NISS

Statistical Disclosure Limitation in the Presence of Edit Rules

Hang J. Kim, Alan F. Karr, Jerome P. Reiter

Technical Report 184
October 2013

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

Statistical Disclosure Limitation in the Presence of Edit Rules

Hang J. Kim

Duke University and National Institute of Statistical Sciences, Durham, NC 27708 (*hangkim@niss.org*)

Alan F. Karr

National Institute of Statistical Sciences, Research Triangle Park, NC 27709 (*karr@niss.org*)

Jerome P. Reiter

Department of Statistical Science, Duke University, Durham, NC 27708 (*jerry@stat.duke.edu*)

ABSTRACT

We articulate and investigate issues associated with performing statistical disclosure limitation (SDL) for data subject to edit rules. The central problem is that many SDL methods generate data records that violate the constraints. We propose and study two approaches. In the first, existing SDL methods are applied, and any constraint-violating values they produce are replaced by means of a constraint-preserving imputation procedure. In the second, the SDL methods are modified to prevent them from generating violations. We present a simulation study, based on data from the Colombian Annual Manufacturing Survey, that evaluates several SDL methods from the existing literature. The results suggest that (i) in practice, some SDL methods cannot be implemented with the second approach, and (ii) differences in risk-utility profiles across SDL approaches dwarf differences across the two approaches. Among the SDL strategies, microaggreggation followed by adding noise and partially synthetic data offer the most attractive risk-utility profiles.

KEY WORDS: Confidentiality, Imputation, Survey, Synthetic data

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant SES–1131897 to Duke University and the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1. INTRODUCTION

Public use microdata offer many benefits, for example, enabling researchers and policy-makers to perform in depth statistical analyses, students to learn skills of data analysis, and citizens to understand their society. However, public use microdata also carry disclosure risks: *intruders* who intend to misuse the information may be able to identify respondents or learn values of sensitive attributes from the public data. Statistical agencies recognize this risk and typically alter the microdata prior to release using one or more statistical disclosure limitation (SDL) techniques. Ideally, the SDL reduces disclosure risk to an acceptable level with low impact on data utility (Willenborg and De Waal 2001).

As collected, microdata often include implausible or impossible values, for example arising from multiple forms of survey error (Groves 1989), such as reporting and measurement error. Agencies prefer not to release such faulty values and so undertake a process usually referred to as "edit and imputation" (De Waal et al. 2011). Typically agencies define faulty values via pre-specified constraints, called *edit rules* or simply *edits*. Examples of the edit rules are *range restrictions* ($V_1 \leq a$), *ratio constraints* ($V_1 \leq bV_2$), *balance constraints* ($V_1 + V_2 = V_3$), and especially for categorical variables *consistency constraints* ($V_1 = a$ is not compatible with $V_2 = b$). Edit violations can be removed by recontacting respondents, manual editing procedures or, as in Fellegi and Holt (1976) and Kim et al. (2013), by imputing constraint-satisfying replacements for violated values in the same way that imputation is performed for missing values.

To date, disclosure review and limitation have been largely disconnected from edit and imputation. Typically editing is performed by one organizational unit, which then transfers the data to another unit that performs SDL. Interaction between the editing and SDL processes is minimal, and sometimes entirely absent. Indeed, those performing the SDL may not even be aware of constraints that the edited data must respect. This paper begins to bridge the gap, by considering what agencies should do when SDL creates altered data records that do not satisfy edit rules.

Taking as given the premise that released data must satisfy the edit constraints, the issue of eliminating edit violations created by the SDL process itself is addressed. As we show, such violations do occur for commonly applied SDL methods, so that the problem is not vacuous. We consider two broad classes of approaches. The first is to apply existing SDL methods and then remove the resulting edit violations, for example using the blank-and-impute methodology in Kim et al. (2013). This assumes that edit violations engendered by SDL can be treated in the same way as those resulting from measurement error. The second approach is to modify SDL methods so that they do not produce edit violations.

To illustrate and evaluate the two approaches, we use a simulation study based on numerical data from the 1991 Colombian Annual Manufacturing Survey. We introduce linear constraints typical of those used to edit business survey data (Winkler and Draper 1996; Thompson et al. 2001; Hedlin 2003). The results of the simulation suggest that, when both are feasible, there is little difference in the risk-utility profiles of SDL-then-edit (first approach) and edit-preserving SDL (second approach) procedures. Indeed, the differences in the profiles across approaches are swamped by differences across SDL methods. We also discuss the relative merits of the SDL techniques, although we view the evidence from the simulations as more suggestive than complete.

There are simpler approaches than the two we propose to investigate. For instance, one could simply delete the post-SDL records with edit violations. This approach has significant shortcomings. Deleting records leads to inefficiencies and even can introduce bias. It is especially problematic when data have survey weights, because there is no clear path to adjusting the weights of the remaining records. Thus, we do not consider this approach further here.

The remainder of the article is organized as follows. In Section 2, several SDL methods and corresponding approaches to generate masked values satisfying edits are described. Section 3 presents results of the simulation study based on the Colombian Manufacturing data which compares the suggested methods under a risk-utility framework. Section 4 concludes with a discussion of future research questions.

