UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M10/05
Methodology

Measurement Error And Statistical Disclosure Control

Natalie Shlomo

Abstract

Statistical agencies release microdata to researchers after applying statistical disclosure control (SDC) methods. Noise addition is a perturbative SDC method which is carried out by adding independent random noise to a continuous variable or by misclassifying values of a categorical variable according to a probability mechanism. Because these errors are purposely introduced into the data by the statistical agency, the perturbation parameters are known and can be used by researchers to adjust statistical inference through measurement error models. However, statistical agencies rarely release perturbation parameters and therefore modifications to SDC methods are proposed that a priori ensure valid inferences on perturbed datasets.

Measurement Error and Statistical Disclosure Control

Natalie Shlomo[1]
[1] Southampton Statistical Sciences Research Institute, University of Southampton,
Highfield, Southampton, SO17 1BJ, United Kingdom
N.Shlomo@Soton.ac.uk

Abstract. Statistical agencies release microdata to researchers after applying statistical disclosure control (SDC) methods. Noise addition is a perturbative SDC method which is carried out by adding independent random noise to a continuous variable or by misclassifying values of a categorical variable according to a probability mechanism. Because these errors are purposely introduced into the data by the statistical agency, the perturbation parameters are known and can be used by researchers to adjust statistical inference through measurement error models. However, statistical agencies rarely release perturbation parameters and therefore modifications to SDC methods are proposed that a priori ensure valid inferences on perturbed datasets.

Keywords: Additive noise, Post-randomisation method, Reliability ratio

1 Introduction

Statistical disclosure control (SDC) methods are becoming increasingly important due to the growing demand for information provided by statistical agencies. More statistical agencies are releasing microdata from social surveys typically under licensing agreements or through data archives. SDC methods aim to prevent sensitive information about individual respondents from being disclosed.

In any released microdata, directly identifying key variables, such as name, address or id numbers, are removed. Disclosure risk arises from attribute disclosure where small counts on cross-classified indirect identifying key variables (such as: age, gender, place of residence, occupation, etc.) can be used to identify an individual and confidential information may be learnt. Identifying key variables are typically categorical since statistical agencies will often coarsen the data before its release. Sensitive variables are continuous (e.g., income) or categorical (e.g., health status). SDC methods can be non-perturbative by limiting the amount of information released or perturbative by altering the data in the microdata. Examples of non-perturbative SDC methods are global recoding, suppression and sub-sampling (see Willenborg and De Waal, 2001). Perturbative methods for continuous variables include adding random noise (Kim, 1986, Fuller, 1993, Brand, 2002), micro-aggregation (replacing values with their average within groups of records) (Anwar 1993, Domingo-Ferrer and Mateo-Sanz, 2002), rounding to a pre-selected rounding base, and rank swapping (swapping values between pairs of records within small groups) (Dalenius and Reiss, 1982, Fienberg and McIntyre, 2005). Perturbative methods for categorical variables include record swapping (typically swapping geography variables) and a more general post-randomization method (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (Gouweleeuw, et al., 1998).

Perturbative methods can be applied to either the identifying key variables or the sensitive variables or both. In the first case identification of a unit is rendered more difficult, and the probability that a unit can be identified is reduced. In the second case, even if an intruder succeeds in identifying a unit by using the values of the indirect identifying key variables, the sensitive variable would hardly disclose any useful information on the particular unit as they have been perturbed.

In this paper, we focus on perturbative SDC methods which purposely introduce measurement errors into the microdata: additive random noise for a continuous variable and misclassification for a categorical variable. Assuming that the SDC parameters are released by the statistical agency, researchers can use these parameters to correct statistical inferences through measurement error models (Fuller, 1987). Following Fuller, 1993, we demonstrate a measurement error model for a simple linear regression on a perturbed dataset. Statistical agencies, however, rarely release SDC parameters due to confidentiality constraints. In this case, statistical agencies need to modify SDC methods so that researchers can make valid inferences on perturbed datasets.

Section 2 focuses on additive random noise to continuous variables and the impact on a simple regression analysis. An SDC method of correlated noise is proposed that preserves the sufficient statistics and allows valid inference from the regression model on the perturbed data. Section 3 focuses on misclassification of a categorical variable through PRAM and the impact on a simple regression model and a chi-square test for independence for a two dimensional table. By placing the property of invariance on the probability mechanism used in PRAM, some statistical inferences can be preserved exactly on the perturbed datasets. We conclude in Section 4 with a discussion on how these SDC methods can be implemented ‘on the fly’ so that they can be tailored specifically to the analysis.

