Simulation Study of the Effectiveness of Masking Microdata with Mixtures of Multivariate Normal Distributions

Sam Hawala
US Census Bureau, Washington DC, SRD/3209/4

Key words: Disclosure Limitation, Masked Microdata, Re-identification

Abstract

Continuous variables in microdata can be masked for protection from disclosure through the use of an additive noise. I consider adding noise that is distributed according to a mixture of normal distributions. There are several parameters involved in constructing the additive noise. The study's purpose is to lay down as a guide a recipe for the choices of these parameters. The proportion of re-identifiable records through the use of Winkler's matching software measures effectiveness of the masking method. Results depend heavily on the matching software used.

Summary

Adding a randomly generated noise to a continuous variable is one possible method for masking a data file. The effectiveness of this method depends on the values assigned to some specific parameters. The reader will get a sense of what these parameters are and how they relate to each other from the details in the paper about a simple case of the method. I simulate data, then apply Winkler's matching software and get the proportion of records that can be re-identified. My goal is to get a direct sense experience of how the percentage of records, that are re-identifiable by Winkler's matching software, varies when I assign different values to the parameters. The data masking procedure is due to Roque (2000). Her paper did not provide full insight on how to choose values for all the parameters involved.

The procedure of adding noise to a continuous variable certainly distorts the values of the variable, but the noise addition is done in such a way as to preserve the means of the universe and of its sub-domains. The correlations between the variables are also preserved. This is consistent with the mission of a statistical data provider, which is to provide data only as statistical totals and not as individual values. The variance of the variable to which noise is added increases by a multiplicative factor *d* of the original variance. This multiplicative factor is one of the parameters of interest. The data provider chooses the value of *d* in such a way that there is maximum protection from disclosure and minimum distortion applied to the data. For relatively low values of *d* the additive-noise method masks the data well without significantly diminishing its analytic utility. A masked data set is analytically valid if results of statistical analyses performed on the original data set can be reproduced using the masked data set.

This research is part of an effort by the US Census Bureau to remain vigilant in protecting respondents' identities while it continues to release microdata to the general public. The public has recently seen a noticeable increase in data availability, in computing power and storage capacity, and an interest and advances in data integration technologies. In light of these changes in the data environment, more powerful methods for disclosure limitation have to be explored and tested. This study is an effort towards that goal.

1. Introduction

The US Census Bureau has a mandate to disseminate the maximum amount of statistical data and must do so in a manner that protects the identities of the respondents. For more than four decades the Bureau has released microdata to the public. Before 1960 the Bureau mainly organized the data in tabular form and released the prepared tables. Microdata consist of data records for individual respondents. Prior to its release, a microdata set is first stripped of direct identifiers such as names, addresses, social security numbers. Then the microdata is edited for confidentiality by subject matter experts, familiar with the multiple uses of the data. The experts work diligently at preserving its analytic validity.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

In effect, there are two conflicting needs: the needs of the Bureau to maintain the confidence of respondents by keeping their data confidential and therefore must mask the data before making the data available to the public, and the needs of various data users for more valid and detailed data to perform complex analyses. For a list of masking methods that data providers use the reader can consult the *Checklist On Disclosure Potential Of Proposed Data Releases*. The Confidentiality and Data Access Committee (CDAC), an interest group of OMB's Federal Committee on Statistical Methodology, publishes this list. Masking methods that have not yet been widely implemented, must go through extensive testing to satisfy the conflicting needs of the data providers and those of the data users.

The method under consideration in this paper uses additive noise to mask data. The main methodology of the additive-noise approach is described in Kim (1986). Some other related papers are Sullivan and Fuller (1989), Fuller (1993), Kim (1990), Kim and Winkler (1995), Roque (2000), and Yancey et al. (2002). The application considered in this paper is a simplified version of Roque's idea of adding noise generated according to a mixture of normal distributions. Winkler's matching software is used for the re-identification. Section 2 describes the masking technique. The results obtained are very much specific to the re-identification tool, which is discussed briefly in section 3. Section 4 describes the simulations and section 5 presents the results.

2. Masking

I discuss the mathematical aspects of the method by assuming that there is one variable X to be masked, and that X has a normal distribution with mean μ and variance σ^2. To mask the data collected on the variable X, generate a random quantity Y according to a probability distribution f_Y, with mean 0 and variance $d\sigma^2$. f_Y is a mixture of normal distributions:

$$f_Y(y) = \sum_{j=1}^{k} w_j f_N(y, \theta_j, \sigma_j) \qquad (1)$$

For each j, f_N is a normal density with mean and variance $\theta_{j,}$ and σ_j^2 respectively. In order for f_Y to be a density we must have:

$$\sum_{j=1}^{k} w_j = 1 \qquad (2)$$

The condition that Y has mean zero is given by

$$\sum_{j=1}^{k} w_j \theta_j = 0 \qquad (3)$$

Since the variance of Y is $d\sigma^2$ we have

$$d\sigma^2 = \sum_{j=1}^{k} w_j (\theta_j^2 + \sigma_j^2)$$

Now define the masked variable Z as

$$Z = X + Y$$

Since the mean of Y, $E(Y)$ is zero, Z has the same mean as X, namely μ, and since X and Y are independent, the variance of Z is $(1+d)\sigma^2$. Moreover, if we let $\sigma_j^2 = d_j \sigma^2$ for $j=1\ldots k$, we now have

$$(d - \sum_{j=1}^{k} w_j d_j)\sigma^2 = \sum_{j=1}^{k} w_j \theta_j^2 \qquad (4)$$

To generate the noise Y, the data provider interested in masking the data, can freely choose the parameters d, the w_j's, θ_j's, and the d_j's, for $j=1\ldots k$, as long as they satisfy equations (2), (3) and (4).

