

# Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data

Robin Mitra and Jerome P. Reiter

Duke University, Durham, NC 27708, USA.  
{rm51, jerry}@stat.duke.edu  
<http://www.stat.duke.edu>

**Abstract.** Statistical agencies alter values of identifiers to protect respondents' confidentiality. When these identifiers are survey design variables, leaving the original survey weights on the file can be a disclosure risk. Additionally, the original weights may not correspond to the altered values, which impacts the quality of design-based (weighted) inferences. In this paper, we discuss some strategies for altering survey weights when altering design variables. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the quality of inferences based on the different strategies.

**Key words:** Disclosure; Multiple imputation; Swapping; Synthetic data; Weights

## 1 Introduction

Survey design variables often contain identifying information, for example race in a survey that over-samples minorities or establishment size in a probability proportional to size sample of businesses. To limit disclosure risks, statistical agencies may need to alter these variables before releasing the data to the public. It also may be necessary to alter the survey weights, which typically are deterministic functions of the design variables. Failure to do so can leave identifying information on the file, effectively defeating the purpose of the masking [1]. For example, an unaltered weight could reveal that a person was part of a minority group or could disclose the size of the establishment. Not altering weights also could affect the quality of data analysts' estimates, because the weights may not be appropriate for making the released sample representative of the population.

In this paper, we discuss some strategies for adjusting survey weights when altering design variables to limit disclosure risks. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the data quality and confidentiality of the different strategies. We also examine the performance of the strategies when swapping identifiers.

## 2 Partially Synthetic Data and Weights

We first review partially synthetic data. Then, we describe some strategies to adjust weights when replacing design variables with synthetic values.

### 2.1 Partial Synthesis

Partially synthetic data comprise the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. Releasing partially synthetic data can preserve confidentiality, since identification of units and their sensitive data can be difficult when some released data are not actual, collected values. Furthermore, using appropriate data generation and estimation methods [2]—based on the concepts of multiple imputation [3] for missing data—analysts can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software, at least for inferences congenial to the model used to generate the data. Provided the agency releases some description of this model, analysts can determine whether or not their questions can be answered using the synthetic data. See [4] and [5] for genuine applications of partially synthetic data.

Following the derivations of [2], we assume that the agency synthesizes some design variables,  $X$ , based on the observed data,  $D = (X, Y_{obs})$ , by drawing new values from the Bayesian posterior predictive distribution of  $(X|D)$ . Imputations are made independently for  $i = 1, \dots, m$  times to yield  $m$  different synthetic data sets. These synthetic data sets are released to the public.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the analyst, seeks inferences about some estimand  $Q$ . In each synthetic data set  $d_i$ , the analyst estimates  $Q$  with some point estimator  $q$  and estimates the variance of  $q$  with some estimator  $v$ . For  $i = 1, \dots, m$ , let  $q_i$  and  $v_i$  be respectively the values of  $q$  and  $v$  in synthetic data set  $d_i$ . The analyst can obtain valid inferences for scalar  $Q$  by using the following quantities:

$$\bar{q}_m = \sum_{i=1}^m q_i/m \tag{1}$$

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

$$\bar{v}_m = \sum_{i=1}^m v_i/m. \tag{3}$$

The analyst can then use the  $\bar{q}_m$  to estimate  $Q$  and  $T_p = b_m/m + \bar{v}_m$  to estimate the variance associated with  $\bar{q}_m$ . For large sample sizes, inferences for scalar  $Q$  can be based on t-distributions with degrees of freedom  $\nu_p = (m-1)(1 + r_m^{-1})^2$ , where  $r_m = (m^{-1}b_m/\bar{v}_m)$  [2].

## 2.2 Survey Weights in Partial Synthesis

In complex surveys, it is well known that analyses that fail to account for the survey design variables can yield biased inferences [6] [7]. To incorporate the design, analysts can use survey-weighted estimation, where the survey weight  $w_i$  for unit  $i$  equals the inverse of the unit's inclusion probability, multiplied possibly by adjustments for nonresponse and calibration. For example, a common survey-weighted estimator of the population mean of  $Y$  based on the sample  $S$  is

$$\bar{y}_w = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}. \quad (4)$$

Weighted estimates exist for regression coefficients, as well as for the variances of these estimators.

