RESEARCH REPORT SERIES
(*Statistics #2003-01)*

Multiplicative Noise for Masking Continuous Data

Jay J. Kim and William E. Winkler

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: April 17, 2003

Multiplicative Noise for Masking Continuous Data

Jay J. Kim and William E. Winkler[1]

Abstract

To protect the identity of the persons or firms on a microdata file, noise is sometimes added to the data before releasing it to the public. There has been conjecture that, rather than adding noise, multiplying noise might better protect the confidentiality. Two forms of multiplicative noise are considered. The first approach is generating random numbers which have mean one and small variance, and multiplying the original data by the noise. The second approach is to take a logarithmic transformation, compute a covariance matrix of the transformed data, generate random number which follows mean zero and variance/covariance c times the variance/covariance computed in the previous step, add the noise to the transformed data and take an antilog of the noise added data. This paper investigates the statistical properties of both methods and shows how well they protect the identity of those on the file via re-identification trials.

Key Words: Microdata, confidentiality, mask, multiplicative noise.

1. Introduction

In 1993, the Department of Health and Human Services (HHS) commissioned the Bureau of the Census to create a microdata file by combining the 1991 March Current Population Survey (CPS) data with income data from the 1990 Internal Revenue Service (IRS) 1040 Income Tax Return file. The file was needed for statistical purposes in analyzing income tax policies and low-income supplemental payments. Income tax and other data needed to be masked in the resultant file so that both IRS and Bureau of the Census confidentiality requirements were met.

To satisfy the two conflicting requirements: (1) protect confidentiality for the people on the file, and (2) maintain analytic properties of the unmasked data, we used an additive noise approach (Kim 1986, 1990; Fuller 1993; Kim and Winkler 1995; Winkler 1998, Roque 2000, Yancey *et al.* 2002). We used the additive noise approach because it is easy to implement and does a good job satisfying both requirements. Alternative approaches that produce synthetic or simulated data (Kennickell 1999) can yield data that satisfy both requirements reasonably well but require considerable skill to implement. To further assure confidentiality of the file that was masked with additive noise, we performed data swapping (Dalenius and Reiss 1982) on a small part of the file (less than one percent) that might possibly be re-identified. Because of minimal data swapping, the main analytic properties of the data were maintained.

---

[1] Statistical Research Division, Bureau of the Census, Suitland, MD 20233. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Because multiplicative noise is often more suitable for economic modeling of income data, Hwang (1986), among others, has conjectured that it might be more suitable than additive noise in some situations. In this paper, we consider two forms of multiplicative noise. The first, called Multiplicative Noise Scheme I, is to generate normal random numbers which have mean 1, and multiply the original data by this noise. The second approach, called Multiplicative Noise Scheme II, is to take a logarithmic transformation on the unmasked data, compute a covariance matrix of the transformed data, generate random numbers which follow a normal distribution with mean $\underline{0}$ and variance/covariance which is c times the variance/covariance obtained in the previous step, add the noise to the transformed data, then take the antilog. The former was once used by the Energy Information Administration in the U.S. Department of Energy. Specifically, to mask the heating (and cooling) degree days, h_j, a random number, r_j, is generated from a normal distribution with mean 1 and variance .0225. The random number is truncated such that the resulting number e_j satisfies $.01 \le |e_j - 1| \le .6$. Note e_j is neither continuous nor discrete, but mixed due to truncation. The masked data $h_j e_j$ were released.

In this paper, we will investigate the statistical properties of both schemes mentioned above (sections 2 and 3) and try the schemes in masking Internal Revenue Service (IRS) income data, calculating the mean and the variance from the masked file in an effort to recover the mean and the variance [2] of the unmasked data (section 4). We also try to match the records in the masked file against those in the unmasked file (section 5).