The w_j's play the role of weights in the distribution of Y given in equation (1). Without any knowledge of how big a role each component in the mixture should play, it is reasonable to begin with equal weights i.e., $w_j = 1/k$, for $j=1\ldots k$. The same can be said about the σ_j's, i.e. I set them all equal, or equivalently let $d_j=c$ for $j=1\ldots k$, for some constant c. In this case (2) is automatically satisfied, (3) and (4) become (5) and (6) respectively:

$$\sum_{j=1}^{k} \theta_j = 0 \qquad (5)$$

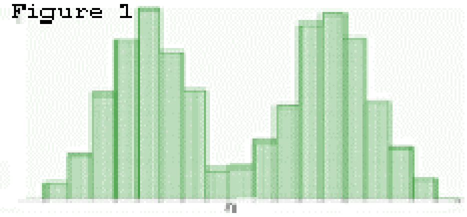
$$k(d-c)\sigma^2 = \sum_{j=1}^{k} \theta_j^2 \qquad (6)$$

The values of the parameters k, d, c, and the θ_j's, $j=1\ldots k$, satisfying (5) and (6) complete the specification of the masking procedure. Equation (5) is a constraint on the θ_j's. This constraint divides the set of possible densities f_Y of the noise variable *Y* into two subsets. One subset has densities that are symmetric about zero. The other subset has densities that are asymmetric about zero. For *k=2* the noise variable *Y* has the following density:

$$f_Y(y) = \frac{1}{2}\left(\varphi_1(y,\theta_1,\sqrt{c}\sigma) + \varphi_2(y,\theta_2,\sqrt{c}\sigma)\right),$$

where $\theta_1 = -\theta_2 = \sqrt{(d-c)}\sigma$, and where $\varphi_i(y,\theta_i,\upsilon)$ is the normal density with mean θ_i and variance υ^2. The density f_Y has two modes at θ_1 and at θ_2 . It is symmetric about zero, as depicted in figure 1.

Figure 1



If this symmetric aspect of the distribution of *Y* is maintained then for other values of *k* we have: when *k* is even, i.e., *k=2h*

$$f_Y(y) = \frac{1}{k}\sum_{j=1}^{k}\varphi_j(y,\theta_j,\sqrt{c}\sigma), \text{ with}$$

$$\theta_j = -\theta_{h+j} = j\sqrt{\frac{12(d-c)}{(k+2)(k+1)}}\sigma \quad \text{for } j=1,\ldots h.$$

When *k* is odd, i.e., *k=2h+1*

$$f_Y(y) = \frac{1}{k}\sum_{j=1}^{k}\varphi_j(y,\theta_j,\sqrt{c}\sigma), \quad \text{with}$$

$$\theta_1 = 0, \text{ and } \theta_j = -\theta_{h+j} = j\sqrt{\frac{12(d-c)}{(k-1)(k+1)}}\sigma$$

for *j=2,…h*.

One particular approach to study the case where the density f_Y is not symmetric about zero is to set

$$\theta_1 = \theta_2 = \ldots = \theta_{k-1} = -\frac{\theta_k}{k-1} \qquad ,$$

which leads to

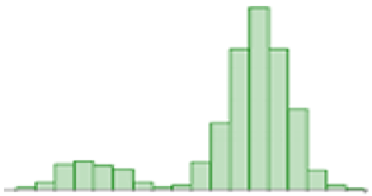
$$f_Y(y) = \frac{k-1}{k}\varphi_1(y,\theta_1,\sqrt{c}\sigma) + \frac{1}{k}\varphi_2(y,\theta_2,\sqrt{c}\sigma)$$

, with $\theta_1 = \sqrt{\frac{d-c}{k-1}}\sigma$ and

$$\theta_2 = \sqrt{(k-1)(d-c)}\sigma$$

This density will a shape depicted in figure 2.

Figure 2



A data provider interested in masking continuous variables by adding noise would calculate the mean *μ*, and variance σ^2 in the original data, then would have to construct the densities of the noise variables by choosing the parameters *k, d,* and *c*. Through data simulation we can generate different data with different values of *μ*, and σ^2, then for each of the values chosen explore several choices of the parameters *k, d,* and *c* and determine the behavior of the re-identification rate.

3. Re-identification

Notwithstanding the masking procedure, the masked data file probably contains records that sophisticated intruders may still be able to re-identify. Winkler's software, which is based on an optimal decision procedure for record linkage introduced by Fellegi and Sunter (1969), provides a very powerful way to decide whether a masked record and a record in the original file are likely matches. The records contain no unique identifiers, so there is no obvious way to decide whether two records correspond to the same individual.

