A METHOD FOR LIMITING DISCLOSURE IN MICRODATA BASED ON RANDOM NOISE AND TRANSFORMATION

Jay Kim, Bureau of the Census

I. Introduction

Survey data is often released as microdata. Survey respondents are thus subjected to the risk of reidentification and disclosure of confidential data, even when identifying information such as name and address is deleted prior to release of data. To avoid this disclosure problem, measures of masking the data have been proposed. They include adding random error, multiplying by random error, microaggregating, data swapping, random rounding, slicing and combining subrecords. Two reseachers compared those measures with respect to their masking capability and impact on key statistics. Specifically, Spruill (1983) performed an empirical study of comparison of additive random noise, multiplicative random error, microaggregation, random rounding and data swapping methods with regard to the effect of masking on key statistics. She also performed a reidentification experiment based on the distance measure of absolute deviation and squared deviation.

Paass(1985) also performed a reidentification experiment based on a refined measure of identification including discriminant analysis. He found from his experiment that the addition of random error is not an effective measure and hence proposed new masking schemes such as slicing and subrecords-combination.

As has been shown in both studies, some measures maintain the unbiased values of summary statistics such as mean and standard deviation but others lose the unbiasedness of the data. Also some schemes preserve the original structural relations and hence original causal relationships. However, others don't. According to Paass, the combination method which is best suitable for masking caused serious distortion of relationships among variables. This squarely puts us in the quandary as to whether or not we opt for protection in spite of grave sacrifice of usefulness of the data. From the users' point of view, maintenance of the usefulness of the data is the abiding requirement for a good masking scheme.

At the Bureau of the Census, we have been faced with masking microdata files. For masking earnings data, a new scheme has been developed. The scheme is a combination of random noise inoculation and transformation. In this paper I will describe this new measure and provide examples of application of the measure on the earnings data. Since multiple regression is the primary use of the earnings data, I will discuss the theoretical effects of masking on the regression.

It should be mentioned that the power of limiting the disclosure by this scheme has not been fully investigated. We are presently planning on performing reidentification experiment using the software developed by Paass' group.

An advantage of the scheme proposed here is, if users are willing to do multiplication to get an unbiased estimate of the second moment of the original (unmasked) variables, then we can compact the data points around the mean while the correlation structure is not hampered. This can be done by using a small "a" value, as to be seen later.

For simplicity, the derivation of formulae is based on the unweighted data.

II. New Scheme

II.1 Transformation on The Variable to Which Random Noise Was Added

As mentioned in section 1, Paass (1985) found that the addition of random noise alone is not sufficient for reducing disclosure risk. He also found that as more data points cluster in a given space it becomes more difficult to reidentify respondents. It implies that as the number of source (original) data points which can be linked to a given masked data point increases, the probability of linking a masked data point to the correct original data point decreases. The new scheme originated from this perspective. That is, by this transformation we try to add an additional layer of protection to persons on the file without harming original interrelationships among the unmasked variables. This is possible since the correlation between variables is invariant under a linear transformation of the variables.

Assume there are p variables, some of which are to be masked. Assuming the i^{th} variable is to be masked, define

x_i: the variable to be masked

e_i: random noise to be added to x_i.

We generate e_i such that e_i are independent of x_i. Let $x_i \sim (\mu_i, \sigma_i^2)$, $e_i \sim (0, c\sigma_i^2)$ and $\mathrm{Cov}(e_i, e_j)_{i \neq j} = c\ \mathrm{Cov}(x_i, x_j)_{i \neq j}$, assuming x_j is also to be masked. In the above c is a constant and the distribution of e_i can be selected from among two distributions; normal distribution and the distribution of x_i.

For the i^{th} variable, define

$$y_{ij} = x_{ij} + e_{ij}, \quad i=1,2,\ldots,p; \quad j=1,2,\ldots,n_i.$$

This is the usual additive random noise model.

