WP. 18
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Improved Variance Estimation for Fully Synthetic Datasets

Prepared by Jörg Drechsler, Institute for Employment Research, Germany

Improved Variance Estimation for Fully Synthetic Datasets

Jörg Drechsler*

* Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg.
(joerg.drechsler@iab.de)

Abstract. Fully synthetic datasets, i.e. datasets that only contain simulated values, arguably provide a very high level of data protection. Since all values are simulated reidentification is almost impossible. This makes the approach especially attractive for the release of very sensitive data such as medical records. However, the established variance estimate for fully synthetic datasets has two major drawbacks. First, it can be positively biased, where the bias is a function of the sampling rate of the original data. Second, it can become negative.

In this paper I illustrate the negative effects of these drawbacks on the estimation of the variance and propose an alternative variance estimate that shows less variability, is always unbiased, and can never be negative. This variance estimate is closely related to the variance estimate for partially synthetic datasets.

1 Introduction

In our data driven world with increasing amounts of information collected on all of us through surveys, electronic health records, internet search logs, etc. adequate protection of privacy is a major concern. On the other hand there are obvious benefits from broad access to the collected data since political and economic decisions can be based on sound information. To balance these two competing objectives, a number of approaches have been suggested in the literature on statistical disclosure control to enable data dissemination without violating confidentiality restrictions (Willenborg and de Waal, 2001). While the first methods developed in the 1980s, such as data swapping (Dalenius and Reiss, 1982) and adding noise (see Brand (2002) for a review), mainly focused on disclosure protection and preserved only some univariate statistics such as the population mean and the variance of a single variable, more sophisticated methods such as post-randomization (Gouweleeuw *et al.*, 1998) or data shuffling (Muralidhar and Sarathy, 2006) have emerged in recent years. This methods explicitly try to optimize the trade-off between analytical validity and disclosure risk for the released datasets. A more radical approach was suggested by Rubin in 1993: generating multiply imputed synthetic datasets. With this approach the original records in the dataset are replaced by multiple synthetic versions, generated by repeatedly drawing from a model fit to the original data. Specifically, Rubin suggested to treat the survey variables for those units from the sampling frame that did not participate in the survey as missing data and multiply impute them according to the multiple imputation framework (Rubin, 1978). Simple random samples from these fully imputed populations are then released to the public. Datasets generated based on this approach are now called fully synthetic datasets in the literature to distinguish the approach from the partially synthetic approach (Little, 1993) for which only sensitive variables and/or variables that bear a high risk of disclosure are replaced with synthetic values (see Drechsler (2011) for a full review of the different approaches to generating synthetic datasets). Since no actual values are released if the fully synthetic approach is applied and the data are generated for units that never actually participated in the survey this approach offers a very high level of data protection and thus is especially attractive for very sensitive data such as medical data. Despite these attractive features, applications of the approach in practice are still limited. Drechsler *et al.* (2008b) apply the approach to a German establishment survey and compare the results obtained from this dataset with results obtainable from partially synthetic datasets (Drechsler *et al.*, 2008a). Reiter (2005a) provides a simulation study that illustrates the advantages but also potential limitations of the approach. Abowd and Vilhuber (2008) suggest how to generate datasets that fulfill ϵ-differential privacy (Dwork, 2006) based on the fully synthetic approach. The synthetic version of the Longitudinal-Business-Database (LBD) (Kinney *et al.*, 2011) can also be considered a fully synthetic dataset since all variables are replaced with synthetic versions. However there is only one synthetic replicate available at the moment and since the LBD is a census, there is no additional protection from generating synthetic values for units that were not included in the original data. Other applications of the fully synthetic approach can be found in Graham *et al.* (2009); Sakshaug and Raghunathan (2010), and Yu (2008).

