2}$$

When the original data are generated by sampling from a finite population, with a sampling design p, the overall variance of T_* is

$$\begin{aligned} V(T_*) &= E_p[V_R(T_* | y_1, \ldots, y_n)] + V_p[E_R(T_* | y_1, \ldots, y_n)] \\ &= \sigma_R^2 E_p\left[\sum\nolimits_{i=1}^{n} y_i^2\right] + V_p(T) \end{aligned}$$

If $\hat{V}_p(T)$ is an unbiased estimator of $V_p(T)$, then an unbiased estimator of $V(T_*)$ is

$$\hat{V}(T_*) = \sigma_R^2 \sum_{i=1}^{n} y_i^2 + \hat{V}_p(T) \tag{3.3}$$

Note that the formula for $\hat{V}(T_*)$ is not usable to data users as it involves all values in the cell. Even in the case of no masking, the variance of a cell total cannot be estimated from the cell total itself, and the variances, i.e., $\hat{V}_p(T)$, need to be calculated and reported by the data agency, which has access to the microdata. In the same way, (3.3) are to be used by data agencies, if they choose to report the estimated variances. Naturally, most data users would like to see both components of $V(T_*)$, in (3.3), instead of just the total, but often the decision as to what information to release rests with the data agency. For example, a data provider may decide to release only the perturbed totals and that too only for the cells which it considers to be nonsensitive.

The fact that T_* is an unbiased estimator of T was noted by Evans et al. (1998). They also observed, through simulations and numerical examples, that cell level noise CVs are generally higher for sensitive cells than for nonsensitive cells. Here, we explain this phenomenon theoretically. From (3.2) we see that the square of cell level noise CV is

$$\psi^2 = \sigma_R^2 \sum_{i=1}^{n} \left(\frac{y_i}{T}\right)^2 = \sigma_R^2 \sum_{i=1}^{n} g_i^2 \tag{3.4}$$

where $g_i = y_i/T$ is the "share" of unit i in the cell total. Recall (from Section 2) that σ_R is the noise CV for each individual value. So, (3.4) gives a simple relationship: cell level noise CV equals unit level noise CV multiplied by $\left[\sum g_i^2\right]^{1/2}$.

For any nonnegative variable Y, $g_i \geq 0$, $i = 1, \ldots, n$ and $g_1 + \cdots + g_n = 1$. It is easy to see that ψ^2, considered as a function of $g_1, \ldots, g_n$, for given n, is permutation symmetric and strictly convex in each argument, which implies that ψ^2 is a Schur-convex function (Marshall and Olkin 1979). This implies the following: (i) ψ^2 or equivalently the cell level noise CV (ψ) increases as $\{g_i\}$, i.e., the shares of the n units, become more heterogeneous, (ii) the maximum possible value of ψ^2 is σ_R^2, which is attained when $g_i = 1$ for some i and 0 for others and (iii) the minimum of ψ^2 is σ_R^2/n, which is attained when $g_1 = \cdots = g_n (= 1/n)$. The inequality in (3.1) essentially says that a cell is sensitive by the $p\%$ rule if the share of the largest unit is very high and hence $g_1, \ldots, g_n$ are hetereogeneous. For nonsensitive cells, $g_1, \ldots, g_n$ are likely to be fairly homogeneous. So, Schur-convexity of ψ^2 implies that noise CV of a cell total is likely to be higher for sensitive cells than for nonsensitive cells (with the same cell frequency, n). As Evans et al. (1998) have noted, this is a desirable property because nonsensitive cells do not pose much disclosure risk and hence don't need to be perturbed much.