For simplicity, assume $n_i = n$, $\forall\, i$. [1/] Here it is proposed to transform y_{ij} by

$$z_{ij} = a\, y_{ij} + b_i \qquad (1)$$

where a and b_i are constants and determined in two different ways. The first approach is to subject a and b_i to two constraints $E(x_i)=E(z_i)$ and $V(x_i)=V(z_i)$. The second approach requires determination of b_i by the first constraint $E(x_i)=E(z_i)$ but determination of a depends on the confidentiality requirement. In this paper we adopt the first approach.

The two constraints are such that the first and second moments of the transformed noise-added-variable are identical to those of the original variable.

First, by subjecting the transformation to the constraint $E(x_i) = E(z_i)$, we obtain $a\,\mu_i + b_i = \mu_i$. Hence

$$b_i = (1 - a)\,\mu_i \qquad (2)$$

By replacing μ_i in (2) by its estimate $\bar{x}_i$ or $\bar{y}_i$ ($\bar{y}_i = \bar{x}_i + \bar{e}_i$), and by substituting $(1\text{-}a)\,\bar{x}_i$ or $(1\text{-}a)\,\bar{y}_i$ for b_i in (1),

$$z'_{ij} = a\, y_{ij} + (1 - a)\,\bar{y}_i \qquad (3)$$

or

$$z_{ij} = a\, y_{ij} + (1 - a)\,\bar{x}_i . \qquad (4)$$

Based on the transformation in (3),

$$V(z_i) = (1 + c)\left\{a^2 + [2a(1 - a) + (1 - a)^2]/n\right\}\sigma_i^2 . \qquad (5)$$

When this equation is solved for "a" under the constraint $V(z_i) = V(x_i)$, one obtains

$$a = \sqrt{\frac{n - 1 - c}{(n - 1)(1 + c)}} \qquad (6)$$

Note that this "a" value will make the coefficient of σ_i^2 (i.e., $V(x_i)$) in equation (5) equal to 1. When n is large,

$$a \tilde{=} 1/\sqrt{1+c}.$$

II.2 Properties of Transformed Variable, z_i. When σ_i^2 Is Known

When σ_i^2 is known, noise is generated using σ_i^2. However, since the generated sample of noise is finite, the estimated variance of $c\,\sigma_i^2$ is to be different from $c\,\sigma_i^2$. The estimated variance is going to be denoted by $c\,s_i^2$.

Usually, microdata is created by taking a sample. Hence even if σ_i^2 is known to the survey takers such as the Bureau of the Census, the microdata users do not know σ_i^2 and hence have to estimate it. The estimate is again s_i^2.

1. $E(z_i) = \mu_i$ (7)

This follows since $E(y_i) = E(x_i + e_i) = \mu_i$, and $E(\bar{y}_i) = \mu_i$.

2. $\bar{z}_i = \bar{y}_i$ (8)

This can be proved easily. This implies that the sample mean of the transformed variable is the same as that of the noise added original variable.

3. $E(\bar{z}_i) = \mu_i$

4. $V(z_i) = (1+c)\{a^2 + [2a(1-a) + (1-a)^2]/n\}\sigma_i^2$

5. $V(z'_i) = \{(1+c)a^2 + [2a(1-a) + (1-a)^2]/n\}\sigma_i^2$ (9)

This follows since $Cov(x_{ij}, \bar{x}_i) = V(\bar{x}_i)$.

6. $Cov_{i \neq j}(z_i,z_j) = (1+c)\left\{a^2 + [2a(1-a) + (1-a)^2]/n\right\} \times Cov(x_i,x_j)$ (10)

This follows from $Cov(z_i,z_j) = Cov(ay_i,ay_j) + Cov[(1-a)\bar{y}_i, (1-a)\bar{y}_j] + 2Cov[ay_i, (1-a)\bar{y}_j], Cov(y_i, \bar{y}_j) = (1+c)\,Cov(x_i,x_j)/n$ and $Cov(\bar{y}_i, \bar{y}_j) = Cov(y_i,y_j)/n$.