Raghunathan *et al.* (2003) developed procedures for obtaining valid inferences from the multiple synthetic datasets. These procedures that are based on combining the point and variance estimates from each synthetic dataset are closely related but differ slightly from the combining rules for multiple imputation for nonresponse (Rubin, 1987). The combining procedures derived in the paper are based on the assumption that the synthetic populations are generated by multiply imputing all the values for those units that did not participate in the survey as proposed by Rubin (1993). Thus, the synthetic populations consist of a combination of a large fraction of imputed values and a small fraction of the originally observed records for the survey respondents. This means that there is a small chance that the released samples from these populations still contain some original records. To avoid this, the authors suggest that "the whole population can be generated based on the posterior predictive distribution of "super" or "future" populations" (Raghunathan et al., 2003, p. 4). Since arguably the main advantage of fully synthetic datasets is based on the notion that all records in the released files are synthetic, all extensions of the original method (Reiter and Drechsler, 2010; Reiter, 2005b; Drechsler and Reiter, 2010) refer to this idea and all the applications of the fully synthetic approach generate datasets that only contain synthetic values. However, the derivations presented in Raghunathan *et al.* (2003) are strictly valid only under the original idea for generating synthetic datasets. If all values in the synthetic data are synthesized, the derivations are only correct under the assumption that the underlying population is infinite since the probability that any original value will be included in the released sample from the synthetic population will tend to zero in this case. In all other cases the variance estimate for the fully synthetic data will be positively biased, where the bias is a function of the sampling rate of the original data.

Another important disadvantage of the variance estimate presented in Raghunathan *et al.* (2003) is that it can be negative. Reiter (2002) proposes an alternative variance estimate that is always positive. However, this variance estimate is conservative and thus adds to the positive bias already present in the original estimate.

In this paper, I present an alternative variance estimate that closely resembles the variance estimate for partially synthetic datasets with the main difference that it adjusts for potentially different sample sizes between the original sample and the synthetic sample. The estimate is always unbiased regardless whether all records are synthesized or whether the originally proposed approach is used. Additionally it shows less variability and can never be negative. The remainder of the paper is organized as follows. In Section 2 I review the inferential methods that were suggested in Raghunathan *et al.* (2003) to obtain inferences from fully synthetic datasets. In Section 3 I derive an alternative variance estimate that overcomes the drawbacks of the originally proposed estimate. Section 4 contains a simulation study that illustrates the drawbacks of the original estimate and the validity of the proposed estimate under various settings. The paper concludes with some final remarks.

2 Traditional procedures for inference from fully synthetic datasets

In their paper, Raghunathan *et al.* (2003) derive the inferential procedures to obtain valid inferences from fully synthetic datasets generated using the design proposed by Rubin (1993). To understand the procedure of analyzing fully synthetic datasets generated under this design, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the mean of a variable, the correlation coefficient between two variables, or a regression coefficient in a linear regression. Inferences for this parameter derived from the original datasets usually are based on a point estimate q, an estimate for the variance of q, u, and a normal or Student's t reference distribution. For analysis of the imputed datasets, let $q^{(i)}$ and $u^{(i)}$ for

$i = 1, ..., m$ be the point and variance estimates for each of the m synthetic datasets. The derivations in Raghunathan *et al.* (2003) implicitly assume that $u^{(i)}$ contains a finite population correction factor (fpc) if necessary. Since the fpc will be important for the alternative variance estimate proposed in this paper, I will assume that $u^{(i)}$ doesn't contain the fpc and explicitly include it in the formulae where necessary. The following quantities are needed for inferences for scalar Q:

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

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

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

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

$$T_f = (1 + m^{-1})b_m - \delta\bar{u}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$, where $\delta = (1 - n_{syn}/N)$ is the finite population correction factor and n_{syn} is the number of records in the released datasets sampled from the synthetic populations. When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_f = (m-1)(1 - \delta\bar{u}_m/((1+m^{-1})b_m))^2$. As discussed previously, a disadvantage of this variance estimate is that it can become negative. For that reason, Reiter (2002) suggests a slightly modified variance estimator that is always positive, $T_f^* = \max(0, T_f) + \gamma(\frac{n_{syn}}{n_{org}}\delta\bar{u}_m)$, where $\gamma = 1$ if $T_f < 0$ and $\gamma = 0$ otherwise. Here, n_{org} is the number of records in the original sample.

