DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 121–138, http://dx.doi.org/10.1515/JOS-2015-0006

Statistical Disclosure Limitation in the Presence of Edit Rules

Hang J. Kim[1], *Alan F. Karr*[2], *and Jerome P. Reiter*[3]

We compare two general strategies for performing statistical disclosure limitation (SDL) for continuous microdata subject to edit rules. In the first, existing SDL methods are applied, and any constraint-violating values they produce are replaced using 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 the performance of the two strategies as applied to several SDL methods. The results suggest that differences in risk-utility profiles across SDL methods dwarf differences between the two general strategies. Among the SDL strategies, variants of microaggregation and partially synthetic data offer the most attractive risk-utility profiles.

Key words: Confidentiality; imputation; survey; synthetic data.

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 in 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; Hundepool et al. 2012).

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). Agencies identify faulty values via prespecified constraints, called *edit rules* or simply *edits*.

[1] Duke University and National Institute of Statistical Sciences, P.O. Box 90251, Durham, NC 27708, U.S.A. Email: hangkim0@gmail.com
[2] RTI International, 3040 East Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709, U.S.A. Email: karr@rti.org
[3] Department of Statistical Science, Duke University, P.O. Box 90251, Durham, NC 27708, U.S.A. Email: jerry@stat.duke.edu
Acknowledgments: This research was supported by the National Science Foundation (SES-11-31897). The authors thank the Editor, the Associate Editor, and the three anonymous referees for their insightful and constructive comments.

© Statistics Sweden

Examples of edit rules for continuous microdata, such as data from economic censuses or surveys, include *range restrictions* ($V_1 \leq a$), *ratio constraints* ($V_1 \leq bV_2$), and *balance constraints* ($V_1 + V_2 = V_3$). When a record fails a set of edits, agencies typically select some fields to replace with imputed values so that all constraints are satisfied (Fellegi and Holt 1976).

To date, assessment of disclosure risks and subsequent SDL have been largely disconnected from edit and imputation in practice. 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 is entirely absent. Indeed, those performing the SDL may not even be aware of constraints that the edited data must respect.

The extant literature offers two general strategies for integrating SDL and editing. The first approach is to apply existing SDL methods and then remove any resulting edit violations; this is illustrated in Shlomo and De Waal (2005; 2008). Essentially, edit violations engendered by SDL are treated in the same way as those resulting from measurement error. The second approach is to use an SDL method that does not produce edit violations; this is illustrated in Torra (2008). Many SDL methods as typically applied do not guarantee edit preservation; however, as we illustrate, some SDL methods can be modified to do so. To our knowledge, these two general strategies have not been compared in terms of impacts on data quality and disclosure risk.

In this article, we make such comparisons by implementing the strategies for several SDL procedures for continuous microdata. We apply the procedures to continuous microdata from the 1991 Colombian Annual Manufacturing Survey. The results of the simulation suggest that, when both strategies are feasible, there is little difference in the risk-utility profiles of edit-after-SDL (first approach) and edit-preserving SDL (second approach) procedures. Indeed, the differences in the profiles across approaches are swamped by differences among 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.

The remainder of the article is organized as follows. In Section 2, we describe several SDL methods and corresponding approaches to generate masked values satisfying edits. In Section 3, we present results of the simulation study and compare the suggested methods under a risk-utility framework. In Section 4, we conclude 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, if it is informative, must be excluded from the final released data. Suppose that $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{ip}\}$ satisfies all constraints or has been corrected to do so prior to SDL. For each $i \in D$, let $\boldsymbol{y}_i$ 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. In general, it is challenging for agencies to determine which variables comprise $\boldsymbol{y}_i^A$ and which comprise $\boldsymbol{y}_i^U$. When this distinction is unclear, arguably the agency should treat all variables as needing disclosure treatment.

2.1. Summary of Selected SDL Methods

In this section, we review the set of SDL methods for continuous microdata that we employ in our simulation, which includes rank swapping, adding noise, variants of microaggregation, and partially synthetic data. We describe each method briefly and refer readers to Hundepool et al. (2012) for further details. Of course, there are more variations on these methods, as well as additional SDL methods. We do not claim that these are a subset of best or most appropriate methods for the data at hand; however, they do serve to help us evaluate the two general strategies for SDL with 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 $0 < \tau_{\text{swap}} < 100$ be a prespecified parameter. Two cases $y_{(i)l}$ and $y_{(j)l}$ are randomly selected, and then swapped only if $|i - j| < n\tau_{\text{swap}}/100$. As τ_{swap} 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 selected 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, $\boldsymbol{\varepsilon}_i \sim N(0, \tau_{\text{noise}}\Sigma^A)$, where Σ^A is the sample covariance of $\{\boldsymbol{y}_1^A, \ldots, \boldsymbol{y}_n^A\}$. The agency sets the parameter τ_{noise} to control the intensity of perturbation. To increase data utility, Shlomo and de Waal (2008) suggest perturbing data within control strata, in which the agency (i) defines Q subgroups of records $\{D_q : q = 1, \ldots, Q\}$, for example, by grouping records into quintiles of some variable, (ii) generates random noise $\boldsymbol{\varepsilon}_i \sim N\big(\boldsymbol{\mu}_q\big(1 - \sqrt{1 - \tau_{\text{cnoise}}^2}\big)/\tau_{\text{cnoise}}, \Sigma_q\big)$ where $\boldsymbol{\mu}_q$ and Σ_q are the sample mean and the sample covariance of records $\{\boldsymbol{y}_j^A : j \in D_q\}$ and $0 < \tau_{\text{cnoise}} \leq 1$ is the parameter to control the amount of random noise, and (iii) replaces $\boldsymbol{y}_i^A$ with $\tilde{\boldsymbol{y}}_i^A = \sqrt{1 - \tau_{\text{cnoise}}^2}\boldsymbol{y}_i^A + \tau_{\text{cnoise}}\boldsymbol{\varepsilon}_i$. We refer to this variation as *controlled adding noise*.

