Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Divisio

Article

Combining synthetic data with subsampling to create public use microdata files for large scale surveys

by Jörg Drechsler and Jerome P. Reiter

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 73-79
Statistics Canada, Catalogue No. 12-001-X

Combining synthetic data with subsampling to create public use microdata files for large scale surveys

Jörg Drechsler and Jerome P. Reiter [1]

Abstract

To create public use files from large scale surveys, statistical agencies sometimes release random subsamples of the original records. Random subsampling reduces file sizes for secondary data analysts and reduces risks of unintended disclosures of survey participants' confidential information. However, subsampling does not eliminate risks, so that alteration of the data is needed before dissemination. We propose to create disclosure-protected subsamples from large scale surveys based on multiple imputation. The idea is to replace identifying or sensitive values in the original sample with draws from statistical models, and release subsamples of the disclosure-protected data. We present methods for making inferences with the multiple synthetic subsamples.

Key Words: Confidentiality; Disclosure; Multiple imputation.

1. Introduction

National Statistical Institutes (NSIs) like the U.S. Census Bureau and Statistics Canada conduct large scale surveys that are highly valued by secondary data analysts, such as the American Community Survey (ACS) and the National Longitudinal Survey of Children and Youth (NLSCY). While these analysts desire access to as much data as possible, the NSI also must protect the confidentiality of survey participants' identities and sensitive attributes. A common strategy for reducing disclosure risks in large scale studies is to release subsamples of the original survey data; for example, the Census Bureau releases a subsample from the collected ACS data comprising 1% of all U.S. households (the collected ACS data comprise 2.5% of all households), and Statistics Canada releases a 20% sample of individuals from the NLSCY. See Willenborg and de Waal (2001) and Reiter (2005) for discussions of the confidentiality protection engendered by sampling. Typically, however, subsampling alone does not eliminate disclosure risks, particularly for units in the subsample with unusual combinations of characteristics. NSIs therefore alter data before dissemination. For example, in the ACS, the Census Bureau performs data swapping, topcoding of selected variables, aggregating of geography, and age perturbation; in the NLSCY, Statistics Canada uses data swapping and suppression.

When implemented with high intensity, as may be necessary to protect confidentiality in highly visible surveys, standard disclosure limitation strategies can seriously distort inferences (Winkler 2007; Elliott and Purdam 2007; Drechsler and Reiter 2010). Further, for many standard techniques it is difficult for data analysts - especially those without advanced statistical training - to properly account for the effects of the disclosure control in estimation. Motivated by these limitations, we propose a new approach for generating public use microdata samples from large scale surveys called subsampling with synthesis. The basic idea is to replace identifying or sensitive values in the original sample with multiple draws from statistical models estimated with the original data file, and release subsamples of the disclosure-protected data. The subsamples can comprise one common set of records, or they can be taken independently.

This approach is a variant of partially synthetic data (Little 1993; Reiter 2003), which has been used in the U.S. to create several public use data products, including the Survey of Income and Program Participation, the Longitudinal Business Database, the Survey of Consumer Finances, the American Community Survey group quarters data, and OnTheMap. The approach proposed here differs from partial synthesis because of the subsampling, which necessitates adjustments to the inferential methods of Reiter (2003); these are presented here. The approach also differs from the methods for creating synthetic public use microdata samples of census data developed recently by Drechsler and Reiter (2010). In subsampling with synthesis, the initial data come from a survey and not from a census; thus, inferences must account for the additional uncertainty that results from the initial sampling.

2. General approach

We now describe the data generation and inferential procedures for the two approaches to subsampling with

1. Jörg Drechsler, Institute for Employment Research, Department for Statistical Methods, Regensburger Straße 104, 90478 Nürnberg, Germany. E-Mail: joerg.drechsler@iab.de; Jerome P. Reiter, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. E-Mail: jerry@stat.duke.edu.

synthesis: releasing different (independent) subsamples, and releasing a common set of records in each subsample. The data generation methods, as well as methods for making valid inferences from the multiple datasets, depend on the subsampling approach. For both approaches, we let D denote the original survey data of n_1 units sampled from a population consisting of N units. We initially assume that the original sampling design is a simple random sample; we later extend to stratified sampling. We assume that all sampled units fully respond in D. Unlike for standard partial synthesis (Reiter 2004), methods have not been developed to handle missing data and synthesis with subsampling simultaneously. We focus here on general descriptions of the approaches and presentation of the inferential methods. We do not discuss synthesis model building strategies; see Drechsler and Reiter (2009) and the references therein for guidance.

