

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 one form of a perturbative SDC method which is carried out by adding 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 SDC methods are proposed that a priori ensure valid inferences on perturbed datasets.

Keywords: additive noise, post-randomization 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 typically 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).

¹

Natalie Shlomo, University of Southampton; email: N.Shlomo@soton.ac.uk

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 probability mechanism (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (Gouweleeuw, Kooiman, Willenborg, and De Wolf, 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 is 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 variable, the sensitive variable would hardly disclose any useful information on the particular unit as they have been perturbed.

In this paper, we focus on a perturbative SDC method which purposely introduces errors into the microdata: additive random noise to a continuous variable and misclassification on 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). 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 should develop more sophisticated SDC methods which allow researchers to 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 additive noise is proposed that will preserve the sufficient statistics and allow 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. 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 in order to ensure 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.

Adding random noise to a continuous variable will not alter the mean of the variable for large datasets but will introduce more variance depending on the variance parameter determined by the statistical agency. 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 an 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:

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

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:

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

since $\text{Cov}(\eta, x) = 0$. 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: $\text{Var}(y) = \text{Var}(y^*) + \text{Var}(\eta)$.

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

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

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:

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

since $\text{Cov}(y, \eta) = 0$. The additive noise on the independent variable biases the slope coefficient downwards which 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:

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

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 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 noise that is correlated with the continuous variable. Correlated noise addition ensures that sufficient statistics (means and variances) of the continuous variable are preserved. One algorithm for generating correlated random noise for a continuous variable y that is easy to implement is as follows:

Procedure (univariate): 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 y . After selecting a δ , calculate: $d_1 = \sqrt{(1 - \delta^2)}$ and $d_2 = \sqrt{\delta^2}$. Now, generate random noise $\mathcal{E}$ 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 y'_i for each record i ($i=1, \dots, n$) as a linear combination: $y'_i = d_1 \times y_i + d_2 \times \mathcal{E}_i$. Note that $E(y') = d_1 E(y) + d_2 [(1 - d_1)/d_2] E(y) = E(y)$ and $\text{Var}(y') = (1 - \delta^2) \text{Var}(y) + \delta^2 \text{Var}(y) = \text{Var}(y)$ since the random noise is generated independently to the original variable y . 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 to demonstrate the effects of adding random noise and correlated random noise to variables in a simple regression model. Assume 1000 records where $x_i \sim N(20, 9)$, $\mathcal{E}_i \sim N(0, 3)$ and model is: $y_i = 3 + 3x_i + \mathcal{E}_i$ (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: $\lambda = 9/10$. Table 1 presents the average regression parameters and their standard errors across 1000 replications.

The attenuation of the slope coefficient in Table 1 is clearly seen when adding noise to the independent variable (from a value of 3.000 to a value of 2.701). By dividing the slope coefficient estimate under measurement error by the reliability ratio, $\lambda = 9/10$, we obtain a consistent estimate, eg. $2.701 \times 10/9 = 3.000$.

Table 1: Simulation study for estimating regression coefficients under measurement error (average across 1000 replications)

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

In Table 1, correlated noise addition to perturb the independent and/or dependent variable (or both), provides estimates for the slope coefficient that are close to the true value. Standard errors are higher which reflects 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 $p_{ij} = p(\text{perturbed category is } j | \text{original category is } i)$ 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} ($j=1, \dots, L$). 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 $E(\mathbf{t}^* | \mathbf{t}) = \mathbf{tP}$. 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.

Statistical agencies can compensate for the measurement error induced by PRAM by placing the condition of invariance on the transition matrix $\mathbf{P}$, i.e. $\mathbf{tP} = \mathbf{t}$. This

releases the researchers of the extra effort to obtain unbiased moment estimates of the original data, since $\mathbf{t}^*$ itself will be an unbiased estimate of $\mathbf{t}$. 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 transition matrix and then drawing a random sample without replacement based on the expected number for changing the values. This procedure ensures exact marginal distributions as well as reduces the additional variance that is induced by the perturbation.

As in all perturbative SDC methods, joint distributions between perturbed and unperturbed variables are distorted which impacts on statistical inference. Variables that typically undergo perturbation are the demographic and geographic identifiers in the microdata which are commonly used in statistical analysis as explanatory variables (e.g., regression modelling and ANOVA). Measurement error models need to be used when undertaking statistical analysis in perturbed datasets. Based on the example in Table 1, instead of the normally distributed x variable, we generate a dichotomous x variable with a probability of 0.6. The residuals are generated as before $\varepsilon_i \sim N(0,3)$ and $y_i = 3 + 3x_i + \varepsilon_i$. We carry out a PRAM procedure on the x variable where the transition matrix $\mathbf{P}$ has the diagonal $p_{11} = 0.8$ for $x = 0$ and $p_{22} = 0.85$ for $x = 1$. The least squares estimate for the slope coefficient under this perturbation scheme was reduced from $\hat{\beta}_1 = 3.000$ to $\hat{\beta}_1 = 1.931$. Calculating the variance of $\mathbf{t}^*$ conditional on the perturbation mechanism based on two independent binomially distributed random variables with parameters (t_0, θ_0) and $(n - t_0, \theta_1)$ respectively, we obtain $V(\mathbf{t}^* | \mathbf{P}) = t_0 p_{11} (1 - p_{11}) + t_1 p_{22} (1 - p_{22}) = 137.2$. The reliability ratio is equal to: $\lambda = 240 / (240 + 137.2) = 0.64$. By dividing the slope coefficient estimate under measurement error by the reliability ratio, $\lambda = 0.64$, we obtain a consistent estimate for the slope coefficient, eg. $1.931 / 0.64 = 3.035$.