2. Multiplicative Noise Scheme I

2.1. Masking Scheme

Let x_{ij} be the value for the i^{th} person's j^{th} characteristic, i = 1, 2,... n; j = 1, 2,... . p. We will denote the noise $e_{i1}, e_{i2} \ldots . e_{ip}$ corresponding to $x_{i1}, x_{i2} \ldots , x_{ip.}$ We let

$$y_{ij} = x_{ij}\, e_{ij}$$

where e_j [3] is a random variable following a normal distribution with mean μ_j and variance σ_j^2 before truncation. The noise is usually doubly truncated such as in the following equation

$$f(e) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2\sigma^2}(e-\mu)^2]}{\frac{1}{\sqrt{2\pi}\sigma}\int_A^B \exp[-\frac{1}{2\sigma^2}(e-\mu)^2]\,de} \qquad \text{for } A < e < B$$

---

[2] After this paper was drafted, a paper (see Muralidhar, et al) dealing with a multiplicative scheme came to the authors' attention. However, our current paper is more comprehensive.

[3] All e_{ij}'s for a given j follow the same distribution. That is, we assume an independent, identical distrubution for all i for a given j.

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(e-\mu)^2\right]}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)} \tag{1}$$

where A and B are the lower and upper truncation points and $\Phi(A)$ stands for the cumulative probability up to A. The above can be reexpressed as

$$\frac{K}{\sigma} Z\left(\frac{e-\mu}{\sigma}\right)$$

where $K = \frac{1}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)}$

and

$$Z(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} x^2\right]$$

The amounts of truncation are $\Phi\left(\frac{A-\mu}{\sigma}\right)$ from below and $1-\Phi\left(\frac{B-\mu}{\sigma}\right)$ from above.

2. 2. Properties of the Masked Data

2.2.1. Expected Value of y_j When $|e_j - \mu_j| \le c$

Since x_j and e_j are independent,

$$E(y_j) = E(x_j)\, E(e_j),$$

where, ignoring subscript j

$$E(e) = K \frac{1}{\sqrt{2\pi}\sigma} \int_A^B e \exp\left[-\frac{1}{2\sigma^2}(e-\mu)^2\right] de$$

$$= \mu + K\left[Z\left(\frac{A-\mu}{\sigma}\right) - Z\left(\frac{B-\mu}{\sigma}\right)\right]\sigma \tag{2}$$

The above integration was achieved by putting $h = \frac{e-\mu}{\sigma}$. Then $e = \sigma h + \mu$ and $de = \sigma dh$. From equation (2),

$$E(x) = \frac{E(y)}{\mu + K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]\sigma}$$

Since the data disseminator will release μ, σ, A and B, users can compute the expected value of the noise. Z(x) is the ordinate of the standard normal curve and Z(-x) = Z(x). If $A = -B$, then the bias of e is zero because $Z(\frac{-B-\mu}{\sigma}) = Z(\frac{B-\mu}{\sigma})$. If $A \neq B$, then the bias can be either positive or negative.

The variance of the noise can be calculated similarly.

$$V(y) = E(y^2) - [E(y)]^2$$

$$= E(x^2)E(e^2) - [E(x)E(e)]^2.$$

Now

$$E(e^2) = K\frac{1}{\sqrt{2\pi}\sigma}\int_A^B e^2 \exp[-\frac{1}{2\sigma^2}(e-\mu)^2]de. \tag{3}$$

Let $h = \frac{e-\mu}{\sigma}$, then $e = \sigma h + \mu$ and $de = \sigma dh$. The above equation becomes

$$E(\sigma h + \mu)^2 = K\frac{1}{\sqrt{2\pi}}\int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\mu}} (\sigma^2 h^2 + 2\sigma\mu h + \mu^2)\exp[-\frac{1}{2}h^2]dh. \tag{4}$$

Equation (4) can be evaluated as the sum of the following three components,

$$\text{Component (1)} \quad - \quad \sigma^2 K\frac{1}{\sqrt{2\pi}}\int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} h^2 \exp[-\frac{1}{2}h^2]dh \tag{5}$$

$$\text{Component (2)} \quad - \quad 2\sigma\mu K\frac{1}{\sqrt{2\pi}}\int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} h \exp[-\frac{1}{2}h^2]dh \tag{6}$$

$$\text{Component (3)} \quad - \quad \mu^2 K\frac{1}{\sqrt{2\pi}}\int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} \exp[-\frac{1}{2}h^2]dh \tag{7}$$