Microaggregation (Defays and Nanopoulos 1993; Domingo-Ferrer and Mateo-Sanz 2002) replaces original values with group averages. Using a clustering algorithm, the original records $\boldsymbol{y}_i$ are partitioned into clusters $\mathcal{G}_g$, each with a fixed size. For each $i \in \mathcal{G}_g$, we replace $\boldsymbol{y}_i^A$ with the group mean $\tilde{\boldsymbol{y}}_{\text{mic},i}^A = \sum_{k \in \mathcal{G}_g} \boldsymbol{y}_k^A / \tau_{\text{mic}}$, where $\tau_{\text{mic}} = |\mathcal{G}_g|$, the cardinality of $\mathcal{G}_g$. Larger cluster sizes result in greater data perturbation. To construct clusters, one can project data onto a single dimension, for example, using the first principal component or the sum of z-scores (Fayyoumi and Oommen 2010). Alternatively, one can find the clusters using a heuristic based on Euclidean distances between records. For example, in *multivariate fixed-size microaggregation* (Domingo-Ferrer and

Mateo-Sanz 2002), the algorithm starts with finding the two records $\boldsymbol{y}_r$ and $\boldsymbol{y}_s$ farthest apart. The first cluster contains $\boldsymbol{y}_r$ and the $\tau_{\text{mmic}} - 1$ records closest to $\boldsymbol{y}_r$, and the second cluster contains $\boldsymbol{y}_s$ and the $\tau_{\text{mmic}} - 1$ records closest to $\boldsymbol{y}_s$. The third and fourth clusters are formed in a similar fashion starting from the two farthest-apart records among the remaining $n - 2\tau_{\text{mmic}}$ records. This repeats until fewer than $2\tau_{\text{mmic}}$ records do not belong to the clusters. These remaining records form a new cluster.

Oganian and Karr (2006) suggest *microaggregation with adding noise*, which blends the clustering and perturbative effects of the two previous techniques. We set $\tilde{\boldsymbol{y}}_i^A = \tilde{\boldsymbol{y}}_{\text{mic},i}^A + \boldsymbol{\delta}_i$, where $\tilde{\boldsymbol{y}}_{\text{mic},i}^A$ is masked by microaggregation and $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^*)$. Oganian and Karr (2006) suggest using $\Sigma^* = \Sigma^A - \tilde{\Sigma}_{\text{mic}}^A$ (if this matrix is positive definite, and otherwise a positive definite approximation to it), where $\tilde{\Sigma}_{\text{mic}}^A$ denotes the sample covariance of $\{\tilde{\boldsymbol{y}}_{\text{mic},1}^A, \ldots, \tilde{\boldsymbol{y}}_{\text{mic},n}^A\}$. A variant of the method is using controlled noise with microaggregation (Shlomo and De Waal 2008): (i) define five subgroups by quintiles D_q where $q = 1, \ldots, 5$, (ii) partition records $i \in D_q$ into cluster $\mathcal{G}_{q,g}$ with size of τ_{cmic}, (iii) replace $\boldsymbol{y}_i^A$ with the group mean $\tilde{\boldsymbol{y}}_{\text{cmic},i}^A = \sum_{k \in \mathcal{G}_{q,g}} \boldsymbol{y}_k^A / \tau_{\text{cmic}}$, and (iv) produce final masked records by adding random noise, $\tilde{\boldsymbol{y}}_{\text{cmicn},i}^A = \tilde{\boldsymbol{y}}_{\text{cmic},i}^A + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^*)$ and Σ^* is the difference between the sample variance of $\{\boldsymbol{y}_j^A : j \in D_q\}$ and the sample variance of $\{\tilde{\boldsymbol{y}}_{\text{cmic},j}^A : j \in D_q\}$. We refer to this method as *controlled microaggregation with adding noise*. We note that the original paper of Shlomo and de Waal (2008) presents microaggregation for data with balance constraints; our version does not use the balance constraints.

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. The additional data sets also offer more information for intruders to attempt identifications; see Reiter and Mitra (2009) and Drechsler and Reiter (2008) for further discussion of this issue.

2.2. *Approaches to SDL in the Presence of Edit Rules*

Both edit-after-SDL and edit-preserving SDL have potentially appealing features. Edit-after-SDL allows agencies to use existing SDL procedures and established edit-imputation procedures, including handling balance edits, without worrying about combining them. This may facilitate production operations when all edits are done in one step. On the other hand, edit-preserving SDL can reduce an agency's workload, since the masked data automatically satisfy the constraints. We now describe how one can implement these two strategies for the SDL methods outlined in Subsection 1. We note that, in some settings, it may be possible to use edit-preserving SDL for some constraints and edit-after-SDL for other constraints (e.g., Shlomo and De Waal 2008); we do not consider such mixed strategies here.