2. SDL METHODS IN THE PRESENCE OF EDIT RULES

As in Reiter (2005), let y_{il} be the collected value of variable l for unit i, for $l = 0, \ldots, p$ and $i \in D$, where D denotes the collected data for the n sampled units. Let y_{i0} be the unique unit identifier, which, of course, must be excluded from the final released data. For each $i \in D$, let $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{ip}\}$ be partitioned as $(\boldsymbol{y}_i^A, \boldsymbol{y}_i^U)$, where $\boldsymbol{y}_i^A$ is a vector of variables available to intruders in external data files, and $\boldsymbol{y}_i^U$ is a vector of variables unavailable to intruders except in the released data file, D^{rel}. To prevent disclosure, the agency uses SDL to alter the values of $\boldsymbol{y}_i^A$ before releasing D^{rel}. Let $\tilde{\boldsymbol{y}}_i^A$ denote the masked values of $\boldsymbol{y}_i^A$, so that D^{rel} after SDL comprises $\tilde{\boldsymbol{y}}_i = (\tilde{\boldsymbol{y}}_i^A, \boldsymbol{y}_i^U)$ for all n records on the file. For simplicity, we assume that the intruder knows $\boldsymbol{y}_i^A$ without any measurement error.

2.1 Summary of Selected SDL Methods

In this section, we look at a set of perturbative SDL methods for microdata, including adding noise, rank swapping and microaggregation, and then introduce partially synthetic data. These methods are described briefly, without considering editing.

Rank swapping (Moore 1996) is a special form of data swapping under which some attribute values are switched between pairs of similar records. Rank swapping is implemented as follows. For each variable l in $\boldsymbol{y}_i^A$, we sort $\{y_{1l}, \ldots, y_{nl}\}$ by its magnitude; let $\{y_{(1)l}, \ldots, y_{(n)l}\}$ denote the ordered values. Let $0 < \zeta < 100$ be a pre-specified parameter. Two cases $y_{(i)l}$ and $y_{(j)l}$ are randomly selected, and then swapped only if $|i - j| < n\zeta/100$. As ζ increases, the intensity of data protection increases but, in general, the data utility decreases.

Adding noise (Kim 1986; Sullivan and Fuller 1990; Tendick 1991) introduces random errors to collected values deemed at high risk of disclosure; for example, set $\tilde{\boldsymbol{y}}_i^A = \boldsymbol{y}_i^A + \boldsymbol{\varepsilon}_i$. A straightforward implementation is to draw random noise from a normal distribution, $\boldsymbol{\varepsilon}_i \sim N(0, c\Sigma^A)$, where Σ^A is the sample covariance of $\{\boldsymbol{y}_1^A, \ldots, \boldsymbol{y}_n^A\}$. The agency sets the parameter c to control the intensity of perturbation.

Microaggregation replaces original values with group averages. Using some clustering algorithm (Fayyoumi and Oommen 2010), the original records $\boldsymbol{y}_i$ are partitioned into groups $\mathcal{G}_j$, each with a fixed size. For each $i \in \mathcal{G}_j$, we replace $\boldsymbol{y}_i^A$ with the group mean $\tilde{\boldsymbol{y}}_{mic,j}^A = \sum_{k \in \mathcal{G}_j} \boldsymbol{y}_k^A / |\mathcal{G}_j|$, where $|\mathcal{G}_j|$ is the cardinality of $\mathcal{G}_j$. Larger cluster sizes results in greater data perturbation.

Microaggregation with adding noise (Oganian and Karr 2006) blends the clustering and perturbative effects of the two previous techniques. We set $\tilde{\boldsymbol{y}}_i^A = \tilde{\boldsymbol{y}}_{mic,i}^A + \boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^*)$. Oganian and Karr (2006) suggest using $\Sigma^* = \Sigma^A - \tilde{\Sigma}_{mic}^A$ (if this matrix is positive definite, and otherwise a positive definite approximation to it), where $\tilde{\Sigma}_{mic}^A$ denotes the sample covariance of the data masked by microaggregation, $\{\tilde{\boldsymbol{y}}_{mic,1}^A, \ldots, \tilde{\boldsymbol{y}}_{mic,n}^A\}$.

Partially synthetic data (Rubin 1993; Little 1993; Reiter 2003) comprise the original n records with sensitive values replaced by multiple imputations. The imputations are generated from models estimated from the original data. The multiple copies enable data analyses to reflect imputation uncertainty appropriately.

2.2 Approaches to SDL in the Presence of Edit Rules

As noted in Section 2, we consider two approaches: allow the SDL process to generate edit violations, but repair them subsequently; and prevent the SDL process from generating violations.

2.2.1 Approach I: Editing After SDL

In this approach, an agency first applies an SDL method to the collected data. Any post-SDL records that violate the constraints are deleted or "repaired" *ex post facto*. The agency treats any SDL-generated edit violations as if they were faulty values. This involves an error localization step, e.g., via the methods of Fellegi and Holt (1976), followed by replacing the localized errors with imputations from some methods that respect constraints. For example, one could use sequential regression imputation (Van Buuren and Oudshoorn 1999; Raghunathan et al. 2001), imputation from joint distributions (Geweke 1991; Tempelman 2007; Coutinho et al. 2011; Kim et al. 2013), or in some settings hot-deck imputation (Bankier et al. 1994; Coutinho and De Waal 2012).

We use the multivariate imputation method proposed by Kim et al. (2013), which is based on mixtures of multivariate normal distributions. This method guarantees that corrected records always lie in the feasible region, i.e., the restricted support of $\boldsymbol{y}_i$ that satisfies all edit rules, while being flexible enough to describe complex distributional features. Let $\mathcal{Y}$ represent the feasible region. Using $K > 1$ mixture components—see Kim et al. (2013) for discussion of setting K—we assume that

$$f(\boldsymbol{y}_i|\Theta_1,\ldots,\Theta_K) \propto \sum_{k=1}^{K} w_k \mathrm{N}(\boldsymbol{y}_i;\boldsymbol{\mu}_k,\Omega_k) I(\boldsymbol{y}_i \in \mathcal{Y}). \tag{1}$$

Here, for each of the $k = 1,\ldots,K$ mixture components, w_k is the probability (or weight) of the component, (μ_k, Ω_k) is the component mean vector and covariance matrix, and $\Theta_k = (w_k, \boldsymbol{\mu}_k, \Omega_k)$. After performing SDL, we identify each record with $\tilde{\boldsymbol{y}}_i \notin \mathcal{Y}$, blank its $\tilde{\boldsymbol{y}}_i^A$, and replace $\tilde{\boldsymbol{y}}_i^A$ with values generated from the posterior predictive distribution, $f(\boldsymbol{y}_i^A|D)$. We refer readers to the Appendix for the specifications of the prior distributions and details of Markov chain Monte Carlo (MCMC) steps.

