Variance Estimation for Combined Probability and Nonprobability Samples

Michael Yang, Nada Ganesh, Edward Mulrow, and Vicki Pineau
NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814

Abstract
Survey researchers have proposed three general approaches to estimation from nonprobability samples: quasi-randomization, superpopulation modeling, and doubly robust (Valliant, 2020). Through case studies and Monte Carlo simulations, the authors have evaluated some commonly used estimation methods associated with these approaches (Ganesh et al., 2017; Yang, et al. 2018, 2019; Mulrow, et al. 2020). Our empirical evaluations show that these methods tend to produce comparable point estimates, but estimates under two of these methods, Propensity Weighting (quasi-randomization) and Small Area Modeling (doubly robust), exhibit superior properties in terms of bias reduction, mean squared error, and confidence interval coverage. Focusing on these two methods, we expand our earlier simulations to explore variance estimation methods. Like our earlier evaluations, the simulation data was generated to mimic the coverage bias exhibited by opt-in online samples for some key characteristics. Our objective is to explore practical variance estimation solutions to guide practitioners who use nonprobability samples but may not have the resources to carry out elaborate variance estimation procedures. Our approach is to simulate Jackknife variances under Propensity Weighting and Small Area Modeling and compare with naïve variances or design variances where we assume that the combined probability and nonprobability sample is a probability sample.

Key Words: nonprobability sample, variance estimation, propensity, small area estimation

1. Introduction

Probability sampling remains the gold standard for survey research. However, as survey costs continue to rise, there has been growing demand for methods that use nonprobability samples and methods that combine probability and nonprobability samples in order to improve the cost efficiency of survey estimation (Baker et al., 2013).

While nonprobability samples provide a lower cost alternative to probability samples, estimates based on nonprobability samples may be biased due to unknown coverage and selection biases. Since there is no known sample design, model-based approaches are required for inferences from nonprobability samples to reduce potential bias. Survey researchers have proposed three general approaches to estimation from nonprobability samples: quasi-randomization, superpopulation modeling, and doubly robust (e.g., Elliott and Valliant, 2017; Valliant, 2020).

Through case studies and Monte Carlo simulations, the authors have evaluated some estimation methods under these approaches (Ganesh et al., 2017; Yang, et al. 2018, 2019; Mulrow, et al. 2020). Our previous evaluations show that these methods produce

comparable point estimates, but two of these methods, Propensity Weighting (quasi-randomization) and Small Area Modeling (doubly robust), exhibit superior properties in terms of bias reduction, mean squared error, and confidence interval coverage. In this paper, we expand our earlier simulations to explore variance estimation under Propensity Weighting and Small Area Modeling. As in prior studies, our simulation data file was generated to mimic the coverage bias exhibited by opt-in nonprobability samples for some key characteristics. Our objective is to explore practical variance estimation solutions to guide practitioners who use nonprobability samples but may not have the resources to implement proper variance estimation procedures. Assuming the combined sample is a probability sample is likely to lead to severe under reporting of variances. In contrast, with some extra computation effort, a replication method like jackknife can implicitly account for the variation due to the fact that the nonprobability sample weights are estimated. Our approach therefore is to simulate jackknife variances under Propensity Weighting and Small Area Modeling and compare with naïve variances or design variances where one assumes that the combined probability and nonprobability sample is a probability sample. The goal is to report more realistic sampling variances when combining probability and nonprobability samples even though jackknife replication in itself is not necessarily a solution because it may still underestimate the true mean squared error (MSE).

2. Propensity Weighting and Small Area Modeling

Valliant (2020) provides a comprehensive review of the three approaches to nonprobability sample estimation, including the assumptions required for each to produce approximately unbiased estimates and methods for variance estimation. Under Quasi-randomization, one estimates the pseudo inclusion probabilities for the nonprobability sample and then carry out design-based estimation using the pseudo weights as if they are design weights. Under Superpopulation Modeling, one develops statistical models for the survey response variables and use these models to project the sample to the population. Finally, Doubly Robust is a combination of Quasi-randomization and Superpopulation Modeling. Doubly Robust estimators are expected to be approximately unbiased and consistent if the pseudo inclusion probability distribution, the superpopulation model, or both are correctly specified (Cao et al., 2009; Elliott and Valliant, 2020; Kang and Schafer, 2007; Kim and Haziza, 2014). In this section, we briefly review our earlier evaluations and then describe in detail how sample weights for the combined probability and nonprobability sample are generated under Propensity Weighting and Small Area Modeling.

2.1 Earlier Evaluation Results

Our earlier evaluations through case studies and simulations compared the following estimation methods for combining a probability and a nonprobability sample:

i. Calibration: Calibrate combined sample weights to reproduce known population benchmarks
ii. Superpopulation Modeling: Use a linear superpopulation model[1] to derive sample weights and population estimates

[1] Our evaluations use a linear model. More generally, superpopulation modeling is not limited to linear models.

iii. Propensity Weighting: Model the inclusion probabilities and weights for the nonprobability sample and combine with probability sample weights
iv. Statistical Matching: Statistically match the nonprobability and probability samples to derive sample weights
v. Small Area Modeling: Use domain-level small area estimation models to derive calibration targets for key survey response variables