Component (1) can be evaluated by integration by parts. To do so, let

$$u = h \quad (\text{hence } du = dh)$$

and

$$dv = \sigma^2 K\frac{1}{\sqrt{2\pi}} h \exp[-\frac{1}{2}h^2]dh \quad (\text{hence } v = -\sigma^2 K\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}h^2]).$$

Equation (5) becomes

$$uv\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} - \int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} v\,du = -\sigma^2 K \frac{1}{\sqrt{2\pi}} h \exp[-\frac{1}{2}h^2]\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} + \sigma^2 K \frac{1}{\sqrt{2\pi}} \int_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} \exp[-\frac{1}{2}h^2]dh$$

$$= -\sigma^2 K h Z(h)\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} + \sigma^2 K \Phi(h)\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}}$$

Equations (6) and (7) reduce to

$$-2\sigma\mu K Z(h)\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} \quad \text{and} \quad \mu^2 K \Phi(h)\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}}, \text{ respectively.}$$

Observe that $\Phi(h)\Big|_{\frac{A-\mu}{\sigma}}^{\frac{B-\mu}{\sigma}} = K^{-1}$. Then

$$E(e^2) = \sigma^2 + \mu^2 + \sigma^2 K[\frac{A-\mu}{\sigma} Z(\frac{A-\mu}{\sigma}) - \frac{B-\mu}{\sigma} Z(\frac{B-\mu}{\sigma})]$$

$$+ 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]$$

If A = -B, the above reduces to

$$E(e^2) = \sigma^2 + \mu^2 + 2\sigma^2 K Z(\frac{A-\mu}{\sigma}).$$

$$[E(e)]^2 = \mu^2 + \sigma^2 K^2 [Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]^2 + 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})].$$

Finally,

$$V(y) = E(x^2)E(e^2) - [E(x)E(e)]^2$$

$$= E(x^2)\{\sigma^2 + \mu^2 + \sigma^2 K[\frac{A-\mu}{\sigma} Z(\frac{A-\mu}{\sigma}) - \frac{B-\mu}{\sigma} Z(\frac{B-\mu}{\sigma})]$$

$$+ 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]\}$$

$$- [E(x)]^2 \{\mu^2 + \sigma^2 K^2 [Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]^2$$

$$+ 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]\}. \qquad (8)$$

Since μ, σ, A and B will be known to users and (the estimate of) $E(x)$ can be easily calculated following the formula in section 2.2.1, the estimate of $V(x)$ also can be obtained.

If A = -B, the variance of y simplifies to

$$V(y) = E(x^2)[\sigma^2 + \mu^2 + 2\sigma^2 K Z(\frac{A-\mu}{\sigma})] - [E(x)]^2 \mu^2$$

and

$$V(x) = \frac{V(y) - \sigma^2 E(x^2)[1 + 2KZ(\frac{A-\mu}{\sigma})]}{\mu^2}. \quad (9)$$

3. Multiplicative Noise Scheme II

3.1. Masking Scheme

We define x_{ij} the same way as before.

Let

$$y_{ij} = \ln x_{ij}$$

$$V(\underline{Y}) = \Sigma,$$

where Σ is the variance/covariance matrix of variables $x_1, x_2, \ldots x_p$. We generate the random numbers following a multivariate normal distribution $N(\underline{0}, c\Sigma)$, where c is a positive number between zero and one. We denote the noise variables $e_1, e_2, \ldots e_p$.

Let $z_{ij} = y_{ij} + e_{ij}$

Thus u_{ij} = antilog of $z_{ij} = \exp(y_{ij} + e_{ij})$

$$= \exp[\ln x_{ij} + e_{ij}] = x_{ij} \exp[e_{ij}] = x_{ij} h_{ij}.$$

The values of some variables such as adjusted gross income can be negative. In that case, to be able to take logarithms on the variable, we suggest adding a small number (same number to all observations) to make all values positive.

3.2. Properties of the Masked Data in Logarithmic Scale

The multiplicative scheme such as $y = a x_1^{\beta_1} x_2^{\beta_2}$ is usually converted to the linear form by taking logarithms on both sides, i.e., $\ln y = \ln a + \beta_1 \ln x_1 + \beta_2 \ln x_2$.