2.2.2 Approach II: Edit Preserving SDL

One way to preserve constraints during SDL that involves randomization is to draw candidate masked values repeatedly until they satisfy all edit rules. This rejection sampling approach can be readily applied for SDL methods based on randomization, particularly when edit rules are based on sets of linear inequalities. For example, an agency that adds noise (or uses microaggregation with noise) to variables can generate $\boldsymbol{\varepsilon}_i$ (or $\boldsymbol{\delta}_i$) repeatedly until the drawn $\tilde{\boldsymbol{y}}_i^A$ satisfies the edit rules. An agency that uses partially synthetic data can generate replacements so that all draws are guaranteed to satisfy the constraints, for example using the imputation engine of Kim et al. (2013) as a synthesizer.

For SDL methods not entailing randomization, rejection sampling is difficult to implement. Rejection sampling is not possible for typical implementations of microaggregation, since no randomization is involved in microaggregation, except possibly in clustering heuristics.

Table 1: Description of variables in the 1991 Colombian Annual Manufacturing Survey with data-derived range restrictions.

Variable	Label	Range restriction
Skilled labor	SL	0.9–400
Unskilled labor	UL	0.9–1,000
Wages paid to skill labor	SW	300–3,000,000
Wages paid to unskilled labor	UW	600–4,000,000
Real value added	VA	50–1,000,000
Real material used in products	MU	10–1,000,000
Capital	CP	5–1,000,000

Table 2: Data-derived ratio edits ($V_1/V_2 \leq b$) for the 1991 Colombian Manufacturing Survey.

	V_2						
V_1	SL	UL	SW	UW	VA	MU	CP
SL	1	20	0.01	0.01	0.1	0.3	2
UL	50	1	0.1	0.005	0.3	5	5
SW	20000	100000	1	50	300	500	1000
UW	66666.7	10000	100	1	200	5000	5000
VA	10000	20000	10	10	1	200	700
MU	50000	100000	33.3	100	100	1	1000
CP	20000	10000	10	16.7	100	100	1

3. SIMULATION STUDY

We use a subset of 6521 establishments from the 1991 Colombian Annual Manufacturing Survey data comprising seven numerical variables: number of skilled employees (SL), number of unskilled employees (UL), wages for skilled employees (SW), wages for unskilled employees (UW), value added (VA), material used in products (MU), and capital (CP). We assume that these records are error-free. Edit rules are introduced including the range restrictions in Table 1 and ratio constraints in Table 2. The introduced constraints are data-derived and hypothetical; they are not actual constraints derived from the domain knowledge of economic experts.

We mask three of the seven variables—number of skilled employees, number of unskilled employees, and capital—and leave the remaining variables unaltered. To facilitate SDL and inference, we work with the natural logarithms of all variables. To avoid new notation, we let $\tilde{\boldsymbol{y}}_i$

Table 3: Numbers of records that violate edit rules across the 20 replications after implementing perturbative SDL methods.

Methods	Mean (%)	SD
`Noise`	157.8 (2.45)	10.1
`Swap`	134.2 (2.09)	6.6
`Mic`	5.0 (0.08)	–
`MicN`	84.1 (1.31)	6.7

represent the vector of natural logarithms of the seven variables. Thus, $\boldsymbol{y}_i^A$ comprises the three log-transformed values $(y_{i\texttt{SL}}, y_{i\texttt{UL}}, y_{i\texttt{CP}})$.

Five SDL methods are implemented. Four of these are "classical:" adding noise with $c = 0.16$ (`Noise`), rank swapping with $\zeta = 10$ (`Swap`), microaggregation with $|\mathcal{G}_j| = 3$ based on principal components clustering (`Mic`), and microaggregation with adding noise (`MicN`). Partially synthetic data (`Synt`) is generated by replacing all of $\boldsymbol{y}_i^A$ with draws from the model of Kim et al. (2013). For `Synt`, we use only only a single draw of the parameters from a converged Markov chain to generate one realization of D^{rel}; in practice, we recommend using multiple draws and releasing multiple data sets to enable variance estimation.

As shown in Table 3 and Figure 1, applying `Noise`, `Swap`, `Mic`, or `MicN` results in edit violations. `Noise` pushes many $\boldsymbol{y}_i$ outside the boundary of $\mathcal{Y}$, resulting in the largest number of edit violations. `Swap` also produces many edit violations, even with a fairly tight swapping range ($\zeta = 10$). `Mic` results in the fewest number of masked records that violate the constraints. If we had applied microaggregation to all of $\boldsymbol{y}_i$, the resulting records always would be inside $\mathcal{Y}$ due to its convexity. Since we replace only each $\boldsymbol{y}_i^A$, we guarantee that $\tilde{\boldsymbol{y}}_i^A$ is in the appropriate subset of the feasible region, but not that $\tilde{\boldsymbol{y}}_i \in \mathcal{Y}$.