Descriptions of these methods may be found in Elliott and Valliant (2017), Bethlehem (2015), D'Orazio et al (2006), and Ganesh et al (2017). In our application, each of these methods produces a set of final weights for the combined probability and nonprobability sample as the final outcome. Of these methods, Small Area Modeling, Propensity Weighting, and Statistical Matching rely on the availability of a probability sample, while Calibration and Superpopulation Modeling do not.[2]

Our earlier empirical evaluations demonstrate that: (1) Calibration and Superpopulation lead to similar results as they tend to rely on the same set of covariates that are available from census data (e.g., American Community Survey, Current Population Survey); (2) Propensity Weighting consistently outperforms the other methods presumably because it is able to use more covariates than the other methods; (3) Small Area Modeling provides the most bias reduction for the modeled response variables, both overall and for subpopulations, especially for response variables that exhibit large biases in the nonprobability sample; and (4) Statistical Matching gives promising results but more evaluations were needed.[3]

Our current evaluations focus on variance estimation under Propensity Weighting and Small Area Modeling, the two methods that produce better estimates based on our earlier comparisons. Below we provide more details of how the combined sample weights are developed under these two methods.

2.2 Propensity Weighting

Propensity Weighting requires the presence of a probability sample, called a reference sample, selected from the target population. The reference sample weights are regular probability sample weights scaled to sum to the target population total. Meanwhile, each nonprobability sample unit is assigned a weight of 1. Here are the steps for developing the combined sample weights under Propensity Weighting:

i. Generate probability sample weights using standard weighting procedures. This typically involves computing sampling or base weights to account for the selection probabilities under the sample design, and weighting adjustments for unknown eligibility, survey nonresponse, and frame coverage. The final probability sample weights are calibrated to known distributions of the target population for a set of demographic variables typically through raking ratio adjustments to census benchmarks. The demographic variables could vary across studies but usually include age, gender, education, race/ethnicity, geography, and the like.

[2] As a general estimation methodology, calibration often uses probability samples.
[3] The authors have since completed further evaluations of Statistical Matching and results are to be reported in a separate publication, Mulrow et al., 2020.

ii. Concatenate the probability sample and the nonprobability sample and create a dichotomous variable *R*, coded 1 for nonprobability sample units and 0 for probability sample units;
iii. Fit a weighted logistic regression model with *R* as the response variable where the probability sample units are weighted by their regular weights and nonprobability sample units assume a weight of 1;
iv. Compute the nonprobability sample weights as the inverse of the inclusion probabilities predicted from the logistic regression model;
v. Calibrate the nonprobability sample weights to the same set of population benchmarks used to calibrate the probability sample;
vi. Combine the probability and nonprobability sample weights through a combination factor that is proportional to the relative size of the probability and nonprobability samples.

Predictor variables in the logistic regression model include demographic (e.g., age, gender, race and ethnicity, marital status), socioeconomic (e.g., education, income, employment), webographic[4], as well as response variables collected from the survey. The final model is validated through cross validation and by examining model diagnostic statistics.

The predicted inclusion probabilities may be sensitive to misspecification of the underlying logistic regression model (Kang and Schafer, 2007, Cao et al., 2009). One may choose to form weighting strata based on the size of the estimated inclusion probabilities and use a common inclusion probability for each stratum, like forming weighting classes from predicted response propensities for nonresponse weighting adjustments (Little, 1986, Valliant et al., 2018). For this investigation, we used the inverse of individual propensities as the weights for the nonprobability sample units.

2.3 Small Area Modeling

To address the issue of potential estimation bias associated with the nonprobability sample, researchers at NORC developed a hybrid calibration weighting method that combines probability and nonprobability samples using small area estimation modeling (Rao, 2003). The resulting weights are calibrated to both standard demographic benchmarks and domain-level estimates for key survey response variables estimated from small area models. Relative to other estimation methods we evaluated, Small Area Modeling generates the most substantial bias reduction across a range of case studies and simulations (Ganesh et al., 2017; Yang et al., 2018, 2019). The implementation of Small Area Modeling involves four major stages, as discussed in the subsections below.

2.3.1 Develop Probability Sample Weights

The same probability sample weighting procedures as described in 2.2 are used.

2.3.2 Develop Nonprobability Sample Weights

[4] Webographic variables are those that are believed to differentiate the online population from the general population.

As there is no known "design" to nonprobability samples, units in the nonprobability sample are given an initial base weight of one. The final nonprobability sample weights may be developed through Calibration, Propensity Weighting, Statistical Matching, or some other methods. The nonprobability sample weights are calibrated to the same known distributions of the population as those used to calibrate the probability sample weights.

2.3.3 Small Area Modeling

Small area estimation models are developed to derive domain-level predicted values for some key response variables, and these predicted values are used as additional raking targets to produce the weights under Small Area Modeling. To start the modeling process, we identify a set of (2 to 4) key response variables from the survey, e.g., using a machine learning approach. Ideally, these key variables are associated with the largest bias in the nonprobability sample and also are highly correlated with other survey response variables.

We then define a set of (20 to 40) domains in the data, where each domain represents a specific and meaningful subgroup for data analysis and reporting. For example, a set of domains may be defined using race, gender, age and educational attainment, and one of which may be African-American males age 18 to 34 with a college degree. The choice of domains should ensure "sufficient" sample size for the probability and non-probability samples per domain, align with analysis and reporting domains, and also capture the variation across domains with respect to substantive survey response variables.