In an additive regression model, when x_1 is exponentially distributed, x_1 is converted to $z_1 = \ln x_1$ and $y = a + \beta_1 z_1 + \beta_2 x_2$ is built. In this case, adding noise to the log-transformed variables makes perfect sense. That is, the properties of the additive noise demonstrated in Kim (1986) and Kim and Winkler (1995) hold in log-scale. The mean is unbiased, the unbiased variance/covariance in log-scale can be recovered and the unbiased subdomain estimates can be

easily obtained from the masked data in log-scale.

3. 3. Properties of the Masked Data

3.3.1. Expected Value of u

We let $\sigma^2 = cV(\ln x)$. Recalling $h = \exp(e)$,

$E(u) = E(x)E(h)$ due to the fact x and h are independent.

$$E[\exp(e)] = \int_{-\infty}^{\infty} \exp(e)\, f(e)\, de$$

$$= \int_{-\infty}^{\infty} \exp(e)\, \frac{1}{\sqrt{2\pi}\sigma}\, \exp[-\frac{1}{2\sigma^2} e^2]\, de$$

$$= \exp[\sigma^2/2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, \exp[-\frac{1}{2\sigma^2}(e - \sigma^2)^2]\, de = \exp[\sigma^2/2].$$

Then $E(u) = \exp[\sigma^2/2]\, E(x)$. (10)

On the average, the mean of the masked variable is $e^{\sigma^2/2}$ times that of the unmasked data. In order to have an unbiased mean of the masked variable, we need the variance of the noise. The variance of noise can be recovered from the masked data by first taking a log-transformation on the masked data, then by computing its variance and multiplying it by $\frac{c}{1+c}$. Then the mean of the unmasked data can be calculated from equation (10) as follows.

Let $\bar{u} = \frac{\sum u_i}{n}$. From equation (10), an unbiased estimator for the mean of unmasked data is

$$\frac{\bar{u}}{\exp[\frac{\sigma^2}{2}]}. \quad (11)$$

3.3.2. Variance of $\bar{u}$

$$V(u) = E(u^2) - [E(u)]^2$$

$$= E(x^2)\, E[\exp(2e)] - \{\exp[\sigma^2/2]\, E(x)\}^2$$

$$= E(x^2)\, E[\exp(2e)] - [E(x)]^2 \exp[\sigma^2].$$

Now

$$E[\exp(2e)] = \int_{-\infty}^{\infty} \exp(2e)\, \frac{1}{\sqrt{2\pi}\sigma}\, \exp[-\frac{e^2}{2\sigma^2}]\, de$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(e_j - 2\sigma^2)^2 + 2\sigma^2]\, de = \exp[2\sigma^2]$$

Then $V(u) = \exp[2\sigma^2]\, E(x^2) - \exp[\sigma^2][E(x)]^2$

and

$$E(x^2) = \frac{V(u)}{\exp[2\sigma^2]} - \frac{E(x)^2}{\exp[\sigma^2]}$$

The variance of x can then be expressed as follows.

$$V(x) = E(x^2) - [E(x)]^2$$

$$= \frac{V(u)}{\exp[2\sigma^2]} - \frac{E(x)^2}{\exp[\sigma^2]} - [E(x)]^2. \qquad (12)$$

3.3.3. Covariance of u_j and $u_{j'}$, $j \neq j'$.

$$Cov(u_j,u_{j'}) = E(u_j,u_{j'}) - E(u_j)E(u_{j'})$$

$$= E(x_j x_{j'} f_j f_{j'}) - E(x_j f_j)E(x_j f_{j'})$$

$$= E(x_j x_{j'}) E(f_j f_{j'}) - E(x_j) E(x_{j'}) \exp[\sigma^2].$$

Now

$$E[\exp(e_j + e_{j'})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[(e_j + e_{j'})] \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_j\sigma_{j'}} \exp[-\frac{1}{2(1-\rho^2)}(\frac{e_j^2}{\sigma_j^2} - 2\rho\frac{e_j}{\sigma_j}\frac{e_{j'}}{\sigma_{j'}} + \frac{e_{j'}^2}{\sigma_{j'}^2})]\, de_j de_{j'}$$

$$= \exp[\frac{\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2}{2}]$$

$$\times \int\int \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_j\sigma_{j'}} \exp[-1/2\{(\frac{e_j}{\sqrt{1-\rho^2}\sigma_j} - \frac{\rho e_{j'}}{\sqrt{1-\rho^2}\sigma_{j'}} - \sigma_j\sqrt{1-\rho^2})^2 + [\frac{e_{j'}}{\sigma_{j'}} - (\sigma_{j'} + \rho\sigma_j)]^2\}]\, de_j\, de_{j'}$$