3 Alternative procedures for inference from fully synthetic datasets

In this section I suggest an alternative variance estimate for fully synthetic datasets that closely resembles the variance estimate for partially synthetic datasets, $T_p = \delta\bar{u}_m + b_m/m$, where $\delta_{org} = (1 - n_{org}/N)$. The point estimate $\bar{q}_m$ is valid irrespective whether all records in the population or only the "missing" records in the population are imputed. The improved variance estimate is given by:

$$T_{alt} = \delta_{org}\frac{n_{syn}}{n_{org}}\bar{u}_m + b_m/m. \tag{5}$$

When n_{syn} is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_{alt} = (m-1)(1 + m\delta_{org}n_{syn}\bar{u}_m/(n_{org}b_m))^2$.

To motivate this alternative variance estimate, we first note that the variance estimate for partially synthetic datasets T_p remains the same irrespective of the fraction of records that are synthesized in the dataset. Thus, in the extreme case we can synthesize all records in the original data and still obtain valid results as long as the sample size remains constant, i.e. $n_{syn} = n_{org}$. We also note that T_{alt} boils down to T_p for $n_{syn} = n_{org}$.

Changing the sample size compared to the original sample size might be beneficial for two reasons: Records that are outliers in the original sample might also show up as outliers in the synthetic sample and an intruder is tempted to assume that he or she identified a specific record if there is a single outlier that shows comparable features to the outlier in the original data. If the sample size increases, it is more likely that records with similar features show up in the synthetic data. Generating large samples can also be advantageous for variables with very skewed distributions for example binary variables for which almost all of the outcomes are either 1 or 0. If only a small number of synthetic datasets is generated and the sample size is small it can happen that non of the rare outcomes will show up in any of the synthetic datasets. This can be avoided by increasing the sample size.

To understand the adjustment of T_p if the sample size of the released data differs from the sample size of the original data, we need to review the derivations for partially synthetic datasets. Generally, the analyst of the synthetic data will be interested in $f(Q|d_{syn})$, where d_{syn} is the set of m released synthetic datasets and Q is the parameter of interest. From Reiter (2003) we know that we can decompose this expression as

$$f(Q|d_{syn}) = \int f(Q|d_{org})f(d_{org}|d_{syn}, B)f(B|d_{syn})dd_{org}dB, \tag{6}$$

where d_{org} is the original sample and $B = Var(q_i|d_{org}, B)$. We assume that large sample approximations hold so that

$$f(Q|d_{org}) \sim N(q_{org}, \delta_{org}u_{org}), \tag{7}$$

where q_{org} and u_{org} are the point estimate and its estimated variance that the analyst would have used, if the original data would have been available. Since we use large sample approximations it is sufficient to determine $f(q_{org}, u_{org}|d_{syn}, B)$ for $f(d_{org}|d_{syn}, B)$. As is standard in the multiple imputation context, we assume that imputations are made so that

$$f(q_i|d_{org}, B) \sim N(q_{org}, B) \tag{8}$$

$$(u_i|D, B) \sim \left(\frac{n_{org}}{n_{syn}}u_{org}, \ll B\right) \tag{9}$$

Note that the factor n_{org}/n_{syn} is an extension compared to the derivations presented in Reiter (2003) that accounts for the potentially different sample sizes of the original and the synthetic dataset. This extension should be valid for all $\sqrt{N}$ consistent estimators, such as the mean or (approximately) the regression coefficient in a linear regression under simple random sampling. Since we assume negligible variance for u_i, we have $u_i \approx \bar{u}_m \approx \frac{n_{org}}{n_{syn}} u_{org}$. Assuming uninformative priors for q_{org} and u_{org} implies that

$$\begin{aligned}
(q_{org}|d_{syn}, B) &\sim N(\bar{q}_m, B/m) \qquad (10)\\
(u_{org}|d_{syn}, B) &\sim \left(\frac{n_{syn}}{n_{org}}\bar{u}_m, \ll B/m\right) \qquad (11)
\end{aligned}$$