2.2.1. Approach I: Edit-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, for example, using the methods of Fellegi and Holt (1976), followed by replacing the localized errors with imputations 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. 2014b), or in some settings hot-deck imputation (Bankier et al. 1994; Shlomo and De Waal 2005; Coutinho and De Waal 2012; Coutinho et al. 2013). As examples of this strategy, Shlomo and De Waal (2008) apply several SDL methods and correct edit-failing records via an edit-imputation procedure based on linear programming; and Cano and Torra (2011) propose adding random noise followed by swapping the noise values of edit-failing records until all records pass edit constraints. We note that neither of these approaches is theoretically guaranteed to preserve all edits.

To implement edit-after-SDL, we propose to use a model-based imputation method which guarantees that all edit-corrections result in records that lie in the feasible region, for example, the restricted support of $\boldsymbol{y}_i$ that satisfies all inequality constraints. Specifically, we adopt the multivariate imputation method proposed by Kim et al. (2014b), which is based on mixtures of multivariate normal distributions and is therefore flexible enough to describe complex distributional features. Let $\mathcal{Y}$ represent the feasible region. Using $K > 1$ mixture components – see Kim et al. (2014b) 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 mixture components, w_k is the probability (or weight) of the component, $(\boldsymbol{\mu}_k, \Omega_k)$ 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, $f(\boldsymbol{y}_i^A|D, \mathcal{Y})$. We refer readers to the Appendix for the specifications of the prior distributions and details of Markov chain Monte Carlo (MCMC) steps. We note that the imputation engine of Kim et al. (2014b) does not automatically extend to handle balance constraints, although it can be modified to do so (Kim et al. 2014a). We also note that agencies can ensure only integer values are released by rounding each imputed value to the nearest integer (we did not do this in our simulation).

2.2.2. Approach II: Edit-Preserving SDL

It is possible to modify some SDL techniques to ensure the masked data satisfy all constraints. A general strategy 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 to variables can generate ε_i (or $\boldsymbol{\delta}_i$) repeatedly until the drawn $\tilde{\boldsymbol{y}}_i$ satisfies the edit rules. We note that rejection sampling approaches can have various negative impacts on data quality. For example, the distribution of the random noise for points near the boundary of the feasible region is not likely to be symmetric, which could result in bias. We also note that balance edits can be difficult to satisfy with rejection sampling.

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. Rejection sampling is generally inappropriate for partially synthetic data, since the model itself should account explicitly for the constrained support (the feasible region). Instead, we use the imputation engine of Kim et al. (2014b), heretofore used exclusively for missing data, as a synthesizer that guarantees the released synthetic values satisfy all edit constraints.

3. Simulation Study

We use a subset of 6,521 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. As edit rules, we introduce linear constraints typical of those used to edit business survey data (Winkler and Draper 1996; Thompson et al. 2001; Hedlin 2003). Table 1 displays the range restrictions, and Table 2 displays the ratio constraints. The introduced constraints are data derived and hypothetical; they are not actual constraints derived from the domain knowledge of economic experts.

To simplify presentation, we mask only three of the seven variables – number of skilled employees, number of unskilled employees, and capital – and leave the remaining variables unaltered. We work with the natural logarithms of all variables. While not necessary, this improves computation in the mixture model used for imputations, as the model needs a smaller number of mixture components. Additionally, log transformations are often useful in statistical inference models with skewed economic data (Petrin and White 2011). To avoid new notation, we let $\boldsymbol{y}_i$ and $\tilde{\boldsymbol{y}}_i$ represent the vectors of natural logarithms of the seven variables in D and D^{rel}, respectively. Thus, $\boldsymbol{y}_i^A$ comprises the three log-transformed values $(y_{i\text{SL}}, y_{i\text{UL}}, y_{i\text{CP}})$.

We use the SDL procedures outlined in Section 2 on the log-transformed values $\boldsymbol{y}_i$, using multiple values of the disclosure parameters when possible. These include adding noise (Noise) with $\tau_{\text{noise}} \in \{0.16, 0.25, 0.36, 0.49\}$, rank swapping (Swap) with $\tau_{\text{swap}} \in \{1, 5, 10\}$, microaggregation based on principal components clustering (Mic)

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 skilled 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

with $\tau_{\text{mic}} \in \{2, 3, 5\}$, microaggregation based on principal components clustering followed by adding noise (MicN), and multivariate fixed-size microaggregation (MMic) with $\tau_{\text{mmic}} \in \{3, 10, 15, 30\}$. We also examined variable-size microaggregation (Solanas and Martnez-Balleste 2006; Domingo-Ferrer et al. 2008); the results were essentially indistinguishable from MMic with $\tau_{\text{mmic}} = 3$ and thus are not reported here. We also use two methods of Shlomo and de Waal (2008), including controlled adding noise (cNoise) with $\tau_{\text{cnoise}} \in \{0.10, 0.30, 0.50\}$ and controlled microaggregation with adding noise based on principal components clustering/subgrouping (cMicN) with $\tau_{\text{cmic}} \in \{2, 3, 5\}$. We generate partially synthetic data (Synt) by replacing all of $\boldsymbol{y}_i^A$ with draws from the model of Kim et al. (2014b). For partially synthetic data, we use 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, provided that doing so does not increase risks unacceptably.

