RESEARCH REPORT SERIES
(Statistics #2018-11)

Post-randomization for Identification Risk Limited Microdata Release from General Surveys

Cheng Zhang[1],
Tapan K. Nayak[2]

[1]Department of Statistics, George Washington University
[2]Center for Statistical Research and Methodology, U.S. Census Bureau
and Department of Statistics, George Washington University

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: October 5, 2018

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Post-randomization for Identification Risk Limited Microdata Release from General Surveys [*]

Cheng Zhang[†] and Tapan K. Nayak[‡]

Abstract

Before releasing survey data, statistical agencies usually perturb the original data to keep each survey unit's information confidential. One significant concern is identity disclosure, which occurs when an intruder correctly identifies the records of a survey unit by matching the values of some key (or pseudo-identifying) variables. Nayak, Zhang and You (2018) developed a post-randomization method for a strict identification risk control in releasing survey microdata. The procedure also well preserves the observed frequencies and hence statistical estimates in case of simple random sampling. We show that in general surveys, the procedure may induce considerable bias in commonly used survey weighted estimators. We propose a modified procedure that better preserves weighted estimates. The procedure is illustrated and empirically assessed with an application to a publicly available U.S. Census Bureau data set.

Key words and Phrases: Identity disclosure; data partitioning; key variable; post-randomization block; survey weighted estimator; total variation distance.

[*] The views expressed in this article are those of the authors and not necessarily those of the U.S. Census Bureau.

[†] Department of Statistics, George Washington University, Washington, DC 20052.

[‡] Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233 and Department of Statistics, George Washington University, Washington, DC 20052.

1. Introduction

Many statistical agencies collect and release data to assist research, planning and policy making. However, agencies also need to keep each survey unit's data confidential. Thus, the released data should inform about the population, but not about any survey participant. In particular, one should not be able to identify the records of any survey unit in the released data. Anonymization by removing all direct identifiers, such as name, social security number and address, is inadequate for avoiding identity disclosure, because an intruder might correctly identify the records of a target unit by matching externally available values of some variables, called pseudo-identifiers or key variables. So, agencies usually release a perturbed version of the original data to control disclosure.

In this paper, we shall consider only identity disclosure by key variable matching. For discussions about other scenarios and forms of disclosure and various data perturbation methods, including grouping, data swapping, cell suppression, random noise mixing and post-randomization, we refer interested readers to the books: Willenborg and de Waal (2001), Duncan et al. (2011) and Hundepool et al. (2012). The perturbed data usually dilute, suppress, and even distort some features of the original data. So, for real application, perturbation methods should be chosen after examining the trade-offs between confidentiality protection and data utility loss; see Duncan and Stokes (2004), Karr et al. (2006) and Cox et al. (2011).

Identity disclosure is one of the most serious forms of confidentiality violation. For categorical key variables, many researchers, e.g., Bethlehem et al. (1990), Greenberg and Zayatz (1992), Chen and Keller-McNulty (1998), Skinner and Elliot (2002) and Shlomo and Skinner (2010), have explored this topic and proposed several identification risk measures. For example, Bethlehem et al. (1990) defined identification risk as the conditional probability that a unit is population unique given that it is sample unique. Shlomo and Skinner (2010) defined it as the probability that for a sample unique unit, a unique match in released data is a correct match. These measures depend on unknown population level frequencies and hence are difficult to use for assessing and controlling disclosure risk.

Recently, Nayak, Zhang and You (2018), henceforth NZY, proposed a novel approach to measuring and controlling identification risk, without having to estimate any unknown parameter. For a brief review, let X denote the cross-classification of all key variables and $c_1, ..., c_k$ denote its categories. Let Z be a randomized version of X, whose values constitute a perturbed data set for public release. Also, let T_j and S_j denote the frequencies of c_j in the original and perturbed data, respectively, and let $\mathbf{T} = (T_1, ..., T_k)'$ and $\mathbf{S} = (S_1, ..., S_k)'$. Suppose an intruder wants to identify the records of a target unit B in the released data. Let $X_{(B)}$ denote B's value of X, and suppose $X_{(B)} = c_j$. NZY assumed that the intruder knows $X_{(B)}$ and that B is in the sample and then identifies B with a randomly selected unit from the units in the released data that match $X_{(B)}$. Note that S_j is the number of matches found for B in released data, and if $S_j = 0$, the intruder does not declare any match. Thus, B can be correctly matched only if $X_{(B)}$ is not changed during perturbation.

Under the above scenario, NZY proposed to assess B's identification risk by

$$R_j(a) = P(CM | X_{(B)} = c_j, S_j = a), \tag{1.1}$$

where CM denotes the event that B is correctly matched and a is a positive integer. Then, they suggested to control (1.1) for all survey units (i.e., all c_j) and all $a > 0$. Thus, as a precise and stringent disclosure control goal, they proposed to ensure that

$$R_j(a) \leq \xi \quad \text{for all } j = 1, \ldots, k, \text{ and all integers } a \geq 1, \tag{1.2}$$

where ξ is specified by the data agency.

One difficulty in dealing with (1.1) is that $R_j(a)$ depends on the unknown population frequencies, and NZY avoided that by further conditioning on $\mathbf{T}$. Specifically, they considered

$$R_j(a, \mathbf{t}) = P(CM | X_{(B)} = c_j, S_j = a, \mathbf{T} = \mathbf{t}), \tag{1.3}$$

which is determined only by the randomization probabilities and thus can be calculated without knowing the population frequencies. NZY ensured (1.2) by guaranteeing

$$R_j(a, \mathbf{t}) \leq \xi \quad \text{for all } a > 0\ ,\ j = 1, \ldots, k \text{ and for all } \mathbf{t}. \tag{1.4}$$

They also argued that in most applications, moderately large values of ξ should be used, as intruders should have substantial evidence for declaring matches and (1.3) is calculated under a very conservative assumption that the intruder knows that his target unit is in the sample.