Now we are ready to fit domain-level small area models for each of the response variables identified earlier using weighted domain-level estimates as input and incorporating external data sources as potential predictors in the models. For this research, a Bivariate Fay-Herriot model (Rao, 2003; Fay and Herriot, 1979) is used to jointly model the domain-level point estimates from the probability sample (y_d^P) and the nonprobability sample (y_d^{NP}):

$$y_d^P = \mathbf{x}_d'\boldsymbol{\beta} + \nu_d + \varepsilon_d^P$$
$$y_d^{NP} = b + \alpha_d^{NP} + \mathbf{x}_d'\boldsymbol{\beta} + \nu_d + \varepsilon_d^{NP}$$

where

d is a domain (e.g. 18-34 year old, male, African American, college degree)
$\mathbf{x}_d$ is a vector of covariates
ν_d's are domain level random effects
b is a fixed effect bias term associated with the nonprobability sample estimate
α_d's are random effect bias terms associated with the nonprobability sample estimate
ε_d^P, ε_d^{NP} are the sampling errors associated with y_d^P and y_d^{NP}, respectively

Once the small area models are finalized, they are used to generate predicted values for each domain and for each response variable using an Empirical Best Linear Unbiased Predictor (EBLUP).

2.3.4 Hybrid Calibration

The final stage, hybrid calibration weights are developed by raking the probability and nonprobability weights to known demographic control totals as well as predicted values derived from the small area models. Before this calibration, the original probability and nonprobability weights are combined through a combination factor that is proportional to the relative sample size. The final weights under Small Area Modeling reproduce the population benchmarks as well as the small area estimates for each domain and each of the key survey response variables.

3. Monte Carlo Simulations

To mimic the type of coverage bias typically exhibited in online opt-in nonprobability samples, we created two sampling frames, one a subset of the other, using survey completes from a large-scale national study about food allergies, as follows:

i. Frame 1, the full population frame, consists of all 40,539 adult survey completes. Random samples selected from Frame 1 are considered probability samples.
ii. Frame 2, a nonrandom subset of Frame 1, consists of 36,917 adult survey completes. To impart coverage bias to Frame 2, we sorted Frame 1 by some key response variables, and then selected 3,622 (9 percent) records for removal to create Frame 2. Random samples selected from Frame 2 are considered nonprobability samples with respect to the population as represented by Frame 1.

Both the probability and nonprobability frames/samples contain a large number of demographic, webographic, and survey response variables. Demographic variables include: age, gender, race/ethnicity, education, employment, marital status, household income, household size, home ownership, household telephone service, and more. Webographic variables include household internet access and early adoption of technology among others. Survey responses variables include self-reported and doctor-diagnosed food allergies, allergy reactions, experiences in allergy treatments, events coinciding with development or outgrowing a food allergy, and perceived risks associated with food allergies.

The Monte Carlo simulations involve 2,500 simulation samples, each consisting of a probability sample of size 400 and a nonprobability sample of size 800 selected using SRSWOR from Frame1 and Frame 2, respectively. For each simulation sample, we generate the combined sample weights, weighted estimates, and other key statistics, as below:

i. Combined sample weights under Propensity Weighting and Small Area Modeling are generated using the respective weighting procedures;
ii. Weighted point estimates under each method are derived for response variables of interest;
iii. Compute design (naïve) variances assuming that the respective combined sample weights are regular probability sample weights;

iv. Compute SRS variances assuming that the respective combined sample is a simple random sample;
v. Compute estimated bias for each response variable as the difference between the known population value and the weighted sample mean.
vi. Define jackknife replicates as follows:
    a. Divide each simulation sample into 50 random and equal-sized groups (jackknife group);
    b. Form 50 jackknife replicates within each simulation sample by deleting a jackknife group from each sample source at a time so that each jackknife replicate contains 392 cases from the probability sample and 784 cases from the nonprobability sample;
vii. Produce jackknife variance estimates as follows:
    a. Ratio adjust the Propensity or Small Area Modeling weights per jackknife replicate to their original full sample weight total
    b. Compute weighted point estimates for each jackknife replicate
    c. Compute weighted point estimate for the full sample
    d. Compute jackknife variance estimates for each simulation sample as $v_{JK}(\hat{\theta}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\theta}_{(g)} - \hat{\theta})^2$, where $\hat{\theta}_{(g)}$ is the estimate from replicate g and $\hat{\theta}$ is the full sample estimate
    e. Compute Jackknife mean squared error (MSE) estimates for each simulation sample as $v_{JK}(\hat{\theta}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\theta}_{(g)} - \theta)^2$, where $\hat{\theta}_{(g)}$ is the estimate from replicate g and θ is the known population value

Finally, all the statistics associated with each simulation is averaged over the 2500 iterations. The simulation results and discussions in the next section are based on these averages.

4. Simulation Results and Discussions

Key simulation results include estimated variances, biases, mean squared errors, and confidence interval coverage associated with the weighted point estimates for a set of 13 survey response variables under Propensity Weighting and Small Area Modeling. All these variables are measured as proportions and their population distributions are known. In particular, the known bias associated with the nonprobability sample ranges from .04 to 8.74 percentage points for the 13 variables. Our evaluations are based on the simulated statistics for the combined sample as well as 18 subdomains.