2 Adding Noise to Continuous Variables

Additive random noise is an SDC method that is carried out on continuous variables. In its basic form, random noise is generated independently and identically distributed with a mean of zero and a positive variance which is determined by the statistical agency. A zero mean ensures that no bias is introduced into the original variable. The random noise is then added to the original variable. There are also more complex mixture models that can be used for adding noise which achieve higher protection levels since it has been found that additive random noise can yield high re-identification risk (Kargupta, et al., 2005).

Adding random noise to a continuous variable will not alter the mean value of the variable for large datasets but will introduce more variance depending on the variance parameter used to generate the noise. This will impact on the ability to make statistical inference, particularly for estimating parameters in a regression analysis.

The ease of analysis in a regression model for a variable subject to additive noise depends on whether the variable is used as the dependent variable or as the independent variable (or both). Standard regression model theory accounts for errors in the dependent variable and therefore adding more noise to the dependent variable should not affect the estimation of the slope parameters.

As an example, assume a simple regression model with a dependent variable y_i that has been subjected to independently generated Gaussian additive noise η_i with a mean of 0 and a positive variance σ_η^2. Assume also an independent variable x_i that is error free. The model is:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1,..., n \\ y_i = y_i^* + \eta_i \end{cases}$$

where y_i^* denotes the true but unobserved value of the dependent variable. If we regress y_i on x_i, then the least squares slope coefficient is:

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y^* + \eta, x)}{Var(x)} = \frac{Cov(y^*, x) + Cov(\eta, x)}{Var(x)} = \frac{Cov(y^*, x)}{Var(x)} \qquad (1)$$

since $Cov(\eta, x) = 0$. The additive noise on the dependent variable does not bias the slope coefficient, however it will increase its standard error due to the increase in the variance: $Var\ (y) = Var\ (y^*) + Var\ (\eta)$.

Complications arise when the random noise η_i is added to the independent variable in the regression model. The model is now:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1,..., n \\ x_i = x_i^* + \eta_i \end{cases}$$

where x_i^* denotes the true but unobserved value of the independent variable. Now regressing y_i on x_i, we obtain for the least squares slope coefficient:

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y, x^* + \eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y, x^*) + Cov(y, \eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y, x*)}{Var(x^*) + Var(\eta)} \qquad (2)$$

since $Cov(y, \eta) = 0$. The additive noise on the independent variable biases the slope coefficient downwards. This is referred to as attenuation. In this case, the researcher needs suitable methodology to deal with the measurement error in the independent variable.

Noting that the estimate for the least squares slope coefficient follows:

$$\hat{\beta} \xrightarrow{p} \frac{Cov\ (y, x^*)}{Var\ (x^*) + Var\ (\eta)} = \frac{\beta\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_\eta^2} = \beta(1 + \sigma_\eta^2 / \sigma_{x^*}^2)^{-1} \qquad (3)$$

Fuller, 1987 defines the term $(1 + \sigma_\eta^2 / \sigma_{x^*}^2)^{-1}$ as the reliability ratio denoted by λ. In a very simple measurement error model, a consistent estimate of the slope coefficient can be obtained by dividing the least-squares estimate from the perturbed dataset by λ.

To calculate the reliability ratio and allow valid inferences, it is assumed that the variance parameter σ_η^2 used to generate the random noise is released by the statistical agency to researchers. This, however, is rarely the case since statistical agencies generally do not reveal parameters of SDC methods. In order to compensate for the measurement error, statistical agencies should employ a different method for adding random noise based on generating noise that is correlated with the original continuous variable. Kargupta, et al., 2005 noted that re-identification is more difficult when adding correlated noise. Correlated noise addition ensures that sufficient statistics (means, variances and correlations) of the original continuous variables are preserved (see also: Kim, 1986 and Tendick and Matloff, 1994). One algorithm for generating correlated random noise for a continuous variable x that is easy to implement is as follows:

Procedure for a univariate case: Define a parameter δ which takes a value greater than 0 and less than equal to 1. When $\delta = 1$ we obtain the case of fully modeled synthetic data. The parameter δ controls the amount of random noise added to the variable *x*. After selecting a δ, calculate: $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$. Now, generate random noise ε independently for each record with a mean of $\mu' = \mu\{(1-d_1)/d_2\}$ and the original variance of the variable σ^2. Typically, a Normal Distribution is used to generate the random noise. Calculate the perturbed variable x_i' for each record *i* (*i=1,..,n*) as a linear combination: $x_i' = d_1 \times x_i + d_2 \times \varepsilon_i$. Note that

$E(x') = d_1 E(x) + d_2[(1-d_1)/d_2]E(x) = E(x)$ and

$Var\ (x') = (1-\delta^2)Var\ (x) + \delta^2 Var\ (x) = Var\ (x)$ since the random noise is generated independently to the original variable *x*. This algorithm can be extended to the multivariate case for simultaneously adding correlated random noise to several variables which preserves the sufficient statistics of each variable as well as the covariance matrix. (see Shlomo and De Waal, 2008).