NZY developed a post-randomization method to ensure (1.4) for $\xi > 1/3$. We shall describe the main parts of the procedure later. They also demonstrated that the procedure affects the relative frequencies of various marginal and joint cells very little. However, it should be noted that comparing relative frequencies based on the original and a perturbed data set is meaningful when the data are collected using simple random sampling (SRS), where sample relative frequencies are standard estimates of corresponding relative frequencies in the population. For general surveys, one should use the survey weights to estimate population level counts and relative frequencies. The main goals of this paper are to examine the NZY procedure's effects on *weighted counts* and propose a modified procedure for better preserving the customary weighted estimates.

In Section 2, we briefly review the central parts the NZY procedure and examine its effects on weighted counts through an example. In Section 3, we discuss certain challenges and ideas for modifying the NZY procedure for application to general surveys. In Section 4, we describe our proposed procedure. In Section 5, we present an illustrative example and an empirical evaluation of our procedure. Section 6 is devoted to some concluding remarks.

2. Effects of the NZY Procedure on Weighted Counts

In this section, we briefly describe the essential parts of the NZY procedure, for $1/3 < \xi < 1/2$, and explore its effects on weighted counts. The NZY procedure has two main parts: creating post-randomization blocks (PRBs) and post-randomizing key variable values within each block. The PRBs are formed by partitioning the data by groups of cells of key variables and then taking all units in all sensitive cells within each partition set. A cell is non-sensitive if its frequency is sufficiently large to make the disclosure risk of the units in that cell less than ξ even if their values are not perturbed. For example, for $1/3 < \xi < 1/2$, any cell with frequency 3 or more is non-sensitive, and all singleton and doubleton cells are sensitive. As all perturbations occur within the PRBs, data partitioning is a vital tool for controlling the nature and magnitude of changes to the original values. A simple approach is to partition the data by the cells of the cross-classification of coarsened versions of some of the key variables. The following example of NZY, which we shall adopt also to evaluate our proposed method, illustrates the technique.

The NZY example concerns the U.S. Census Bureau's 2013 person-level Public Use Microdata Sample (PUMS) for the state of Maryland, available at
https://www.census.gov/programs-surveys/acs/data/pums.html. The sample size is 59,033 and the data set contains the values of several demographic and economic variables. NZY used five key variables: gender (2), age (92), race/ethnicity (9), marital status (5) and Public Use Microdata Area (PUMA) (44), where the number of categories of each variable is given in parentheses. These five variables shall be denoted $X_1, \ldots, X_5$, respectively. Their cross-classification yields 364,320 cells. The PUMS data set yields only 25,406 nonempty cells, of which 13,662 are singleton and 4,777 are doubleton cells.

In NZY, the data are partitioned using the cross-classification of the following three (coarsened) variables: $X_1^* =$ gender, $X_2^* =$ age given in the 7 intervals 0 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, and 65 and above, and $X_3^* =$ race shown in the three categories white, black and 'other races.' The cross-classification gives 42 partition sets. For example, all females of 'other races' in the age group 18 - 24 constitute one partition set. All units in the singleton and doubleton cells of the original key variables in each partition set form one PRB. The number of cells in the 42 PRB's ranged between 124 and 1480. The post-randomization step, described below, keeps each perturbed value within its PRB. Consequently, the above partitioning preserves gender, allows marital status and PUMA to change freely, keeps age within the broader categories of X_2^* and retains race, if it is white or black.

The NZY procedure perturbs data using post-randomization. The general idea was introduced in Gouweleeuw et al. (1998) and has been investigated further by Van den Hout and Van der Heijden (2002), Van den Hout and Elamir (2006), Cruyff et al. (2008), Shlomo and De Waal (2008), Nayak and Adeshiyan (2016) and others. To describe the NZY post-randomization, which is applied independently to each PRB, let m denote the number of cells in a PRB and for notational simplicity, suppose the cells are $c_1, \dots, c_m$. Typically, m is much smaller than k, the number of cells in complete cross-classification. Then, the true category of each unit is randomized independently with transition probabilities

$$p_{ij} = P(Z = c_i | X = c_j) = \begin{cases} 1 - \frac{\theta}{t_j}, & \text{if } i = j; \\ \\ \frac{\theta}{(m-1)t_j}, & \text{if } i \neq j, \end{cases} \tag{2.1}$$

where θ is a design parameter whose value is chosen to meet our disclosure control goal. We shall use $P = ((p_{ij}))$ to denote transition probability matrix. NZY called the above scheme IFPR (inverse frequency post-randomization), as the probability that it changes the category of a unit is inversely proportional to the frequency of the unit's category. When a category is changed, the new category is chosen at random from the remaining categories in the PRB.

To satisfy (1.4), for $1/3 < \xi < 1/2$, the NZY procedure uses a suitable value of θ, based on the following results; see NZY for derivations and proofs. Under any post-randomization (including none), $R_j(a, \mathbf{t}) \leq 1/3$ if $t_j \geq 3$ or $a \geq 3$. So, for $1/3 < \xi < 1/2$, only the units in singleton and doubleton cells need perturbation, and (1.4) needs to be guaranteed only for $a = 1, 2$. Also, for a general P,

$$R_j(1, \mathbf{t}) = \left[t_j + \frac{1 - p_{jj}}{p_{jj}} \sum_{i \neq j}^{m} \frac{p_{ji}}{1 - p_{ji}} t_i \right]^{-1} \tag{2.2}$$

and for IFPR, $R_j(1,\mathbf{t}) \leq h(\theta)$, where

$$h(\theta) = \begin{cases} \frac{1-\theta}{1-\theta+\theta^2}, & \text{if } \theta \leq \frac{2}{3}, \\ \frac{2-\theta}{4-2\theta+\theta^2}, & \text{if } \theta > \frac{2}{3}, \end{cases}$$

which is a strictly decreasing function of θ, with $h(0) = 1$ and $h(1) = 1/3$. Moreover, for any given θ, $R_j(2,\mathbf{t}) \leq R_j(1,\mathbf{t})$ if $m \geq (1-\theta)^{-1}$. Let $\theta_0(\xi)$ denote the solution of $h(\theta) = \xi$. The NZY procedure satisfies (1.4) by using $\theta = \theta_0(\xi)$ and including at least $\lceil (1-\theta_0(\xi))^{-1} \rceil$ cells in each PRB.