7. $Corr(z_i,z_j) = Corr(x_i,x_j)$ (11)

This follows since the coefficients of $V(z_k)$, $k = i,j$ and $Cov(z_i,z_j)$ are identical.

8. Let t be an unmasked variable, then

$$Cov(z_i,t) = [a + (1-a)/n]\,Cov(x_i,t) \qquad (12)$$

This follows since $Cov(y_i,t) = Cov(x_i,t)$ and $Cov(\bar{y}_i, t) = Cov(x_i,t)/n$.

9. $Corr(z_i,t) = [a + (1-a)/n]\,Corr(x_i,t)$ (13)

This follows since $V(z_i) = V(x_i)$, but $Cov(z_i,t) = [a+(1-a)/n]\,Cov(x_i,t)$.

10. $Corr(z_i,t) < Corr(x_i,t)$

This follows from $a < 1$ and hence $[a + (1-a)/n] < 1$.

11. $Corr(y_i,t) = [1/\sqrt{(1+c)}]\,Corr(x_i,t)$ (14)

This follows since $Cov(y_i,t) = Cov(x_i,t)$ and $V(y_i) = (1+c)\,V(x_i)$.

The correlation between y_i and t is always less than the correlation between x_i and t.

12. Correlation between z_i and t is asymptotically the same as the correlation between y_i and t.

This follows since when n is large, $Corr(z_i,t) \tilde{=} a\,Corr(x_i,t)$ and $a \tilde{=} 1/\sqrt{1+c}$. Note that when n is large, $(n-1-c)/(n-1) \tilde{=} 1$ and hence from equation (6), $a \tilde{=} 1/\sqrt{1+c}$.

II.3 Properties of Transformed Variable, z, When σ^2 is Unknown

In practice, σ^2 is not known, hence σ^2 is estimated from the sample (estimate denoted by s^2 as usual), and then random noise is generated using this s^2. If s^2 is calculated from the noise e, it will not be exactly the same as s^2 due to the sampling variability of e. Hence, we denote the estimated variance of noise by s^2.

Taking a respondent sample and adding noise can be interpreted as a two stage sampling.

In Stage 1, n respondents are selected from a population of size N and observations are made. From the observations, sample statistics are calculated (This is the usual situation, but if a whole population is observed, population statistics are calculated).

In Stage 2, noise sample is generated using the statistics and noise is added to the observed values. From this perspective, the mean and variance can be interpreted as follows. Denoting the respondent sample by $\mathcal{S}$.

$E(y_i) = E[E(x_i + e_i)|\mathcal{S})] = E(x_i) = \mu_i$.
The above follows since $E(e_i|\mathcal{S}) = 0$. Hence,

$$E(z_i) = E[E(z_i|\mathcal{S})] = E[ax_i + (1-a)\bar{x}_i] = \mu_i$$

$$V(y_i) = V[E(x_i + e_i|\mathcal{S})] + E[V(x_i + e_i|\mathcal{S})]$$

$$= V(x_i) + E(cs_i^2) = (1+c)\sigma_i^2,$$

which follows since $V(x_i + e_i|\mathcal{S}) = V(e_i|\mathcal{S}) = cs_i^2$. Therefore,

$$V(z_i) = V\left\{E[ay_i + (1-a)\bar{y}_i\ |\mathcal{S}]\right\} + E\left\{V[ay_i + (1-a)\bar{y}_i\ |\mathcal{S}]\right\}$$

$$= V[ax_i + (1-a)\bar{x}_i] + E\left\{V[ae_i + (1-a)\bar{e}_i]\right\}.$$

But the second term in the above expression reduces to

$$E\left\{cs_i^2a^2 + cs_i^2[2a(1-a) + (1-a)^2]/n\right\}$$

$$= \left\{a^2 + [2a(1-a) + (1-a)^2]/n\right\}c\sigma_i^2.$$

Hence the above variance of z reduces to the variance

in (6).