Table 1 presents a simulation study which demonstrates the effects of adding random noise and correlated random noise to variables in a simple regression model. Each row in the table represents a different scenario consisting of the type of noise added (random or correlated) and whether the noise was added to the dependent variable, independent variable or both. We generate 1000 records where $x_i \sim N(20,9)$, $\varepsilon_i \sim N(0,3)$ and the model is: $y_i = 3 + 3x_i + \varepsilon_i$ (the true intercept is 3 and the true slope coefficient is 3). We generate Gaussian random noise: $u_i \sim N(0,1)$ as well as correlated noise according to the procedure described above with $\delta = 0.1$. Note that in this case, the reliability ratio is: $\lambda = 9/10$. We repeat for 1000 replications and present in Table 1 the average regression parameters and their standard errors.

The attenuation of the slope coefficient in Table 1 when adding random noise to the independent variable can be seen (from a value of 3.000 to a value of 2.701). We divide the slope coefficient that was estimated from the perturbed data by the reliability ratio, $\lambda = 9/10$ and obtain a consistent estimate for the slope, eg. $2.701 \times 10/9 = 3.000$. The intercept can then be consistently estimated.

Table 1. Simulation study for estimating regression coefficients from data subjected to additive and correlated noise (average across 1000 replications)

Model	Intercept		Slope	
	Estimate	SE	Estimate	SE
Original model	2.997	0.363	3.000	0.018
Additive Random Noise on:				
Dependent variable	3.008	0.438	3.000	0.022
Independent variable	8.976	0.672	2.701	0.033
Both dependent and independent variables	6.985	0.512	2.801	0.025
Correlated Noise on:				
Dependent variable	3.285	0.413	2.986	0.020
Independent variable	3.299	0.409	2.985	0.020
Both dependent and independent variable (multivariate method)	3.010	0.444	2.999	0.022

As can be seen in Table 1, adding correlated noise to the independent variable, the dependent variable, or both variables provides estimates for the slope coefficient and intercept that are close to the true value. Standard errors are higher which reflect the added uncertainty due to the noise addition.

3 Misclassification of Categorical Variables

As described in Shlomo and De Waal (2008), we examine the use of the Post-randomization Method (PRAM) (Gouweleeuw, et al., 1998) to perturb a categorical variable. This method is a more general case of record swapping. Willenborg and De Waal (2001) describe the process as follows:

Let $\mathbf{P}$ be a $L \times L$ transition matrix containing conditional probabilities $p_{ij} = p(\text{perturbed category is } j \mid \text{original category is } i)$ for a categorical variable with L categories. Let $\mathbf{t}$ be the vector of frequencies and $\mathbf{v}$ the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/n$, where n is the number of records in the microdata. For each record of the data, the category of the variable is changed or not changed according to the prescribed transition matrix $\mathbf{P}$ and the result of a random draw from a multinomial distribution with parameters p_{ij} (j=1,…,L). If the j-th category is selected, category i is moved to category j. When i = j, no change occurs.

Let $\mathbf{t}^*$ be the vector of the perturbed frequencies. $\mathbf{t}^*$ is a random variable and $\mathrm{E}(\mathbf{t}^* \mid \mathbf{t}) = \mathbf{tP}$. Assuming that the transition matrix $\mathbf{P}$ has an inverse $\mathbf{P}^{-1}$, this can be used to obtain an unbiased moment estimator of the original data: $\hat{\mathbf{t}} = \mathbf{t}^*\mathbf{P}^{-1}$. Statistical analysis can be carried out on $\hat{\mathbf{t}}$. In order to ensure that the transition matrix has an inverse and to control the amount of perturbation, the main diagonal of $\mathbf{P}$ is dominant, i.e. each entry on the main diagonal is over 0.5. The risk of re-identification under PRAM can generally be high and depends on the values of the diagonal of $\mathbf{P}$. The method introduces ‘uncertainty’ into the true values and this adds to the protection level.