2.1 Releasing different random subsamples

2.1.1 Summary of approach

To begin, the NSI creates m partially synthetic datasets, $D_{syn} = \{D_i: i = 1, \ldots, m\}$, for the original survey following the approach of Reiter (2003). Specifically, the NSI replaces identifying or sensitive values in D with multiple imputations. Synthesis models are estimated using only the records whose values will be synthesized. The synthesis is done independently m times, resulting in D_{syn}. The NSI then takes a simple random subsample of $n_2 < n_1$ records from each D_i. These m subsamples, $d_{syn} = \{d_i: i = 1, \ldots, m\}$, are released to the public.

The analyst of d_{syn} seeks inferences about some estimand Q, such as a population mean or regression coefficient. In each d_i, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator u, where the analyst specifies q and u acting as if d_i were the collected data. Here, u is specified ignoring any finite population correction factors; for example, when q is the sample mean, $u = s^2/n_2$, with s^2 being the sample variance. For $i = 1, \ldots, m$, let q_i and u_i be the values of q and u in d_i. The following quantities are needed for inferences.

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m \quad (1)$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2 / (m-1) \quad (2)$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m. \quad (3)$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_d = (n_2/n_1 - n_2/N)\,\bar{u}_m + b_m/m \quad (4)$$

to estimate the variance of $\bar{q}_m$. Derivations of these estimates are presented in Section 2.1.2. We note that without subsampling, *i.e.*, $n_2 = n_1$, (4) equals the variance estimate for standard partial synthesis (Reiter 2003). For large n_2, inferences are based on a t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_d}(0, T_d)$, with degrees of freedom $\nu_d = (m-1)(1 + (n_2/n_1 - n_2/N)\, m\bar{u}_m / b_m)^2$.

The inferential methods can be extended to stratified samples in which the NSI uses the same strata for the subsample and original sample. Let N_h be the population size in stratum h, where $h = 1, \ldots, H$. For each h, let $\bar{q}_{mh}$ and T_{dh} be the values of (1) and (4) computed using only the records in d_{syn} in stratum h. These estimates are used in inferences for population quantities in stratum h. For inferences about the entire population mean, the point estimate of Q is $\bar{q}_m = \sum_h (N_h/N)\,\bar{q}_{mh}$, and its estimated variance is $T_d = \sum_h (N_h/N)^2 T_{dh}$. Point and variance estimates for nonlinear functions of means can be derived using Taylor series expansions. We note that NSIs should release the values of n_{2h}/n_{1h} for all strata to enable variance estimation.

2.1.2 Derivation of inferences for the different random subsamples approach

The analyst seeks $f(Q \mid d_{syn})$, which can be written as

$$f(Q \mid d_{syn}) = \int f(Q \mid D_{syn}, d_{syn})\, f(D_{syn} \mid d_{syn})\, dD_{syn}. \quad (5)$$

For all derivations in Section 2.1.2, we assume that the analyst's distributions are identical to those used by the NSI for creating D_{syn}. We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which we derive using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

Let Q_i and U_i be the point estimate of Q and its variance that the analyst would compute with D_i (which is not available to the analyst). Let $\bar{Q}_m$, $\bar{U}_m$, and B_m be defined as in (1) - (3) but using Q_i and U_i. From standard partial synthesis results (Reiter 2003), we have $(Q \mid D_{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m/m)$. We assume that $(q_i \mid D_i) \sim N(Q_i, (1 - n_2/n_1)u_i)$ and, as is typical in multiple imputation contexts, that $u_i \approx \bar{u}_m$. Thus, using standard Bayesian theory, we have $(\bar{Q}_m \mid d_{syn}) \sim N(\bar{q}_m, (1 - n_2/n_1)\bar{u}_m/m)$ and $((m-1)b_m/(B_m + (1 - n_2/n_1)\,\bar{u}_m) \mid d_{syn}) \sim \chi^2_{m-1}$. Hence, we have $f(Q \mid d_{syn}, B_m, \bar{U}_m) = N(\bar{q}_m, \bar{U}_m + B_m/m + (1 - n_2/n_1)\,\bar{u}_m/m)$.