Let $w = \frac{e_j}{\sqrt{1-\rho^2}\sigma_j} - \frac{\rho e_{j'}}{\sqrt{1-\rho^2}\sigma_{j'}} - \sigma_j\sqrt{1-\rho^2}$ and

$$v = \frac{e_{j'}}{\sigma_{j'}} - (\sigma_{j'} + \rho\sigma_j).$$

Then $dw = \frac{de_j}{\sqrt{1-\rho^2}\sigma_j}$ and $dv = \frac{de_{j'}}{\sigma_{j'}}$

The above equation becomes

$$\exp\left(\frac{\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2}{2}\right)\int\int \frac{1}{2\pi}\exp[-1/2(w^2+v^2)]\,dw\,dv = \exp\left(\frac{\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2}{2}\right).$$

Then

$$Cov(u_j, u_{j'}) = \exp\left(\frac{\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2}{2}\right)E(x_j, x_{j'}) - \exp\left(\frac{\sigma_j^2 + \sigma_{j'}^2}{2}\right)E(x_j)E(x_{j'}). \quad (13)$$

The multiplier of $E(x_j x_{j'})$ is different from that of $E(x_j)E(x_{j'})$ in the above. The covariance of x_j and $x_{j'}$ can be computed as follows.

$$Cov(x_j, x_{j'}) = \left\{\frac{\sum u_{ij}u_{ij'}}{\exp[(\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2)/2]} - \frac{n\bar{u}_j\bar{u}_{j'}}{\exp[(\sigma_j^2 + \sigma_{j'}^2)]}\right\}/(n-1). \quad (14)$$

The correlation coefficient ρ can be obtained from the noise-added variables. As the noise was generated to maintain the same correlation structure, the correlation between the noise-added variables will be on the average the same as that between the unmasked variables in log-scale. If $\rho = 0$, the covariance formula in (13) reduces to

$$Cov(u_j, u_{j'}) = \exp\left(\frac{\sigma_j^2 + \sigma_{j'}^2}{2}\right)Cov(x_j, x_{j'}).$$

4. A Numerical Example

4.1 Data to be masked

The data to be masked are eight income fields from the 1991 Internal Revenue Service (IRS) 1040 Tax Return File. The eight fields are i) Wage and Salary Income, ii) Taxable Interest Income, iii) Dividend Income, iv) Rental Income, v) Non-Taxable Interest Income, vi) Social Security Income, vii) Total Income and viii) Adjusted Gross Income.

4.2 Numerical Example of Scheme I.

We tried the scheme that EIA once used. That is, random numbers, e_j are generated from a normal distribution with mean 1 and variance .0225. The generated random numbers are truncated such that the resulting numbers e_j satisfy $.01 \le |e_j - 1| \le .6$. This translates into i). $.4 \le e_j \le .99$ or ii). $1.01 \le e_j \le 1.6$. The density function of e_j is

$$f(e_j) = \frac{\frac{1}{.15\sqrt{2\pi}} \exp[-\frac{1}{.045}(e_j - 1)^2]}{\Phi(\frac{1.6-1}{.15}) - \Phi(\frac{1.01-1}{.15}) + \Phi(\frac{.99-1}{.15}) - \Phi(\frac{.4-1}{.15})}$$

Following equation (2),

$$E(e_j) = 1 + .15\frac{z(4) - z(.0667) + z(-.0667) - z(-4)}{\Phi(4) - \Phi(.0667) + \Phi(-.0667) - \Phi(4)}$$

Since Z(-x) = Z(x), the numerator becomes zero and $E(e_j)$ becomes 1. Thus the mean is unbiased. The following table shows the means from the unmasked and masked data.