For procedures involving randomness, we generate 20 masked data sets from different random seeds. For the microaggregation procedures (Mic and MMic), we use only one masked data set since these methods are deterministic. As evident in Table 3 and illustrated in Figure 1, all the perturbative SDL methods except MMic3 and MMic10 result in edit violations when applied without edit-preserving modifications. Adding noise with the larger values of τ_{noise} pushes many $\boldsymbol{y}_i$ outside the boundary of $\mathcal{Y}$, resulting in the largest number of edit violations. Rank swapping also produces many edit violations, even with the fairly tight swapping range of $\tau_{\text{swap}} = 10$. Microaggregation and multivariate

Table 3. Numbers of records that violate edit rules across the 20 replications (or single realizations for Mic and MMic) after implementing perturbative SDL methods

Method	Mean	%	%	Mean	%	Method	Mean	%
Noise16	157.8	2.5	Mic3N	84.1	1.3	Mic2	4.0	0.1
Noise25	255.4	4.0	Mic5N	116.2	1.8	Mic3	5.0	0.1
Noise36	406.2	6.3	cMic2N	54.8	0.8	Mic5	15.0	0.2
Noise49	614.8	9.6	cMic3N	83.1	1.2	MMic3	0.0	0.0
cNoise10	7.6	0.1	cMic5N	116.1	1.8	MMic10	0.0	0.0
cNoise30	27.9	0.4	Swap01	5.6	0.1	MMic15	1.0	0.02
cNoise50	48.1	0.7	Swap05	45.1	0.7	MMic30	2.0	0.03
Mic2N	53.5	0.8	Swap10	134.2	2.1			

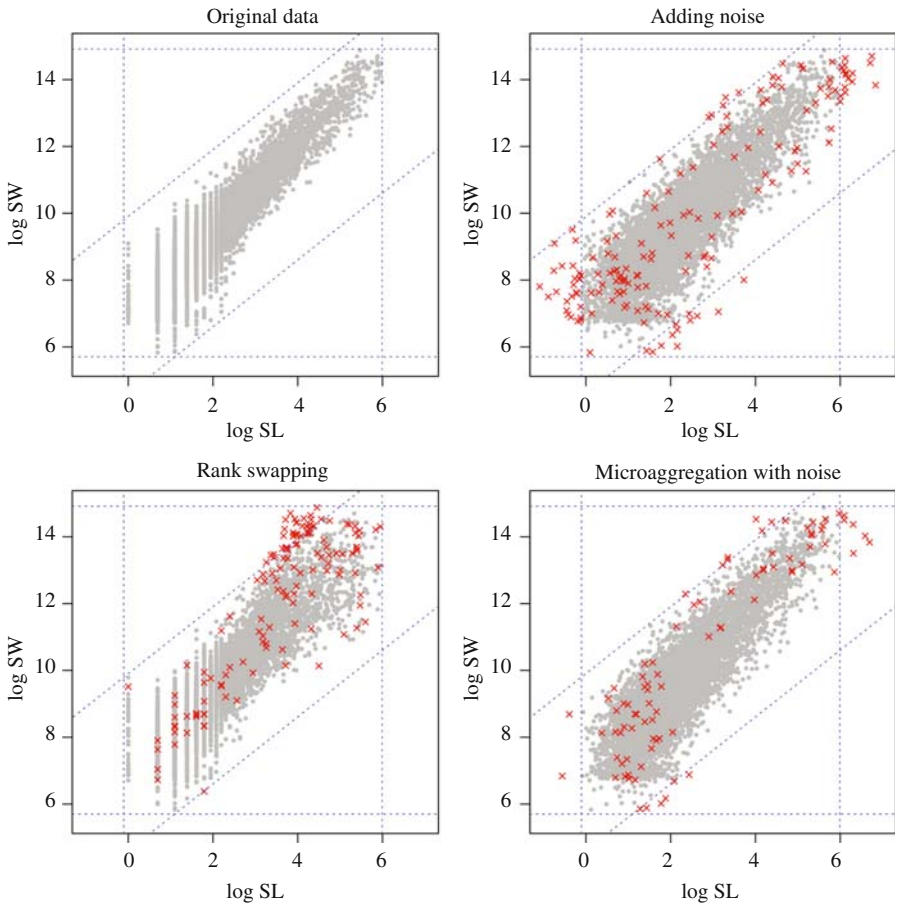

Fig. 1. Illustrative example of how SDL can result in violations of linear constraints. Top-left panel shows pre-SDL data for the log(SL) and log(SW). The variables SL, UL, and CP are masked by adding noise with $\tau_{noise} = 0.16$ (Noise16, top-right panel), rank swapping with $\tau_{swap} = 10$ (Swap10, bottom-left panel), and microaggregation of $\tau_{mic} = 3$ with adding noise (Mic3N, bottom-right panel). Solid circles indicate records that satisfy edit rules and " × " indicate records that violate constraints, i.e., $\tilde{\mathbf{y}}_i \notin Y$

fixed-size microaggregation result in only a few masked records that violate the constraints. This is because microaggregation generally moves values away from boundaries and hence towards the feasible region. In fact, if we had applied microaggregation to all variables in $\mathbf{y}_i$, the resulting records always would be inside $\mathcal{Y}$ due to its convexity. Since we replace only each $\mathbf{y}_i^A$, we cannot guarantee that $\tilde{\mathbf{y}}_i \in Y$. As a general conclusion, we note that the number of edit violations increases with the amount of perturbation for every class of SDL methods.

We next seek to correct any edit violations using the two general strategies. For edit-after-SDL, we replace all values of $\mathbf{y}_i^A$ of edit-failing records with draws from the imputation model outlined in Subsection 2.2.1. For edit-preserving SDL, we use the rejection sampling scheme of Subsection 2.2.2 for all methods involving randomness. For rank swapping with $\tau_{\text{swap}} = 10$, we did not obtain a D^{rel} without edit violations even after 1,000 independent replications of swapping. 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 for this procedure in this simulation design is hopeless.