Edit-preserving SDL was implemented with `Noise` and `MicN` via a rejection sampling scheme. We attempted to use a rejection sampling scheme for `Swap`; however, in 1000 generations of possible D^{rel} contained edit violations. Each D^{rel} had at least 99 out of 6,521 records that violated the constraints, suggesting that waiting for a constraint-preserving, rank-swapped data set in this simulation design is hopeless. Therefore, we do not include `Swap` in evaluations of the

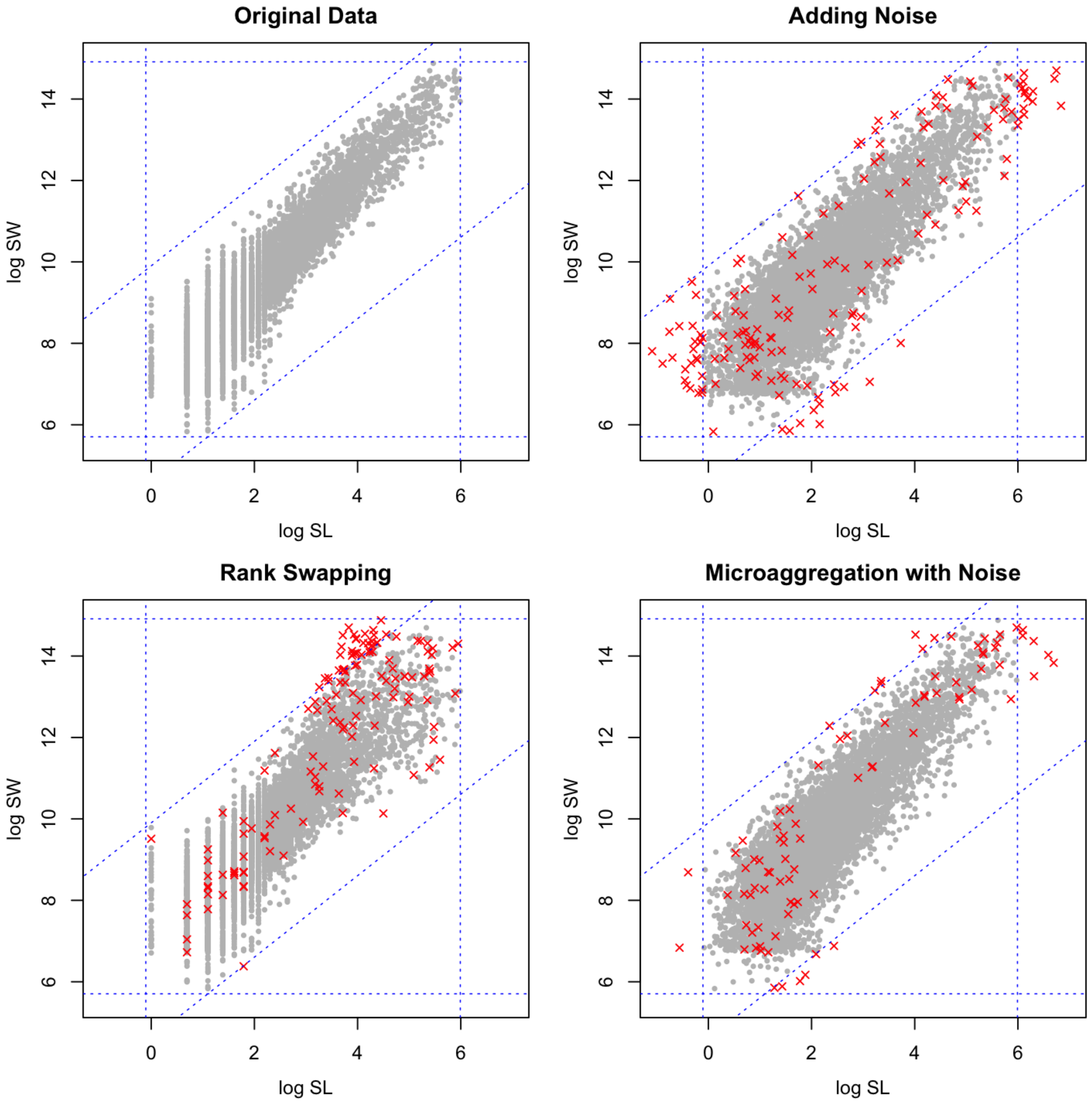

Figure 1: Illustrative example of SDL methods with linear constraints. Top-left panel shows pre-SDL data in terms of variables log `SL` and log `SW`. Three variables, `SL`, `UL`, and `CP` are masked by adding noise (`Noise`, top-right panel), rank swapping (`Swap`, bottom-left panel), and microaggregation with adding noise (`MicN`, bottom-right panel). Solid circles indicate records that satisfy edit rules and "×" indicate the violated records, i.e., $\tilde{\boldsymbol{y}}_i \notin \mathcal{Y}$.

edit-preserving SDL approach. (We have not considered sub-sampling in this paper, but sub-sampling in such a manner that no chosen record violates the edit constraints seems likely to have the same hopelessly unlikely problem as data swapping. If the measurement errors that cause edit violations are correlated with true data values, rejection-based sub-sampling would introduce bias.)

As a measure of disclosure risk, we use the *percentage of linked* criterion of Domingo-Ferrer et al. (2001). First, we compute the distances

$$d_{i,j} = \sqrt{\sum_{l} (y_{il}^A - \tilde{y}_{jl}^A)^2}, \qquad \forall\ i, j = 1, \ldots, n,$$

where $l \in$(`SL`, `UL`, `CP`). For each i, we find the record j that achieves the minimum value of $d_{i,j}$. When $y_{i0} = y_{j0}$, i.e., the record in D^{rel} can be linked correctly to D based on matching the available variables, we let $t_i = 1$ and otherwise let $t_i = 0$. The risk measure is $PL = \sum_{i=1}^{n} t_i/n$.