One might expect the effect of noise on a cell total to diminish as the cell frequency increases. We note that if a new value y_{n+1} is added to a cell that already has n values $y_1, \ldots, y_n$, the noise variance, σ_C^2 given by (3.2), increases but noise CV (ψ) decreases unless the added value is very large. Specifically, it can be seen that the noise CV decreases, with the addition of the value y_{n+1}, if and only if

$$y_{n+1} < \frac{2\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} y_i^2\right)}{\left(\sum_{i=1}^{n} y_i\right)^2 - \sum_{i=1}^{n} y_i^2} = 2\left(\sum_{i=1}^{n} y_i\right)\left[\frac{1}{\gamma^2} - 1\right]^{-1} \tag{3.5}$$

where $\gamma^2 = \sum_{i=1}^{n}(y_i/T)^2$ (with $T = \sum_{i=1}^{n} y_i$) is a measure of the heterogeneity of $y_1, \ldots, y_n$. As a numerical example, if $n = 5$ and $y_1, \ldots, y_5$ are 10, 6, 3, 2 and 1, then the right side of (3.5) is 19.76, which is about twice the largest value in the cell. Note that, in general, $1/n \leq \gamma^2 \leq 1$, which implies that $[(1/\gamma^2) - 1]^{-1} \geq 1/(n-1)$ and hence (3.5) holds if $y_{n+1} < [2/(n-1)]\sum_{i=1}^{n} y_i \approx 2\bar{y}$.

3.2. Disclosure Control

We now examine the efficacy of noise multiplication for confidentiality protection. Suppose a cell contains n units with values $y_1, \ldots, y_n$ and let T and T_* denote the true and noisy cell totals. What can an intruder infer about the value of a specific (target) unit, say the value y_1, from a reported noisy total? Assume that the intruder has full knowledge about the masking procedure, i.e., the noise distribution is revealed to the public. The intruder's uncertainty about y_1, after learning a noisy total T_*, depends on his prior knowledge about $y_1, \ldots, y_n$. Logically, his uncertainty should be expressed using his posterior distribution derived using the Bayes theorem, where the noise distribution determines the likelihood function. It may be noted that derivation of the posterior distribution requires the intruder's prior information about $y_1, \ldots, y_n$; a prior distribution of y_1 alone is not sufficient (see Lambert 1993). Thus, a proper Bayesian updating of an intruder's knowledge is usually very difficult. Also, there are many intruders, who have different prior information and hence would gain different amounts of knowledge from the reported T_*. Which intruder's information gain should a data agency consider for assessing disclosure risk? A further complication is that the information gain also depends on the target unit.

We shall instead consider a conservative situation: the intruder knows all original values in the cell except y_1, and has no information about y_1. Consider the natural estimator of y_1, given by

$$\hat{y}_1 = T_* - \sum_{i=2}^{n} y_i$$

Letting $e_1 = \hat{y}_1 - y_1$ denote the estimation error, it can be seen easily that the mean and variance of e_1, for given $y_1, \ldots, y_n$, are 0 and

$$V(e_1) = V(\hat{y}_1 - y_1) = \sigma_R^2 \sum_{i=1}^{n} y_i^2 \tag{3.6}$$

More realistically, an intruder may know the true total (T_c) of a coalition of units and have an estimate (guess) $\tilde{T}_r$ for the total of the remaining units, excluding y_1. Such an intruder may calculate (estimate) y_1 as

$$\tilde{y}_1 = T_* - T_c - \tilde{T}_r$$

It can be seen that the error of this estimator, viz., $e_1^* = \tilde{y}_1 - y_1$, has mean $(T_r - \tilde{T}_r)$, where T_r is the true total of the remaining units, and $Var\left(e_1^*\right) = \sigma_R^2 \sum_{i=1}^n y_i^2$ (both mean and variance are with respect to the noise distribution). Also, if the noise distribution is symmetric, then e_1^* is also symmetrically distributed. In addition, if the distribution of e_1^* is continuous and unimodal, then for any given $k, P(|e_1^*| < k)$ is a decreasing function of $|T_r - \tilde{T}_r|$ and consequently $\tilde{y}_1$ is most accurate when $\tilde{T}_r = T_r$.