To get $f(Q \mid d_{syn})$, we need to integrate out B_m and $\bar{U}_m$ from this distribution. We do so by substituting B_m and $\bar{U}_m$ with their approximate expected values. To approximate

$E(B_m \mid d_{syn})$, we use $b_m - (1 - n_2/n_1)\,\bar{u}_m$. To approximate $E(\bar{U}_m \mid d_{syn})$, we note that

$$\begin{aligned}\mathrm{Var}(Q \mid d_i) &= E[\mathrm{var}(Q \mid D_i) \mid d_i] + \mathrm{var}[E(Q \mid D_i) \mid d_i] \\ &= E(U_i \mid d_i) + \mathrm{var}(Q_i \mid d_i). \qquad (6)\end{aligned}$$

Here, $\mathrm{var}(Q \mid d_i) = (1 - n_2/N)u_i$. Solving (6), we have $E(\bar{U}_m \mid d_{syn}) \approx (n_2/n_1 - n_2/N)\bar{u}_m$. After substitution of these expected values, we have $\mathrm{var}(Q \mid d_{syn}) = T_d$.

Since we use an estimated variance for Q, we approximate $f(Q \mid d_{syn})$ with a t-distribution with mean $\bar{q}_m$ and variance T_d. The degrees of freedom, ν_d, is derived by matching the first two moments of $(\nu_d T_d) / \{(n_2/n_1 - n_2/N)\bar{u}_m + B_m/m + (1 - n_2/n_1)\bar{u}_m/m\}$ to those of a $\chi^2_{\nu_d}$ distribution.

2.2 Releasing the same random subsample

At first glance, releasing a common set of records in each subsample looks like standard partial synthesis. However, Reiter's (2003) variance estimator can be positively biased in this context. To illustrate, suppose that D comprises one variable with sample mean $\bar{x}_1$. Also suppose that we create D_{syn} by replacing all values of x, and we randomly select a common set of n_2 records for the subsample. Let $m = \infty$, and let Q be the population mean of x. If replacements are simulated from the correct model, which is estimated with D, then $\bar{q}_\infty = \bar{x}_1$. Hence, $\mathrm{var}(\bar{q}_\infty)$ is identical to the variance of $\bar{x}_1$, which is $(1 - n_1/N)\,s_1^2/n_1$. However, Reiter's (2003) variance estimate includes $\bar{u}_m$ based on $(1 - n_2/N)\,s_2^2/n_2$, where $E(s_2^2) = s_1^2$. Hence, in general Reiter's (2003) variance will have positive bias for subsamples with synthesis.

In place of standard partial synthesis, we adopt the approach taken by Reiter (2008) for multiple imputation for missing data when records used for imputation are not used or disseminated for analysis. This setting is akin to subsampling the same records in each d_i because the models for the synthesis are estimated with D, but the analyst only has d_{syn} for analysis; that is, not all records used for imputation are disseminated for analysis.

For convenience, we summarize the methodology of Reiter (2008) here but do not include the derivations. First, as in standard partial synthesis, the NSI estimates the synthesis models using only the records whose values will be synthesized. Let θ be the parameters that govern the distribution of the synthetic data models. Second, the NSI samples m values of θ from its posterior distribution. Third, for each drawn $\theta^{(l)}$ where $l = 1, \ldots, m$, the NSI draws a replacement dataset $D^{(l,\,p)}$ from the synthesis models based on $\theta^{(l)}$. The NSI repeats this process r times for each $\theta^{(l)}$. Finally, the NSI releases the collection of $M = mr$ subsamples from these datasets, $d^* = \{d^{(l,\,p)}: l = 1, \ldots, m;\ p = 1, \ldots, r\}$. Each $d^{(l,\,p)}$ includes an index of its nest l.