From this, we can derive the posterior mean and variance of $(Q|d_{syn}, B)$ as

$$\begin{aligned}
E(Q|d_{syn}, B) &= E(E(Q|d_{org})|d_{syn}, B) = \bar{q}_m \qquad (12)\\
Var(Q|d_{syn}, B) &= E(Var(Q|d_{org})|d_{syn}, B) + Var(E(Q|d_{org})|d_{syn}, B)\\
&= \delta_{org}\frac{n_{syn}}{n_{org}}\bar{u}_m + B/m \qquad (13)
\end{aligned}$$

To obtain $f(Q|d_{syn})$ we would need to integrate $f(Q|d_{syn}, B)$ over $f(B|d_{syn})$. Instead we use the approximations presented in Rubin (1987, pp. 90–92) to obtain

$$(Q|d_{syn}) \sim t_{\nu_{alt}}(\bar{q}_m, T_{alt}) \qquad (14)$$

The degrees of freedom ν_{alt} are obtained by matching the first two moments of $\delta_{org}\frac{n_{syn}}{n_{org}}\bar{u}_m + B/m$ to a mean-square random variable.

4 Simulation studies

To evaluate the validity of the alternative variance estimate and to illustrate the drawbacks of the original variance estimate I use a repeated simulation design. For the simulation I generate a population of $N = 10,000$ records consisting of one standard normal variable Y. I repeatedly draw simple random samples of different sample size (1%, 5%, 10%, and 20%) from this population and consider this fraction of the data as the original survey data. Assuming uninformative priors, synthetic versions of these survey data can be generated in three steps.

1. Draw $\sigma^2 \sim (n-1)\hat{\sigma}^2\chi^{-2}_{n-1}$.
2. Draw $\mu \sim N(\bar{y}, \sigma^2/n)$.
3. Draw $y_{syn} \sim N(\mu, \sigma^2)$,

where $\hat{\sigma}^2$ is the variance of Y estimated from the survey sample. To illustrate the difference between the originally proposed approach and the way fully synthetic datasets are generated in practice, I generate two different versions of synthetic data. In the first version, I generate a synthetic population by appending the n

records form the original sample with $N - n_{org}$ imputed values for the originally "missing" values of the population. In the next step I draw $n_{syn} = 2n_{org}$ records from this synthetic population, i.e. the sample size of the synthetic datasets is always twice the sample size of the original sample. These are the records that should be released to the public. I call this the Raghunathan-Reiter-Rubin (RRR) approach. In the second version, I draw n_{syn} directly as outlined above. Since this is the approach that would normally be applied in practice, I call this the practical approach. Using both approaches I generate $m = 5, 20, 100$ synthetic datasets from each original sample. I assume the analyst is interested in estimating the mean of Y from the synthetic data. I compute the estimated variance for the synthetic sample mean based on the original variance estimate T_f and on the alternative variance estimate T_{alt} under both synthesis designs. For the simulation study I repeat the whole process of sampling from the population, generating synthetic datasets, and analyzing the synthetic datasets 5,000 times. Results for different m and different sample sizes are presented in Figure 1 and Tables 1 and 2. Figure 1 presents box plots of the ratio of the estimated variance over the true variance of $\bar{q}_m$ across the 5,000 simulation runs. In each panel, these ratios are computed (from left to right) for

- the originally proposed variance estimate applied to the synthetic data generated based on the originally suggested design (T_f^{RRR}),
- the originally proposed variance estimate applied to the synthetic data generated based on the design used in practice(T_f^{prac}),
- the alternative variance estimate applied to the synthetic data generated based on the originally suggested design (T_{alt}^{RRR}),
- the alternative variance estimate applied to the synthetic data generated based on the design used in practice (T_{alt}^{prac}).