Comparing (3.6) with (3.2), we see that multiplicative noise induces the same level of uncertainty (noise variance) about any specific value as about the total of the cell containing that value. This is not surprising because in our context, the knowledge of y_1 is equivalent to knowledge of the cell total. Actually, since a cell total is larger than any specific value in the cell, in terms of CV, uncertainty about any individual value is larger than the uncertainty about the cell total. Also note that the expression in (3.6) is a symmetric function of $y_1, \ldots, y_n$, and hence it can be used to assess uncertainty about any one of the cell values $y_1, \ldots, y_n$ when the remaining ones are known.

How should we choose the noise distribution? To answer this question, we should take both data quality and confidentiality into account. For privacy protection we may require

$$2\sigma_R \left(\sum_{i=1}^n y_i^2 \right)^{1/2} \geq y_i \left(\frac{p}{100} \right) \text{ for } i = 1, \ldots, n \tag{3.7}$$

so that approximate 95% error bounds for each value y_i are at least $p\%$ away from its actual value. As our assumption that the intruder knows all values except y_i is rather conservative, we believe a modest value of p would be reasonable in practical applications. Note that (3.7) is satisfied if and only if the inequality holds for the largest value in the cell, i.e.,

$$1 + \left(\frac{y_2}{y_1} \right)^2 + \cdots + \left(\frac{y_n}{y_1} \right)^2 \geq \frac{1}{4\sigma_R^2} \left(\frac{p}{100} \right)^2$$

where $y_1 \geq \cdots \geq y_n$ are the ordered values in the cell. From the data quality perspective, and in view of (3.2), it seems logical to require

$$2\sigma_R \left(\sum_{i=1}^n y_i^2 \right)^{1/2} \leq T \left(\frac{p_*}{100} \right) \tag{3.8}$$

so that approximate 95% bounds for the cell total are no more than $p_*\%$ away from the true total. Naturally, one would want (3.7) to hold for a "large" p and also (3.8) for a "small" p_*, which may not be possible. It is easy to see that both (3.7) and (3.8) can be satisfied if and only if $p_* \geq (y_1/T)p$. The optimum combinations of attainable (p, p_*) are determined by

$$p_* = \left(\frac{y_1}{T} \right) p \quad \text{or} \quad \frac{p_*}{p} = \frac{y_1}{T} \tag{3.9}$$

So, one needs to choose p and p_* satisfying (3.9) and then determine the corresponding value of σ_R. This approach can be applied to each cell when the goal is to publish only one table. Note that in this approach, σ_R^2 would be different for different cells.

If many tables are to be published based on the masked data, as is usually the case, then satisfying (3.7) for all cells in all tables is a more challenging task. One possibility is to use $\sigma_R = p/200$ so that (3.7) is satisfied for all $n \geq 1$ and all $y_1, \ldots, y_n$. This approach protects all values at the unit level using a conservatively large value of σ_R, for the given p, which may not be attractive from the data quality perspective. The following may be a better compromise: use a common noise distribution with tolerable σ_R, perhaps around .02 or .03, and then publish only those tables whose cells satisfy (3.7). Clearly, this approach may require redefining the cells of a table.

4. Properties of a Balanced Noise Method

The disclosure risk from publishing the observed total of a cell is small if the cell has several fairly homogeneous contributors. Generally, the need for perturbing a cell total decreases as the cell frequency increases. However, as (3.2) shows, in independent noise masking, the cell-level noise variance increases as more contributors join a cell. For altering the cell totals differently for sensitive and nonsensitive cells, Massell and Funk (2007a, b) proposed a balanced noise procedure, where the direction of change of a value is determined by the preceding perturbations within the cell. For balanced noise, one must select and use a specific table for balancing noise factors, but as one traverses the cells in the table, one assigns noise factors to the microdata. The procedure can be described as follows (for simplicity, we consider independent noise factors and do not require that all establishment values within a company be changed in the same direction).