For $l = 1, \ldots, m$ and $p = 1, \ldots, r$, let $q^{(l,\,p)}$ and $u^{(l,\,p)}$ be the estimate of Q and its estimated variance computed with $d^{(l,\,p)}$. Here, $u^{(l,\,p)}$ includes the finite population correction factor. The following quantities are used for inferences:

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} q^{(l,\,p)}/(mr) = \sum_{l=1}^{m} \bar{q}_r^{(l)}/m, \qquad (7)$$

$$\bar{w}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} (q^{(l,\,p)} - \bar{q}_r^{(l)})^2 / \{m(r-1)\} = \sum_{l=1}^{m} w_r^{(l)}/m, \qquad (8)$$

$$b_M = \sum_{l=1}^{m} (\bar{q}_r^{(l)} - \bar{q}_M)^2 / (m-1), \qquad (9)$$

$$\bar{u}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} u^{(l,\,p)}/(mr). \qquad (10)$$

The analyst can use $\bar{q}_M$ to estimate Q and $T_s = \bar{u}_M - \bar{w}_M + (1+1/m)\,b_M - \bar{w}_M/r$ to estimate the variance of $\bar{q}_M$. When r is large, inferences are based on a t-distribution, $(\bar{q}_M - Q) \sim t_{\nu_s}(0, T_s)$, with degrees of freedom

$$\nu_s = \left(\frac{\{(1+1/m)b_M\}^2}{(m-1)T_s^2} + \frac{\{(1+1/r)\bar{w}_M\}^2}{\{m(r-1)\}T_s^2} \right)^{-1}. \qquad (11)$$

It is possible that $T_s < 0$, particularly for small m and r. Instead, analysts can use the always positive but conservative variance estimator, $T_s^* = \lambda T_s + (1-\lambda)(1+1/m)b_M$, where $\lambda = 1$ when $T_s > 0$ and $\lambda = 0$ otherwise. Motivation for this estimator is provided in Reiter (2008). Generally, negative values of T_s can be avoided by making m and r large. When $T_s < 0$, inferences are based on a t-distribution with $(m-1)$ degrees of freedom, which comes from using only the first term and T_s^* in (11).

For stratified designs, the point estimate for whole population quantities is $\bar{q}_M = \sum_h (N_h/N)\bar{q}_{Mh}$, and its estimated variance is $T_s = \sum_h (N_h/N)^2 T_{sh}$, where $\bar{q}_{Mh}$ and T_{sh} are the point estimate and its variance in stratum h. The degrees of freedom in the t-distribution for stratified sampling is

$$\nu_{st} = \left\{ \frac{\left\{\sum_h ((N_h/N)^2(1+1/m)b_{Mh})\right\}^2}{(m-1)\sum_h (N_h/N)^2 T_{sh}^2} + \frac{\left\{\sum_h ((N_h/N)^2(1+1/k)\bar{w}_{Mh})\right\}^2}{\{m(r-1)\}\sum_h (N_h/N)^2 T_{sh}^2} \right\}^{-1}. \qquad (12)$$

This is derived by moment matching to a χ^2 random variable.

3. Illustrative simulations using a stratified sampling design

In this section, we investigate the analytical properties of the inferential procedures for subsampling with synthesis for stratified simple random sampling. We generate a population of $N = 1{,}000{,}000$ records comprising five variables, $Y_1, \ldots, Y_5$, in $H = 4$ strata. Y_1 is a categorical variable with ten categories generated according to the distribution in Table 1. The distributions for $(Y_2, \ldots, Y_5)$ are displayed in Table 2, along with the stratum sizes.

Table 1
Empirical distribution of Y_1 in the generated population

	1	2	3	4	5	6 7		8	9	10
percentage	24.77	32.63	16.38	15.06	7.13	2.53	0.95	0.33	0.15	0.09

To create D, we randomly sample $n_{1h} = 7{,}500$ records from each stratum. Each subsample comprises $n_{2h} = 5{,}000$ records for each stratum. In practice, the NSI might use proportional allocation to set each n_{1h} and choose smaller sampling rates to set n_{2h}. We use a common sample size and large sampling fractions to illustrate that the variance formulas for subsampling with synthesis correctly handle non-trivial finite population correction factors, *e.g.*, 50% of the records are sampled in stratum 4.

We consider Y_4 and Y_5 to be the confidential variables and illustrate two synthesis scenarios. In the first, we synthesize all records' values of Y_4 and Y_5. To do so, in each stratum we simulate Y_{4h} using a regression of Y_{4h} on (Y_{1h}, Y_{2h}, Y_{3h}) estimated with D, and we simulate Y_{5h} using a regression of Y_{5h} on $(Y_{1h}, Y_{2h}, Y_{3h}, Y_{4h})$ estimated with D. Predictions of Y_{5h} are based on the synthesized values of Y_{4h}. In the second approach, in each stratum we replace Y_{4h} and Y_{5h} only for all records with $Y_{3h} > p_h$, where p_h is the 90^{th} percentile of Y_3 in the population in stratum h. We generate replacement values by sampling from regression models; however, the models in each stratum are estimated only with those records satisfying $Y_{3h} > p_h$.