Table 1. Mean of Masked (Based on Scheme I) and Unmasked Data, n=59,315

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Masked	23,821	1,825	583	1,189	337	945
Unmasked	23,799	1,825	587	1,190	342	947

As seen in the table, the estimates of the means from the masked data are all close to those from the unmasked data.

$$E(e_j^2) = V(e_j) + [E(e_j)]^2 = .0225 + 1 = 1.0225$$

From equation (8),

$$V(y_j) = E(x_j^2)(\sigma_j^2 + \mu_j^2) - [E(x_j)]^2\mu_j^2$$

$$= \mu_j^2 V(x_j) + \sigma_j^2 E(x_j^2)$$

$$= \mu_j^2 V(x_j) + \sigma_j^2\{V(x_j) + E[(x_j)]^2\}.$$

Since $E(x_j) = E(y_j)$,

$$V(x_j) = \frac{1}{\mu_j^2 + \sigma_j^2}\{V(y_j) - \sigma_j^2[E(y_j)]^2\}$$

which is

$$\frac{1}{1.0225}\{V(y_j) - .0225[E(y_j)]^2\}$$

Using the above expression, the standard deviation of Wage, Taxable Interest, Dividend, Non-Taxable Interest, Rent and Social Security Income is calculated. In the table below, these estimates and the standard deviations of the unmasked data are shown.

Table 2 Standard Deviation of Masked (Based on Scheme I) and Unmasked Data

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Masked	40,423	8,069	6,131	22,089	15,568	3,202
Unmasked	44,221	7,982	6,378	21,986	17,007	3,205

The standard deviations for four items obtained from the masked data are close to those from the unmasked data. However, for the remaining two items (Wage and Non-Taxable Interest), the standard deviation of the masked data are close to 9 percent off from that of the unmasked data.

4.3. Numerical Example of Scheme II

The masking scheme was applied with the c-values (as shown in section 3.1) of .01 and .10. Since many income fields have zero entry and logarithm cannot be taken on zero, 1.0 was added to every entry in the data set and the resulting data are masked. The variance and covariance (hence correlation) are location-invariant. We need to subtract one (1) from the masked mean to retrieve the mean of the original data. The means recovered from the masked data are as follows.

Table 3. Mean of Masked (Based on Scheme II) and Unmasked Data, c=.01

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Masked	23,787	1,846	588	1,162	337	952
Unmasked	23,799	1,825	587	1,190	342	947

The mean estimates from the masked data with c=.01 are all very close to those from the unmasked data.

Table 4 shows similar data for the standard deviation.

Table 4 Standard Deviation of Masked (Based on Scheme II) and Unmasked Data, c=.01

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Masked	29,887	8,101	6,262	15,600	15,080	2,944
Unmasked	44,221	7,982	6,378	21,986	17,007	3,205
Difference	-32.4 %	1.5 %	-1.8 %	-29.1 %	-11.3 %	-8.1 %

The standard deviation is severely underestimated for Wage and Rent (32.1 and 29.1 percent, respectively). The estimated standard deviation for Non-Taxable Interest and Social Security income is substantially low (11.3 and 8.1 percent, respectively).

Table 5. Mean of Masked (Based on Scheme II) and Unmasked Data, c=.10

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Masked	24,266	1,901	581	1,137	322	957
Unmasked	23,799	1,825	587	1,190	342	947

The means obtained from the masked data using c=.10 are in a fairly close range of those from the unmasked data. In comparison with those with c=.01, they are much farther off from the mean of the unmasked data. However, this can be expected as the new data have ten times higher noise in the log-scale.

Table 6 Standard Deviation of Masked (Based on Scheme II) and Unmasked Data, c=.10

	Wage	Taxab Int	Dividend	Rent	N_TAX Int	SS Inc
Masked	74,732	8,122	4,936	10,388	10,324	3,000
Unmasked	44,221	7,982	6,378	21,986	17,007	3,205
Difference	69.0 %	1.8 %	-22.6 %	-52.8 %	-39.3 %	-6.4 %

Except for Taxable Interest (and probably Social Security income), the masked data have the standard deviation wildly different from the standard deviation of the unmasked data. Sometimes, the difference is more than 50 percent of the standard deviation of the unmasked data. This comparison is not in log-scale.

As noted before, the users are interested in these data in log-scale. Thus, it is proper to compare the unmasked and masked data in log-scale. We will compare them when c = .01.