We use two measures of data utility: an approximate Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) of D^{rel} from D, and the propensity score (U$_{\text{prop}}$) utility measure suggested by Woo et al. (2009). For KL, we use a closed-form expression based on a normality assumption,

$$KL = \frac{1}{2}\left[\text{tr}\left\{(\Sigma^{rel})^{-1}\Sigma\right\} + \left(\bar{\boldsymbol{y}}^{rel} - \bar{\boldsymbol{y}}\right)^T (\Sigma^{rel})^{-1}\left(\bar{\boldsymbol{y}}^{rel} - \bar{\boldsymbol{y}}\right) - p - \ \log\left(\frac{|\Sigma^{rel}|}{|\Sigma|}\right)\right], \qquad (2)$$

where $\bar{\boldsymbol{y}}$ and Σ are the sample mean and the sample covariance of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ in D, and $\bar{\boldsymbol{y}}^{rel}$ and Σ^{rel} are the corresponding statistics of $\{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n\}$ in D^{rel}. For U$_{\text{prop}}$, we first concatenate D^{rel} and D, and add an indicator variable whose values equal one for all records in D^{rel} and equal zero for all records in D. Using the concatenated data, we estimate the logistic regression of the indicator variable on all seven variables (after log transformations), including main effects and all interactions up to third order; that is, we fit

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \sum_{a=1}^{7}\beta_a \log Y_{ia} + \sum_{a,b} \log Y_{ia} \log Y_{ib} \\ &\quad + \sum_{a,b,c}\beta_{abc} \log Y_{ia} \log Y_{ib} \log Y_{ic}.\end{aligned}$$

For $i = 1, \ldots, 2n$, we compute the set of predicted probabilities $\hat{p}_i$. The risk measure is

$$\mathrm{U}_{\mathrm{prop}} = \frac{1}{2n}\sum_{i=1}^{2n}\left(\hat{p}_i - \frac{1}{2}\right)^2.$$

Values of $\mathrm{U}_{\mathrm{prop}}$ near zero represent high data utility, since they imply we are not able to distinguish between D^{rel} and D.

For each method we generate 20 masked datasets, each from different random seeds. Note that all 20 datasets for `MIC` are identical, since this methods is deterministic. Table 4 displays the average values of KL, $\mathrm{U}_{\mathrm{prop}}$ and PL over the 20 replicates for each method. For methods that can be implemented with both approaches, namely `Noise` and `MicN`, the risk-utility profiles are very similar. This suggests that, for qualifying methods, the decision to deal with edits after or during SDL has little impact on disclosure risk and data quality. However, the risk-utility profiles are quite different across SDL techniques. In particular, according to both utility measures, `Synt` preserves the distribution of the original data most faithfully, with `Swap` a somewhat distant second place. In terms of disclosure risks, `MicN` has the smallest value of PL, although the values of PL are fairly low for all methods in this simulation.

Figure 2 displays a risk-utility (R-U) map (Duncan and Stokes 2004; Gomatam et al. 2005; Cox et al. 2011) of the results, using KL as the utility measure. The figure includes values for all 20 realizations of $D^{rel,m}$. Smaller values of PL and KL represent higher levels of data protection and data utility, so the risk-utility frontier consists of candidate releases with no other candidate to their "'southwest." The R-U frontier includes `MicN`, which among these methods has the minimum level of disclosure risk, and `Synt`, which among these methods has the maximum level of data utility and a low level of disclosure risk.