Under PRAM, joint distributions between perturbed and unperturbed variables are distorted which impacts on statistical inference. Variables that typically undergo PRAM are the demographic and geographic identifiers in the microdata which are commonly used in statistical analysis as explanatory variables, for example in regression models. If the statistical agency releases the probability transition matrix **P** then measurement error models can be used. As an example, instead of generating the normally distributed *x* variable for 1000 records in Section 2, we generate a dichotomous z variable obtaining a value of 1 with a probability of 0.6 and 0 otherwise. Note that $Var\ (z) = 240$ in the dataset. The residuals are generated as before with $\varepsilon_i \sim N(0,3)$ and $y_i = 3 + 3z_i + \varepsilon_i$. We carry out a PRAM procedure on the z variable where the probability matrix P has the diagonal $p_{00} = 0.8$ for $z = 0$ and $p_{11} = 0.85$ for $z = 1$. The average least squares estimate for the slope coefficient after 1000 replications is reduced from $\hat{\beta}_1 = 3.000$ in the original dataset to $\hat{\beta}_1 = 1.931$ in the perturbed dataset. In order to calculate the reliability ratio defined in Section 2 to compensate for the measurement error, we need to calculate the additional variance to z due to PRAM, $Var\ (z^* \mid \mathbf{P})$, where z^* is the perturbed categorical variable. This is based on two independent binomially distributed random variables with parameters (z_0, p_{00}) and $(n - z_0, p_{11})$ respectively, where $z_0 = \sum I(z_i = 0)$ and n=1000*:*

$V(\mathbf{z}^* \mid \mathbf{P}) = z_0 p_{00}(1 - p_{00}) + (n - z_0)p_{11}(1 - p_{11}) = 137.2$ The reliability ratio is equal to: $\lambda = 240/(240 + 137.2) = 0.64$. Dividing the slope coefficient estimated from the perturbed dataset by the reliability ratio, $\lambda = 0.64$, we obtain a consistent estimate for the slope, eg. $1.931/0.64 = 3.035$. The calculation of the reliability ratio for the measurement error model depends on the release of the probability matrix **P.** As mentioned, statistical agencies do not generally release SDC parameters. For a regression model, the method of correlated noise addition described in Section 2 can also be applied to a categorical dummy variable to ensure consistent estimation of regression parameters and valid inferences in the perturbed dataset.

Categorical variables imply other types of statistical analysis, such as the chi-square test for independence. Statistical agencies can compensate for the measurement error induced by PRAM by ensuring that the marginal frequency counts of the perturbed variable are approximately equal to the marginal frequency counts of the original variable. This is done by placing the condition of invariance on the transition matrix $\mathbf{P}$, i.e. $\mathbf{tP} = \mathbf{t}$ where **t** is the vector of frequencies. The property of invariance means that the expected values of the marginal distribution of the variable under perturbation are preserved. In order to obtain the exact marginal distribution, we propose using a “without” replacement strategy for selecting the categories to change (or not change). This is carried out by calculating the expected number of categories to change according to the probability matrix and then drawing a random sample without replacement of those categories and changing their values. This procedure ensures exact marginal distributions as well as reduces the additional variance that is induced by the perturbation.

For the purpose of carrying out a chi- square test for independence on a frequency table, the variables spanning the table should be perturbed as a single variable by cross-classifying the categories. For example, if we are interested in analyzing associations in health status with 2 categories and ethnicity with 7 categories, we combine the two variables to obtain a single variable with 14 categories. This single variable is perturbed using an invariant probability matrix of size 14×14 and drawing samples of categories to change without replacement. The resulting chi-square statistic from the perturbed dataset will be equal to the chi-square statistic of the original dataset. To demonstrate, we again generate a dichotomous z variable obtaining a value of 1 with a probability of 0.6 and zero otherwise and $\varepsilon_i \sim N(0,3)$ for 1000 records. We define

$u_i = \exp(3z_i + \varepsilon_i)/(1+\exp(3z_i + \varepsilon_i))$ and classify into a dichotomous variable q obtaining the value of 1 if $u_i \geq 0.7$ and the value of zero otherwise. We are interested in a chi-square statistic for the two dimensional table spanned by z and q. Tables 2a to 2e contain the results of one realization out of a 1000 replications where Table 2a presents the counts and chi-square statistic from the original data. The other tables were calculated as follows:

- Table 2b: PRAM procedure on the z variable and an independent mechanism for changing categories, denoted by z^*,
- Table 2c: PRAM procedure on the z variable under the property of invariance and the without replacement strategy for selecting categories to change, denoted by z^{*I},
- Table 2d: Similar to Table 2b with PRAM applied on the combined single variable obtained by cross-classifying z and q , denoted z^* and q^*,
- Table 2e: Similar to Table 2c with PRAM applied on the combined single variable obtained by cross-classifying z and q, denoted z^{*I} and q^{*I} .

The diagonals of the probability matrices are dominant between 0.8 and 0.85.

