Multiplicative Noise for Masking Continuous Data

Jay J. Kim and William E. Winkler, Bureau of the Census
Jay J. Kim, 4401 Silver Hill Road, Suitland, MD 20233 jay.jong.ik.kim@census.gov

Key Words : Additive Noise, Masking, Multiplicative Noise, Reidentification

I Introduction

Around 1994, the Bureau of the Census was commissioned by the Department of Health and Human Services (HHS) to create a microdata file by combining the 1991 March Current Population Survey (CPS) data which has income components with the 1990 Internal Revenue Service (IRS) 1040 Income Tax Return file. The Treasury Department allowed us to use their data for creating this file, but with one condition. We had to mask their income component which was composed of eight income variables. On the other hand, HHS was going to use the file to conduct their research on welfare reform. Thus we had to meet two somewhat conflicting requirements : (1) protect confidentiality for the people on the file, (2) maintain analytic properties of the unmasked data.

Additive noise approach (Kim 1986, 1990; Fuller 1993; Kim and Winkler 1995; Winkler 1998 and Roque 2000) can satisfy both requirements and is easy to implement. Synthetic or simulated data approach (Kennickell 1997, 1999) meets the requirements well, but requires good skill to implement. At the time we used additive noise plus data swapping approach for masking this file.

Additive noise is used more often than others in masking microdata files. However, there has been conjecture that, the multiplicative noise approach might do a better job protecting the confidentiality. Two forms of multiplicative noise are considered in this paper. The first is generate random numbers which are around mean 1, and multiply the original data by the noise (which will be called Multiplicative Noise Scheme I). The second approach is to take a logarithmic transformation on the unmasked data, compute a covariance matrix of the transformed data, generate normal random numbers which follow mean $\underline{0}$ and c times the variance/covariance computed in the previous step, add the noise to the transformed data and take antilog of the noise added data (which will be called Multiplicative Noise Scheme II). The former was used by the Energy Information Administration in the U.S. Department of Energy. Specifically, to mask the heating (and cooling) degree days, h, a random number, e, is generated from a normal distribution with mean 1 and variance .0225. The random number is truncated such that the resulting number satisfies

$.01 \le |e-1| \le .6$. Then masked data were released.

In this paper, we will investigate statistical properties of both schemes mentioned above (section II and III) and try the schemes in masking IRS income data mentioned above, calculating mean and variance in an effort to recover the original mean and variance[1] (section IV). We also try to match the records in the masked file against those in the unmasked file (section V).

II. Multiplicative Noise Scheme I

II.1. Masking Scheme

Let x_{ij} be the value for the i^{th} person's j^{th} variable, i = 1, 2,.., n; j=1, 2,.. .p;

We will denote the noise variables $e_{.1}, e_{.2}, , , e_{.p}$, corresponding to unmasked variables $x_{.1}, x_{.2}, , , x_{.p}$.

Let $y_{.j} = x_{.j} e_{.j}$, where $e_{.j}$ follows normal distribution with mean μ_j and variance σ_j^2 before truncation. For convenience, we will drop subscript j from μ_j and σ_j^2. We also ignore the dot in the subscript. Note the noise is usually doubly truncated such as

$$f(e_j) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(e_j-\mu)^2}}{\frac{1}{\sqrt{2\pi}\sigma}\int_A^B e^{-\frac{1}{2\sigma^2}(e_j-\mu)^2} de_j}$$

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(e_j-\mu)^2}}{\Phi(\frac{B-\mu}{\sigma}) - \Phi(\frac{A-\mu}{\sigma})} \quad .... (1)$$

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

$$\frac{K}{\sigma} Z(\frac{e_i-\mu}{\sigma})$$

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

II. 2. Properties of the Masked Data

II.2.1. Expected Value of y_j When $|e_j - \mu| \le c$

$E(y_j) = E(x_j)E(e_j)$ due to the fact x_j and e_j are independent.

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

---

[1] 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 much more comprehensive

From the equation

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

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

The variance of noise can be calculated similarly.