Ideally, the box plots should be centered around one with small variability. From the box plots it is obvious that the alternative variance estimate should be preferred even if the originally proposed variance estimate would be unbiased since the alternative estimate shows far less variability than the original estimate for all simulation settings especially for small m for which the estimated variance can be more than 10 times the true variance in some of the simulation runs for the original variance estimate. The panels also illustrate the risk of negative variance estimates for T_f. Only for $m = 100$ we do not obtain any negative variance estimates in any of the simulation runs. Finally, the panels illustrate the positive bias of T_f for the synthesis design that is usually applied in practice once the sampling rate is 5% or higher. As the sampling rate increases so does the bias. The risk of negative variance estimates and the potential bias of T_f are further illustrated in Table 1. The table contains

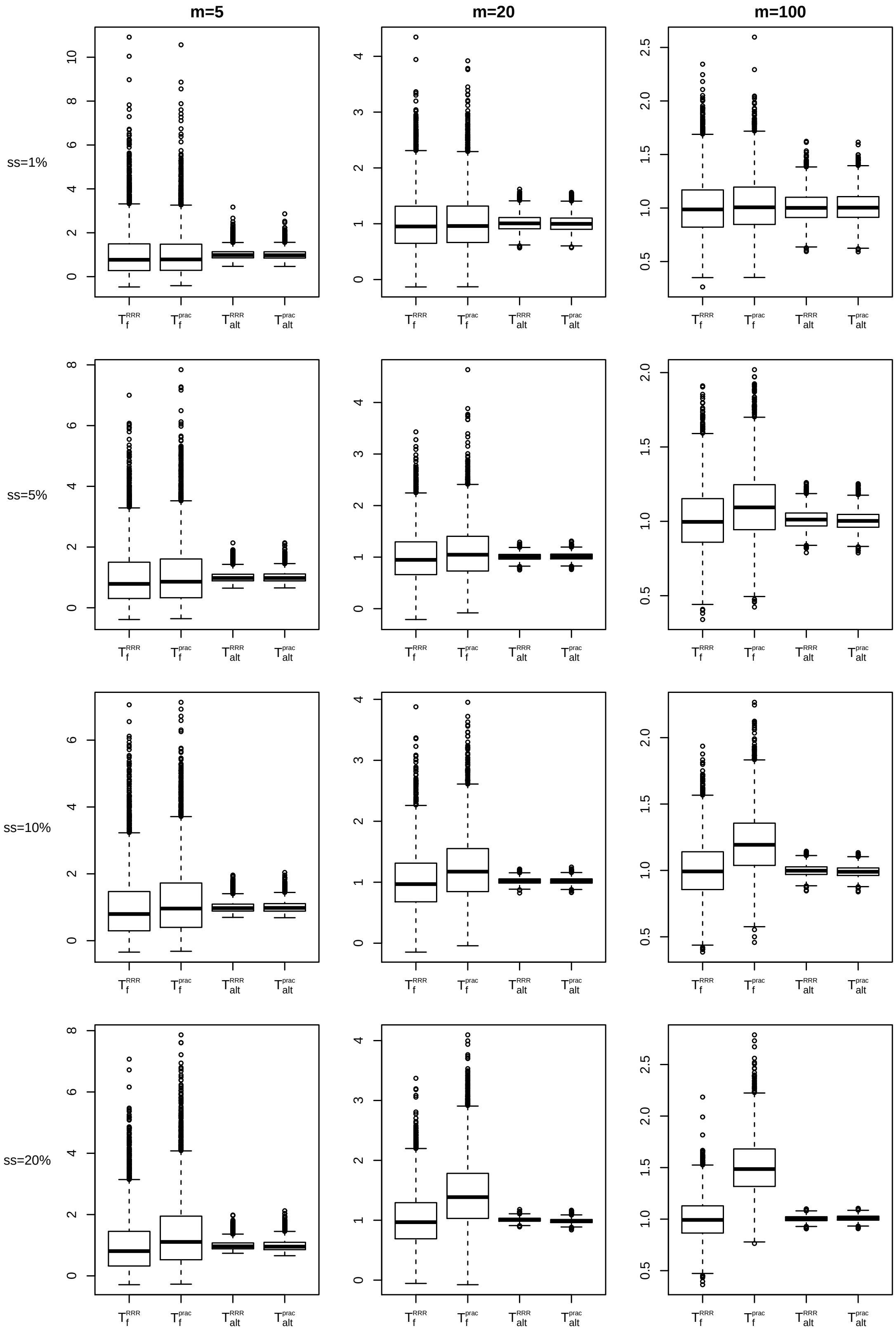

Figure 1: Ratio of the estimated variance over the true variance of $\bar{q}_m$ across the 5,000 simulation runs for different sample sizes (ss) and different number of imputations m.

Table 1: Simulation results based on originally proposed variance estimate

samp.	m	$(\bar{T}_f)/var(\bar{q}_m)$		$\%T_f < 0$		$(\bar{T}^*_f)/var(\bar{q}_m)$		CI cover (T^*_f)	
frac.		RRR	prac	RRR	prac	RRR	prac	RRR	prac
1%	5	1.03	1.02	10.26	10.06	1.12	1.11	98.72	98.80
	20	1.02	1.03	0.20	0.22	1.02	1.03	97.02	96.70
	100	1.01	1.03	0.00	0.00	1.01	1.03	94.96	95.06
5%	5	1.01	1.09	10.30	8.98	1.10	1.17	98.68	98.76
	20	1.01	1.10	0.20	0.10	1.01	1.10	96.32	97.02
	100	1.01	1.11	0.00	0.00	1.01	1.11	95.54	96.28
10%	5	1.02	1.19	9.60	7.34	1.09	1.25	98.66	98.58
	20	1.03	1.23	0.16	0.04	1.03	1.23	96.56	97.50
	100	1.00	1.21	0.00	0.00	1.00	1.21	94.40	96.50
20%	5	1.01	1.37	7.78	4.66	1.06	1.40	98.60	99.24
	20	1.02	1.45	0.06	0.02	1.02	1.45	96.00	97.86
	100	1.00	1.50	0.00	0.00	1.00	1.50	95.24	98.66