Figure 1 shows the regressions of design standard errors on SRS standard errors under Propensity Weighting (left) and Small Area Modeling (right). Each data point corresponds to a variable and domain combination. To compute design standard errors, we assume that the final weights under the two methods are regular probability sample weights; and to compute SRS standard errors, we assume that the samples are simple random samples. As expected, design standard errors track very closely with SRS standard errors. Note that the vertical scales are not the same for the two graphs: the Small Area regression line actually has a steeper slope (1.16 vs. 1.04), indicating that design standard errors are larger under Small Area Modeling than under Propensity Weighting. As we have also reported in earlier investigations, Propensity Weighting typically results in less weight variation than Small

Area Modeling even when predicted propensities are directly used to derive the pseudo weights (Yang et al., 2018, 2019).

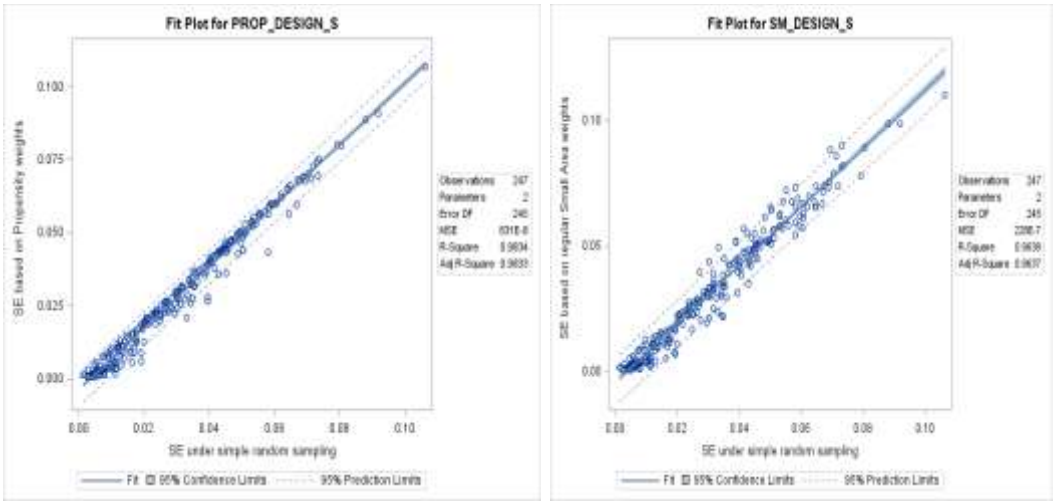

Figure 1: Regressions of design standard errors on SRS standard errors

Design standard errors are likely to underestimate the underlying sampling variation of the sample statistics. Although no formal proof exists for combined probability and nonprobability samples, we expect replication variance estimators to produce more realistic variance estimates. The delete-a-group jackknife replication procedures implemented here implicitly reflect the estimated weights and weighting adjustments under each method. Figure 2 shows the regressions of Jackknife standard errors on design standard errors. With a slope of 1.02, Propensity Jackknife SEs are only slightly greater than the corresponding design SEs. On the other hand, Small Area Jackknife SEs are about twice as large as Small Area design SEs (slope=2.05). Using Small Area weights as design weights for variance estimation is likely to severely underestimate the true variance.