Suppose a cell contains n values, $y_1 \geq \cdots \geq y_n$. The balanced noise procedure changes them sequentially to $y_1^*, \ldots, y_n^*$, where

$$y_i^* = (1 + W_i U_i) y_i, \; i = 1, \ldots, n \tag{4.1}$$

$U_1, \ldots, U_n$ are iid random variables with a common pdf $f_U(.)$ whose support is a subset of $[0, \infty)$, W_1 is 1 or -1 with equal probability and for $i \geq 2$, $W_i = 1$ if $\sum_{j=1}^{i-1}(y_j^* - y_j) < 0$, $W_i = -1$ if $\sum_{j=1}^{i-1}(y_j^* - y_j) > 0$ and W_i is 1 or -1 with equal probability if $\sum_{j=1}^{i-1}(y_j^* - y_j) = 0$. For simplicity, we shall assume that $\sum_{j=1}^{i-1}(y_j^* - y_j) \neq 0$ with probability 1. From (4.1) we see that W_i determines the direction of change of y_i and U_i determines the magnitude. The direction of change of the largest value (y_1) is randomly selected and the subsequent values (i.e., $y_2, \ldots, y_n$) are increased or decreased depending on the sign of the cumulative effect of the preceding changes. Thus, perturbation magnitudes $U_1, \ldots, U_n$ are determined independently, but the directions are dependent. The distribution $f_U(.)$ is known and is selected by the data agency. Note that the noise factors are $R_i = 1 + W_i U_i$, $i = 1, \ldots, n$, and they are not independent.

In the balanced noise method, starting with the second-largest value, each perturbation aims to undo in part the cumulative effect of the previous changes on the cell total. Intuitively, if n is moderately large and $y_1, \ldots, y_n$ are fairly uniform, the cell total is expected to change little. However, due to dependencies among the noise factors, distributional properties of the change in a cell total are not obvious. In the following, we present some theoretical results for the balanced noise procedure. Our main conclusions are contained in Proposition 4.1 and Theorem 4.2.

To investigate statistical properties of the balanced noise procedure, let

$$T_i = \sum_{j=1}^{i} y_j, \quad T_{i*} = \sum_{j=1}^{i} y_j^* \quad \text{and} \quad D_i = T_{i*} - T_i = \sum_{j=1}^{i} W_j U_j y_j$$

for $i = 1, \ldots, n$. Note that $T_* = T_{n*}$ and $W_i = -sign\,(D_{i-1})$, $i = 2, \ldots, n$. For given $y_1, \ldots, y_n$, note that D_1 is a function of (W_1, U_1) and for $i \geq 2$, W_i is a function of $(W_1, U_1, \ldots, U_{i-1})$ and D_i is a function of $(W_1, U_1, \ldots, U_i)$. So, let's write $D_1 = D_1(W_1, U_1)$ and for $i \geq 2$, $W_i = W_i(W_1, U_1, \ldots, U_{i-1})$ and $D_i = D_i(W_1, U_1, \ldots, U_i)$.

Lemma 4.1 *The functions $D_1, \ldots, D_n$ and $W_2, \ldots, W_n$ are skew-symmetric in W_1, that is, for all $u_1, \ldots, u_n$,*

$$D_i(1, u_1, \ldots, u_i) = -D_i(-1, u_1, \ldots, u_i),\ i = 1, \ldots, n \tag{4.2}$$

$$W_i(1, u_1, \ldots, u_{i-1}) = -W_i(-1, u_1, \ldots, u_{i-1}),\ i = 2, \ldots, n \tag{4.3}$$