All of the chi-square statistics in Tables 2b to 2e are significant and it is reassuring that none of the perturbed tables provided an erroneous conclusion of independence compared to the original table, but this may not always be the case. It is clear that only Table 2e can give the exact value for the chi-square statistic under the property of invariance and the without replacement strategy for selecting categories to change on the combined cross-classified variable.

4 Discussion

Statistical agencies prepare microdata for release by applying SDC methods according to their disclosure control standards and policies for data protection. The protected microdata are then typically delivered to a data archive where approved

Tables 2. Study of chi-square tests for independence under PRAM (one realization out of 1000 replications)

Table 2a: Original counts

$\chi^2 = 420.7$

q	z		
	0	1	Total
0	307	112	419
1	58	523	581
Total	365	635	1000

Table 2b: z perturbed randomly

$\chi^2 = 168.8$

q	z*		
	0	1	Total
0	258	180	438
1	107	455	562
Total	365	635	1000

Table 2c: z perturbed under invariance

$\chi^2 = 181.0$

q	z^{*I}		
	0	1	Total
0	254	165	419
1	111	470	581
Total	365	635	1000

Table 2d: z and q perturbed randomly

$\chi^2 = 287.6$

q*	z*		
	0	1	Total
0	288	133	421
1	91	488	579
Total	379	621	1000

Table 2e: z and q perturbed under invariance

$\chi^2 = 420.7$

q^{*I}	z^{*I}		
	0	1	Total
0	307	112	419
1	58	523	581
Total	365	635	1000

researchers can download the data to their personal computers. Since the microdata has many variables, the protection afforded by pre-defined SDC methods is limited. In addition, we have shown that when statistical agencies do not release the parameters of the SDC methods, it is almost impossible to develop measurement error models for analysis. We demonstrated how the analytical properties of the data can be preserved for a regression analysis and a chi-square test of independence by modifying standard SDC methods. However, developing SDC methods that a priori preserve the analytical properties of the data for all types of statistical analysis is a hard problem. Two possible ways of solving this problem are:

- Develop a remote analysis server where software code is submitted and run on the original data and the outputs checked for disclosure risk prior to their release.
- Develop specialized software that can tailor SDC methods applied to the microdata before its release according to the type of analysis specified. The SDC methods are applied 'on-the-fly' in the software package. The software would also include flexible table generation since aggregated data is a non-perturbative SDC method for microdata.

Implementing ‘on the fly’ SDC methods would not only increase the utility in the microdata for the specified analysis but would also reduce disclosure risk.

Statistical agencies need to carefully consider whether releasing SDC parameters, such as the variance used to generate additive noise, actually increases disclosure risk. While SDC methods can be modified to preserve some analytical properties of the perturbed microdata, it is only through the release of SDC parameters that researchers can compensate for measurement error and ensure correct inferences.

References

1. Anwar, N.: Micro-Aggregation-The Small Aggregates Method. Informe Intern. Luxembourg, Eurostat (1993)
2. Brand, R.: Micro-data Protection Through Noise Addition. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, Vol. 2316 , pp. 97--116. Springer, New York (2002)
3. Dalenius, T. and Reiss, S.P.: Data Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 7, 73--85 (1982)
4. Domingo-Ferrer, J. and Mateo-Sanz, J.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Transactions on Knowledge and Data Engineering, 14, Issue 1, 189--201 (2002)
5. Fienberg, S.E. and McIntyre, J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. In: Domingo-Ferrer, J. and Torra, V. (eds.) Privacy in Statistical Databases. LNCS, Vol. 3050, pp. 14--29. Springer, New York (2004)
6. Fuller, W.A.: Measurement Error Models. Wiley, New York (1987)
7. Fuller, W.A.: Masking Procedures for Micro-data Disclosure Limitation. Journal of Official Statistics, 9, 383--406 (1993)
8. Gourweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463--478 (1998)
9. Kargupta, H., Datta, S., Wang, Q., and Ravikumar, K.: Random Data Perturbation Techniques and Privacy Preserving Data Mining. Knowledge and Information Systems, 7 (4), 387—414 (2005)
10. Kim, J.J.: A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation. ASA Proceedings of the Section on SRM, 370--374 (1986)
11. Shlomo, N. and De Waal, T.: Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. Journal of Official Statistics, 24, No. 2, 1--26 (2008)
12. Tendick, P. and Matloff, N.: A Modified Random Perturbation Method for Database Security. ACM Transactions on Database Systems, 19 (1), 47—63 (1994)
13. Willenborg, L.C.R.J. and De Waal, T.: Elements of Statistical Disclosure Control in Practice. LNS Vol. 155. Springer, New York (2001)