Evaluating Estimation Methods for Combining Probability and Nonprobability Samples through a Simulation Study

Yang, Michael, Nada Ganesh, Edward Mulrow, and Vicki Pineau

NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814

Abstract

There is growing demand for survey estimation methods to combine probability and nonprobability samples for improved cost efficiency and more timely data dissemination. Researchers have proposed a range of estimation methods involving nonprobability samples, four of which are: (1) Calibration—calibrate total estimates to known population totals; (2) Superpopulation Modeling—use a superpopulation model to derive estimates; (3) Propensity Weighting—model the propensity of inclusion in a nonprobability sample to derive a pseudo weight; and (4) Small Area Modeling—a small area estimation approach developed at NORC (Ganesh et al., 2017).

We have previously evaluated these methods empirically using data from a food allergy survey and an NORC internal AmeriSpeak® study. Results showed that the different estimation approaches produce different pseudo weights but largely comparable point estimates (Yang et al., 2018). This paper expands our earlier study by presenting comparative evaluations of these methods through a Monte Carlo simulation study. The simulation data was generated to mimic the coverage bias exhibited by opt-in online panel samples in terms of key characteristics. We compare the properties of composite estimates derived from using different estimation approaches for combining probability and nonprobability samples.

Key words: survey estimation, nonprobability sample, composite estimation

1. Introduction

While probability sampling remains the gold standard for survey research, 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.

Nonprobability samples may provide a lower cost alternative to probability samples; however, the quality of the data is oftentimes low, and estimates based on nonprobability samples may be biased due to unknown coverage biases. Model-based approaches are required for inferences with nonprobability samples, alone or in conjunction with probability samples, to reduce the potential bias.

In this paper, we present some key results from our comparative analysis based on a simulation study. Specifically, we evaluate four different estimation methods that combine the probability and nonprobability sample estimates by comparing the properties of the resulting composite estimates. Key properties of the composite estimates compared include: bias reduction relative to

known bias, ratio of mean squared error to known variance, and true confidence interval coverage associated with the composite estimates.

2. Methods Evaluated

Survey researchers and practitioners have proposed and experimented with a range of estimation methods based on nonprobability samples for decades. More recently, there has been increased interest in estimation methods that use both probability and nonprobability samples. We conducted a literature review to identify and delineate methods reported in journals, workshops, and conferences. Our focus was on reported studies in recent years that represent research developments and/or empirical results related to nonprobability sample estimation. We documented the underlying statistical models, the model-based estimators, and the properties of the estimators, to the extent relevant information is available.

In this paper, we evaluate three commonly used methods along with a method developed by NORC.[1]

- **Calibration:** Calibrate total estimates to known population benchmarks.
- **Superpopulation Modeling (Superpopulation):** Use a superpopulation model to derive population estimates.
- **Propensity Weighting (Propensity):** Model the propensity to be included in a nonprobability sample.
- **Small Area Modeling**: Developed by NORC's AmeriSpeak® team (Ganesh, et al, 2017), this method begins with calibration, and relies on methods commonly used in Small Area Estimation to adjust weights so that small domain estimates are improved.

The four methods provide different approaches to modeling the weights for the nonprobability sample. Of these methods, Small Area Modeling and Propensity Weighting rely on the availability of a probability sample, while Calibration and Superpopulation Modeling do not.[2] However, the latter two methods, and Small Area Modeling, rely on known population totals for covariates that are used for calibration or in the superpopulation model. We now briefly describe these methods as implemented in this study.

2.1 Calibration

The Calibration method involves calibrating the sample weights such that weighted estimates of totals from the sample match known population benchmarks. The weights are developed through iterative proportional fitting (e.g., raking). Our approach is based on DiSogra et al. (2011) and Fahimi et al. (2015), and involves additional raking variables that presumably differentiate the nonprobability sample from probability samples. When the nonprobability sample is an opt-in

---

[1] Statistical Matching and Doubly Robust Estimation are other proposed methods used by some researchers (see Chen et al., for example). Future investigations may include these methods.

[2] As a general estimation methodology, calibration often uses probability samples.

online sample, these additional "webographic" raking variables, such as "early adopter" variables, attempt to correct for the nonprobability sample's higher likelihood of including individuals who are more likely to be users of new products and technology.