In general, survey weights are used to estimate population frequencies and proportions. For a given characteristic A, i.e., a subset of the sample space of the survey variables, let $F(A)$ and $\pi(A)$ denote the frequency and proportion, respectively, of A in the population. Let n denote the sample size and w_i denote the survey weight of unit u_i. The Horvitz-Thompson estimator uses the inverse of the selection probability of u_i for w_i. In practice, w_i's are also calibrated to account for nonresponse. Then, customary estimators of $F(A)$ and $\pi(A)$ are:

$$\hat{f}(A) = \sum_{i=1}^{n} w_i I(u_i \in A) \quad \text{and} \quad \hat{\pi}(A) = \hat{f}(A) \div \sum_{i=1}^{n} w_i. \tag{2.3}$$

We shall refer to $\hat{f}(A)$ as a weighted count (or frequency). In SRS, all weights are equal and consequently $\hat{\pi}(A) = (1/n)\sum_{i=1}^{n} I(u_i \in A)$.

As the NZY procedure does not alter the survey weights, it may affect the weighted counts more than the unweighted counts. In the NZY example, we examined this for some variables, and the results for 'race' from one perturbed data set are given in Table

1. Recall that due to our data partitioning, race remains unchanged if it is white or black. So, we do not include those two categories. We also omit the 'Alaska native alone' category as the original data contains only 1 unit in that category. In Table 1, t and f denote the raw and weighted counts in the original data, whereas t^* and f^* denote the corresponding counts in a perturbed data set. The table also gives the relative absolute differences: $\Delta_t = |t - t^*|/t$ and $\Delta_f = |f - f^*|/f$.

Table 1: Effects of NZY procedure on raw and weighted counts of *Race* categories.

Race	t	t^*	Δ_t	f	f^*	Δ_f
American Indian alone	97	92	0.0515	11235	10589	0.0356
American Indian and Alaska native	42	46	0.0952	4820	3210	0.2689
Asian	3461	3345	0.0335	354215	376384	0.0564
Native Hawaiian and other Pacific	20	21	0.0500	2443	3400	0.0794
Some other race alone	1349	1337	0.0089	210331	177567	0.1380
Two or more races	1623	1652	0.0179	168669	180626	0.0660

In Table 1, the values of Δ_f are mostly larger than the corresponding values of Δ_t. It is seen that the NZY procedure affects the weighted counts substantially although the raw counts do not change much. The most drastic change occurred for the 'America Indian and Alaska native' category, where the raw count increased by 9.52% but the weighted count decreased by 26.89%. For the 'some other race alone' category, which contained 1349 units, the perturbed weighted frequency deviated by 13.8%, whereas the raw frequency changed by only 0.89%. We also observed similar phenomena for other key variables. These findings indicate that the NZY procedure should be modified suitably for applying to general surveys.

3. Preserving Weighted Counts

The NZY procedure attempts to retain data utility in two ways, by data partitioning and using an unbiased post-randomization. Specifically, IFPR implies that $E(S_j|\mathbf{T}) = T_j, j = 1, \ldots, k$, i.e., the expected frequency of any cell after perturbation is the same as its frequency in the original data. For any cell that is included in any PRB, the expected number of units that move out of the cell is the same as the expected number of units that move into the cell, both being θ. This greatly helps to preserve raw frequencies. However, even if the same number of units move out and move in, their total weights may differ substantially. Consequently, IFPR is expected to be less effective for preserving weighted counts.

If the survey weights differ only due to stratification, the matter can be handled easily by applying the NZY procedure within each stratum. For general surveys, a natural idea is to use a post-randomization procedure that changes the weighted counts unbiasedly, as Gouweleeuw et al. (1998) suggested. Suppose a PRB consists of m cells $c_i (i = 1, \ldots m)$ with raw and weighted frequencies t_i and f_i, respectively. Also let $\mathbf{t} = (t_1, \ldots, t_m)'$ and $\mathbf{f} = (f_1, \ldots, f_m)'$. Then, P is unbiased with respect to $\mathbf{f}$ if $P\mathbf{f} = \mathbf{f}$. If P chosen satisfying this condition, the expected weighted frequency of any cell after post-randomization will equal its original weighted frequency. Analogous to (2.1), a solution of $P\mathbf{f} = \mathbf{f}$ is

$$p_{ij} = P(Z = c_i|X = c_j) = \begin{cases} 1 - \frac{\theta}{f_j}, & \text{if } i = j; \\ \frac{\theta}{(m-1)f_j}, & \text{if } i \neq j. \end{cases} \tag{3.1}$$

Here, the design parameter θ can be assigned any value in $[0, \min_j\{f_j\}]$. However, as we show next, it may not be possible to guarantee (1.4) using any P of the above form.

Consider the situation where all m cells in a PRB are singletons, i.e., $t_i = 1, i = 1, \ldots, m$. This implies that $f_i = w_i, i = 1, \ldots, m$. For notational simplicity, suppose $w_1 \leq \ldots \leq w_m$. Let $\mathbf{t}_* = (1, \ldots, 1)'$. Then using (3.1) in (2.2), we find that under (3.1), the identification risk of the unit in c_m, when it is matched uniquely (i.e., $a = 1$), is

$$\begin{aligned} R_m(1, \mathbf{t}_*) &= \Big[t_m + \frac{1 - p_{mm}}{p_{mm}} \sum_{i=1}^{m-1} \frac{p_{im}}{1 - p_{im}} t_i\Big]^{-1} \\ &= \Big[1 + \frac{\theta}{w_m - \theta} \sum_{i=1}^{m-1} \frac{\theta}{(m-1)w_i - \theta}\Big]^{-1}. \end{aligned}$$

We can verify that $R_m(1, \mathbf{t}_*)$ decreases as θ increases and it increases with each w_i. So, we obtain a lower bound for $R_m(1, \mathbf{t}_*)$ by taking $\theta = w_1$ (the largest possible value of θ) and $w_i = w_1$ for $i \neq m$, viz.

$$\begin{aligned} R_m(1, \mathbf{t}_*) &\geq \Big[1 + \frac{w_1}{w_m - w_1} \sum_{i=1}^{m-1} \frac{w_1}{(m-1)w_1 - w_1}\Big]^{-1} \\ &= \Big[1 + (\frac{w_m}{w_1} - 1)^{-1} \frac{m-1}{m-2}\Big]^{-1}. \end{aligned}$$