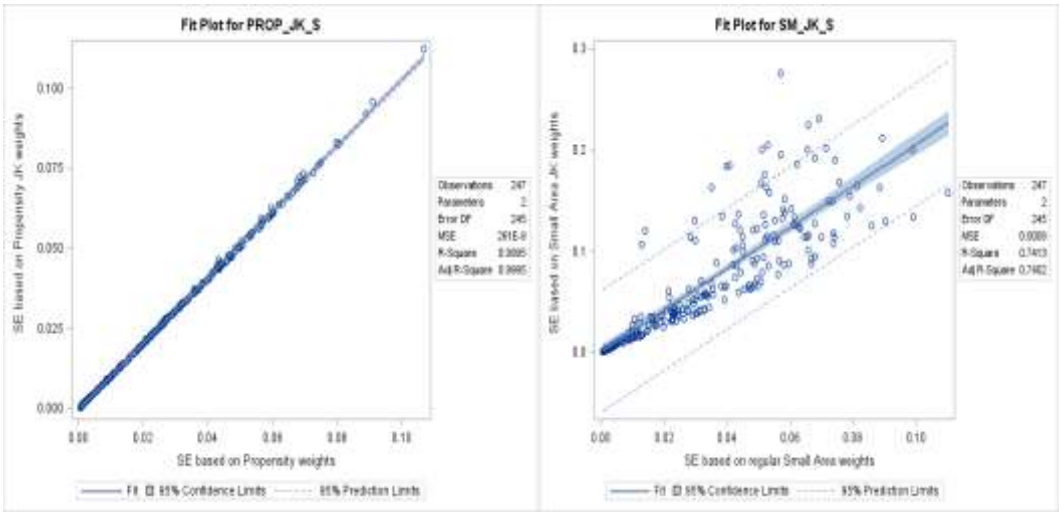

Figure 2: Regressions of Jackknife standard errors on design standard errors

Figure 3 shows regressions of jackknife RMSE on jackknife SE under each method. Note that the horizontal scales are different between the two graphs. The design and jackknife standard errors may be small under Propensity Weighting, but their jackknife RMSEs are much larger than jackknife SEs (slope=5.76), suggesting that the total error contains a substantial bias component under Propensity Weighting. On the other hand, Small Area RMSEs are only about 90% larger than Small Area jackknife SEs (slope=1.9). Figure 2 also shows that the residuals of the regression tend to be smaller for Propensity than for Small Area, which indicates that the correlation between jackknife RMSEs and SEs are high and that the bias is not only substantial but also consistent across analysis domains and response variables. On the other hand, the regression of jackknife RMSEs on jackknife SEs under Small Area is less linear, the residuals much larger, and there are also signs of heteroscedasticity across the domains and response variables. Large residuals are typically associated with variables that contain more bias in the nonprobability samples. Although Small Area Modeling is more effective in reducing bias for variables that are more biased, as discussed later, it is not able to remove all the bias for such variables.

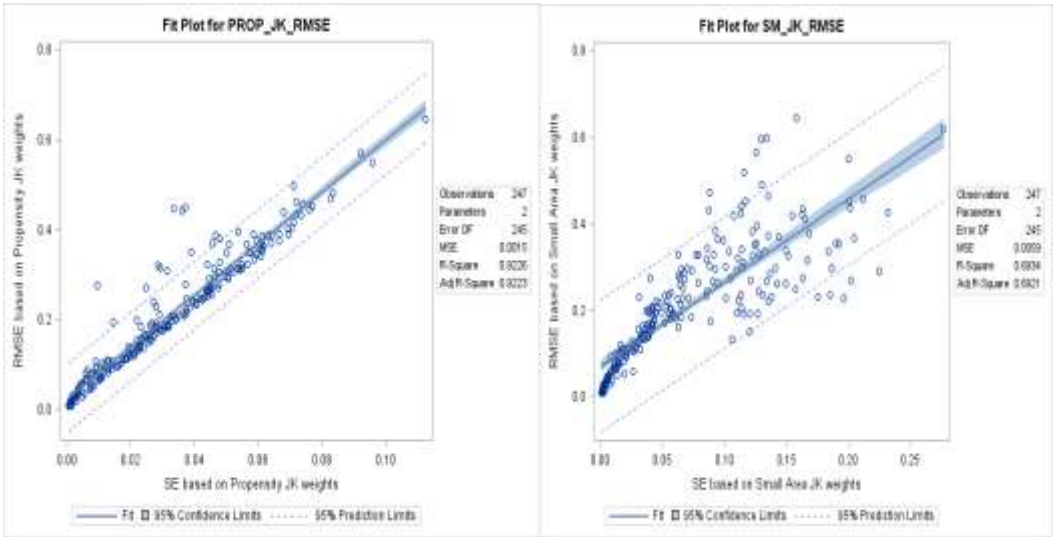

Figure 3: Regression of Jackknife RMSE on JK SE

Figure 4 shows the regression of Small Area jackknife RMSEs on Propensity jackknife RMSEs. Small Area RMSEs are on average smaller, or about 90% of the Propensity RMSEs (slope=.92). So, if we use RMSE as the ultimate standard, Small Area performs slightly better on average based on these simulations, although the difference is not large. An examination of the residuals show that the relative strength of the two methods also could vary across analysis domains and response variables. In general, no strong evidence thus far from these current simulations challenges our earlier conclusion that both Propensity Weighting and Small Area Modeling are viable estimation methods for combining probability and nonprobability samples.

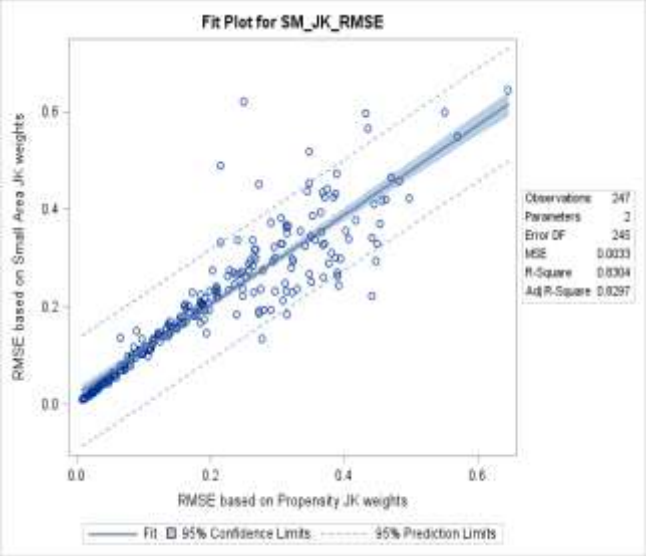

Figure 4: Regression of Small Area jackknife RMSEs on Propensity jackknife RMSEs

Turning to the examinations of estimated bias, Figure 5 shows the distribution of bias per domain for estimates based on the simulated probability samples only. As expected, both the average and the median of the simulated bias associated with the probability samples is about 0. Weighted estimates associated with some simulated samples are slightly biased, but even the outlier biases are extremely small. Therefore, any bias in the combined sample estimates is originated from the inclusion of the nonprobability sample.