The calibration weights for the nonprobability sample are developed via the following steps:

- Assign an initial weight of 1 to all nonprobability sample units.
- Rake the initial weights to known demographic control totals from the Current Population Survey (CPS) as well as webographic control totals.

The CPS demographic variables include age, gender, education, race/ethnicity, geography, and income. The webographic control totals are estimated using NORC's AmeriSpeak® Panel, a large probability sample representing the U.S. household population.

2.2 Superpopulation Modeling

This method is referred to as the superpopulation approach in Elliot and Valliant (2017). Other important references for superpopulation modeling approaches include Dever and Valliant (2010, 2016). Under this approach, one fits a model for the dependent Y variable using the sample data, and then use the model to predict Y for the entire population. The underlying model adopted for this study is a linear one. Under this model, the predictor of the population total can be expressed as a weighted function of the observed Y variable. Note that this approach could also be used with a probability sample, as discussed in Valliant, Dorfman, and Royall (2000). The difference is that design-based inference is not an option for a nonprobability sample.

To introduce the modeling approach for a nonprobability sample, consider estimating a finite population total. The general idea in model-based estimation is to sum the responses for the sample cases and add to them the sum of predictions for nonsample cases. The key to unbiased estimation is that the variables to be analyzed for the sample and nonsample follow a common model and that this model can be discovered by analyzing the sample responses. When both the sample and nonsample units follow the same model, model parameters can be estimated from the sample and used to make predictions for the nonsample cases. An appropriate model usually includes covariates which are known for each individual sample case. The covariates are usually unknown for individual nonsample cases.

Suppose that the mean of a variable y_i follows a linear model:

$$E_M(y_i|X_i) = X_i'\boldsymbol{\beta},$$

where the subscript M denotes the expectation with respect to the model, X_i is a vector of p covariates for unit i and $\boldsymbol{\beta}$ is a parameter vector. Given a sample s, an estimator of the slope parameter is

$$\widehat{\boldsymbol{\beta}} = (X_s'X_s)^{-1}X_s'y_s$$

where $\boldsymbol{X}_s$ is the $n \times p$ matrix of covariates for the sample units, and $\boldsymbol{y}_s$ is the n-vector of sample y's.

A predictor of the y population total is:

$$\hat{t} = \sum_{i \in s} y_i + (\boldsymbol{t}_{Ux} - \boldsymbol{t}_{sx})' \widehat{\boldsymbol{\beta}}$$

where $\boldsymbol{t}_{Ux}$ and $\boldsymbol{t}_{sx}$ are vectors of X totals for the population and sample, respectively. The estimated total $\hat{t}$ can be written as the weighted sum of the observed y's where the weights are:

$$w_i = 1 + (\boldsymbol{t}_{Ux} - \boldsymbol{t}_{sx})'(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}X_i$$

This estimator is equal to the GREG estimator (Särndal, Swensson and Wretman, 1992) if the inverse selection probabilities in that estimator are all set to 1. Note that the weights depend only on the x's and not on y. As a result, the same set of weights can be used for all sample estimates. A single set of weights will not be equally efficient for every y, but this situation is also true for design-based weights.

2.3 Propensity Weighting

This is the propensity weighting or quasi-randomization approach as discussed in Elliot and Valliant (2017). It requires the presence of a probability sample, called a reference sample, selected from the target population. Under this approach, one fits a logistic regression model to estimate the inclusion probability of the nonprobability units, and then use the predicted probabilities to derive the nonprobability sample weights or pseudo weights. Here are the steps for developing the propensity weights:

- Concatenate the probability sample and the nonprobability sample;
- Create a dichotomous variable, R, which is coded 1 for nonprobability sample units and 0 for probability sample units;
- Fit a logistic regression model with R as the response variable;
- Use the predicted propensities as the estimated inclusion probabilities for the nonprobability sample units;
- Compute the nonprobability sample weights as the inverse of the predicted inclusion probabilities.

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, and some response variables collected from the survey. The final model is validated through cross validation and by examining model diagnostic statistics. As we will see later, the ability to include response variables from the survey turns out to be a major advantage of this approach.