Now, it follows that for any $0 < \xi < 1$, $R_m(1, \mathbf{t}_*) > \xi$ if

$$\frac{w_m}{w_1} > 1 + \Big(\frac{m-1}{m-2}\Big)\Big(\frac{\xi}{1-\xi}\Big). \tag{3.2}$$

The right side of (3.2) is a decreasing function of m and the inequality implies, for example, $R_m(1, \mathbf{t}_*) > 0.5$ if $(w_m/w_1) > 3$ and $m \geq 3$. For the PUMS data set used in our example, the survey weights differ substantially and (3.2) holds frequently even for $\xi = 0.5$. Actually, the ratio of the largest to smallest survey weights was much larger than

3 in many PRBs. Thus, while (3.1) is useful for perturbing weighted counts unbiasedly, it is not adequate for ensuring (1.4) for practical values of ξ, say $\xi \leq 0.5$. The difficulty stems from the facts that unbiasedness is with respect to weighted counts whereas identification risks depend on raw counts.

We propose to use (2.1) to control disclosure risks, but refine data partitioning to better preserve the weighted counts. A natural idea is to further split the data by survey weights so that the survey weights are fairly homogeneous within each PRB. However, it causes cell splitting, as discussed next, which increases perturbation rates and reduces disclosure risk. In the NZY procedure, the PRBs are defined essentially by the singleton and doubleton cells of the key variables. These cells are grouped and all units falling in the cells within a group form one PRB. As a result, the two units in a doubleton cell fall in one PRB and each unit's category is changed with probability $\theta/2$. If a PRB is further divided into multiple PRBs by survey weights, two units in a doubleton cell (originally) may fall in two different PRBs, each appearing as a singleton unit within its own PRB, in which case the category of each unit will be changed with probability θ (instead of $\theta/2$ in the NZY procedure).

Next, we use an example to elaborate cell splitting and its effects. Suppose that the left panel of Table 2 shows one PRB in the NZY procedure. It consists of 11 units, falling in 8 cells, $c_1, \ldots, c_8$. For each unit, the left panel shows its original category (X), survey weight (w) and cell frequency (t). The survey weights form two well separated clusters, centered around 100 and 200. To form weight homogeneous PRBs, the original PRB is split into two PRBs, by weight clusters, as shown in the right panel of Table 2, where t' shows cell frequency within the PRB. Consider post-randomizing the original categories within each of the two new PRBs using (2.1). Here, the originally doubleton units 2

Table 2: An example of PRB splitting by survey weight.

unit	X	w	t
1	c_1	100	1
2	c_2	101	2
3	c_2	200	2
4	c_3	200	1
5	c_5	99	2
6	c_5	199	2
7	c_6	100	2
8	c_6	102	2
9	c_7	201	1
10	c_8	101	2
11	c_8	200	2

unit	X	w	t'
1	c_1	100	1
2	c_2	101	1
5	c_5	99	1
7	c_6	100	2
8	c_6	102	2
10	c_8	101	1

unit	X	w_i	T'
3	c_2	200	1
4	c_3	200	1
6	c_5	199	1
9	c_7	201	1
11	c_8	200	1

and 3 (in cell c_2) appear as uniques within their own PRBs, and thus move out of c_2 with probability θ. However, units 7 and 8 in c_6 fall in one PRB and will be changed with probability $\theta/2$. Perturbing some doubleton units with probability θ instead of $\theta/2$ further reduces identification risk. Thus, the choice of θ as the solution of $h(\theta) = \xi$ is expected to be conservatively large, as we shall see in our example in Section 5.

4. Proposed Method

The five-step procedure proposed below guarantees (1.2) in a general survey for any given $\xi \in (1/3, 1/2)$ and a set of categorical key variables, specified by the data agency. Essentially, we modify the PRB forming process in the NZY procedure, to reduce its effects on statistical estimates. In particular, the first three steps of the two procedures are the same, which we state concisely and refer interested readers to NZY for additional discussion.

Step 1. For given ξ, calculate $\theta_0(\xi)$ by solving $h(\theta) = \xi$, and $m_0 = \lceil\{1 - \theta_0(\xi)\}^{-1}\rceil$. These two values are used in steps 4 and 5 below.

Step 2. Choose a set $\mathcal{C}$ of categorical survey variables for post-randomization. This set should contain all key variables. It may also contain some non-key variables to avoid edit failures. Let $X_{\mathcal{C}}$ denote the cross-classification of all variables in $\mathcal{C}$.

Step 3. Partition the data set by groups of similar cells of $X_{\mathcal{C}}$ such that the number of singleton and doubleton cells (of $X_{\mathcal{C}}$) in each partition set is at least m_0. One convenient approach, suggested by NZY, is to suitably coarsen each variable in $\mathcal{C}$ and use their cross-classification. This amounts to a rectangular partition of the sample space of the variables in $\mathcal{C}$. However, the subsequent steps of the proposed procedure work as well for any data partitioning.

Step 4. We form the PRBs in this step. Within each partition set, we take all units falling in singleton and doubleton cells of $X_{\mathcal{C}}$ and divide those units into relatively weight homogeneous PRBs, each containing at least m_0 cells. For simplicity, we suggest a rank-based splitting procedure, but other approaches may also be used. First, we rank the units by ascending survey weights, breaking ties at random so that each rank corresponds to only one unit. Then, from the ranked list, take the first $2m_0$ units to form one PRB, next $2m_0$ units to form another PRB and so on until we are left with less than $4m_0$ units, which are put in the last PRB. Thus, each PRB contains at least $2m_0$ units and hence at least m_0 cells.

Step 5. Post-randomize the $X_{\mathcal{C}}$ values within each PRB using (2.1) and cell frequencies within the PRB. Specifically, if a PRB contains m non-empty cells, say $c_1, ..., c_m$, and the frequency of c_i $(i = 1, \ldots, m)$ within the PRB is t'_i, then we apply the transition probabilities $p_{jj} = 1 - \theta_0/t'_j$ and $p_{ij} = \theta_0/[(m-1)t'_j]$ for $i \neq j$, where θ_0 is as calculated in Step 1.