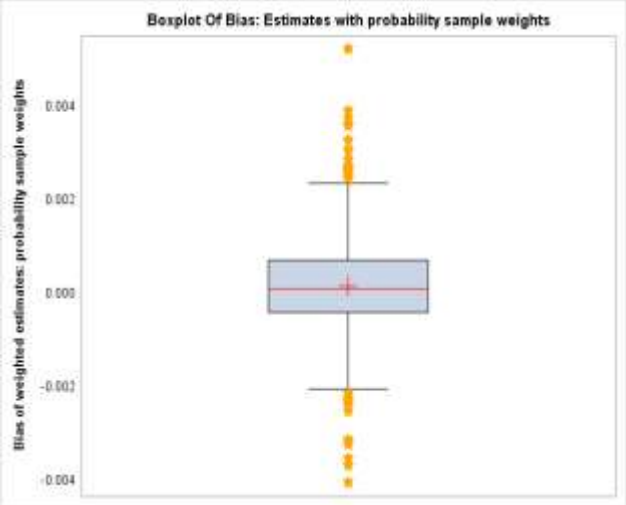

Figure 5: distribution of simulated bias for the probability sample

Figure 6 shows the bias of combined sample estimates under Propensity Weighting and Small Area Modeling. As expected, combined sample estimates contain some level of bias under both Propensity and Small Area methods. These method may help to reduce bias but they generally cannot eliminate the bias introduced by the nonprobability sample.

Therefore, using pseudo weights as regular design weights in estimation could lead to erroneous inferences about the underlying population distribution of the response variable.

Boxplots of biases are shown separately for modeled and non-modeled variables under Small Area Modeling. Based on the interquartile range, the boxplots show that estimated biases under Small Area Modeling tend to have smaller variance than those under Propensity Modeling. The boxplots also show that estimates of modeled variables contain more bias than estimates of non-modeled variables. Although a little counterintuitive given how hybrid calibration works under Small Area Modeling, this is actually not unexpected because response variables with the largest biases are typically chosen to be the model variables. With the more targeted bias reduction inherent in the Small Area Modeling method, variables with largest biases benefit most from this approach. Modeling variables with little bias in the nonprobability sample would be both ineffective and unnecessary. On the other hand, larger biases associated with modeled response variables demonstrate that, although Small Area Modeling achieves substantial bias reduction, significant bias may still remain in the final estimates. The amount of bias will depend on the degree to which (1) the initial combined sample weights are estimated accurately, and (2) the small area models are specified correctly.

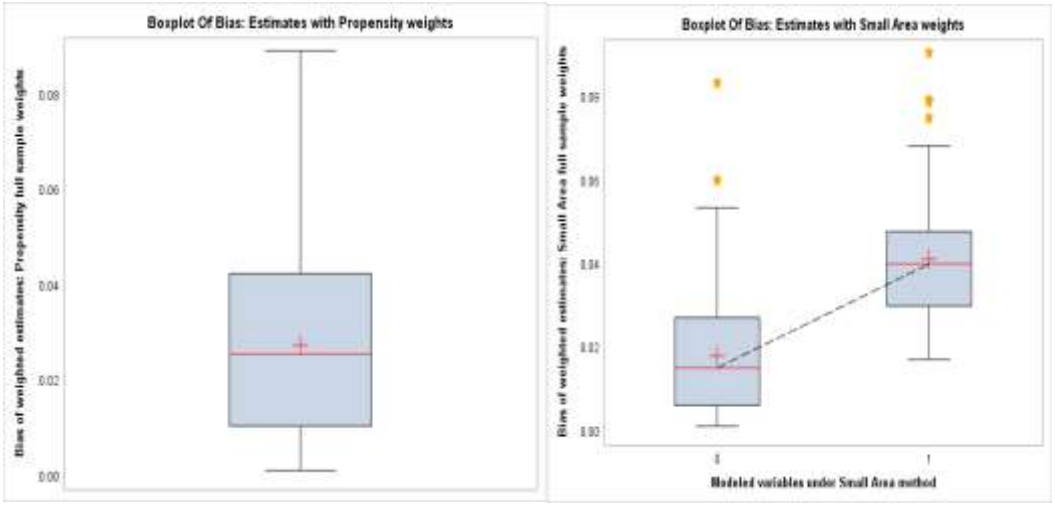

Figure 6: Bias under Propensity Weighting and Small Area Modeling

Figure 7 shows the boxplots of 95 percent confidence interval coverage under Propensity Weighting, using the design standard errors and the jackknife standard errors. The coverage rate is the proportion of the confidence intervals, built around the weighted point estimates using the estimated standard errors, that contains the known population proportion for each variable and domain. Confidence interval coverage is similar between using Propensity design SEs and Propensity jackknife SEs, reflecting the earlier discovery that the design SEs and jackknife SEs track each other closely. In both cases, the median coverage rate is lower than 90%, the average around 70%, and even the maximum is below 95%. On the other end of the distribution, for a quarter of the domains/variables, the coverage rate is below 56%.