As measures of disclosure risk, we use the *percentage of linked* criterion of Domingo-Ferrer, Mateo-Sanz, and Torra (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 $y_{i0} = y_{j0}$, that is, the record in D^{rel} can be linked correctly to D based on matching the available variables, we let $t_i^{(1)} = 1$ and otherwise let $t_i^{(1)} = 0$. We then define one risk measure as PL1 $= \sum_{i=1}^n t_i^{(1)}/n \times 100$. Similarly, we let $t_i^{(2)} = 1$ when the correct link for record i in D has either the smallest or second smallest value among all the $d_{i,j}$, and $t_i^{(2)} = 0$ otherwise. We define a second risk measure as PL2 $= \sum_{i=1}^n t_i^{(2)}/n \times 100$, the percentage of records for which the correct link is among the two closest matches. Finally, we define a third risk measure, PL3, as the percentage of records for which the correct link is among the three closest matches.

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_{prop}) utility measure suggested by Woo et al. (2009). For KL, we use a closed-form expression based on a normality assumption,

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

where $\bar{\mathbf{y}}$ and Σ are the sample mean and the sample covariance of $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ in D, and $\bar{\mathbf{y}}^{\text{rel}}$ and Σ^{rel} are the corresponding statistics of $\{\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n\}$ in D^{rel}. For U_{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

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

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

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

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

Table 4 displays the average values of KL, U_{prop} and PL1 — PL3 over the replicates for each method. When methods are implemented with both strategies, the risk-utility profiles are fairly similar across the two strategies. This is not overly surprising, since these SDL methods typically generate only a modest number of edit violations in these data. Nonetheless, for these methods, the edit-after-SDL version does slightly outperform the edit-preserving SDL version, generally offering both lower risk and higher utility. This results largely from the imputations, which are generally of higher quality than the repeated draws from the rejection sampling scheme.

In Table 4, the differences in the risk-utility profiles across the two ways of dealing with edit violations are dwarfed by differences in the profiles across the classes of SDL methods. This suggests that the choice of SDL method is more important than the strategy for correcting edit violations.

Figure 2 displays a risk-utility (R-U) map (Duncan and Stokes 2004; Gomatam et al. 2005; Cox et al. 2011) for all realizations of D^{rel} and the most competitive procedures, using U_{prop} as the utility measure and PL1 as the risk measure. The risk-utility frontier consists of candidate releases with no other candidate to their "southwest." The R-U frontier includes the variants of microaggregation with adding noise (MicN), which have the lowest levels of disclosure risk, and partially synthetic data (Synt), which has the maximum level of data utility and a low level of disclosure risk. Several variants of MMic are close to the frontier (and would be on the frontier but for Synt and Swap10), generally having high utility for reasonable disclosure risks.

4. Concluding Remarks

Based on our studies, there appear to be no appreciable differences between the strategies of edit-after-SDL and edit-preserving SDL, at least when both are possible. Hence, arguably, agencies can choose an SDL procedure without too much consideration of how they will ensure the released data satisfy all edits, at least when the SDL method does not generate a large number of edit violations. Microaggregation with adding noise, multivariate fixed-size microaggregation and partially synthetic data were the most effective strategies in our simulations. The last method 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. We believe that in the long run, the most desirable approach is one that fully integrates editing, imputation and SDL.

Table 4. Measured data utility and disclosure risk. Entries include the averages of KL, U_{prop}, PL1, PL2, and PL3 from 20 replications of each method, except Mic and MMic which have only single replicates. Note that the three risk measures are highly correlated (all correlations are at least 0.988), so that they offer similar conclusions about the risk-utility profiles of the different SDL methods

	Inverse Utility				Risk					
	KL		U_{prop} (×100)		PL1		PL2		PL3	
Methods	I	II	I	II	I	II	I	II	I	II
Noise16	0.34	0.35	2.2	2.2	1.91	2.12	3.42	3.75	4.74	5.20
Noise25	0.50	0.52	2.9	3.0	1.11	1.26	2.05	2.26	2.90	3.20
Noise36	0.64	0.67	3.4	3.5	0.74	0.82	1.37	1.52	1.94	2.12
Noise49	0.75	0.81	3.5	3.8	0.51	0.60	0.96	1.10	1.38	1.55
cNoise10	0.0007	0.0007	0.002	0.001	48.21	48.34	67.32	67.45	77.47	77.59
cNoise30	0.04	0.04	0.05	0.05	8.78	8.85	14.91	15.07	19.83	20.00
cNoise50	0.16	0.16	0.2	0.3	2.56	2.60	4.63	4.65	6.47	6.45
Mic2N	0.50	0.51	2.9	3.0	0.57	0.58	1.10	1.12	1.61	1.64
Mic3N	0.64	0.66	4.0	4.2	0.35	0.35	0.67	0.66	0.99	0.97
Mic5N	0.75	0.78	4.8	5.2	0.29	0.24	0.54	0.45	0.76	0.66
cMic2N	0.51	0.51	2.9	3.0	0.57	0.59	1.10	1.11	1.61	1.64
cMic3N	0.64	0.66	4.1	4.3	0.34	0.32	0.69	0.65	0.99	0.97
cMic5N	0.75	0.79	4.9	5.3	0.27	0.24	0.52	0.45	0.78	0.65
Swap01	0.002	0.002	0.004	0.004	63.55	63.81	79.79	80.22	85.91	86.27
Swap05	0.08	–	0.05	–	6.22	–	12.01	–	17.26	–
Swap10	0.24	–	0.1	–	0.94	–	1.86	–	2.91	–
Mic2	0.59	–	2.7	–	1.38	–	3.05	–	4.18	–
Mic3	1.34	–	4.6	–	0.67	–	1.43	–	2.19	–
Mic5	2.71	–	6.3	–	0.34	–	0.75	–	1.24	–
MMic3	–	0.01	–	0.01	–	7.73	–	15.83	–	26.83
MMic10	–	0.05	–	0.1	–	2.12	–	4.22	–	6.17
MMic15	0.08	–	0.2	–	1.37	–	2.75	–	4.28	–
MMic30	0.16	–	0.3	–	0.74	–	1.56	–	2.37	–
Synt	–	0.02	–	0.06	–	0.59	–	1.13	–	1.66