the ratio of the average estimated variance over the true variance before (T_f) and after (T^*_f) adjustments for negative variance estimates. Furthermore, the fraction of negative variance estimates for T_f and the coverage rates of the 95% confidence intervals for T^*_f are reported. The coverage rate represents the percentage of the 5,000 synthetic 95% confidence intervals that cover the true population mean. All results are presented for the originally proposed synthesis design (RRR) and for the design used in practice (prac). We note that T_f is unbiased if the RRR design is used. However, for $m = 5$ the fraction of negative variance estimates varies between 7.8% and 10.3% leading to an adjusted variance estimate T^*_f that is 6% to 12% too large even under the RRR design. As a consequence coverage rates vary around 98.8% for small m even though the original estimate T_f is unbiased. For the synthesis design that is usually used in practice T_f is biased for most scenarios. Only if the sampling fraction is 1% no bias can be observed. Once the sampling rate increases, the true variance is overestimated. This overestimation increases from 10% for a 5% sample to up to 50% for a 20% sample. As a consequence all coverage rates are too high even if a high number of synthetic datasets ($m = 100$) guarantees that no negative variance estimates occur. Table 2 presents the same results based on the alternative variance estimate. The variance estimates are unbiased for all scenarios leading to actual coverage rates that are very close to there nominal coverage.

I also evaluated the effect of changing the synthetic sample size relative to the original sample size. Increasing the synthetic sample size will reduce the risk of negative variance estimates and stabilize the original variance estimate. It has only small effects on the alternative variance estimates, since only the variance between the datasets b_m is reduced if the sample size is increased. The variance reducing effect is negligible for the alternative variance estimate once $m = 20$ or more datasets are generated. But even if the synthetic sample is four times larger than the original sample, the alternative variance estimate still shows far less variability and up to 4.6% of the original variance estimates are negative for $m = 5$.