We want to discuss a few points about the above procedure. As noted earlier, one may use other data partitioning methods in Step 3. A sequential approach with varying segmentation may be useful in some application. For instance, in the NZY example, one might first partition the data by gender and three race groups, white, black and 'other races' and then futher divide each group by suitable age intervals for the group. The data set contains 37201 whites, 15239 blacks and 6593 'other races.' The 7 age intervals of X_2^* may be appropriate for the two smallest groups, characterized by males and females of 'other races.' The black male and black female groups are larger and one may use smaller age intervals. For white by gender groups, one may use even smaller age intervals. Alternatively, one may partition each group by wide age intervals and then further divide large subgroups by other variables, e.g., marital status.

There is a trade-off between steps 3 and 4. The scope for forming weight homogeneous PRBs, in Step 4, decreases as the resulting partition set sizes in Step 3 decrease. Recall that in our procedure, each PRB contains at least $2m_0$ units. A finer partition in Step 3 yields a greater control on the magnitude of possible changes to the original data values, but it reduces the scope for achieving homogeneity of survey weights due to smaller partition sizes. For instance, the data partitioning in the NZY example guarantees that any original age between 0 and 17 will remain within this interval after perturbation. A finer partition, say by dividing the interval 0 to 17 into 0 to 9 and 10 to 17, yields a tighter control on possible changes to age, but also creates two smaller partition sets, which may not be further divided by small survey weight intervals while meeting the minimum PRB size requirement. In practice, one may optimize steps 3 and 4 through some experimentation.

One may also consider interchanging steps 3 and 4, i.e., first partition the data by fixed or quantile intervals of survey weights and then by coarsened versions of key variables. However, we found this approach problematic in our empirical investigation. Specifically, it yielded some PRBs with less than m_0 cells. The problem can be avoided by partitioning the data appropriately (and differently) in each survey weight class. However, that requires much human intervention, which is costly and inconvenient.

5. An Example

In this section, we present results from an application of our procedure to the 2013 Maryland PUMS, described in Section 2. For direct comparison, we adopt the disclosure control goal and data partitioning of NZY, as reviewed in section 2. Specifically, we take the same five key variables (gender, age, race, marital status and PUMA) and $\xi = .395$, and correspondingly $\theta_0 = 0.8$ and $m_0 = 5$. We shall examine effects of our procedure on both identification risks and survey weighted statistical estimates.

First, we describe the distribution of survey weights of the 59,033 persons in the data set. The average weight is about 100 and the total weight is 5,928,814, which is also an estimate of the population size. The weight distribution is highly positively skewed. Figure 1 shows the histogram of all weight that are 500 or less, which account for 99.7% of all sampling units. The modal class of the histogram is $(60, 70)$. We also found that 84.44% of all weights are between 50 and 150, and 91.74% are below 200. The boxplot in Figure 2 depicts the 177 weights that are above 500. These constitute 0.3% of all units, and their total weight is 107,300, which represents 1.8% of the population. The wide variation and high skewness of the survey weights contributed to the weak performance of the NZY procedure, noted in Section 2.

Figure 1: Histogram of survey weights that are 500 or lower.

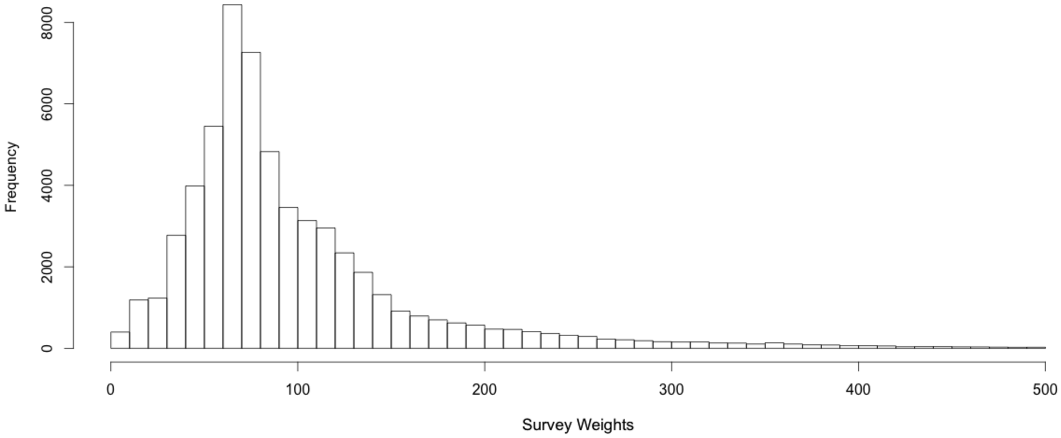
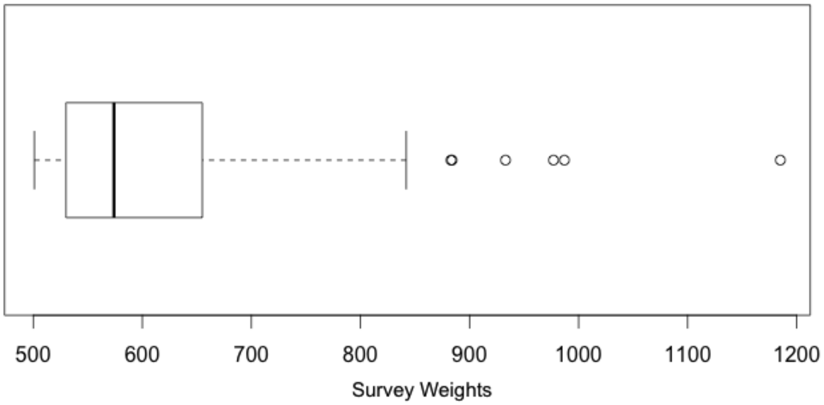


Figure 2: Boxplot of survey weights above 500.