When synthesizing some sampling design variables  $X$ , it is necessary to adjust weights to reflect the new values. We consider two approaches: (i) recalculate the weights (RCAL) to be consistent with the synthetic values, effectively making the synthetic sample representative of the population, and (ii) copy and paste (CPP) the original weights of records whose original design variables match the synthetic ones. The RCAL method preserves some properties of the original sampling weights that CPP does not; for example, the sum of the RCAL weights equals the sum of the observed weights, whereas the sums are not necessarily equal for the CPP weights. Additionally, the CPP cannot be applied unless exact matches are available.

## 3 Simulation Studies

In this section, we use simulation studies to investigate the implications for data quality of using survey-weighted analyses based on weights from the RCAL and CPP methods. For comparisons, we also consider using unweighted (UNW) estimates and survey-weighted estimates based on the old weights (OLDW). The simulations include stratified sampling, probability proportional to size sampling, and two stage cluster sampling. We also apply the procedures on genuine data from the Survey of Youth in Custody [6]. Unless stated otherwise, all estimates and standard errors are calculated using the survey package in the  $R$  statistical software.

### 3.1 Stratified Sampling Simulation

We first generate a stratified population of size 20000. The four strata are formed by crossing two binary variables,  $X_1$  and  $X_2$ . There are approximately 1000, 2000, 10000, and 7000 units in stratum one through four, respectively. We generate two survey variables,  $Y_1$  and  $Y_2$ , from the following distributions:

$$y_{1h} \sim N(\mu_h, \sigma_h^2) \quad (5)$$

$$y_{2h} \sim N(\alpha_h + \beta_h y_{1h}, \tau_h^2) \quad (6)$$

where  $\mu_h$ ,  $\sigma_h^2$ ,  $\alpha_h$ ,  $\beta_h$ , and  $\tau_h^2$  differ for each stratum  $h$ . The observed data are a stratified sample of 250 units from each stratum, so that the weight for all observations in stratum one equals 4, in stratum two equals 8, in stratum three equals 40, and in stratum four equals 28.<sup>1</sup>

For each sample, we synthesize  $(X_1, X_2, Y_2)$  from their joint posterior predictive distribution conditional on  $Y_1$ , which remains unaltered. To do so, we first simulate replacement values for  $X_1$  by using a logistic regression conditional on  $Y_1$ . Second, we simulate replacement values for  $X_2$  by using a logistic regression on  $(Y_1, X_1)$ , using the synthetic  $X_1$  for predictions. Third, we simulate replacement values for  $Y_2$  using a linear regression conditional on  $(Y_1, X_1, X_2)$ , using the synthetic  $X_1$  and  $X_2$  for predictions. We repeat this process independently  $m = 5$  times to obtain five partially synthetic datasets for each  $D$ .

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of  $Y_1$  and  $Y_2$ ; the percentages of values of  $Y_1$  greater than the population 50th, 80th and 95th percentiles, and likewise for  $Y_2$ ; the two regression coefficients from the linear regression of  $Y_2$  on  $Y_1$ , the four regression coefficients from the linear regression of  $Y_2$  on  $(Y_1, X_1, X_2)$ ; and, the eight regression coefficients from the linear regression of  $Y_2$  on  $(Y_1, X_1, X_2)$  and their interactions. The synthetic 95% confidence intervals are based on the methods in Section 2.1, with  $\bar{v}_m$  equal to the design-based variance estimate as computed in the *R* software. Because of how *R* computes variances, the  $\bar{v}_m$  is the same for the RCAL and CPP methods, although the point estimates of  $Q$  differ.