For the different subsamples approach, we generate $m = 5$ synthetic surveys as outlined in Section 2.1. For the same subsample approach, we first draw $m = 5$ values of θ, the regression coefficients and variances. For each $\theta^{(l)}$, we generate $r = 5$ synthetic datasets for every first stage nest.

For all scenarios, we repeat the process of (i) creating D by sampling from the population and (ii) generating subsamples with synthesis a total of 5,000 times. For each of these 5,000 runs, we obtain inferences for fifty quantities, including the population means and within-stratum means of Y_4 and Y_5, the coefficients from a regression of Y_3 on all other variables, and the coefficients from a regression of Y_5 on all other variables. The regressions are estimated separately in each stratum.

Table 2
Parameters for drawing $(Y_2, \ldots, Y_5)$ for the population

	Stratum size	Model	Distribution of the error term
Stratum 1	750,000	$Y_2 = Y_1 + e$	$e \sim N(0, 5)$
		$Y_3 = Y_1 + Y_2 + e$	
		$Y_4 = Y_1 + Y_2 + Y_3 + e$	
		$Y_5 = Y_1 + Y_2 + Y_3 + Y_4 + e$	
Stratum 2	200,000	$Y_2 = 2Y_1 + e$	$e \sim N(0, 10)$
		$Y_3 = 2Y_1 + 0.5Y_2 + e$	
		$Y_4 = 2Y_1 + 0.5Y_2 + Y_3 + e$	
		$Y_5 = 2Y_1 + 0.5Y_2 + 0.5Y_3 - 0.25Y_4 + e$	
Stratum 3	40,000	$Y_2 = -3Y_1 + e$	$e \sim N(0, 30)$
		$Y_3 = -3Y_1 - 1.5Y_2 + e$	
		$Y_4 = -3Y_1 + Y_2 - 1/3Y_3 + e$	
		$Y_5 = -3Y_1 + Y_2 - 1/3Y_3 + 1/9Y_4 + e$	
Stratum 4	10,000	$Y_2 = -2Y_1 + e$	$e \sim N(0, 20)$
		$Y_3 = -Y_1 - 1.5Y_2 + e$	
		$Y_4 = -2Y_1 + Y_2 + 1/4Y_3 + e$	
		$Y_5 = 2Y_1 - Y_2 - 1/4Y_3 + 1/16Y_4 + e$	

Figure 1 displays key results of the simulations. The left panel displays the ratios of the simulated average of T_d (and T_s) over the corresponding simulated var($\bar{q}_m$) for the fifty estimands. The median ratios are close to one in all scenarios, and the averages of T_d (and T_s) never differ by more than 10% from their actual variances. Thus, both T_s and T_d appear to be approximately valid variance estimators.

The middle panel of Figure 1 summarizes the percentages of the 5,000 synthetic 95% confidence intervals based on T_d (and on T_s) that cover their corresponding Q. The coverage rates are close to 0.95 except for the regression coefficients for the same subsampling approach with 100% synthesis. For these coefficients, $T_s < 0$ in up to 38% of the simulation runs, so that confidence intervals are based on the conservative T_s^*. The highest fraction of negative variances occurs in the smallest stratum which has a sampling rate of 50%. All variance estimates are positive when only 10% of the records are synthesized.

The right panel of Figure 1 displays the ratios of the simulated root mean squared error (RMSE) of $\bar{q}_m$ over the simulated RMSE from the subsamples without any synthesis. For the same subsampling approach, the RMSEs of the synthetic subsamples tend to be smaller than the RMSEs based on the subsamples without any synthesis, particularly for the 100% synthesis. The smaller RMSEs result because the synthesis models are determined with D, *i.e.*, the survey data before taking the subsample, so that they carry additional information that is not in the subsamples without synthesis. For the different synthetic subsamples, the RMSE ratios typically exceed one. Here, increased synthesis leads to greater loss in efficiency. We note that the RMSEs from the different sample and same sample approaches in Figure 1 are not directly comparable because they are based on different denominators.