Similarly, the covariance between z_i and z_j, and the covariance between z_i and t in this case reduce to the covariances in the previous case.

II.4 Impacts of Masking on Regression

Without loss of generality, the variances can be assumed to be homogeneous since heterogeneous variances can be changed to the homogeneous by proper transfomration of variables (see p. 221 of reference 4).

Case 1. When σ^2 is known

Smith considered the effects of masking by random noise on the regression when all the variables are masked and when σ^2 is known for generating random noise, but the variance estimate s^2 is used for the regression. Here we deal with the problem from a broader perspective under the same set of conditions.

Define $X' = (x_1, x_2, \ldots, x_p)$, a vector of p variables, X^* is a realization of X. Define Y, Y^* and Z, similarly. Note that Y is a vector of variables, at least some of which are masked and Z is a vector of transformed variables. Define $E(X) = \mu$ and $V(X) = V$. Then $E(Y) = \mu$ and $V(Y) = (1+c)\,V$.

To build a regression of x_1 on $X_2' = (x_2, x_3, \ldots, x_p)$, X, X^*, and V are partitioned as follows.

$$X = \begin{pmatrix} x_1 \\ X_2 \end{pmatrix}, \qquad X^* = \begin{pmatrix} x_1^* \\ X_2^* \end{pmatrix},$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

Then

$$E(x_1 | X_2 = X_2^*) = \mu_1 + V_{12} V_{22}^{-1} (X_2^* - \mu_2)$$

In the above $\mu_1 - V_{12} V_{22}^{-1} \mu_2$ is the intercept and $V_{12} V_{22}^{-1}$ is the vector of coefficients.

Theorem 1. If all the masked variables have the same first two moments as the unmasked variables then the regression coefficients and intercept based on the masked data are on the average identical to those of the unmasked data.

Proof. Proof follows from the fact that the intercept $\mu_1 - V_{12} V_{22}^{-1} \mu_2$ and the coefficients $V_{12} V_{22}^{-1}$ remain the same throughout masking.

The above theorem applies to the data masked by our scheme.

Theorem 2. If all the masked variables have the same first moments but their second moments are proportional (at the same rate) to those of the unmasked, the regression coefficients and intercept based on the masked variables only are on the average identical to those of the unmasked data.

Proof. $E(Y) = E(X) = \mu$ and the new variance and covariance matrix is k V, where k is a constant. The new coefficients are $kV_{12}^* (kV_{22})^{-1} = V_{12} V_{22}^{-1}$ which is the same as that for the unmasked.

This theorem applies to the data masked by the random noise approach.

If the covariance between two unmasked variables is maintained after masking, but the covariance between the masked and unmasked is not, then an adjustment of the covariance to make it unbiased is required to preserve the same correlation coefficients, on the average. The covariance between the masked and unmasked variables in our scheme is [a + (1-a)/n] times the covariance between unmasked variables. Thus the covariance must be adjusted to insure unbiasedness of the coefficients and intercept. On the other hand, the corresponding covariance in random noise approach is unbiased [2], but the variance of the masked variable is (1 + c) times that of the unmasked variable. Hence, this variance needs to be adjusted.

Lemma 1. If the unbiased variance-covariance structure is maintained after masking, but the means lose unbiasedness, then the regression coefficients based on the data including masked variables, on the average, would remain identical to those based on unmasked data, but the intercept would not.

Proof. By inspection of the formula of intercept.

This lemma can be applied to the data generated by the random noise approach. If the masked data is adjusted before inputting in the computer to make the sample variance and covariance unbiased, the resulting sample means will become biased. Thus, the intercept of the regression based on this data will be biased.

The above two lemmas can be combined and rephrased in terms of correlation coefficients.