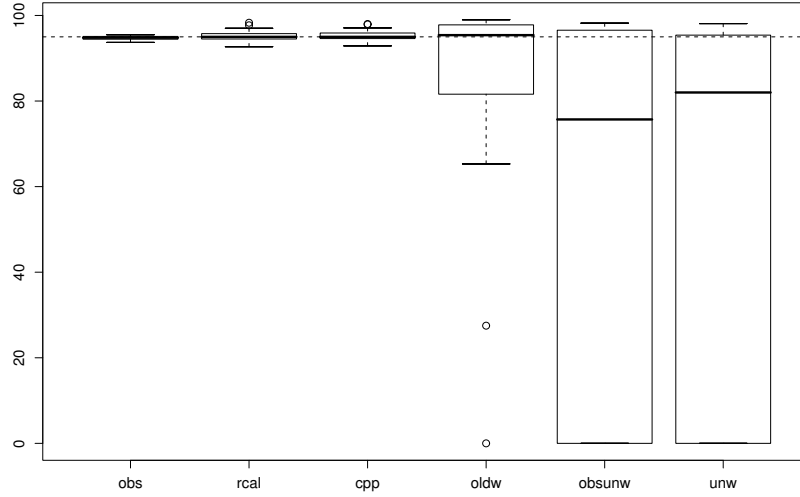
To illustrate RCAL and CPP in this setting, suppose one synthesized dataset comprises 200 records in each of stratum one and stratum two, and 300 records in each of stratum three and stratum four. Using RCAL, the new weight of all records in stratum one equals 5, in stratum two equals 10, in stratum three equals 100/3, and in stratum four equals 70/3. Using CPP, the weight of all records in stratum one remains at 4, in stratum two remains at 8, in stratum one remains at 40, and in stratum four remains at 28.

Figure 1 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. As expected, the coverage rates based on the observed data (OBS) are around 95%. Those based on RCAL and CPP also are near 95%. Coverage rates based on method OLDW do not match those based on the observed data. OLDW is particularly problematic for analyses involving  $Y_2$ . Coverage rates based on method UNW are too low for the means and proportions. This is not surprising, since unweighted means and percentages are known to be biased in unequal probability samples. Method UNW does provide coverage rates like those for unweighted analyses based on the observed data (OBSUNW).

As a check on the amount of alteration in the strata, for each sample we compare the modes of the  $m$  imputed values of the records' synthetic stratum indicators to their actual stratum indicators in the observed data. Approximately

---

<sup>1</sup> The weights actually are slightly different from the integer values because the strata sizes are not precisely 1000, 2000, 10000, and 7000.



**Fig. 1.** Box plots of coverage rates for the twenty-three estimands in the stratified sampling simulation. The coverage rates based on *RCAL* or *CPP* are closest to those based on the observed data.

45% of records can be placed in their original stratum by using this strategy, indicating a sizeable number of re-allocations of stratum memberships.

### 3.2 PPS Sampling Simulation

We generate a population of size 20000 in which the design variable  $X$  is a size variable. We generate  $X$  from a log-normal distribution and add a constant so that all values are far from zero. The minimum size in the population equals 50, and the maximum size equals 412. The total of the size values equals 1,328,252. We then generate the survey variables  $Y_1$  and  $Y_2$  from

$$y_1 \sim N(1.3x, 32^2) \quad (7)$$

$$y_2 \sim N(1.2x + 0.9y_1, 32^2). \quad (8)$$

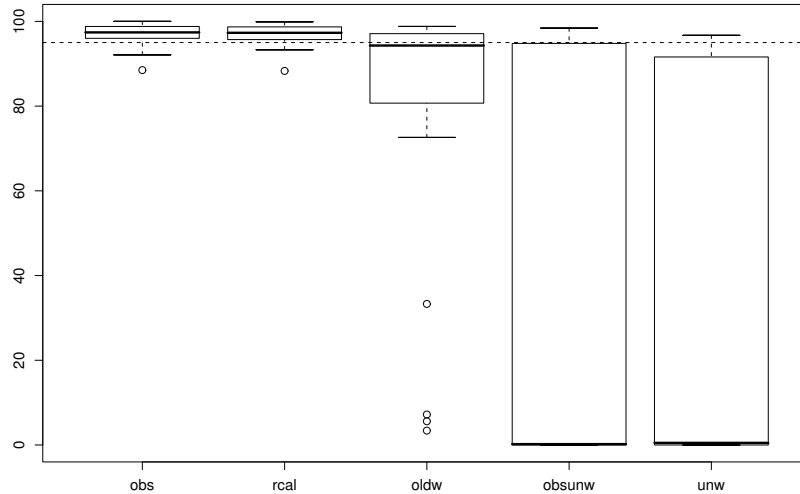
This results in correlations between  $X$  and  $Y_1$  of 0.63, between  $X$  and  $Y_2$  of 0.76, and between  $Y_1$  and  $Y_2$  of 0.82. We sample 1000 records from this population with probability proportional to the size variable  $X$  using the Hartley-Rao algorithm [8] For any observation  $i$ , the weight is  $w_i = 1328252/(1000x_i)$ .