To enable comparisons across the methods, as well as to illustrate the losses in efficiency from subsampling, we repeat the simulation design using $m = 25$ for the independent subsamples approach and $mr = 25$ for the same subsamples approach. The left panel of Figure 2 displays the simulated RMSE ratios for the fifty estimands in the different scenarios, where the denominators are the average RMSEs based on the original data before any confidentiality protection. The right panel of Figure 2 displays the ratios of simulated average lengths of the 95% confidence intervals, where the denominators are the average lengths based on the original data before any confidentiality protection. Based on the left panel, for a given total number of released datasets and given synthesis percentage, the independent sample approach results in more efficient estimates than the same sample approach. The right panel tells a similar story, although it is harder to see because of the scaling. Here, the same sample approach with 100% synthesis results in high fractions of negative variance estimates, so that the adjusted variance T_s^* is often used, thereby inflating the interval lengths. Figure 2 also includes results from synthesis without any subsampling, which generally provides more efficient estimates than either subsampling approach.

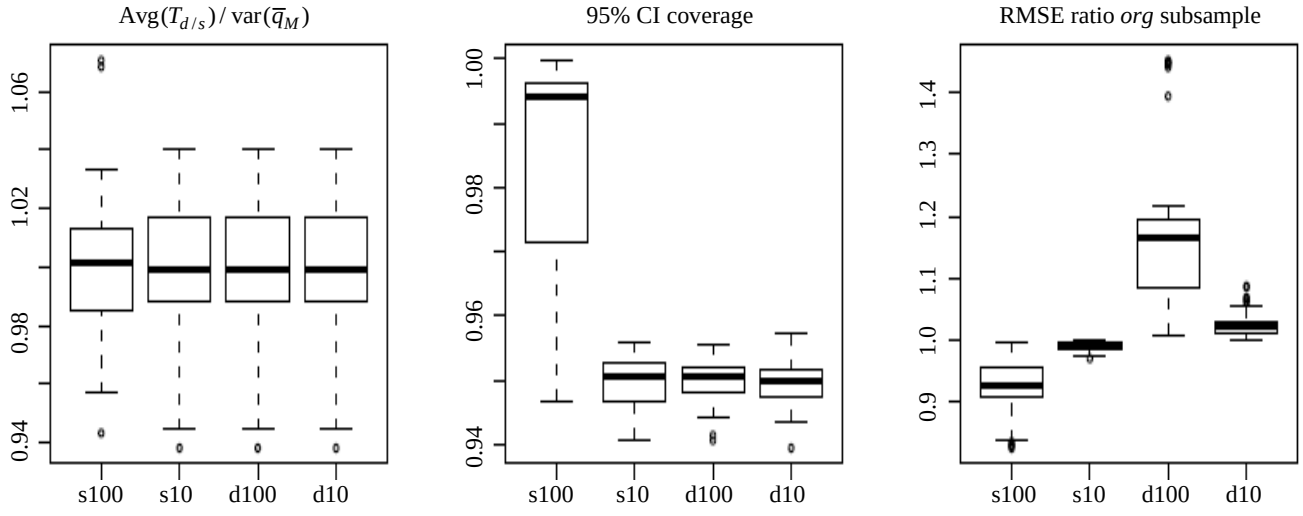

Figure 1 Simulation results for the stratified sampling design. In the labels, s and d indicate the same subsample and the different subsamples approach. The numbers indicate the percentage of records that are being synthesized. The denominators of the RMSE are based on the point estimates from the subsamples without synthesis. For the different subsamples approach, the RMSE is computed from the average of the m point estimates. Each box plot comprises fifty estimands