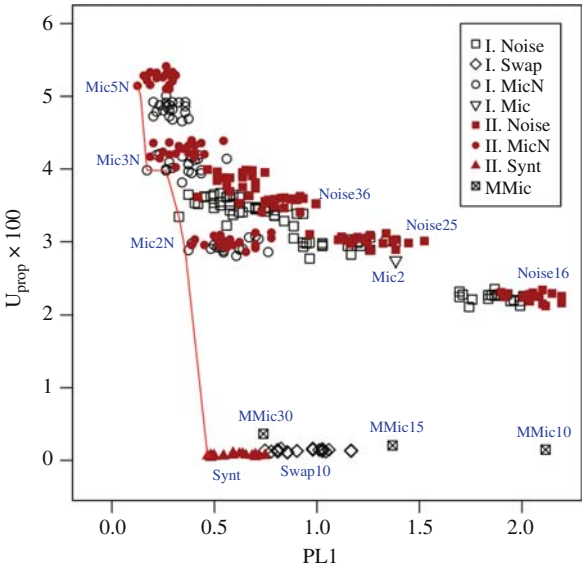

Fig. 2. Risk-utility map with the SDL methods. The solid line indicates the risk-utility frontier. The open symbols represent edit-after-SDL approaches, and the solid symbols represent edit-preserving SDL approaches. Smaller values of PL1 and U_{prop} represent the higher levels of data protection and data utility. Note that the plot does not include cMicN's because the results are very similar to those of MicN. The other methods whose results are not shown in the plot have high risk and/or low utility

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. Indexed microaggregation (Cox et al. 2011) is able to protect risky weights. However, by any measure, much more work remains than has been carried out so far.

Appendix: The Joint Multivariate Imputation Using Normal Mixtures

For imputations of faulty values, we use the joint multivariate normal method developed in Kim et al. (2014b) and described in Section 2. The likelihood function in (1) can be re-expressed with latent variables z_i by

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

and

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

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

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

where $\Phi = diag(\phi_1, \ldots, \phi_p)$, and $\phi_j \sim \mathrm{Gamma}(a_\phi, b_\phi)$ for $j = 1, \ldots, p$. Here, IW denotes the inverse Wishart distribution and Gamma(a,b) denotes the Gamma distribution with mean a/b. For flexible modeling of the component weights, we adopt the stick-breaking representation of a truncated Dirichlet process (Sethuraman 1994; Ishwaran and James 2001):

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

In the simulation study, we follow Kim et al. (2014b) and set $\boldsymbol{\mu}_0 = 0$, $h = 1$, $\zeta = p + 1$, $a_\phi = b_\phi = 0.25$, $a_\alpha = b_\alpha = 0.25$ 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 supposes 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

$$f(Y_N \,|\, \Theta_1, \ldots, \Theta_K) = \prod_{i=1}^{N} \sum_{k=1}^{K} w_k \mathrm{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \Omega_k),$$

where $\Theta_k = (\mu_k, \Omega_k, w_k)$. Given the augmented sample Y_N, the parameters $\Theta_k = (w_k, \mu_k, \Omega_k)$ can be sampled via Gibbs sampling. Setting $f(N) \propto 1/N$ 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

$$N - n \,|\, n, \Theta_1, \ldots, \Theta_K, \mathcal{Y} \sim \text{Negative Binomial}\big(n, 1 - h_\Theta(\mathcal{Y})\big),$$

where

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

The MCMC algorithm for sampling from this distribution relies on the following steps.

1. For $k = 1, \ldots, K$, draw $\Omega_k \sim \mathrm{IW}(\zeta_k, \Phi_k)$ and $\boldsymbol{\mu}_k \sim \mathrm{N}(\boldsymbol{\mu}_k^*, \Omega_k/(N_k + h))$ where $\boldsymbol{\mu}_k^* = (N_k \bar{\boldsymbol{y}}_k + h\boldsymbol{\mu}_0)/(N_k + h)$, $\zeta_k = \zeta + N_k$,$\Phi_k = \Phi + S_k + (\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)'/(1/N_k + 1/h)$. We calculate the sample mean $\bar{\boldsymbol{y}}_k$ and the sample covariance S_k from the error-free, pre-SDL values $Y_n = \{\boldsymbol{y}_i, \mathrm{i} = 1, \ldots, n\}$ 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 $k = 1, \ldots, K-1$, draw $v_k \sim \mathrm{Beta}\Big(1 + N_k, \alpha + \sum_{g>k} N_g\Big)$. Set $v_K = 1$. Compute $w_k = v_k \prod_{g<k} (1 - v_g)$.