Table 7. Mean of Masked (Based on Scheme II) and Unmasked Data in Log-Scale, c=.01

	Wage	Taxab Int	Dividend	Rent	N_TAX Int	SS Inc
Masked	8.296	3.928	1.246	0.766	0.269	0.912
Unmasked	8.297	3.928	1.248	0.768	0.270	0.911

The means of the masked data in log-scale are almost identical with those of the log-transformed unmasked data.

Table 8 Standard Deviation of Masked (Based on Scheme II) and Unmasked Data, in Log-Scale c=.01

	Wage	Taxab Int	Dividend	Rent	N_TAX Int	SS Inc
Masked	3.552	3.306	2.600	2.445	1.437	2.707
Unmasked	3.569	3.321	2.623	2.462	1.445	2.719
Difference	-0.48 %	- 0.45 %	- 0.88 %	- 0.690%	- 0.55%	- 0.44%

The standard deviations of the masked data in log-scale are again almost identical with those of the unmasked in log-scale.

The covariances between the masked variables in log-scale are very close to those between the unmasked variables. The correlation structure of the unmasked data carries over to the masked data in log-scale. Thus, this masked data set almost perfectly satisfies the users' requirements.

5. Re-identification of the Records in the File

As part of our original work with additive noise (Kim and Winkler 1995), we had to maintain statistics such as means and covariances on pre-specified subdomains determined by age, race, and sex. The most easily re-identified records are those that are outliers in the point cloud determined by the quantitative variables. These types of outliers are typically both sample and population outliers. Our re-identification rates provide an upper bound

Our matching metrics for individual quantitative variables are determined by the types of noise that are added. If additive noise is used, then the deviations between masked and original variables are on the additive scale. If multiplicative noise is used, then the deviations are on the log scale. The metrics adjust agreement weights downward as the masked variables differ from the original (or intruders') values by greater amounts. Depending on the characteristics of the data being matched, deviations between 10 and 20 percent typically get a full disagreement

weight. If there are eight quantitative variables being matched, then it is possible that only three variables may be needed to re-identify some outliers.

Our re-identification rates provide an upper bound on the re-identification rates an actual intruder might obtain. We match the masked sample file against the unmasked sample file because we do not have the original population files of unmasked IRS data. An intruder might construct a file using various public and semi-public data sources and attempt re-identification. The additive or multiplicative noise are typically capable of effectively masking virtually all points in the interior of the point clouds. Few can be re-identified. If we had the original population files the re-identification rates would be much lower for the exceptionally small proportion of interior points that are re-identified. Most of the re-identifications are of outliers that can be re-identified because they are population outliers. Depending on the quality of the external files available to an intruder, we expect an intruder's re-identification rates to be slightly to substantially lower than our re-identification rates.

During matching, we used a d-metric for quantitative variables when additive noise was applied and an l-metric when multiplicative noise was applied. The software allows the user to specify a value between 0.001 and 0.999 with the default being 0.20. A full agreement weight is adjusted downward toward the full disagreement weight as the proportional difference between the two values being compared increases (see e.g., Yancey *et al.* 2002). The EM algorithm is used to get the optimal probabilities for separating matches (re-identifications) from non-matches (non-re-identifications). More details are given in Kim and Winkler (1995). The other measure is the same as the first, but difference is in log-scale. The former is called d-metric and the latter l-metric. An efficient linear sum assignment algorithm forces 1-1 matching in a manner that further increases the re-identification rate (see e.g., Winkler 1998). The match rate is summarized as follows.

Table 7. Match Rate

	d-metric	l-metric
Scheme I	-	41 %
Scheme II with c = .01	8 %	8 %
Scheme II with c = .10	4 %	10 %

Scheme I has the highest match rate using l-metric, which is 41 percent. This is probably predictable since around 49.5 percent of the noise multiplied to the unmasked data lies within the range of .9 and 1.1.

It is surprising, concerning Scheme II, to find out that adding more noise does not necessarily protect the file better. That is, using l-metric we could re-identify the masked records more often with c=.10 than with c=.01. This is likely an artifact of how the actual sample of noise affects

the masked data. With a different seed number for the random number generator, we would expect the re-identification rate with higher amounts of noise to be somewhat lower. The match rate for the file masked by additive noise was 0.8 percent and with a combination of additive noise and swapping of easily re-identified records was less than 0.1 percent (Kim and Winkler 1995).