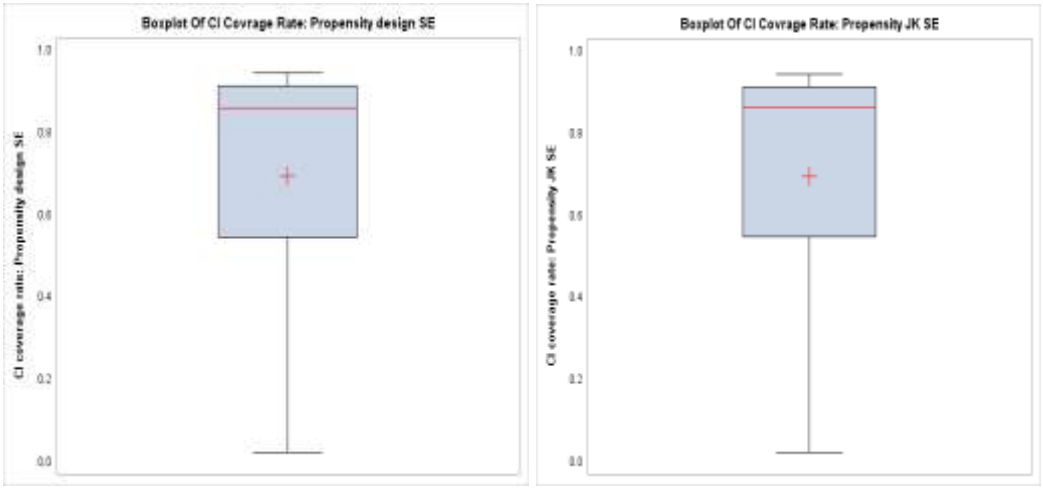

Figure 7: Confidence interval coverage under Propensity Weighting

Figure 8 shows the boxplots of 95 percent confidence interval coverage under Small Area Modeling, again using the design standard errors and the jackknife standard errors but also separately for modeled and other variables. Similar patterns emerge here under Small Area Modeling, except that confidence interval coverage is much better for modeled variables than for other variables. This is especially true under jackknife SEs, where the median coverage rate is over 95%. For the non-modeled variables, on the other hand, the median coverage rate is about 75%, with a quarter of the intervals below 20%.

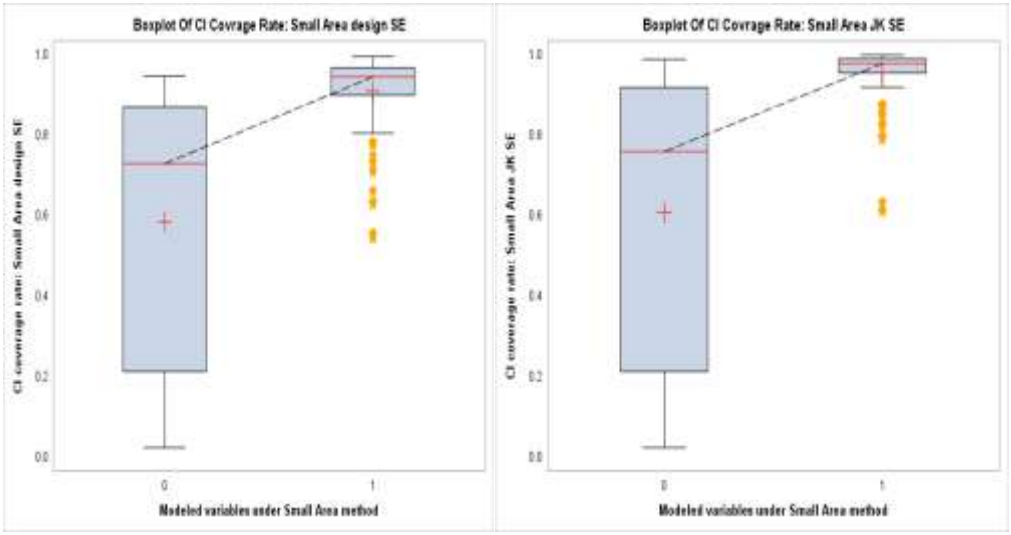

Figure 8: Confidence interval coverage under Small Area Modeling

To summarize our main findings through these simulations: The combined probability and nonprobability sample estimates still contain some level of bias. Combined sample estimates under Small Area Modeling tend to have smaller bias but larger variance than those under Propensity Weighting. Larger variances under Small Area originates from the additional weight calibration to small area estimates that usually increases the variation of

the combined sample weights. Such increase could be substantial if the bias in the modeled response variables is large and therefore greater weighting adjustments are needed to reduce such bias. When estimated mean squared errors are used for comparisons, we found that estimates under Small Area Modeling tend to perform better although the difference is not very large. Under Small Area Modeling, the modeled variables typically have the largest biases and require the most adjustments to reduce biases. Small Area Modeling achieves the largest bias reduction for these modeled variables but it does not remove the bias. Confidence interval coverage rate under Propensity Weighting is low and never achieves the nominal 95% coverage rate, which may be an indication that Propensity Weighting tends to underestimate the sampling variance and ineffective in bias reduction. In contrast, with generally larger standard errors and smaller bias, confidence intervals have much better coverage under Small Area Modeling.

5. Concluding Remarks

Our earlier evaluations indicate that Propensity Weighting and Small Area Modeling present two viable alternatives for estimation from combined probability and nonprobability samples. This study extends our earlier work by exploring variance estimation under these two methods. Results from our evaluation of a third promising alternative, Statistical Matching, is being reported in a separate publication (Mulrow et al., 2020).