We synthesize  $(X, Y_2)$  from their joint Bayesian posterior predictive distribution conditional on  $Y_1$ , which remains unaltered. To do so, we first simulate replacement values for  $X$  by using a generalized additive model (GAM) condi-

tional on  $Y_1$ .<sup>2</sup> Second, we simulate replacement values for  $Y_2$  by using a linear regression on  $(Y_1, X)$ , using the synthetic  $X$  for predictions. We repeat this process independently  $m = 5$  times to obtain five partially synthetic datasets for each  $D$ .

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of  $X$ ,  $Y_1$ , and  $Y_2$ ; the percentages of values of  $Y_1$  greater than the population 50th, 80th and 95th percentiles, and likewise for  $X$  and  $Y_2$ ; the six regression coefficients from the linear regression of  $Y_2$  on  $Y_1$ ,  $Y_2$  on  $X$ , and  $Y_1$  on  $X$ ; and, the three regression coefficients from the linear regression of  $Y_2$  on  $(Y_1, X)$ . The synthetic 95% confidence intervals are based on the methods in Section 2.1, with  $\bar{v}_m$  equal to the design-based variance estimate as computed in the *R* software.

The CPP method is not applicable here, because the simulated sizes do not match exactly with original sizes. The RCAL simply involves plugging in the synthesized values of the  $x_j$  in  $1328252/(1000x_j)$ .



**Fig. 2.** Box plots of coverage rates for the twenty-one estimands in the PPS simulation. The coverage rates based on *RCAL* are closest to the those based on the observed data.

Figure 2 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates based on the observed data are slightly higher than 95%, because we did not specify the finite population correction in variance estimates. The coverage rates based on

<sup>2</sup> For more details and code to implement this procedure, contact the second author.

RCAL closely match those based on the observed data. Those based on OLDW do not match the observed data coverage rates, especially for analyses involving the size variable. Method UNW tends to have poor coverages for means and proportions, producing biased estimates of the population quantities as expected.

As a check on the amount of alteration in the size measures, for each record we compute the average of the five synthetic sizes. We then find the record in the population with the closest actual size to that average. Using this approach, approximately 0.2% of the respondents are correctly re-identified from the synthetic data. In contrast, releasing the old weights completely undoes the protection of the synthesis of size, since the original size can be backed out of the original weight.

### 3.3 Two Stage Cluster Sampling

We generate a population of 20000 units in which the data are grouped in 200 clusters. Twenty clusters have size 200; forty clusters have size 150; sixty clusters have size 100; and, eighty clusters have size 50. We generate the survey variables  $Y_1$  and  $Y_2$  from

$$y_1 \sim N(20 + \omega_c, 3^2) \tag{9}$$

$$y_2 \sim N(1.2y_1 + \delta_c, 20^2) \tag{10}$$

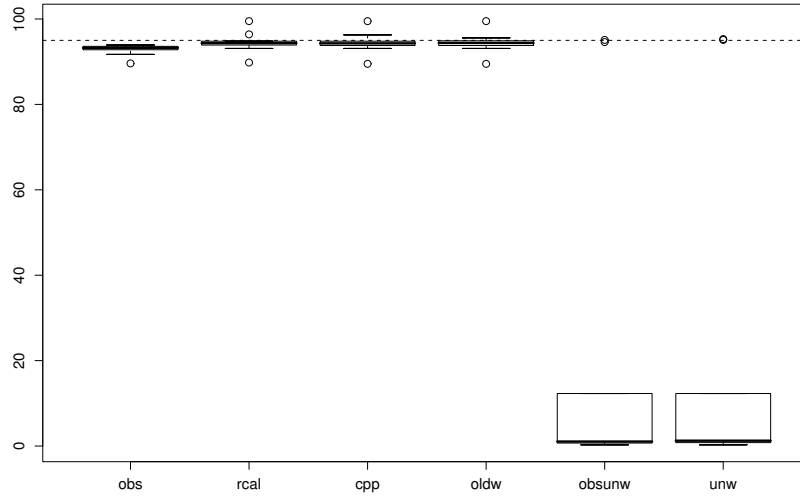
where  $\omega_c$  and  $\delta_c$  differ for each cluster  $c$ . We select observed data from this population using two-stage cluster sampling. We take a simple random sample of 40 clusters from the population of 200. For each sampled cluster, we take a simple random sample of 25 observations.