Proof Note that (4.3) follows from (4.2) as $W_i = -sign(D_{i-1})$. So, we only need to prove (4.2). Clearly, $D_1(1, u_1) = u_1 y_1 = -[-u_1 y_1] = -D_1(-1, u_1)$, and hence $W_2(1, u_1) = -W_2(-1, u_1)$, as $W_i = -sign(D_{i-1})$. We can now use induction to prove the lemma. Suppose (4.2) holds for $i = 1, \ldots, k-1$ (and hence (4.3) holds for $i = 2, \ldots, k$). Note that, for $l = \pm 1$,

$$D_k(l, u_1, \ldots, u_k) = D_{k-1}(l, u_1, \ldots, u_{k-1}) + W_k(l, u_1, \ldots, u_{k-1}) u_k y_k \tag{4.4}$$

The proof can now be completed easily using the induction hypothesis on (4.4) and the fact that (4.2) holds for $i = 1$. □

Theorem 4.1 *For all $i \geq 1$, (i) D_i is symmetrically distributed around 0 and (ii) the marginal distribution of W_i is uniform over {−1,1}, i.e.,*

$$P(W_i = 1) = P(W_i = -1) = 0.5 \tag{4.5}$$

Proof Considering the joint distribution of $D_1, U_1, \ldots, U_i$ and generically denoting relevant densities by $p(.)$, we see that for all $\mathbf{u}_i = (u_1, \ldots, u_i)$,

$$p(1, \mathbf{u}_i) = p(1)p(\boldsymbol{u}_i) = \frac{1}{2}p(\mathbf{u}_i) = p(-1)p(\mathbf{u}_i) = p(-1, \mathbf{u}_i) \tag{4.6}$$

Take any fixed interval $[a, b]$. For $k = -1, 1$, let $A(k) = \{\mathbf{u}_i : a \leq D_i(k, \mathbf{u}_i) \leq b\}$ and $A^*(k) = \{\mathbf{u}_i : -b \leq D_i(k, \mathbf{u}_i) \leq -a\}$. By Lemma 4.1, $A(k) = A^*(-k)$ and

$$\begin{aligned} P(a \leq D_i \leq b) &= P(W_1 = 1)P[\boldsymbol{U}_i \in A(1)] + P(W_1 = -1)P[\boldsymbol{U}_i \in A(-1)] \\ &= P(W_1 = -1)P[\boldsymbol{U}_i \in A^*(-1)] + P(W_1 = 1)P[\boldsymbol{U}_i \in A^*(1)] \\ &= P(-b \leq D_i \leq -a) \end{aligned} \tag{4.7}$$

as $P(W_1 = -1) = P(W_1 = -1) = 1/2$. Since (4.7) holds for all $a \leq b$, D_i is symmetrically distributed around 0. The second part of the theorem follows from part (i) and the fact that $W_i = -sign(D_{i-1})$. □

Since the distribution of U_i does not depend on any of the other variables, including W_i, part (ii) of Theorem 4.1 yields the following:

Corollary 4.1 *Marginally, each noise factor $R_i = 1 + W_i U_i$ is symmetrically distributed with $E(R_i) = 1$ and $V(R_i) = E[V(R_i|W_i)] + V[E(R_i|W_i)] = \sigma_U^2 + \mu_U^2$, where μ_U and σ_U^2 are the mean and variance of $f_U(.)$.*

Theorem 4.1 also implies that a perturbed cell total (T^*) is symmetrically distributed around the observed total (T) and hence T^* is an unbiased estimator of T. As the balanced noise procedure is applied at cell level, the cells must be predefined. However, in practice, data agencies are obliged to prepare and publish many different tables based on the same data set. Thus, while balanced noise masking of a microdata set must be done with a "reference" table, it is important to assess its effect on cell totals of other tables. By Corollary 4.1, the magnitude of perturbation of any value, i.e., $(y_i - y_i R_i)$, is symmetrically distributed around 0 and hence $\sum_{i \in A}(y_i - y_i R_i)$ is also symmetrically distributed around 0, for any set of units A. The main practical implication of this discussion is the following:

Proposition 4.1 *For any set of units A, the noisy total $\sum_{i \in A} y_i R_i$ is symmetrically distributed around the corresponding total in the original data set, i.e., $\sum_{i \in A} y_i$. So, for any cell in any table, the noisy total is an unbiased estimator of the true total.*

We shall now examine the noise variance for the total of a reference cell and the gain in data quality from the balancing procedure.