Estimates from nonprobability samples, along or in combination with probability samples, are most likely to be biased due to unknown coverage and selection biases. Removing such biases remains a challenge. In theory, the biases may be removed if the sample inclusion probability models or the superpopulation models are correctly specified. In reality, however, such models are unlikely to be exactly correct. In general, based on our simulation results, using the modeled sample weights as regular design weights will likely underestimate the variances and lead to erroneous inferences.

For complex estimation based on combined probability and nonprobability samples, replication variance estimation methods may be used to implicitly to account for the extra sampling variation due to modeled nonprobability sample weights. For studies that lack sufficient resources, it might be prudent to report confidence intervals in addition to weighted point estimates, where the z-multiplier should be greater than 1.96 for the 95 percent confidence interval when the modeled weights are considered design weights under regular design-based estimation. Our simulations suggest a multiplier that is about twice as large under Small Area Modeling. Different simulation data may lead to different results and more research is need in this area. Meanwhile, we acknowledge that jackknife variance estimation in itself cannot be a full solution as long as the combined estimates remain biased. Given that the true MSE is likely to be larger than the jackknife standard error, estimated errors should be interpreted with caution and transparency.

In terms of future research, we intend to include Statistical Matching in our exploration of variance estimation methods. The current simulations feature a 2:1 nonprobability sample to probability sample size ratio. Future simulations will consider different sample size

ratios. Finally, to take advantage of the superior bias reduction abilities of the Small Area Modeling method, we will continue to explore hybrid methods to integrate Propensity Weighting and Statistical Matching with Small Area Modeling.

Acknowledgements

The authors acknowledge the analytical support provided by Julia Batishev.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013), " Report of the AAPOR Task Force on Non-probability Sampling," *Journal of Survey Statistics and Methodology*, 1, 90-143.

Bethlehem J. (2015) "Solving the nonresponse problem with sample matching?" *Social Science Computer Review*, Vol. 34, Issue 1, pp. 59 – 77.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data." *Biometrika*, 96(3):732–734.

Chen, Yilin, Pengfei Li, Changbao Wu. (2019). "Doubly Robust Inference with Non-probability Survey Samples," Paper presented at the Joint Statistical Meetings, Denver, Colorado.

Dever, J. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*. 36 45–56.

Dever, J. and Valliant, R. (2016). GREG estimation with undercoverage and estimated controls. *Journal of Survey Statistics and Methodology*. 4 289–318.

DiSogra, C., Cobb, C., Dennis, J.M. and Chan, E. 2011. Calibrating nonprobability Internet samples with probability samples using early adopter characteristics. Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM). Miami Beach, FL.

D'Orazio M., Di Zio M., Scanu M. (2006) *Statistical matching: Theory and practice*. Wiley, Chichester.

Elliot, M. R., Valliant, R. (2017). "Inference for Nonprobability Samples," *Statistical Science* 2017, Vol. 32, No. 2, 249–264.

Fahimi, M., Barlas, F.M., Thomas, R.K. and Buttermore, N. (2015). "Scientific surveys based on incomplete sampling frames and high rates of nonresponse," *Survey Practice*, v.8 (5).

Fay, R.E., and Herriot, R.A. (1979). "Estimates of income for small places: An application of James-Stein procedures to Census data," *Journal of the American Statistical Association*, v. 74 (366), pp. 269-277.

Ganesh, N., Pineau, V., Chakraborty, A., Dennis, J.M., (2017). "Combining Probability and Non-Probability Samples Using Small Area Estimation." *Joint Statistical Meetings 2017 Proceedings.*

Kang, J.D.Y., and Schafer, J.L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523-539.

Kim, J.K., and Haziza, D. (2014), "Doubly Robust Inference with Missing Data in Survey Sampling," *Statistica Sinica*, 24, 375-394.

Little, R. J. A. (1986). "Survey nonresponse adjustments for estimates of means." *International Statistical Review*, 54(2):139–157.

Mulrow, Edward, Nada Ganesh, Vicki Pineau, and Michael Yang. (2020). "Using Statistical Matching to Account for Coverage Bias When Combining Probability and Nonprobability Samples," *Joint Statistical Meetings 2020 Proceedings* (Forthcoming).

Rao, J.N.K. (2003). *Small Area Estimation*, John Wiley & Sons, Inc.

Rivers, D. (2007), "Sampling for Web Surveys", Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 1-26.

Sarndal, C. E., Swensson B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer, New York.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York, 2nd edition.

Valliant, R. (2020). "Comparing Alternatives for Estimation from Nonprobability Samples," *Journal for Survey Statistics and Methodology*, vol. 8, 231-263.

Yang, Y. Michael, Nada Ganesh, Ed Mulrow, and Vicki Pineau. (2018). "Estimation Methods for Nonprobability Samples with a Companion Probability Sample," *Joint Statistical Meetings 2018 Proceedings*.

Yang, Y. Michael, Nada Ganesh, Edward Mulrow, and Vicki Pineau. (2019). "Evaluating Estimation Methods for Combining Probability and Nonprobability Samples through a Simulation Study," *Joint Statistical Meetings 2019 Proceedings*.

Yang, Y. Michael, Nada Ganesh, Edward Mulrow, and Vicki Pineau. (2019). "Estimation Methods for Combining Probability and Nonprobability Samples." Paper presented at the 75th Annual Conference of the American Association for Public Opinion Research.