Lemma 2. If the means and correlations of the masked variables are unbiased, (naturally or by adjustment), then the regression coefficients and the intercept of the model fitted on the wholly or partly masked data will be, on the average, the same as those obtained from the unmasked data.

Theorem 3. If all variables in the regression model are masked and the second moments of the masked variables are the same as those of the unmasked, then the residual error variance of the regression will be the same as that of the regression based on the unmasked.

Proof. $V(x_1 | X_2) = V_{11} - V_{12} V_{22}^{-1} V_{21}$
Since all the variances of the masked variables are identical to those of the unmasked, $V(z_1 | Z_2)$ will be the same as $V(x_1 | X_2)$.

This theorem applies to the data masked by our scheme, but it does not apply to the data masked by the random noise approach. The residual error variance based on the latter is (1+c) times that based on the unmasked data.

Lemma 3. If the variables in the regression, all or part of which are masked, have the same second moments as the unmasked, then the residual error variance based on the data will be the same as that based on the unmasked data.

Proof. Omitted.

This applies to the data, on the average, part of which is masked by our scheme and the covariance between masked and unmasked variables is adjusted to be unbiased.

Define

$$X = \begin{pmatrix} x_{21} - \bar{x}_2 & \cdots & x_{p1} - \bar{x}_p \\ x_{22} - \bar{x}_2 & \cdots & x_{p2} - \bar{x}_p \\ \vdots & \vdots & \vdots \\ x_{2n} - \bar{x}_2 & \cdots & x_{pn} - \bar{x}_p \end{pmatrix}$$

Theorem 4. Under the same condition as in Lemma 3, the standard errors of regression coefficients based on the data will be the same as those based on the unmasked data.

Proof. In general, the variance of the coefficient can be expressed as

$$\sigma^2 (X'X)^{-1}$$

where X is defined above. The proof follows from inspection of the above variance.

Theorem 5. The variance of the intercept based on the data having the same variance-covariance and means as the unmasked data is identical to the variance of intercept based on the unmasked data.

Proof. The variance of the intercept is

$$\sigma^2 \left[1/n + \bar{x}'(X'X)^{-1}\bar{x}\right]$$

The proof follows from inspection of the above variance formula.

The data masked by our scheme satisfies the above theorem but the data generated by the random noise approach never, even with adjustment, does.

Case 2. When σ^2 is not known

When σ^2 is not known, s^2 is used to generate noise, hence due to the sampling error of noise, the actual variance will be different from s^2. Only repeated generation of noise and calculation of s^2 (more precisely cs^2) infinitely many times will result in their average equal to s^2. This has significant implications on the regression coefficients.

Denote the regression coefficients based on σ^2 by β, s^2 by b and $\hat{s}^2$ by $\hat{b}$. Note that $\hat{b}$ will be calcultated using sample variance which is a mixture of s^2 and $\hat{s}^2$. For example in the random noise case, if noise whose variance is 1/2 of s^2 is used then the resulting sample variance will be

$$s^2 + .5\,\hat{s}^2 .$$

Hence the regression coefficients will be estimated based on this type of variance. If the conditions for unbiasedness of b seen in the previous theorems and lemmas are met, then

$$E(\hat{b}) = E[E(\hat{b} \mid \mathcal{S})] = E(b) = \beta .$$

This means under the suitable conditions $\hat{b}$ can be unbiased.

III. Examples of Application of the Scheme

The scheme proposed here was tried for masking earnings data. A separate scheme was also investigated, namely, the addition of random normal noise with zero mean and standard deviation equal to 1/2 the observation. This scheme is also included in the comparison.

Random numbers were generated using a subroutine in IMSL called GGNSM. This routine generates standard normal multi-variates which follow a specified correlation structure among the variables. Using these variates random noise was generated. Also RLMUL in IMSL was used to run regression. Box-and-Whisker plot was obtained by using IMSL, too.