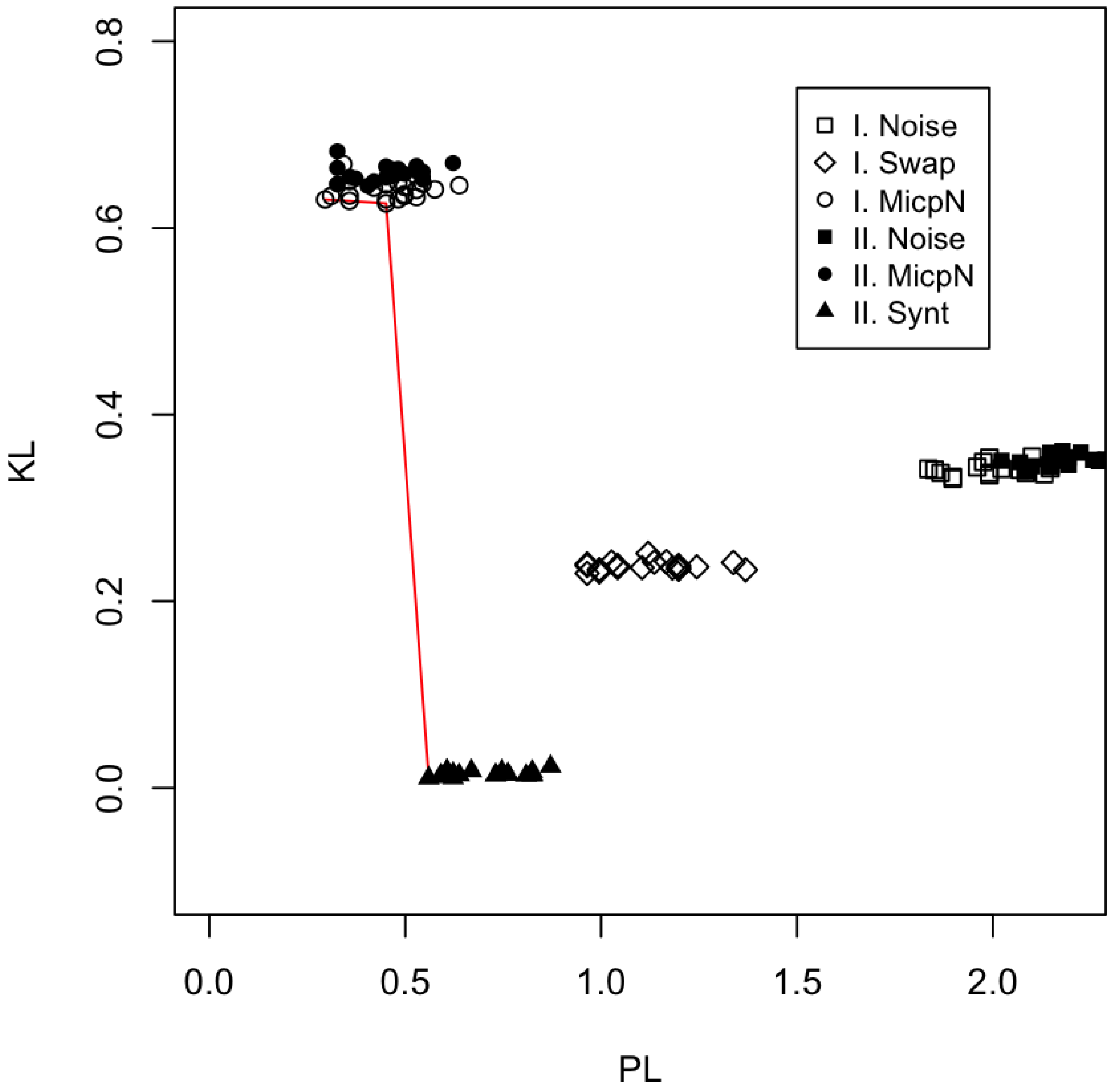

Figure 2: Risk-utility map with the SDL methods. The solid line indicates the risk-utility frontier. The open symbols represent Approach I and the solid symbols represent Approach II. The smaller values of PL and KL represent the higher levels of data protection and data utility.

Table 4: Measured data utility and disclosure risk. Entries include the averages of KL, $\mathrm{U_{prop}}$ and PL from 20 replications of each method.

	Approach	`Noise`	`Swap`	`Mic`	`MicN`	`Synt`
KL	I	.34	.24	1.34	.64	–
	II	.35	–	–	.66	.02
$\mathrm{U_{prop}}$	I	.0225	.0013	.0463	.0406	–
	II	.0225	–	–	.0425	.0007
PL	I	2.05	1.12	.78	.45	-
	II	2.26	–	–	.45	.70

4. CONCLUDING REMARKS

We have shown how it is possible to perform SDL for data subject to edit rules. Consistent with other studies, microaggregation followed by additive noise and partially synthetic data seem to be particularly effective strategies. The latter has the additional advantage that the synthesis methodology can be used to impute missing data values and implement edit-preserving SDL simultaneously, following the two-stage approach described in Reiter (2004).

An intriguing aspect of the editing–SDL "disconnect" is whether edited values should be protected in the same way as original reported data. This point, perhaps, is more subtle than it may seem initially. One interpretation is that a statistical agency promises to protect whatever information the subjects provide, even if that information is believed, or known to be, erroneous. Under this logic, edited and imputed values are not respondent information (i.e., they have been imputed rather than reported) and therefore might be treated differently during SDL. Another view is that the agency is also charged with protecting its best estimate of actual values, as opposed to reported values, which implies that edited and imputed values do require SDL. To our knowledge this issue remains unresolved, and, indeed, largely unaddressed. As noted in Section 1, we believe that in the long run, the most desirable approach is one that fully integrates editing, imputation and SDL.

Finally, we note two somewhat technical issues. First, some statistical agencies do not always include edit and imputation flags in released data. The risk and utility consequences of doing this are unexplored. The underlying issue is one of transparency (Karr 2009; Cox et al. 2011). Second, our research to date has not touched the role of weights, which was addressed to some extent in Cox et al. (2011). Weights themselves may pose disclosure risk (e.g., of unreleased values of design variables), but are generally ignored in all three of the editing, imputation and SDL processes. Some editing procedures, such as seeking additional information from "large" and low–weight respondents, consider weights implicitly. Some implementations of data swapping can accommodate weight constraints. For example, indexed microaggregation of Cox et al. (2011) is able to protect risky weights. However, by any measure, much more remains to be done than has been done.

APPENDIX: THE JOINT MULTIVARIATE IMPUTATION USING NORMAL MIXTURE

As described in Section 2, the joint multivariate imputation method developed in Kim et al. (2013) is used. The likelihood function in (1) can be re-expressed with latent variables z_i by

$$f(\boldsymbol{y}_i \mid z_i, \mu, \Omega) \propto \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \Omega_{z_i}) I(\boldsymbol{y}_i \in \mathcal{Y})$$

and

$$P(z_i = k) = w_k,\ k = 1, \ldots, K.$$

Following Lavine and West (1992), we assume the prior distributions,

$$\boldsymbol{\mu}_k \mid \Omega_k \sim \mathrm{N}(\boldsymbol{\mu}_0, h^{-1}\Omega_k), \quad \Omega_k \sim \mathrm{IW}(f, \Phi)$$

where $\Phi = diag(\phi_1, \ldots, \phi_p)$, and $\phi_j \sim \Gamma(a_\phi, b_\phi)$ for $j = 1, \ldots, p$. Here, IW denotes the inverse Wishart distribution and Γ denotes the Gamma distribution with mean a_ϕ/b_ϕ. For flexible modeling of the component weights, we adopt the stick-breaking representation of a truncated

Dirichlet process (Sethuraman 1994; Ishwaran and James 2001):

$$\begin{aligned} w_k &= v_k \prod_{g<k}(1-v_g) \;\; \text{for } k=1,\ldots,K \\ v_k &\sim \text{Beta}(1,\alpha) \;\; \text{for } k=1,\ldots,K-1; \;\; v_K = 1 \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha). \end{aligned}$$

In the simulation study, we follow Kim et al. (2013) and set $\boldsymbol{\mu}_0 = 0$, $h = 1$, $f = p+1$, $a_\phi = b_\phi = 0.25$, $a_\alpha = b_\alpha = 0.25$ and $K = 40$.