Public use microdata usually contain a number of fields that are common attributes, such as age, sex, and race. These common fields are useful for matching. The data is blocked according to the values of these fields. Not all fields, however, contain an equal amount of information, and error rates vary. For example a field such as sex, which has only two (value) states, neither of which has a low frequency, could not impart enough information to identify a match uniquely. The field race imparts a little more information. It has a few more value states but may be more incorrectly recorded.

Weights are used to measure the contribution of each field to the probability of making an accurate classification. For any record pair, a composite weight can be computed by summing the individual field weights. A record pair is classified as a match if the composite weight is above a threshold value, a mismatch if the composite weight is below another threshold value, and an undecided situation if the composite weight is between these two thresholds. The threshold values can be calculated given the acceptable probability of false matches and the probability of false mismatches.

4. Simulations

The simulated data contain eight independent normally distributed variables with coefficient of variation (*CV*) ranging from 0.01 to 100. There are no variables to use for blocking, so there is one record per block. The parameter d is most important when we consider the quality of the masked data. With higher values of d the data gets highly distorted and we expect the re-identification rate to be small. I used the value $d=.2445$ which was used by Roque (2000) and Yancey (2002). I also used other values for d ranging from *.1* to *.4.* for comparison. The parameter c must be such that $c < d$.

The simulation procedure generates different data every time it is run since the random seeds are in turn generated randomly. Running Winkler's software and determining the number of matches n, among the 1500 pairs of simulated and masked records take about 50 seconds each time. The re-identification rate is $r = n/1500$.

5. Results

The reader is reminded again here that the results of this kind of study depend on the re-identification or matching tool used.

- For fixed d and c, when the variance of the original data is large compared to the mean, less data distortion is needed to mask the data to avoid re-identification. Thus, in some ways the relative dispersion in the original data serves as a deterrent against re-identification. For example when d is fixed at 0.2445 and c at .025, and the magnitude of the standard deviation (σ) is tenfold that of the mean, or $CV=10$, I obtained re-identification rates of the order of 2%. A rate of 2% is an acceptable rate according to experts in the field of matching and re-identification. Also important is the fact that for higher values of CV, the data provider may choose smaller values for d, leading to less distortion of the data. The variability of the data serves as a deterrent to re-identification, as long as the mean is not too large, and smaller amounts of noise have to be added to protect against disclosure when the coefficient of variation is 10 or more.

- The results also show that the symmetry of the noise density f_Y leads to lower rates of re-identification. When all the other parameters are held fixed, the difference in re-identification rates between employing a density that is symmetric and a density that is asymmetric about zero is significant. This result was obtained at all values of μ, and σ^2 and thus all values of *CV* that I tried. See table 1, where $k=3$, $d=.2445$, and $c=.025$. The re-identification rates are consistently higher when the density f_Y is asymmetric.

- Among the symmetric distributions there were no significant changes in the re-identification rate when I changed the number k of mixture components and held d, c, μ ,and σ^2 constant. The results for $k=2$, and $k=7$, when $d=.2445$ and $c=.025$ are given in table 2. The results with $k=4$, are not shown but they are very similar.

- Among the parameters k, d, and c, d has the greatest effect on the rate r. There is a negative association between the parameter d and the re-identification rate r. For a fixed coefficient of variation $CV = 1$ and fixed $c=.025$ the re-identification rate r decreases from around 16% to 2% as d increases from *0.1* to *0.4,* see Table 3.

- There were no noticeable changes in the re-identification rate r when the parameter c changes.

<u>Table 1.</u> Comparing results for symmetric and asymmetric distributions with
$k=3$, $d=.2445$, and $c=.025$

CV	*Symm.*	*Asym.*	Dif %
0.01	11%	15%	4
0.1	18%	24%	6
1	6%	7%	1
10	1%	2%	1
100	1%	1%	0

Table 2. Comparing results for *k* = 2 and *k* = 7 with *d=.2445,* and *c=.025*

CV	***k* = 2**	***k* = 7**	Diff. (%)
0.01	11%	12%	1
0.1	18%	17%	1
1	6%	7%	1
10	1%	2%	1
100	1%	1%	0

Table 3. Comparing results for d, with *CV=1, k=2, c=.025,* f_Y symmetric.

d	***r***
0.1	16%
0.2	8%
0.2445	5%
0.3	3%
0.4	2%

6. References

Fellegi, I.P., and Sunter, A.B. (1969) "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210.

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

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

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. J., Winkler W. E. (1995) "Masking Microdata Files." *American Statistical Association, Proceedings of the Section on Survey Research Methods,* 114-119.

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

Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 802-807.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 435-439.

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox (ed.) *Business Survey Methods,* New York: J. Wiley, 355-384.

Yancey, W. E., Winkler, W. E., Creecy, R. H. (2002), "Disclosure risk assessment in perturbative microdata protection." *Inference Control in Statistical Databases,* Springer, 135-152