Table 2: Simulation results based on alternative variance estimate

samp.	m	$(\bar{T}_{alt})/var(\bar{q}_m)$		CI cover (T_{alt})	
frac.		RRR	prac	RRR	prac
1%	5	1.02	1.01	94.86	94.54
	20	1.01	1.01	95.04	94.72
	100	1.01	1.01	94.82	94.86
5%	5	1.01	1.02	95.24	95.08
	20	1.01	1.01	95.28	95.10
	100	1.01	1.00	95.50	95.50
10%	5	1.01	1.02	94.88	94.96
	20	1.02	1.02	95.24	95.10
	100	1.00	0.99	94.56	94.56
20%	5	1.00	1.00	94.64	95.16
	20	1.01	0.99	95.08	95.06
	100	1.00	1.01	95.56	95.58

5 Conclusions

Releasing fully synthetic datasets can be a viable data dissemination strategy for highly sensitive data for which traditional SDC methods do not offer sufficient protection. If all records in the released dataset are synthetic the risk of disclosing sensitive information is very low. However, with the initially proposed strategy to generate these datasets there is a small chance that some originally observed records are still included in the released files since only those units that did not participate in the survey are synthesized. Even though it has been argued that all records in the population could be synthesized, the traditional point and variance estimates for fully synthetic datasets are derived under the assumption that the originally observed records remain unchanged. In this paper I illustrate that this can lead to a biased variance estimate if the sampling rate of the original sample is high (5% or higher). I derive an alternative variance estimate that has three advantages over the originally proposed variance estimate. First, it is always unbiased irrespective whether all records or only those records that did not participate in the survey are synthesized. Second, it can never be negative. Third it has less variability than the originally proposed variance estimate. The potential bias for the original variance estimate might be negligible in many practical situations where sampling rates are 1% or even lower and negative variance estimates can be avoided by increasing m although Drechsler *et al.* (2008a) obtained some negative variance estimates even for $m = 100$. Still, the alternative estimate shows far less variability in the simulations presented in this paper and there are no obvious drawbacks from using this estimate instead of the original one. However, it must be noted that the derivations presented in this paper are based on the assumption that a simple random sample was used for the original sample and for the synthetic sample. Most surveys are conducted under more complex sampling designs. Generalizing the results presented here for other sampling designs is an important area for future research. Furthermore, the variance estimate is only valid for $\sqrt{N}$ consistent estimates. If the estimate converges with a different rate, the variance estimate has to be adjusted. However, these adjustments should be straightforward if the rate of convergence is known.

References

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 239–246. New York: Springer.

Brand, R. (2002). Microdata protection through noise addition. In *Inference Control in Statistical Databases, From Theory to Practice*, 97–116, London, UK, UK. Springer-Verlag.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control–Theorey and Implementation.* New York: Springer.

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy* **1**, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008b). A new approach for disclosure control in the IAB Establishment Panel – multiple imputation for a better data access. *Advances in Statistical Analysis* **92**, 439–458.

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.

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *ICALP 2006*, 1–12. New York: Springer.

Gouweleeuw, J., Kooiman, P., Willenborg, L., and de Wolf, P. P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* **14**, 463–478.

Graham, P., Young, J., and Penny, R. (2009). Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. *Journal of Official Statistics* **25**, 407–426.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. Tech. rep., Center for Economic Studies (CES), CES-WP-11-04.

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

Muralidhar, K. and Sarathy, R. (2006). Data shuffling–a new masking approach for numerical data. *Management Science* **52**, 658–670.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

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

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica* **20**, 405–421.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 20–34. Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Sakshaug, J. and Raghunathan, T. E. (2010). Synthetic data for small area estimation. In J. Domingo-Ferrer and E. Magkos, eds., *Privacy in Statistical Databases*, vol. 6344 of *Lecture Notes in Computer Science.* Springer.

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

Yu, M. (2008). *Disclosure Risk Assessments and Control.* Ph.D. thesis, University of Michigan, Program in Survey Methodology.