Theorem 4.2 *Suppose a cell in the reference table has n units with ordered values $y_1 \geq \cdots \geq y_n$. Then, the conditional variance of the perturbed total $T_* = y_1 R_1 + \cdots + y_n R_n$, given the original data, has the following representation:*

$$V(T_*) = \sigma_R^2 \sum_{i=1}^{n} y_i^2 - 2\mu_U \sum_{i=1}^{n-1} y_{i+1} E[|D_i|] \tag{4.8}$$

where $\sigma_R^2 = \mu_U^2 + \sigma_U^2$, and μ_U and σ_U^2 *are the mean and variance of $f_U(.)$.*

Proof From preceding definitions and discussions it can be verified easily that for $i = 2, \ldots, n$, (i) $D_i = D_{i-1} + W_i U_i y_i$, (ii) $E(W_i) = 0$ and $W_i^2 = 1$ with probability 1 and (iii) $D_{i-1} W_i = -|D_{i-1}|$ with probability 1. From these and the fact that $\{U_i\}$ are independent of all other variables we get

$$\begin{aligned} V(T_*) = V(D_n) &= V[D_{n-1} + W_n U_n y_n] \\ &= V(D_{n-1}) + V(W_n U_n y_n) + 2cov(D_{n-1}, W_n U_n y_n) \\ &= V(D_{n-1}) + E\left(U_n^2\right) y_n^2 + 2\mu_U y_n E[D_{n-1} W_n] \\ &= V(D_{n-1}) + \sigma_R^2 y_n^2 - 2\mu_U y_n E[|D_{n-1}|] \end{aligned} \tag{4.9}$$

The proof can now be completed easily by expanding the recurrence relation in (4.9) and noting that $V(D_1) = \sigma_R^2 y_1^2$. □

Comparing (4.8) with (3.2), we see that balanced noises reduce the noise variance of a cell total (in the reference table) by $2\mu_U \sum_{i=1}^{n-1} y_{i+1} E[|D_i|]$. Also, unlike in the case of independent noise multiplication where the variance of a perturbed total always increases with the addition of an extra value, here the variance of a perturbed total with increasing number of components may increase or decrease depending on the actual values being added as well as on the mean and the variance of the noise distribution. Specifically,

$$\begin{aligned} V(T_{(n+1)*}) - V(T_{n*}) &= \sigma_R^2 y_{(n+1)}^2 - 2\mu_U y_{n+1} E[|D_n|] \\ &= y_{(n+1)}\left\{\sigma_R^2 y_{(n+1)} - 2\mu_U E[|D_n|]\right\} \end{aligned}$$

which can be positive or negative.

5. Discussion

In this article we have presented some theoretical properties of multiplicative noise masking procedures for preserving confidentiality of private information in statistical databases. We showed that the sample moments and correlations based on the original data can be recovered unbiasedly from the masked data, and unbiased polynomial estimators based on the original data can be adapted easily for the masked data. These results are important from the data analysis perspective. We believe our results and discussions are helpful in clarifying the effects of multiplicative noise on tabular magnitude data. In particular, the results that the Evans et al. (1998) procedure has little effect on the total of a nonsensitive cell and that the balanced noise procedure of Massell and Funk (2007a, b) is unbiased are reassuring.

For assessing disclosure risk and choosing a noise distribution, in connection with the Evans et al. (1998) procedure, we considered a rather conservative scenario, where the intruder knows all values in a cell except that of the target unit. It would be useful to consider other and more realistic scenarios. One inherent difficulty in ascertaining disclosure risk is that different intruders have different target units as well as different prior information. We believe further research on modeling intrusion behavior and developing an aggregate measure of disclosure risk would be of much practical value.