As mentioned, parameters of statistical disclosure control methods are rarely released by statistical agencies. This means that statistical agencies need to implement SDC methods that a priori ensure proper inference on perturbed datasets. For example, to carry 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. Adding the property of invariance to the transition matrix and drawing samples without replacement of values to change, we obtain the exact chi-square statistic as in the original dataset. To demonstrate this, we again generate a dichotomous x variable with a probability of 0.6 and $\varepsilon_i \sim N(0,3)$ for 1000 records.

We define $y_i = \exp(3x_i + \varepsilon_i) / (1 + \exp(3x_i + \varepsilon_i))$ which is classified into a dichotomous variable q obtaining the value of 0 if $y_i < 0.7$ and the value of 1 otherwise. In the first case, we carry out a PRAM procedure on the x variable and in the second case, we carry out a PRAM procedure on the cross-classified x and q variables. The PRAM procedure is first carried out using a random transition matrix $\mathbf{P}$ and an

independent mechanism for determining whether a category is changed or not. In the second step we place the property of invariance on the transition matrix and use the without replacement strategy for selecting categories to change which ensures exact margins. The diagonals of the transition matrices are dominant between 0.8 and 0.85.

Table 2a contains the original counts of the 2 by 2 table. Tables 2b to 2e contain the perturbed counts for one realization of the simulation. Tables 2b and 2c contain the perturbed counts when only the x variable is misclassified where Table 2c places the property of invariance on the transition matrix. Similarly, Tables 2d and 2e contain the perturbed counts when both the x variable and q variable are simultaneously misclassified and Table 2e places the property of invariance on the transition matrix. All of the chi-square statistics 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 inference for the chi-square test of independence.

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

Table 2a: Original counts $\chi^2 = 420.7$				Table 2b: x perturbed randomly $\chi^2 = 168.8$			
q	x		Total	q	x^*		Total
	0	1			0	1	
0	307	112	419	0	258	180	438
1	58	523	581	1	107	455	562
Total	365	635	1000	Total	365	635	1000

Table 2c: x perturbed under invariance $\chi^2 = 181.0$				Table 2d: x, q perturbed randomly $\chi^2 = 287.6$			
q	x^{*I}		Total	q^*	x^*		Total
	0	1			0	1	
0	254	165	419	0	288	133	421
1	111	470	581	1	91	488	579
Total	365	635	1000	Total	379	621	1000

Table 2e: x, q perturbed under invariance $\chi^2 = 420.7$			
q^{*I}	x^{*I}		Total
	0	1	
0	307	112	419
1	58	523	581
Total	365	635	1000

4 Discussion

Statistical agencies prepare microdata for release by applying SDC methods according to their disclosure control practices and policies for data protection. The protected

microdata are then typically delivered to a data archive where approved researchers can download the microdata to their personal computers. Since the microdata may have hundreds of variables, the protection afforded by pre-defined SDC methods may likely still contain disclosure risks. As an alternative to downloading the disclosure controlled microdata, we propose developing a specialized software package which can implement SDC methods 'on the fly'. This package would also include the specification and protection of tabular data. The general idea is that researchers would specify the variables of interest and the types of analysis that they require through the software interface. The researcher's input would determine which SDC methods to apply to the microdata prior to its release that are specifically tailored to the analysis.

Implementing 'on the fly' SDC methods would not only increase the utility in the microdata for the specified analysis but would also raise the level of protection. As an example, suppose the researcher is interested in carrying out log linear modeling based on a set of explanatory variables. By ensuring that the marginal counts are not perturbed for the main effects, 2-way and 3-way interactions, the estimated expected counts will not be affected.

Most statistical inferences however will need to take into account the measurement error arising from the perturbation of the microdata and measurement error models need to be developed and incorporated into the analysis. Statistical agencies need to carefully consider whether releasing SDC parameters, such as the variance used to generate additive noise, actually increases the disclosure risk. Without these SDC parameters, researchers are not able to compensate for the measurement errors which may lead to erroneous inference.

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 Inference Control in Statistical Databases (ed. J. Domingo-Ferrer), LNCS Vol. 2316, 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 Privacy in Statistical Databases (eds. J. Domingo-Ferrer and V. Torra), LNCS Vol. 3050, 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. Gourwelleuw, 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. 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)
10. 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)
11. Willenborg, L.C.R.J. and De Waal, T.: Elements of Statistical Disclosure Control in Practice. LNS Vol. 155. Springer, New York (2001)