To facilitate the estimation of μ and Ω, we use a data augmentation technique developed by O'Malley and Zaslavsky (2008). The data augmentation adopts a larger, hypothetical sample $Y_N = \{Y_n, Y_{N-n}\}$ where Y_n is the set of $\boldsymbol{y}_i \in \mathcal{Y}$ following the likelihood in Equation (1) and Y_{N-n} consists of the values from outside of $\mathcal{Y}$, so that

$$p(Y_N \mid \Theta_1,\ldots,\Theta_K) \;=\; \prod_{i=1}^{N}\sum_{k=1}^{K} w_k \text{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \Omega_k).$$

where $\Theta_k = (\boldsymbol{\mu}_k, \Omega_k, w_k)$. Given the augmented sample Y_N, the parameters $\Theta_k = (w_k, \boldsymbol{\mu}_k, \Omega_k)$ can be sampled via Gibbs sampling. Setting $p(N) \propto 1/N$ as suggested by Meng and Zaslavsky (2002) and O'Malley and Zaslavsky (2008), the conditional density of the size of Y_{N-n} is distributed as

$$N - n \mid n, \Theta_1, \ldots, \Theta_K, \mathcal{Y} \;\sim\; \text{NegativeBinomial}\left(n, 1 - h_\Theta(\mathcal{Y})\right).$$

where

$$h_\Theta(\mathcal{Y}) = \int_{\{\boldsymbol{y}:\boldsymbol{y}\in\mathcal{Y}\}} \sum_{k=1}^{K} w_k \text{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \Omega_k) d\boldsymbol{y}.$$

The following MCMC steps after initialization are implemented.

1. For each k, draw $\Omega_k \sim \text{IW}(f_k, \Phi_k)$ and then draw $\boldsymbol{\mu}_k \sim \text{N}\left(\boldsymbol{\mu}_k^*, \Omega_k/h\right)$ where $\boldsymbol{\mu}_k^* = (N_k\bar{\boldsymbol{y}}_k + h\boldsymbol{\mu}_0)/(N_k + h)$, $f_k = f + N_k$, $\Phi_k = \Phi + S_k + (\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)'/(1/N_k + 1/h)$. We calculate the sample mean $\bar{\boldsymbol{y}}_k$ and the sample covariance S_k from the error-free, pre-SDL

values $Y_n = \{\boldsymbol{y}_i, i = 1, \ldots, n\}$ and the drawn auxiliary values Y_{N-n} by $\bar{\boldsymbol{y}}_k = \sum_{\{i:z_i=k\}} \boldsymbol{y}_i / N_k$ where $N_k = \sum_{i=1}^{N} I(z_i = k)$ and $S_k = \sum_{\{i:z_i=k\}} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)'$.

2. For each k, draw $v_k \sim \text{Beta}\left(1 + N_k, \alpha + \sum_{g>k} N_g\right)$. Set $v_K = 1$ and calculate $w_k = v_k \prod_{g<k}(1 - v_g)$.

3. For each $j = 1, \ldots, p$, draw $\phi_l \sim \Gamma\left(a_\phi + K(p+1)/2, b_\phi + \sum_{k=1}^{K} \Omega_{k(r,r)}^{-1}/2\right)$ where $\Omega_{k(r,r)}^{-1}$ is the r-th diagonal element of Ω_k^{-1}.

4. Draw α from $\Gamma\left(a_\alpha + K - 1, b_\alpha - \log w_K\right)$.

5. For each $i = 1, \ldots, n$, sample $z_i \sim \text{Categorical}(w_{i1}^*, \ldots, w_{iK}^*)$ where

   $w_{ik}^* = w_k \text{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \Omega_k) / \left[\sum_{g=1}^{K} w_g \text{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_g, \Omega_g)\right]$.

6. Sample (N, Z_{N-n}, Y_{N-n}) jointly from their full conditional distribution starting with $c_{in} = c_{out} = 0$.

   6.1. Draw $z^* \sim \text{Categorical}(w_1, \ldots, w_K)$.

   6.2. Draw $\boldsymbol{y}^* \sim \text{N}(\boldsymbol{\mu}_{z^*}, \Omega_{z^*})$.

   6.3. If $\boldsymbol{y}^* \in \mathcal{Y}$, set $c_{in} = c_{in} + 1$.

   If $\boldsymbol{y}^* \in \mathcal{Y}^c$, set $c_{out} = c_{out} + 1$, $\boldsymbol{y}_{n+c_{out}} = \boldsymbol{y}^*$, and $z_{n+c_{out}} = z^*$.

   6.4. Repeat 6.1 through 6.3 until $c_{in} = n$.

   6.5. Let $N = n + c_{out}$.

7. To correct post-SDL records with edit violations, a special type of Metropolis-Hastings, called the Hit-and-Run sampler (Chen and Schmeiser 1993), is adopted. In the initialization step, we propose any starting value $\tilde{\boldsymbol{y}}_i^{A(0)}$ such that $(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(0)}) \in \mathcal{Y}$ by using rejection sampling or extreme points approach (see Kim et al. 2013). Then, the following steps update $\tilde{\boldsymbol{y}}_i^{A(t)}$ which will replace $\tilde{\boldsymbol{y}}_i^A$ which violates edit rules.