To apply our procedure, we first partitioned the data, as in NZY, into 42 sets by the cross-classification of gender, X_2^* and X_3^*, as described in Section 2. Within each partition set, we took all units in singleton and doubleton cells of $X_{\mathcal{C}}$. Recall that for $\xi = 0.395$, only those units need perturbation. The number of singleton and doubleton cells in the

42 partition sets ranged between 124 and 1480. As described in step 4, we arranged all singleton and doubleton units within each partition set by increasing order of survey weight, breaking ties at random. Then, we formed PRBs by taking 10 consecutive units, starting with the first, until less than 20 units remained left, which were assigned to one PRB. Thus, each PRB contains at least 10 units and hence at least 5 cells (as no cell contains more than 2 units). Recall that (1.2) is guaranteed to hold for $\xi = 0.395$ if each PRB contains at least 5 cells and $\theta_0 = 0.8$. Finally, we post-randomized all $X_{\mathcal{C}}$ values within each PRB, as stated in Step 5, with $\theta_0 = 0.8$. As in Sholomo and Skinner (2010) and NZY, some results from one perturbed data set are reported below. However, we observed similar results when we repeated the post-randomization in Step 5.

5.1. Empirical Identification Risk

Our procedure changed the $X_{\mathcal{C}}$-category of 10,938 (or 80.06%) of the 13,662 singleton units and 7,542 (or 78.94%) of the 9,554 doubleton units. In contrast, the NZY procedure changed the $X_{\mathcal{C}}$-category with probabilities 0.8 and 0.4 for singleton and doubleton units, respectively. In our case, the perturbation rate (78.94%) for doubleton units is much larger, as many of those units are treated as singleton units (within their PRBs) due to further partitioning by survey weights. As one would expect, and seen below, the increased perturbation rate lowered identification risks.

For presenting the empirical identification risks, we shall use τ and τ^* to denote the number of matches for a unit in the original and perturbed data, respectively. In a perturbed data set, a unit's probability of correct match is 0 if its category is changed, and $1/\tau^*$ otherwise. (Note that $\tau^* \geq 1$ if its category remains unchanged.) We calculated this probability for all 23,216 originally singleton and doubleton units. Table 3 gives their averages over certain classes. The corresponding values from NZY are given in parentheses.

Table 3: Empirical identification risks.

	$\tau = 1$	$\tau = 2$	
$\tau^* = 1$	0.2168 (0.2315)	0.1258 (0.3933)	0.1882 (0.2849)
$\tau^* = 2$	0.1997 (0.1961)	0.1032 (0.3477)	0.1505 (0.2827)
	0.1362 (0.1348)	0.0965 (0.3027)	

In Table 3, the row and column headings describe the cases considered. Thus, for example, empirical identification risks are 0.1362 for originally singleton units ($\tau = 1$), 0.1882 for uniquely matched units in perturbed data ($\tau^* = 1$), and 0.1258 for originally doubleton units with unique matches in perturbed data ($\tau = 2, \tau^* = 1$). The empirical risks under NZY and the new procedure in the column $\tau = 1$ are very similar. This is not surprising because both procedures change the category of each singleton unit with the same probability (0.8). The risks for doubleton units, in the column $\tau = 2$, are substantially lower under our procedure. The same is also seen for $\tau^* = 1$ and $\tau^* = 2$. The values in Table 3 also show that the NZY procedure better controls the risks for singleton units than doubleton units. In contrast, under our procedure, identification risks for singleton units are larger than that for doubleton units. Also, for our procedure, the largest risk in Table 3 is 0.2168 (when $\tau = 1, \tau^* = 1$), which is much smaller than $\xi = .395$. Thus, one may use a suitably smaller θ to improve data utility while meeting the disclosure control goal.

5.2. Effects on Estimates of Population Frequencies

We shall now examine effects of our procedure on (weighted) estimates of population level frequencies, as described in Section 2, and also on estimates of their sampling variance, which are calculated as follows. For each person i in the PUMS data set, the U.S. Census Bureau gives one survey weight (w_i) and 80 replicate weights, $w_{ij}, i = 1, \ldots, n, j = 1, \ldots, 80$. The population total, τ_Y, of a survey variable Y is estimated by $\hat{\tau}_Y = \sum_{i=1}^{n} w_i Y_i$. The variance of $\hat{\tau}_Y$ is estimated by

$$\hat{V}(\hat{\tau}_Y) = \frac{1}{80} \sum_{i=1}^{80} (\hat{\tau}_{Yj} - \hat{\tau}_Y)^2, \tag{5.1}$$

where $\hat{\tau}_{Yj} = \sum_{i=1}^{n} w_{ij} Y_i$. For categorical variables, in above formulas Y_i is replaced by appropriate indicator functions as in (2.3). For further details about calculation of survey weights, replicate weights and variance estimation we refer the reader to U.S. Census Bureau (2006), Fay (1984), Fay and Train (1995), Wolter (2007) and Ash (2014).

Table 4 gives some results for six *race* categories. We omit the 'White' and 'Black' categories as those are kept unchanged during perturbation and the 'Alaska native alone' category as it contained only 1 unit. In Table 4, f represents the (weighted) estimates of population frequencies from the original data, as reported earlier in Table 1, and the estimates under our procedure are denoted by $\tilde{f}$. Similarly, $\hat{\sigma}$ and $\tilde{\sigma}$ denote the estimated standard deviations, based on the replicate weight method as described by (5.1), calculated from the original and our perturbed data sets, respectively.

The values in Table 4 show that $|f - \tilde{f}|/f$ is noticeably smaller than $|f - f^*|/f$, given in Table 1, for all but the 'American Indian Alone' category. Thus, our procedure improves the accuracy of statistical estimates. For the three large race categories (with f

Table 4: Estimated frequencies of Race categories

Race	f	$\tilde{f}$	$\frac{\lvert f-\tilde{f}\rvert}{f}$	$\hat{\sigma}$	$\tilde{\sigma}$
American Indian alone	11235	9882	0.1204	742	545
American Indian and Alaska native	4820	5582	0.1581	422	476
Asian	354215	359347	0.0145	1620	3115
Native Hawaiian and other Pacific	2443	2474	0.0123	321	334
Some other race alone	210331	206418	0.0186	5139	3894
Two or more races	168669	168073	0.0035	3791	2354

over 160,000), f and $\tilde{f}$ are quite close. Our procedure (and also NZY) affects categories with small frequencies substantially, as almost all units in such categories are single or doubleton and are randomized. As Table 1 shows, the three smallest frequencies are 20, 42 and 97, and two of their corresponding categories have fairly large relative difference, $|f - \tilde{f}|/f$, in Table 4.