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

$$= E(x_j^2)E(e_j^2) - [E(x_j)E(e_j)]^2$$

Now

$$E(e_j^2) = K\frac{1}{\sqrt{2\pi}\,\sigma}\int_A^B e_j^2 \exp[-\frac{1}{2\sigma^2}(e_j-\mu)^2]de_j \quad \ldots . (3)$$

which is, after some algebra

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

Note if A = -B, the above reduces

$$E(e_i^2) = \sigma^2 + \mu^2 + 2\sigma^2 KZ(\frac{A-\mu}{\sigma}).$$

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

Finally,

$$V(y_i) = E(x_i^2)\{\sigma^2+\mu^2+\sigma^2 K[\frac{A-\mu}{\sigma}Z(\frac{A-\mu}{\sigma}) - \frac{B-\mu}{\sigma}Z(\frac{B-\mu}{\sigma})]$$
$$+ 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]\}$$
$$-[E(x_i)]^2\{\mu^2+\sigma^2K^2[Z(\frac{A-\mu}{\sigma})-Z(\frac{B-\mu}{\sigma})]^2$$
$$+ 2\sigma\mu K[Z(\frac{A-\mu}{\sigma}) - Z(\frac{B-\mu}{\sigma})]\} \quad .. (4)$$

Since μ, σ, A and B will be known to users and (the estimate of) $E(x_j)$ can be easily calculated following the formula in section II.2.1. $V(x_j)$ also can be captured.
If A = -B, the variance of y_j simplifies to

$$V(y_j) = E(x_j^2)[\sigma^2+\mu^2+2\sigma^2 KZ(\frac{A-\mu}{\sigma})] - [E(x_j)]^2\mu^2$$

Thus

$$V(x_j) = \frac{V(y_j) - \sigma^2 E(x_j^2)[1+2KZ(\frac{A-\mu}{\sigma})]}{\mu^2}$$

III. Multiplicative Noise Scheme II

III.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 $x_{.1}, x_{.2}, , , x_{.p}$

We generate the random number following multivariate normal distribution $N(\underline{0}, c\Sigma)$, where c is a positive number but less than 1. We denote the noise variables $e_{.1}, e_{.2}, , , , e_{.p}$, corresponding to unmasked variables. $x_{.1}, x_{.2}, , , , x_{.p}$.

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

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

$$= e^{\ln x_{ij} + e_{ij}} = x_{ij}e^{e_{ij}}$$

Note values of some variables such as rental income can be zero. In that case, to be able to take logarithm of the variable, we suggest adding a number to make all values positive.

III. 2. Properties of the Masked Data in Logarithmic Scale

The multiplicative scheme such as $y = ax_1^B x_2$ is usually converted to the linear form by taking logarithms on both sides, i.e., $\ln y = \ln a + \beta \ln x_1 + \ln x_2$. 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 can be recovered and unbaised subdomain estimates can be obtained from the masked data in log-scale.

III. 3. Properties of the Masked Data

III.3.1. Expected Value of u_j

For convenience, we will let $\sigma^2 = cV(\ln x_{.j})$.
$E(u_{.j}) = E(x_{.j})E(f_{.j})$ due to the fact x_j and f_j are independent. Again we will ignore the dot in the subscript.

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

Thus $E(u_j) = e^{\sigma^2/2} E(x_j)$

Thus on the average the mean of the masked variable is $e^{\sigma^2/2}$ times that of the unmasked data.

Note in order to have an unbiased mean of the masked variable, we need the variance of noise. The variance of noise can be recovered from the masked data by first taking log-transformation on the masked data, compute its variance and multiply it by $\frac{c}{1+c}$. The mean of the unmasked data can be calculated as follows.

Let $\bar{u}_j = \frac{\sum \bar{u}_{ij}}{n}$. Then

$$\bar{x}_j = \frac{\bar{u}_j}{e^{\frac{\sigma^2}{2}}}$$

III.3.2. Variance of $\bar{u}_j$

$$V(u_j) = e^{2\sigma^2} E(x_j^2) - e^{\sigma^2}[E(x_j)]^2$$

and

$$E(x_j^2) = \frac{V(u_j)}{e^{2\sigma^2}} - \frac{E(x_j)^2}{e^{\sigma^2}}$$

The variance of x_j can then be expressed as follows.