6. Concluding Remarks

Two forms of multiplicative noise have been examined. The first is based on generating random numbers that have a truncated normal distribution with mean 1 and small variance and multiplying the original data by the numbers. The second approach is to take a logarithmic transformation, compute a covariance matrix of the transformed data, generate random numbers which follow mean $\underline{0}$ and variance/covariance c times the variance/covariance computed in the previous step, add this noise to the transformed data and take the antilog of the noise-added data. Both schemes were tried on IRS income data.

Table 8. Comparison of Means for the Schemes

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Scheme I	23,821	1,825	583	1,189	337	945
Scheme II, c=.01	23,787	1,846	588	1,162	337	952
Scheme II, c=.10	24,266	1,901	581	1,137	322	957
Unmasked	23,799	1,825	587	1,190	342	947

The above table shows that the first scheme has, in general, means closer to the means of the unmasked data. Means using Scheme II with c=.01 are always closer to the means of the unmasked data than those from Scheme II with c=.10.

Table 9. Comparison of Standard Deviations for the Schemes

	Wage	Taxab Int	Dividend	Rent	N_Tax Int	SS Inc
Scheme I	40,423	8,069	6,131	22,089	15,568	3,202
Scheme II, c=.01	29,887	8,101	6,262	15,600	15,080	2,944
Scheme II, c=.10	74,732	8,122	4,936	10,388	10,324	3,000
Unmasked	44,221	7,982	6,378	21,986	17,007	3,205

Among the three schemes above, except for Dividend, Scheme I has the best standard deviations.

Comparing Scheme II with c=.01 to Scheme II with c=.10, we can notice that Scheme II with c=.01 is better except for Social Security income.

In terms of mean and variance, Scheme I looks best among the three schemes considered. The variance for some items for Scheme II is too unreliable. However, the mean and variance in log-scale for Scheme II are very close to those of the unmasked. Thus if the users are interested in the statistics in log-scale, then Scheme II is excellent in retaining the data utility.

In terms of match rate, Scheme I is worst. This may be to a limited degree overcome if we use normally distributed random numbers having a mean more than 20 percent from 1. However, the resulting numbers would be more different than the current ones from the unmasked, which some users might not like.

In conclusion, Scheme I may be good if the data disseminator wants to make minor changes to the original data. However, this is in exchange for data security. On the surface, this scheme seems to change the data more than the additive noise mode, but by taking logarithms on the data, it turns into an additive noise scheme. Scheme II destroys data utility for some items. It should be noted, however, that Scheme II maintains the data utility well in log-scale.

References

Dalenius, T. and Reiss, S.P. (1982), "Data Swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, **6**, 73-85.

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

Hwang, J. T. (1986), "Multiplicative Error-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy," *Journal of the American Statistical Association,* **81** (395), 680 - 688.

Johnson, N. and Kotz, S. (1970), *Continuous Univariate Distributions - 1, Distributions in Statistics*, John Wiley & Sons: New York.

Kennickell, A. B. (1999), "Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 248-267 (available at http://www.fcsm.gov).

Kim, J. (1986), "A method for limiting disclosure in microdata based on random noise and transformation," American Statistical Association 1986 Proceedings of the Section on Survey Research Methods, 370-374.

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," *American Statistical Association*,

Proceedings of the Section on Survey Research Methods, 456-461.

Kim, J. and Winkler, W. (1995), "Masking IRS Income Data on A Merged File between 1990 CPS File and IRS Income Tax Return File," *American Statistical Association*, *1995 Proceedings of the Section of Survey Research Methods*, 114-119.

Muralidhar, K., Batrah, D. and Kirs, P.J. (1995), "Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach.," *Management Science*, **41** (9), 1549-1584.

Roque, G.M. (2000), "Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. Dissertation, Department of Statistics, University of California at Riverside.

Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, 50-69.

Yancey, W.E., Winkler, W.E., and Creecy, R.H. (2002), "Disclosure Risk Assessment in Pertubative Microdata Protection," In (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.