We synthesize the observed cluster indicators  $X$  and  $Y_2$  from their joint posterior predictive distribution conditional on  $Y_1$ . First, we simulate replacement values for  $X$  by using a multinomial logistic regression model conditional on  $Y_1$ . Only the observed clusters in each sample are used to fit the model. Second, we simulate replacement values for  $Y_2$  from a normal linear regression on  $(Y_1, X)$ , using the synthetic  $X$  for predictions. We repeat this process independently  $m = 5$  times to obtain five partially synthetic datasets for each  $D$ .

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of  $Y_1$ , and  $Y_2$ ; the percentages of values of  $Y_1$  greater than the population 50th, 80th and 95th percentiles, and likewise for  $Y_2$ ; the regression coefficient of  $Y_1$  from the linear regression of  $Y_2$  on  $Y_1$ ; and, the regression coefficient of  $Y_1$  from the linear regression of  $Y_2$  on  $(Y_1, X)$ . The synthetic 95% confidence intervals are based on the methods in Section 2.1, with  $\bar{v}_m$  equal to the design-based variance estimate as computed with (i) our own variance estimation code for means and percentages and (ii) the *R* survey package for regression coefficients.

For any record in cluster  $c$ , the weight equals the product of five and the inverse of the fraction of records sampled in that cluster. To apply RCAL, only the second term in the multiplication changes, depending on the new fraction of records in each cluster. To apply CPP, we use the same process illustrated

in Section 3.1. As in the stratified sampling simulation, variance estimates of means and proportions are the same for methods RCAL and CPP.



**Fig. 3.** Box plots of coverage rates for the ten estimands in the cluster sampling simulation. The coverage rates based on *RCAL*, *CPP*, and *OLDW* are close to the those based on the observed data.

Figure 3 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates based on the observed data are slightly lower than 95%, whereas the coverage rates based on *RCAL* and *CPP* are nearly 95%. The coverage rates based on *OLDW* are like those based on *RCAL* and *CPP*. In this constructed population, many weights do not change substantially after applying *RCAL* and *CPP*—even though cluster memberships change—due to the identical second stage sampling rate and the existence of many clusters of the same population size. This also explains why *RCAL* and *CPP* result in similar coverage rates. Coverage rates based on *UNW* are very low for means and proportions but close to 95% for the regression coefficients.

As a check on the amount of alteration in the cluster indicators, we compare the units’ modal synthesized cluster indicators to their corresponding observed indicators, like the strategy used in the stratified simulation,. Approximately 15% of units can be placed in their original cluster, indicating a sizeable number of re-allocations of cluster memberships.

### 3.4 Survey of Youth in Custody

We now examine the performance of RCAL, the method that performs best across all the simulations, on genuine data from the 1987 Survey of Youth in Custody. The survey interviewed youths in juvenile institutions about their family background, previous criminal history, and drug and alcohol use. The sampling frame comprises 206 facilities. The eleven facilities (strata 6 to 16) with more than 360 youths were treated as strata. The remaining facilities were divided in five strata (strata 1 to 5) based on size. These facilities were sampled with probability proportional to size, and residents within sampled facilities were sampled with predetermined sampling fractions. The sample contains 50 facilities and 2,621 youths.