   7.1. Draw a direction $\boldsymbol{d}^*$ uniformly from the surface of the $|\tilde{y}_i^A|$-dimensional unit sphere centered at the origin.

7.2. Draw a signed distance λ^* from the uniform distribution on Ξ,

$$\Xi = \left\{ \lambda : \left(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(t)} + \lambda \boldsymbol{d}^* \right) \in \mathcal{Y} \right\}$$

7.3. Accept or reject the proposal $\tilde{\boldsymbol{y}}_i^{A*} = \tilde{\boldsymbol{y}}_i^{A(t)} + \lambda^* \boldsymbol{d}^*$ with the acceptance probability ρ_i, where

$$\rho_i = \min \left[1, \frac{f(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A*} | \Theta_{z_i})}{f(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(t)} | \Theta_{z_i})} \right].$$

REFERENCES

Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1994). Imputing Numeric and Qualitative Variables Simultaneously. In American Statistical Association Proceedings of the Survey Research Method Section, pp. 242–247.

Chen, M.-H., and Schmeiser, B. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers. Journal of Computational and Graphical Statistics, 2, 251–272.

Coutinho, W., and De Waal, T. (2012). Hot Deck Imputation of Numerical Data Under Edit Restrictions. Discussion Paper 2012243, Statistics Netherlands.

Coutinho, W., De Waal, T., and Remmerswaal, M. (2011). Imputation of Numerical Data Under Linear Edit Restrictions. Statistics and Operations Research Transactions, 35, 29–62.

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, 160–183.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), Handbook of Statistical Data Editing and Imputation, Wiley.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001). Comparing SDC Methods for

Microdata on the Basis of Information Loss and Disclosure Risk. In Pre-proceedings of ENK-NTTS, pp. 807–826.

Duncan, G. T., and Stokes, S. L. (2004). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. Chance, 17, 16–20.

Fayyoumi, E., and Oommen, B. J. (2010). A Survey on Statistical Disclosure Control and Micro-Aggregation Techniques for Secure Statistical Databases. Software: Practice and Experience, 40, 1161–1188.

Fellegi, I. P., and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Geweke, J. (1991). Efficient Simulation from the Multivariate Normal and Student-T Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. In Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, pp. 571–578.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk–Utility Framework for Remote Access Analysis Servers. Statistical Science, 20, 163–177.

Groves, R. M. (1989), Survey Errors and Survey Costs, New York: Wiley.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics, 19, 177–199.

Ishwaran, H., and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. Journal of the American Statistical Association, 96, 161–173.

Karr, A. F. (2009). The Role of Transparency in Statistical Disclosure Limitation. Presented at the Joint Unece/Eurostat Work Session on Statistical Data Confidentiality. Bilbao, Spain, 2–4 December 2009. Available At *http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.41.e.pdf.*

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2013). Multiple Imputation of Missing or Faulty Values Under Linear Constraints. Technical Report 182, National Institute of Statistical Sciences.

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

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22, 79–86.

Lavine, M., and West, M. (1992). A Bayesian Method for Classification and Discrimination. Canadian Journal of Statistics, 20, 451–461.

Little, R. J. A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Meng, X.-L., and Zaslavsky, A. M. (2002). Single Observation Unbiased Priors. Annals of Statistics, 30, 1345–1375.

Moore, R. A. (1996). Controlled Data-Swapping Techniques for Masking Use Microdata Sets. Statistical Research Report 96/04, US Bureau of the Census, Statistical Research Division. Available at *http://www.census.gov/srd/www/byyear.html.*

Oganian, A., and Karr, A. F. (2006). Combinations of SDC Methods for Microdata Protection. In Privacy in Statistical Databases 2006, Lecture Notes in Computer Science, eds. J. Domingo-Ferrer and L. Franconi, Berlin: Springer, pp. 102–113.

O'Malley, A. J., and Zaslavsky, A. M. (2008). Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse. Journal of the American Statistical Association, 103, 1405–1418.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, 27, 85–95.

Reiter, J. P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29, 181–188.

——— (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30, 235–242.

——— (2005). Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association, 100, 1103–1112.

Rubin, D. B. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. Statistica Sinica, 4, 639–650.

Sullivan, G., and Fuller, W. A. (1990). The Use of Measurement Error to Avoid Disclosure. In American Statistical Association Proceedings of the Survey Research Method Section, pp. 802–807.

Tempelman, C. (2007). Imputation of Restricted Data. Ph. D. dissertation, University of Groningen.

Tendick, P. (1991). Optimal Noise Addition for Preserving Confidentiality in Multivariate Data. Journal of Statistical Planning and Inference, 27, 341–353.

Thompson, K. J., Sausman, K., Walkup, M., Dahl, S., King, C., and Adeshiyan, S. (2001). Developing Ratio Edits and Imputation Parameters for the Services Sector Censuses Plain Vanilla Ratio Edit Module Test. Economic Statistical Methods Report ESM-0101, US Bureau of the Census, Wahsington, DC.

Van Buuren, S., and Oudshoorn, K. (1999). Flexible Multivariate Imputation by MICE. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden, Netherlands.

Willenborg, L., and De Waal, T. (2001), Elements of Statistical Disclosure Control, New York: Springer–Verlag.

Winkler, W. E., and Draper, L. R. (1996). Application of the SPEER Edit System. Research Report RR96/02, Statistical Research Division, US Bureau of the Census, Washington, DC.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. Journal of Privacy and Confidentiality, 1, 111–124.