We may also note that for a small category, the absolute difference between the estimates of its relative frequency based on the original and our perturbed data is very small. For example, our estimates of the proportion of persons in the population falling in the 'American Indian Alone' category are $\hat{p} = 11235 \div 5928814 = 0.0019$ based on the original data, and $\tilde{p} = 9882 \div 5928814 = 0.0017$ based on our perturbed data, as the total weight is 5928814. Here, the absolute difference is $|\hat{p} - \tilde{p}| = 0.0002$.

In Table 4, the values of $\hat{\sigma}$ and $\tilde{\sigma}$ differ noticeably, even for large categories. Thus, data perturbation has a stronger effect on variance estimates. Theoretically, the variance of an estimator based on perturbed data is larger than that of a corresponding estimator based on the original data. We leave variance estimation from perturbed data as a future research project, which may also require a close examination of replicate weights and the stability of (5.1).

Table 5 reports analogous results for the categories of marital status. There, all relative differences, $|f - \tilde{f}|/f$, are quite small and the estimated standard deviations, $\hat{\sigma}$ and $\tilde{\sigma}$, are also fairly close. We believe that a primary reason for this feature is the absence of very small categories. We also compared estimated age distributions based on the original and perturbed data sets. The plots of the two distributions were nearly identical and hence are not shown here.

Table 5: Estimates of marital status frequencies

Marital Status	f	$\tilde{f}$	$\frac{\lvert f-\tilde{f}\rvert}{f}$	$\hat{\sigma}$	$\tilde{\sigma}$
Married	2,250,297	2,256,196	0.0026	7396	6796
Widowed	270,321	266,229	0.0151	2127	2232
Divorced	480,035	472,311	0.0161	4368	4147
Separated	114,783	120,778	0.0522	1842	2304
Never married	2,813,378	2,813,300	0.0000	6237	5605

We also examined effects of our procedure on estimates of joint distributions. For direct comparison, we explored the 14 joint distributions that were examined by NZY. Those correspond to certain sets of variables chosen from the five key variables (sex, age, race, marital status (mar) and puma) and two non-key variables: class of workers (work) and education level (edu), which have 9 and 8 categories, respectively. We measure the effects of data perturbation using the total variation distance (TVD). For a given set of variables, let p_i and r_i denote the estimates of the relative frequency of the i-th cell based on the original and a perturbed data set, respectively. The TVD is a global measure of the difference between the two estimated distributions and it is given by

$$TVD = \frac{1}{2}\sum_i |p_i - r_i|,$$

where the sum is over all cells.

Table 6: Total variation distance showing effects of data perturbation on estimates of joint distributions

Variables	NZY	Our	Cells	Variables	NZY	Our	Cells
race, mar	0.0138	0.0043	45	puma, work	0.0433	0.0376	396
race, puma	0.0296	0.0157	396	puma, edu	0.0412	0.0355	352
race, edu	0.0128	0.0086	72	sex, race, mar	0.0152	0.0072	90
race, work	0.0106	0.0037	81	sex, race, edu	0.0134	0.0095	144
mar, edu	0.0153	0.0122	40	mar, race, edu	0.0295	0.0220	360
mar, work	0.0162	0.0095	45	sex, race, work	0.0091	0.0043	162

The first and fifth columns of Table 6 state the variable considered. The TVDs between the estimated distributions (i.e., weighted relative frequencies) based on the original and NZY perturbed data are given under 'NZY' columns. The corresponding values for our procedure, i.e., the TVDs between estimated distributions based on the original and our perturbed data, are reported under 'Our' columns. The 'Cells' columns show the number of cells in the cross-classification of the variables considered. In all cases, our TVD values are quite small and they are smaller, often substantially, than corresponding NZY values. As expected, TVD increases when a new variable is added to a variable set. For example, for both procedures, the TVD for {race, mar} is smaller than the TVDs for {sex, race, mar} and {mar, race, edu}. Table 6 also shows that TVD tends to increase with the number of cells.

5.3. Effects on a Logistic Regression

The released data are often used for modeling and prediction. To get a sense of effects of data perturbation on such analysis, we examined a logistic regression. Specifically, we modeled employment status (Y, with $Y = 1$ if employed and 0 otherwise) using gender (X_1), age (X_2) and education (W) as covariates. We let $X_1 = 1$ if a unit is female and 0 otherwise. Education is an ordinal variable with 8 categories: grade 6 or less, grade 7-12 but no high school diploma, high school diploma, some college but not degree, Associate degree, Bachelor's degree, Master's or professional degree, and Doctorate degree. We treated it as a discrete variable, assigning $W = 0$ for grade 6 or less through $W = 7$ for Doctorate degree. For model fitting we used data from all units with $X_2 \geq 18$, i.e., from all work eligible persons. Recall that due to our data partitioning, a perturbed age falls between 0 and 17 if and only if the original age is in that interval. Thus, work eligible people in the original and perturbed data are the same.

Let $\vec{X} = (X_1, X_2, W)$ and $\pi(\vec{X}) = P(Y = 1|\vec{X})$ denote the conditional probability of a person being employed given X. We assume the logistic regression model

$$\log\Big(\frac{\pi(\vec{X})}{1 - \pi(\vec{X})}\Big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 W.$$

Let $\hat{\beta}_i$ and $\tilde{\beta}_i$ denote the estimated coefficient of X_i from original and perturbed data, respectively. By applying the 'glm' function in R with weights, we obtained:

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-1.946, -0.3793, -2.025 \times 10^{-3}, 0.8276),$$
$$(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3) = (-1.961, -0.3822, -0.924 \times 10^{-3}, 0.8197).$$

Clearly, the estimates obtained from the original and perturbed data are very close. Logistic regression parameters are interpreted conveniently in terms of odds and odds ratio (see e.g., Agresti, 2011). In our example, for given $\vec{X}$, the odds of 'employed' is

$\pi(\vec{X})/(1-\pi(\vec{X}))$. One interpretation of β_i, is that e^{β_i} is the proportional increment of odds of 'employed' as the variable associated with β_i increases by one unit while all other covariates remain unchanged. Table 7 gives the odds ratio estimates from the original and perturbed data, respectively. Not surprisingly, our data perturbation affects odds ratio estimates negligibly. Note that gender and education affect employment status substantially, as the corresponding odds ratios are markedly different from 1. Recall that gender was coded as $X_1 = 1$ if female and 0 if male. So, the odds of employment for females is about 68% of the odds for males.