3. Update $\Phi = diag(\phi_1, \ldots, \phi_p)$ by drawing $\phi_j \sim \text{Gamma}\big(a_\phi + \zeta K/2, b_\phi + \sum_{k=1}^{K} \Omega^{-1}_{k(j,j)}/2\big)$ for each $j = 1, \ldots, p$, where $\Omega^{-1}_{k(j,j)}$ is the jth diagonal element of Ω_k^{-1}.
4. Draw α from $\text{Gamma}\big(a_\alpha + K - 1, b_\alpha - \log w_K\big)$.
5. For $i = 1, \ldots, n$, sample $z_i \sim \text{Categorical}(w^*_{i1}, \ldots, w^*_{iK})$ where

$$w^*_{ik} = w_k \text{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \Omega_k) / \Big[\sum\nolimits_{g=1}^{K} w_g \text{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_g, \Omega_g)\Big].$$

6. Sample (N, Z_{N-n}, Y_{N-n}) jointly from their full conditional distribution as follows. Let $c_{\text{in}} = c_{\text{out}} = 0$.
   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 $c_{\text{in}} = c_{\text{in}} + 1$.
   6.4. If $\boldsymbol{y}^* \in \mathcal{Y}^c$, set $c_{\text{out}} = c_{\text{out}} + 1$, $\boldsymbol{y}_{n + c_{\text{out}}} = \boldsymbol{y}^*$, and $z_{n+c_{\text{out}}} = z^*$.
   6.5. Repeat 6.1 through 6.3 until $c_{\text{in}=n}$.

   Let $N = n + c_{\text{out}}$. Now, $Y_{N-n} = \{\boldsymbol{y}_n + 1, \ldots, \boldsymbol{y}_n + c_{\text{out}}\}$ and $Z_{N-n} = \{z_{n+1}, \ldots, z_{n+c_{\text{out}}}\}$.
7. To update the replacement draws of the faulty values, we use a Hit-and-Run sampler (Chen and Schmeiser 1993). In the initialization step, we propose a starting value $\tilde{\boldsymbol{y}}_i^{A(0)}$ such that $(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(0)}) \in \mathcal{Y}$, for example by using rejection sampling or an extreme-points approach (see Kim et al. 2014b). At any MCMC iteration $t \geq 0$, we update the current value $\tilde{\boldsymbol{y}}_i^{A(t)}$ (which replaces the faulty $\tilde{\text{y}}_i^A$) with the following steps.
   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,

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

   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

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

5. References

Bankier, M., M. Luc, C. Nadeau, and P. Newcombe. 1994. "Imputing Numeric and Qualitative Variables Simultaneously." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 242–247. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1994_036.pdf. (accessed February 2015).

Cano, I. and V. Torra. 2011. "Edit Constraints on Microaggregation and Additive Noise." In *Privacy and Security Issues in Data Mining and Machine Learning*, edited by

C. Dimitrakakis, A. Gkoulalas-Divanis, A. Mitrokotsa, V.S. Verykios, and Y. Saygin, 1–14. Berlin: Springer.

Chen, M.H. and B. Schmeiser. 1993. "Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers." *Journal of Computational and Graphical Statistics* 2: 251–272. DOI: http://dx.doi. org/10.2307/1390645.

Coutinho, W. and T. de Waal. 2012. *Hot Deck Imputation of Numerical Data Under Edit Restrictions*. Discussion Paper 2012243, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/6C97F296-EE33-4F26-A813-6432ED530249/0/201223x10pub.pdf. (accessed February 2015).

Coutinho, W., T. de Waal, and M. Remmerswaal. 2011. "Imputation of Numerical Data Under Linear Edit Restrictions." *Statistics and Operations Research Transactions* 35: 29–62.

Coutinho, W., T. de Waal, and N. Shlomo. 2013. "Calibrated Hot-Deck Donor Imputation Subject to Edit Restrictions." *Journal of Official Statistics* 29: 299–321. DOI: http://dx.doi.org/10.2478/jos-2013-0024.

Cox, L.H., A.F. Karr, and S.K. Kinney. 2011. "Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act." *International Statistical Review* 79: 160–183. DOI: http://dx.doi.org/10.1111/j.1751-5823.2011.00140.x.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: Wiley.

Defays, D. and P. Nanopoulos. 1993. "Panels of Enterprises and Confidentiality: The Small Aggregates Method." In Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, November 2–4, 1992, 195–204. Ottawa, Ontario, Canada. Available at: http://www.researchgate. net/publication/243784453_Panels_of_enterprises_and_confidentiality_the_small_aggregates_ method. (accessed February 2015).

Domingo-Ferrer, J. and J.M. Mateo-Sanz. 2002. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering* 14: 189–201. DOI: http://dx.doi.org/10.1109/69.979982.

Domingo-Ferrer, J., F. Sebe, and A. Solanas. 2008. "A Polynomial-Time Approximation to Optimal Multivariate Microaggregation." *Computers and Mathematics with Applications* 55: 714–732. DOI: http://dx.doi.org/10.1016/j.camwa.2007.04.034.

Domingo-Ferrer, J., J.M. Mateo-Sanz, and V. Torra. 2001. "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk." In Pre-proceedings of ENKNTTS, 807–826. Available at: http://neon.vb.cbs.nl/casc/NTTSJosep.pdf. (accessed February 2015)

Drechsler, J. and J.P. Reiter. 2008. "Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and Y. Saygin, 227–238. New York: Springer.

Duncan, G.T. and S.L. Stokes. 2004. "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding." *Chance* 17: 16–20. DOI: http://dx.doi.org/10.1080/09332480.2004.10554908.