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

$$= \frac{V(u_j)}{e^{2\sigma^2}} - \frac{E(x_j)^2}{e^{\sigma^2}} - [E(x_j)]^2$$

III.3.3. Covariance of u_j and $u_{j'}$, j $\neq$ j'.

$$Cov(u_j, u_{j'}) = e^{\frac{\sigma_j^2 + 2\rho\sigma_j\sigma_{j'} + \sigma_{j'}^2}{2}} E(x_j, x_{j'})$$

$$- e^{\frac{\sigma_j^2 + \sigma_{j'}^2}{2}} E(x_j) E(x_{j'})$$

Note the multiplier of $E(x_j x_{j'})$ is different from that of $E(x_j)E(x_{j'})$ in the above.

Covariance of x_j and $x_{j'}$ can be computed as follows.

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

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. Note if $\rho = 0$, the covariance formula above reduces to

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

IV. A Numerical Example

IV.1 Data to be masked

The data to be masked is eight income fields from the 1991 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.

IV.2 Numerical Example of Scheme I.

We tried the scheme EIA used. That is, a random number, e_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 $.01 \leq |e_j - 1| \leq .6$. This translates into i). $.4 \leq e_j \leq .99$ or ii). $1.01 \leq e_j \leq 1.6$. Thus the density function of e_j is

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

Following equation (2),

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

Since Z(-x) = Z(x), the numerator becomes zero and $E(e_j)$ becomes 1. Thus the mean is unbiased. Table 1 shows the means from the unmasked and masked data. As seen in the table, estimates of 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 (4),

$$V(y_j) = E(x_j^2)(\sigma^2 + \mu^2) - [E(x_j)]^2\mu^2$$

$$= \mu^2 V(x_j) + \sigma^2 E(x_j^2)$$

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

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

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

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 Table 2, these estimates and those from the unmasked data are shown. 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 is close to 9 percent off from that of the unmasked data.

IV.3. Numerical Example of Scheme II

The masking scheme was twice applied to the same data set as before. C value (as shown in section III.1) of .01 and .10 was used for masking. Since many income fields have zero entry and logarithm cannot be taken on zero, one (1) was added to every entry in the data set and the resulting data is masked. Note variance and covariance (hence correlation) are location-invariant. However, we subtract one (1) from the mean to retrieve the mean of the original data. The means recovered from the masked data are in Tables 3 and 5. The means estimated 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. 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).

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

Except for Taxable Interest (and probably Social Security income), the masked data has the standard deviation (with c=.10) wildly different from the standard deviation of the unmasked data (see Table 6). Sometimes, the difference is more than 50 percent of the standard deviation of the unmasked data.

V. Re-identification of the Records in the File

Records in the masked file were matched against thosein the unmasked file. The re-identification rates provide an upper bound on the re-identification rate that might be obtained using the public-use data and external files. Two measures were used for matching. The first is based on the proportional difference between the masked and unmasked values. 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. 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. Additionally, 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 in Table 7.

Our matching rule is, roughly speaking, if the difference is within 20 percent of the smaller of the masked and unmasked values, it is declared a match. The l-metric applies this rule to the log-transformed data. Scheme 1 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 bigger 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.

Note that 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).

VI. Concluding Remarks

Two forms of multiplicative noise have been examined. The first is based on generating random numbers which are around 1, 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 $\underline{0}$ and variance/covariance c times the variance/covariance computed in the previous step, add the noise to the transformed data and take antilog of the noise added data. Both schemes were tried on IRS income data.

The numerical examples 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.

Among three schemes above, except 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 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. 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.

Normal assumption was made for the unmasked data. Lognormal may be more realistic one. We also can develop the scheme without distributional assumption. Covariance formula is available for Scheme II, but not for Scheme I. So our future task is investigate schemes under the alternative (or no) assumption.

VII. References

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*, Vol. 81, No. 395, pp 680 - 688.

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

Kim, J (1985), A method for limiting disclosure in microdata based on random noise and transformation. *American Statistical Association 1986 Proceedingsof the Section on Survey Research Methods*, pp 370-374.
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*, pp. 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* Vol. 41, No. 9. pp 1549-1584
Winkler, W. E. (1998), Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, *Research in Official Statistics*, 50-69.

Appendix

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

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

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

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 %

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

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 %

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 %

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

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