To simplify the illustration, we deleted four facilities for which size was unknown and ignored the small amount of unit nonresponse. We re-specified the original survey weights to reflect the smaller number of facilities and clusters in this reduced dataset. We filled in the small number of missing item values using univariate re-sampling.<sup>3</sup>

We consider facility membership to be potentially identifying information. Therefore, we generate new facility identifiers for all records in the dataset. For strata 6 to 16, we synthesize the stratum value for each observation using a multinomial regression estimated with records in strata 6 to 16 only. For purposes of illustration, we include main effects for all twenty-two predictors in the regression model except for race, education, and who the youth lived with before being institutionalized. These variables are excluded to enable the model to be identifiable, since there is multi-collinearity in the data. For strata 1 - 5, we synthesize the facility indicators using another multinomial regression estimated with records in strata 1 to 5 only. This model excludes from the synthesis model the youth's race, education, who they lived with, whether anyone in the family served time, the type of crime, and their alcohol use. More terms are dropped because the sample sizes in these facilities are small. A potentially more accurate synthesis model would incorporate informative prior distributions on the parameters of the logistic regression with all twenty-two predictors.

We create  $m = 5$  partially synthetic data sets. We then recalculate the survey weights using the RCAL method, which involves recalculations like described in the cluster sampling simulation. Table 1 displays the observed and synthetic point estimates and 95% confidence intervals for a variety of estimands. All results are based on survey-weighted estimation based on the design. Generally, the observed and synthetic point estimates and confidence intervals are similar. The one possible exception is the regression coefficient for the indicator variable corresponding to Asian race; however, the observed and synthetic confidence intervals are relatively wide and overlap to an extent.

To check on the amount of alteration in the facility indicators, we apply the strategy used in the cluster sampling simulation—place the youth in its modal

---

<sup>3</sup> We recommend using the principled approach of multiple imputation to handle missing data, but the univariate re-sampling is adequate to illustrate the performance of the RCAL approach.

imputed facility—and find that approximately 17% of youths can be placed in their original facility.

**Table 1.** Point and interval estimates based on observed and synthetic data for Survey of Youth in Custody

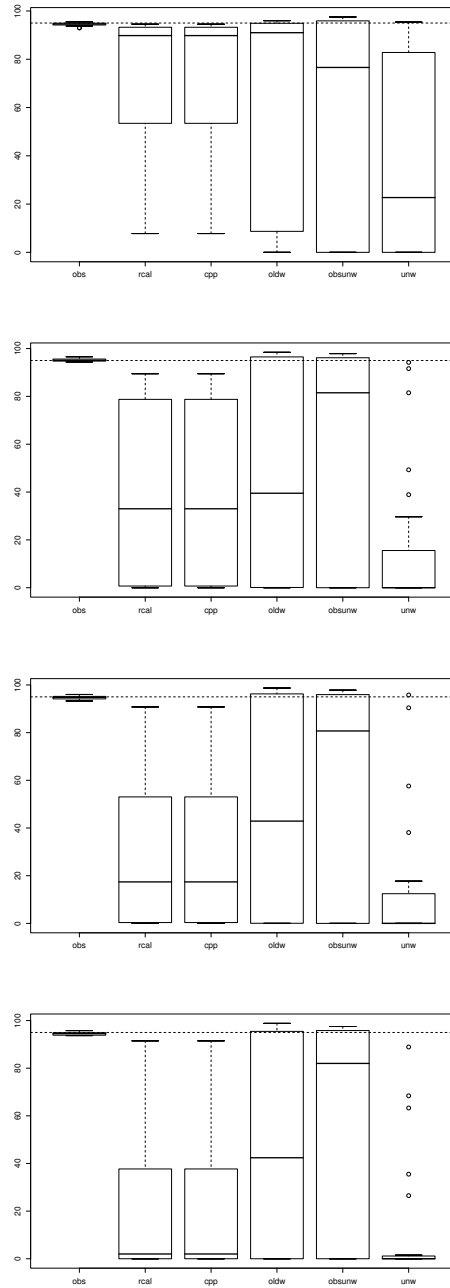
Variable	$q_{obs}$	Obs. 95% CI	$\bar{q}_5$	Syn. 95% CI
Avg. age	16.7	(16.6, 16.8)	16.8	(16.7, 16.9)
Avg. age at first arrest: Hispanics	13.0	(12.7, 13.2)	13.0	(12.6, 13.2)
Avg. age at first arrest: non-Hispanics	13.0	(12.9, 13.1)	13.0	(12.8, 13.1)
% with age at first arrest < 15	73.4	(71.3, 75.5)	73.1	(70.8, 75.4)
% with age at first arrest > 18	.39	(.16, .62)	.40	(.15, .64)
% used drugs	25.4	(23.4, 27.3)	25.2	(23.2, 27.1)
% females	7.4	(6.1, 8.6)	7.5	(6.1, 9.0)
Coefficients in logistic regression of ever violent on				
Intercept	1.36	(.80, 1.93)	1.33	(.73, 1.92)
Age at first arrest	-.083	(-.126, -.041)	-.082	(-.127, -.037)
Black	.46	(.25, .67)	.48	(.27, .69)
Asian	.33	(-.72, 1.38)	.76	(-.28, 1.79)
American Indian	-.014	(-.551, .523)	-.088	(-.726, .549)
Other	1.35	(.56, 2.15)	1.21	(.42, 2.00)