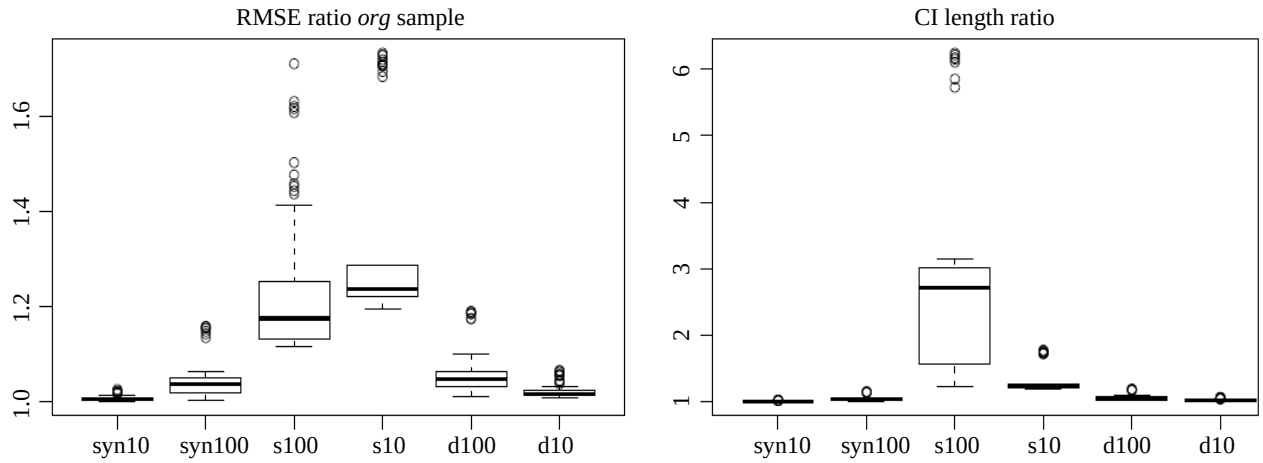

Figure 2 Efficiency comparisons for the stratified sampling design. In the labels, *org* and *syn* indicate the original sample and the synthetic sample before subsampling; and, s, d, and the numbers are as in Figure 1. The denominators of the RMSE are based on the point estimates from the original sample without synthesis. Each box plot comprises fifty estimands

4. Concluding remarks

The different subsamples and same subsamples approaches have competing advantages. For a fixed number of released datasets M, the different subsamples approach enables estimation with greater efficiency than the same subsamples approach - as evident in Figure 2 - since the released subsamples are independent rather than correlated. The different subsamples approach also guarantees positive variance estimates; the same subsample approach does not. However, with large M the different subsamples approach weakens the confidentiality protections of subsampling, since the combined datasets are likely to contain most of the records from the original survey. Hence, unless the subsampling rate is small (*e.g.*, 1% or 2%), the NSI may have to make m modest (*e.g.*, $m = 5$) to use the different subsamples approach. Because of this, the different samples approach is not viable when the original sample size is modest.

As an alternative to subsampling with synthesis, agencies could release partially synthetic data that include all records from the original sample, assuming that they are willing to release files of that size. Partial synthesis on the original data generally engenders estimates with lower variances than subsampling with synthesis - as evident in Figure 2 - since more records are released. However, partial synthesis on the original data generally engenders higher disclosure risks than subsampling with synthesis, since more at risk records are in the released data and since the additional protection from subsampling is absent. Agencies can compare the two options on disclosure risks using the methods of Drechsler and Reiter (2008), which account for the protection afforded by sampling, and on data utility by comparing inferences for representative analyses.

It is also possible that the process of subsampling may engender sufficient additional protection to enable lesser amounts of synthesis than would be necessary in a partial synthesis of the entire original dataset. Evaluating the data utility for subsampling with synthesis versus synthesis only for given disclosure risks is beyond the scope of this short note, but it is an interesting area for future research.

We have not developed subsampling with synthesis approaches for sampling designs other than (stratified) simple random samples. For the different subsamples approach, appropriate inferential methods require an approximately unbiased estimate of the variance from the first phase of sampling that can be computed from the subsample alone. This is elusive for complicated designs. For the same subsample approach, we conjecture that analysts can use the inferential methods presented in Section 2.2, provided that $\bar{u}_M$ appropriately accounts for the two phases of sampling. We note that the formulas for $\bar{w}_M$ and b_M remain the same for other designs. Evaluating this conjecture is a subject of future research.

Acknowledgements

This research was supported by U.S. National Science Foundation grant SES-0751671.

References

Drechsler, J., and Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer and Y. Saygin), New York: Springer, 227-238.

Drechsler, J., and Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25, 589-603.

Drechsler, J., and Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

Elliott, M., and Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39, 1101-1118.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.

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 identification risks in microdata. *Journal of the American Statistical Association*, 100, 1103-1113.

Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 933-946.

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

Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Tech. rep., Statistical Research Division, U.S. Bureau of the Census, Washington, DC.