The balanced noise method of Massell and Funk (2007a, b) is a useful procedure as it retains unbiasedness and at the same time reduces noise variances of the cell totals in the reference table. Intuitively, we expect the gain from balancing to depend on the choice of the reference table. This aspect, as well as how to choose the reference table, deserves further investigation. Other balancing methods, e.g., randomly order the units in each cell and then apply the procedure, may also be explored and compared.

Multiplicative noise masking is a useful tool for preserving confidentiality of private information in statistical databases. One attractive feature of multiplicative noise, for positive quantitative variables, is that it provides uniform record level protection to all values, as the noise CV is constant (same as the noise variance). However, multiplicative noise perturbation is not a panacea. Obviously, the procedure is not applicable to qualitative variables. Also, while moments and correlations can be estimated easily, estimation of other population parameters, such as quantiles, and adapting standard nonpolynomial estimators for applying to the perturbed data may be difficult. We hope to address some of these issues in future communication.

6. References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Brand, R. (2002). Microdata Protection through Noise Addition. In Inference Control in Statistical Databases, J. Domingo-Ferrer (ed.). Berlin: Springer, 97–116.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (Ed) (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: Elsevier.

Duncan, G.T. and Fienberg, S.E. (1999). Obtaining Information while Preserving Privacy: A Markov Perturbation Method for Tabular Data. In Eurostat Statistical Data Protection '98 Lisbon. Luxemburg: Eurostat, 351–362.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination. Journal of the American Statistical Association, 81, 10–28.

Duncan, G.T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Duncan, G.T. and Stokes, S.L. (2004). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. Chance, 17, 16–20.

Evans, T., Zayatz, L., and Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. Journal of Official Statistics, 4, 537–551.

Federal Committee on Statistical Methodology (2005). Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22 (2nd revision). U.S. Office of Management and Budget, Washington, DC.

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

Greenberg, B. and Zayatz, L. (1992). Strategies for Measuring Risk in Public Use Microdata Files. Statistical Neerlandica, 46, 33–48.

Hedayat, A.S. and Sinha, B.K. (1991). Design and Inference in Finite Population Sampling. New York: John Wiley.

Hwang, J.T. (1986). Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy. Journal of the American Statistical Association, 81, 680–688.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and Sanil, A.P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60, 224–232.

Keller-McNulty, S., Nakhleh, C.W., and Singpurwalla, N.D. (2005). A Paradigm for Masking (Camouflaging) Information. International Statistical Review, 73, 331–349.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 303–308.

Kim, J.J. and Winkler, W.E. (1995). Masking Microdata Files. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 114–119.

Kim, J.J. and Winkler, W.E. (2003). Multiplicative Noise for Masking Continuous Data. Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., April.

Lambert, D. (1993). Measure of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Marshall, A.W. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Application. New York: Academic Press.

Massell, P. and Funk, J. (2007a). Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata. Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III), Montreal, Canada.

Massell, P. and Funk, J. (2007b). Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection. Proceedings of the Research Conference of the Federal Committee on Statistical Methodology, Arlington, Virginia.

Reiter, J.P. (2005). Estimating Identification Risk in Microdata. Journal of the American Statistical Association, 100, 1101–1113.

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

Skinner, C.J. and Elliot, M.J. (2002). A Measure of Disclosure Risk for Microdata. Journal of the Royal Statistical Society, Series B, 64, 855–867.

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

Willenborg, L.C.R.J. and De Waal, T. (2001). Elements of Statistical Disclosure Control. New York: Springer.

Yancey, W.E., Winkler, W.E., and Creecy, R.H. (2002). Disclosure Risk Assessment in Perturbative Microdata Protection. Inference Control in Statistical Databases, J. Domingo-Ferrer (ed.). Berlin: Springer, 135–152.

Received March 2010
Revised August 2010