Table 7: Estimated odds ratios

	Intercept	Gender	Age	Education
Original	0.1429	0.6844	0.9980	2.2877
Perturbed	0.1407	0.6823	0.9990	2.2700

6. Discussion

The NZY paper developed a novel approach to measuring and controlling identification risk in releasing microdata. In this paper, we demonstrated that while that procedure preserves raw cell frequencies fairly well, it may distort survey weighted frequencies considerably, which are commonly used to estimate population frequencies. This is mainly due to the fact that the NZY procedure is unbiased with respect to raw frequencies, but not weighted frequencies. On the other hand, as we showed, an unbiased post-randomization with respect to weighted frequencies may not be adequate for limiting identification risks. We presented a post-randomization method that limits identification risk rigorously and better preserves survey weighted statistical estimates. We illustrated the procedure with application to a real data set.

As we discussed earlier, the proposed procedure may be overly conservative as it treats many doubleton units as singletons. So, one may reduce the perturbation probabilities suitably to enhance data utility while meeting the disclosure control goal. For example, in the application in Section 5, Table 3 shows that the largest risk (0.2168) is much smaller than $\xi = .395$, the specified upper bound. Thus, the choice of $\theta_0 = 0.8$ is overly conservative. A natural approach to calculating a suitably smaller value is to choose θ such that the identification risk in the worst case does not exceed ξ. In our application, the worst case is $\tau = 1, \tau* = 1$, for which NZY proved that identification risks are bounded above by $h_*(\theta) = (1-\theta)/(1-\theta+\theta^2)$. So, a practical value of θ may be obtained by solving $h_*(\theta) = \xi$. For $\xi = .395$, this yields $\theta = 0.69$.

We should mention that De Waal and Willenborg (1997) discussed a scenario for target matching using survey weights, which postulates that each unit is uniquely indexed by survey weight and the intruder is able to reconstruct that mapping. However, as Fienberg (2010) noted, that scenario is very unlikely due to the complexity of weight calculations in most practical surveys. Also, in stratified sampling (where survey weights are easily calculated), we suggest to use the NZY procedure within each stratum and then combine stratum level estimates appropriately. Thus, we believe that keeping the survey weights unchanged should not cause much disclosure concern.

Acknowledgment. We thank Dr. Martin Klein and Dr. Eric Slud for giving us some helpful comments on an earlier draft.

References

[1] Agresti, A. (2013). *Categorical Data Analysis*, third edition, Wiley.

[2] Ash, S. (2014). Using successive difference replication for estimating variances, *Survey Methodology*, 40(1), 47-59.

[3] Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association*, 85(409), 38-45.

[4] Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, 14(1), 79-95.

[5] Cox, L.H., Karr, A.F., and Kinney, S.K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. *International Statistical Review*, 79(2), 160-183.

[6] Cruyff, M. J., Van Den Hout, A., and Van Der Heijden, P. G. (2008). The analysis of randomized response sum score variables, *Journal of the Royal Statistical Society, Ser. B*, 70(1), 21-30.

[7] De Waal, A. G. and Willenborg, L. C. R. J. (1997). Statistical disclosure control and sampling weights, *Journal of Official Statistics*, 13(4), 417.

[8] Duncan, G.T., Elliot, E. and Juan Jose Salazar, G. (2011). *Statistical Confidentiality: Principles and Practice*, New York: Springer.

[9] Duncan, G T., and Stokes, S. L. (2004). Disclosure risk vs. data utility: The RU confidentiality map as applied to topcoding, *Chance*, 17(3), 16-20.

[10] Fay, R.E. (1984). Some Properties of Estimates of Variance Based on Replication Methods, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 495-500.

[11] Fay, R.E., and Train, G.F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties, *Proceedings of the Section on Government Statistics, American Statistical Association*, 154-159.

[12] Fienberg, S. E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses, *Journal of Privacy and Confidentiality*, 1(2), 4.

[13] Gouweleeuw, J. M., Kooiman, P., and de Wolf, P. P. (1998). Post randomisation for statistical disclosure control: Theory and implementation, *Journal of Official Statistics*, 14(4), 463-478.

[14] Greenberg, B. V., and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files, *Statistica Neerlandica*, 46(1), 33-48.

[15] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P-P. (2012). *Statistical Disclosure Control*, New York: Wiley.

[16] Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician*, 60(3), 224-232.

[17] Nayak, T. K. and Adeshiyan, S. A. (2016). On invariant post-randomization for statistical disclosure control, *International Statistical Review*, 84, 26-42.

[18] Nayak, T. K., Zhang, C. and You, J. (2018). Measuring Identification Risk in Microdata Release and Its Control by Post-randomisation, *International Statistical Review*, 86(2), 300-321.

[19] Shlomo, N., and De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure, *Journal of Official Statistics*, 24(2), 1-26.

[20] Shlomo, N., and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata, *The Annals of Applied Statistics*, 4(3), 1291-1310.

[21] Skinner, C. J., and Elliot, M. J. (2002). A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society, Ser. B*, 64(4), 855-867.

[22] U.S. Census Bureau (2006). Technical Paper 66, "Design and Methodology: Current Population Survey", October 2006.

[23] Van den Hout, A., and Elamir, E. A. (2006). Statistical disclosure control using post randomisation: Variants and measures for disclosure risk, *Journal of Official Statistics*, 22(4), 711-731.

[24] Van den Hout, A. and Van der Heijden, P. (2002). Randomized response, statistical disclosure control and misclassification: a review, *International Statistical Review* 70(2), 269-288.

[25] Willenborg, L.C.R.J. and De Waal, T. (2001). *Elements of Statistical Disclosure Control*, New York: Springer.

[26] Wolter, K. (2007). *Introduction to variance estimation*, Springer Science & Business Media, Chicago.