## 4 Extension to Data Swapping

In this section, we examine the performance of the weight adjustment procedures when swapping design variables. We use the stratified sampling simulation. Rather than synthesizing new stratum indicators, we assign some percentage of the stratum indicators to be randomly swapped, creating one masked dataset per observed dataset. Stratum indicators might be swapped with a like value, resulting in no change for the unit’s stratum in the masked data.

We follow the simulation design in Section 3.1, except that we leave  $Y_2$  unaltered. We consider swapping rates of 5%, 30%, 50% and 100%. Figure 4 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates get progressively worse as the degree of swapping increases. For the 5% swapping simulation, coverage rates based on RCAL or CPP are better than those based on OLDW or UNW, but they remain inadequate. With this version of swapping, the methods RCAL and CPP yield exactly the same weights and hence estimates, since there always are 250 records in each stratum.

Comparing swapping to partial synthesis, the coverage rates based on RCAL and CPP are much closer to nominal in the partially synthetic data than in the swapped data, even though we replaced the values of  $Y_2$  in the former but not the latter. Using a swapping rate of 50% leaves approximately 45% of records’



**Fig. 4.** Box plots of coverage rates for swapping percentages of 5%, 30%, 50%, and 100%, going from top panel to bottom panel. None of the methods based on the swapped data have satisfactory coverage properties.

stratum indicators unchanged, which is the same percentage of records that can be placed in their correct stratum when using partially synthetic data. But, the partial synthesis clearly is more effective at preserving the statistical properties of the data.

## 5 Conclusions

The simulations in this paper illustrate the importance of survey weights when altering design variables to limit disclosure risks. Releasing the original weights can lead to biased inferences or compromise identity of respondents. At least for partially synthetic data, recalculating the weights to be consistent with released values can improve design-based estimation. Unfortunately, this approach does not appear to improve inferences sufficiently when using data swapping of design variables. Further research is needed to investigate the viability of the recalculation approach for more complicated multi-stage sampling schemes. Additionally, research is needed to see how adjustments for non-response and calibration interact with the recalculation approach.

**Acknowledgments.** This research was funded by the National Science Foundation grant ITR-0427889.

## References

1. De Waal, A.G., Willenborg, L.C.R.J.: Statistical Disclosure Control and Sampling Weights. *Journal of Official Statistics*, **13** (1997) 417–434
2. Reiter, J.P.: Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, **29** (2003) 181–189
3. Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York (1987)
4. Kennickell, A.B.: Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In: Alvey, W., Jamerson, B. (eds.): *Record Linkage Techniques, 1997*. National Academy Press, Washington, D.C. (1997) 248–267
5. Abowd, J.M., Woodcock, S.D.: Disclosure Limitation in Longitudinal Linked Data. In: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (eds.): *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies North-Holland, Amsterdam* (2001) 215–277
6. Lohr, S.L.: *Sampling: Design and Analysis*. Duxbury Press, New York (1999)
7. Reiter, J.P., Zanutto, E.L., Hunter, L.W.: Analytical Modeling in Complex Surveys of Work Practices. *Industrial Labor Relations Review*, **59** (2005) 82–100.
8. Valliant R., Dorfman A.H., Royall R.M.: *Finite Population Sampling and Inference*. John Wiley & Sons, New York (2000) 72–73.