Fayyoumi, E. and B.J. Oommen. 2010. "A Survey on Statistical Disclosure Control and Microaggregation Techniques for Secure Statistical Databases." *Software: Practice and Experience* 40: 1161–1188. DOI: http://dx.doi.org/10.1002/spe.992.

Fellegi, I.P. and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association* 71: 17–35. DOI: http://dx.doi.org/10.1080/01621459.1976.10481472.

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, April 21–24, 1991. 571–578. Seattle, Washington. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.27.568&rep = rep1&type = pdf. (accessed February 2015).

Gomatam, S., A.F. Karr, J.P. Reiter, and A.P. Sanil. 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.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control*. West Sussex, UK: John Wiley & Sons.

Ishwaran, H. and L.F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96: 161–173. DOI: http://dx.doi.org/10.1198/016214501750332758.

Karr, A.F. 2009. *The Role of Transparency in Statistical Disclosure Limitation*. Presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.41.e.pdf. (accessed February 2015).

Kim, H.J., L.H. Cox, A.F. Karr, J.P. Reiter and Q. Wang. 2014a. *Simultaneous Edit-Imputation for Contineous Microdata*. Technical Report 189, National Institute of Statistical Sciences, Research Triangle Park, NC. Available at: https://www.niss.org/sites/default/files/tr189_updated.pdf (accessed February 2015).

Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox, and A.F. Karr. 2014b. "Multiple Imputation of Missing or Faulty Values Under Linear Constraints." *Journal of Business & Economic Statistics* 32: 375–386. DOI: http://dx.doi.org/10.1080/07350015.2014.885435.

Kim, J.J. 1986. "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 370–374. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1986_069.pdf. (accessed February 2015).

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

Lavine, M. and M. West. 1992. "A Bayesian Method for Classification and Discrimination." *Canadian Journal of Statistics* 20: 451–461. DOI: http://dx.doi.org/10.2307/3315614.

Little, R.J.A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9: 407–426.

Meng, X.L. and A.M. Zaslavsky. 2002. "Single Observation Unbiased Priors." *The Annals of Statistics* 30: 1345–1375.

Moore, R.A. 1996. *Controlled Data-Swapping Techniques for Masking Use Microdata Sets*. Research Report RR96/04, Statistical Research Division, U.S. Bureau of the Census, Washington, DC. Available at: https://www.census.gov/srd/papers/pdf/rr96-4.pdf. (accessed February 2015).

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

O'Malley, A.J. and A.M. Zaslavsky. 2008. "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse." *Journal of the American Statistical Association* 103: 1405–1418. DOI: http://dx.doi.org/10.1198/016214508000000724.

Petrin, A. and T.K. White. 2011. "The Impact of Plant-Level Resource Reallocations and Technical Progress on U.S. Macroeconomic Growth." *Review of Economic Dynamics* 14: 3–26. DOI: http://dx.doi.org/10.1016/j.red.2010.09.004.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger. 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.

Reiter, J.P. 2004. "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation." *Survey Methodology* 30: 235–242.

Reiter, J.P. 2005. "Estimating Risks of Identification Disclosure in Microdata." *Journal of the American Statistical Association* 100: 1103–1112. DOI: http://dx.doi.org/10.1198/016214505000000619.

Reiter, J.P. and R. Mitra. 2009. "Estimating Risks of Identification Disclosure in Partially Synthetic Data." *Journal of Privacy and Confidentiality* 1: 99–110.

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.

Shlomo, N. and T. de Waal. 2005. *Preserving Edits When Perturbing Microdata for Statistical Disclosure Control*. S3RI Methodology Working Paper M05/12, Southampton Statistical Sciences Research Institute. Available at: http://eprints.soton.ac.uk/14725/1/14725-01.pdf. (accessed February 2015).

Shlomo, N. and T. de Waal. 2008. "Protection of Micro-Data Subject to Edit Constraints Against Statistical Disclosure." *Journal of Official Statistics* 24: 229–253.

Solanas, A. and A. Martnez-Balleste. 2006. "V-MDAV: A Multivariate Microaggregation With Variable Group Size." In Proceedings of the 17th IASC Symposium on Computational Statistics, August 28–September 1, 2006. 917–925. Rome, Italy. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.1680&rep=rep1&type=pdf. (accessed February 2015).

Sullivan, G. and W.A. Fuller. 1989. "The Use of Measurement Error to Avoid Disclosure." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 802–807. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1989_148.pdf. (accessed February 2015).

Tempelman, C. 2007. *Imputation of Restricted Data*. Ph. D. dissertation, University of Groningen. Available at: http://dissertations.ub.rug.nl/faculties/eco/2007/d.c.g.tempelman. (accessed February 2015).

Tendick, P. 1991. "Optimal Noise Addition for Preserving Confidentiality in Multivariate Data." *Journal of Statistical Planning and Inference* 27: 341–353. DOI: http://dx.doi.org/10.1016/0378-3758(91)90047-I.

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

Torra, V. 2008. "Constrained Microaggregation." *Transactions on Data Privacy* 1: 86–104.

Van Buuren, S. and K. Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden, Netherlands. Available at: http://www. stefvanbuuren.nl/publications/Flexible%20multivariate%20-%20TNO99054%201999.pdf. (accessed February 2015)

Willenborg, L. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Winkler, W.E. and L.R. Draper. 1996. *Application of the SPEER Edit System*. Research Report RR96/02, Statistical Research Division, U.S. Bureau of the Census, Washington, DC. Available at: https://www.census.gov/srd/papers/pdf/rr96-2.pdf. (accessed February 2015).

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

Received October 2013
Revised May 2014
Accepted September 2014