Table 1 shows means of 3 unmasked as well as masked variables. In our scheme, three versions were tried by varying the amount of variance of noise, i.e., 25%, 50% and 100% of the variance of the unmasked variable were tried.

Due to the sampling variability of the mean of noise the sample means in the table are all different from the sample mean of the unmasked.

Correlation coefficients between variables were calculated (see Table 2). None of the correlations obtained from our data is significantly different from the original ones. However, both coefficients obtained from the other scheme are significantly different.

Table 3 has the results of the multiple regression in which all the variables were masked. Our data (with V(e) = .25 V(x)) provides more reliable results.

Tables 4 provides the MSE, F values and the variance of the dependent variable explained by the regression (R^2). Our data gives MSE values close to those of the unmasked, but the other data does not. The percentage of variance is higher in our results than under the other scheme.

IV. Concluding Remarks

So far properties of the new masking scheme have been considered and some examples of application have been shown. However, the power of limiting the disclosure by this scheme has not been tested. We are planning on embarking on the experiment using Paass' software which was developed for his reidentification study. It should be noted that as far as our scheme is concerned, the probability of reidentification can be manipulated by using the "a" value. It is possible since by lowering the "a" value, we can shift the weight toward the mean and thus reduce the reidentifiability of the respondents. However, since the correlation structure can be maintained, if necessary by adjustment, the regression can be run on the data without adverse effect.

A lot more questions remain to be answered concerning this scheme. These will be investigated as soon as time permits.

1/ If $n_i \neq n$, then the a values in (6) would be different for each i.

2/ This is the model Spruill used for her experiment. However, this can be changed by using correlated noise, which ensures unbiasedness of the correlation. This latter approach was used in my experiment.

References

Draper, N.R. and Smith, H., Applied Regression Analysis, John Wiley & Sons, Inc., 1966

Graybill, F.A., An Introduction to Linear Statistical Model, Vol. I, McGraw-Hill Book Company Inc., 1961

Paass, G., Disclosure Risk and Disclosure Avoidance for Microdata, presented at the International Association of Social Science for Information Service and Technology, 1985

Rao, C.R., Linear Statistical Inference and Its Application, John Wiley and Sons, Inc., 1973

Searle, S.R., Linear Models, John Wiley & Sons, Inc., 1971

Smith, P.J., Effect of Addition of Random Error to Perturb Sensitive Data and to Preserve Confidentiality of Response, internal memorandum, Census Bureau

Spruill, N.L., Confidentiality and Analytic Usefulness of Masked Business Microdata, the Public Research Institute, Alexandria, Va., 1983

APPENDIX

Table 1—Comparison of Means, n=2000

Item	Var 1	Var 2	Var 3
Unmasked	852.44	903.90	1099.68
Our Scheme 1*	876.77	939.48	1134.33
Our Scheme 2*	886.85	954.22	1148.71
Our Scheme 3*	901.12	975.03	1169.0
Other Scheme	838.06	962.47	1091.67

Our scheme 1,2,3 corresponds to the scheme with V(e) is 25%, and 100% of V(x) in that order.

Table 2—Correlation Coefficients

Item	Var 1 vs Var 2	Var 3 vs Var 4
Unmasked	.74	.76
Our Scheme 1	.74	.76
Our Scheme 2	.74	.76
Our Scheme 3	.74	.76
Other Scheme	.68	.82

Table 3—Regression Coefficients - All Vars Masked

Item	Unmasked	Ours*	Other
x_1	-.01	-.02	.01
x_2	.18	.18	.23
x_3	.68	.69	.65
Slope	972	937	951

* Our scheme 1 is used

Table 4—ANOVA Based on Regression - All Variables Masked

Item	Unmasked	Ours	Other
MSE	6796.5	6781.5	7287.1
F	508.16	522.42	411.24
R^2	43.30	43.98	38.20