2.4 Small Area Modeling

Small area estimation methods are used to jointly model domain-level estimates for one or more key survey variables from the probability and the nonprobability sample (Ganesh et al., 2017). The model includes a set of covariates (X), fixed and random bias terms, and domain-level random effects. The nonprobability sample weights are developed via the following steps:

- 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}) :

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

  - d is a demographic group (e.g. 18-34 year old, male, Hispanic).
  - $\mathbf{x}_d$ is a vector of covariates.
  - v_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.
  - $\varepsilon_d^P, \varepsilon_d^{NP}$ are the sampling errors associated with y_d^P, y_d^{NP}.
- Predicted small area estimates for each domain are obtained using an Empirical Best Linear Unbiased Predictor (EBLUP).
- Nonprobability sample weights are derived such that combined sample estimates (using the weights) match the small area estimates for each domain for one or more key survey variables.

The small domains are defined by cross-classifying a set of demographic variables that are of interest:

- Age (18-34 years, 35-49 years, 50-64 years, 65+ years),
- Education (Some college or less, college graduate or higher),
- Race/Hispanic ethnicity (Hispanic, non-Hispanic Black, non-Hispanic All Other), and
- Gender (male, female)

The choice of domains was motivated by "sufficient" sample size for the probability and non-probability samples for each domain, and also to capture the variation in the substantive estimates across domains.

3. Monte Carlo Simulation Setup

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 adult survey completes from a large-scale national study about food allergies, as follows:

- Frame 1, the full population frame, consists of all 40,539 adult survey completes. Random samples selected from Frame 1 are considered probability samples.
- Frame 2 is a subset of Frame 1, and consists of 36,917 adult survey completes. To impart coverage bias to Frame 2, we sort Frame 1 by some key variables, and then selected cases for removal. We selected integers from a binomial distribution with n = 40,538 and p = 0.25, added 1 to each integer so that they range was from 1 to 40,539, and removed records with row numbers matching the selected values. Due to multiple selection of the same integers, only 3,622 (9 percent) of Frame 1 records were removed to create Frame 2. Random samples selected from Frame 2 are considered nonprobability samples with respect to Frame 1.

Both the probability and nonprobability frames/samples contain a large number of demographic and webographic variables. Demographic variables include: age, gender, race/ethnicity, education, employment, marital status, household income, household size (including children), home ownership, household telephone service, and more. Webographic variables include household internet access among others. The food allergy study survey responses used for the Frames also contain self-reported and doctor-diagnosed food allergies, both current and outgrown, allergy reactions, experiences in allergy treatments, events coinciding with development or outgrowing a food allergy, and perceived risks associated with food allergies.

The number of Monte Carlo iterations is 2,500. For each iteration:

- A probability sample of size 400 is selected using SRSWOR from Frame 1, and standard design weight is computed for each sample case;
- A nonprobability sample of size 800 is selected using SRSWOR from Frame 2, and pseudo weight is derived for each case using each of the four estimation methods under evaluation;
- Weighted estimates from each sample are derived for response variables of interest. For the nonprobability sample, four different estimates are derived corresponding to each of the estimation methods;
- Composite estimation is used to derive the combined estimates for response variables based on both probability and nonprobability samples, where the combination factor is proportional to the probability and nonprobability sample size;
- Bias, mean squared error (MSE), and true confidence interval coverage associated with the composite estimates are derived for each response variable under each estimation method.

Final summary statistics for each evaluation variable and estimation method combination are computed by averaging over the 2,500 iterations.

4. Comparisons of Composite Estimates

For each response variable, known bias associated with the nonprobability frame (Frame 2) under-coverage is calculated using the difference of population proportions between Frames 1 and 2. That is,

$$B_{pop} = P_{Frame\ 1} - P_{Frame\ 2}$$

where $P_{Frame\ 1}$ and $P_{Frame\ 2}$ is the population proportion computed from the probability and nonprobability frame for a response variable, respectively. The magnitude and direction of known bias differs by response variable. Figure 1 shows the size of known bias associated with the six variables that have the largest absolute bias[3]. The six variables are ordered by the size of the absolute bias in this and the other figures, with the first variable having the largest absolute bias and the last variable having the smallest absolute bias. Our subsequent evaluations will focus on these six response variables only.
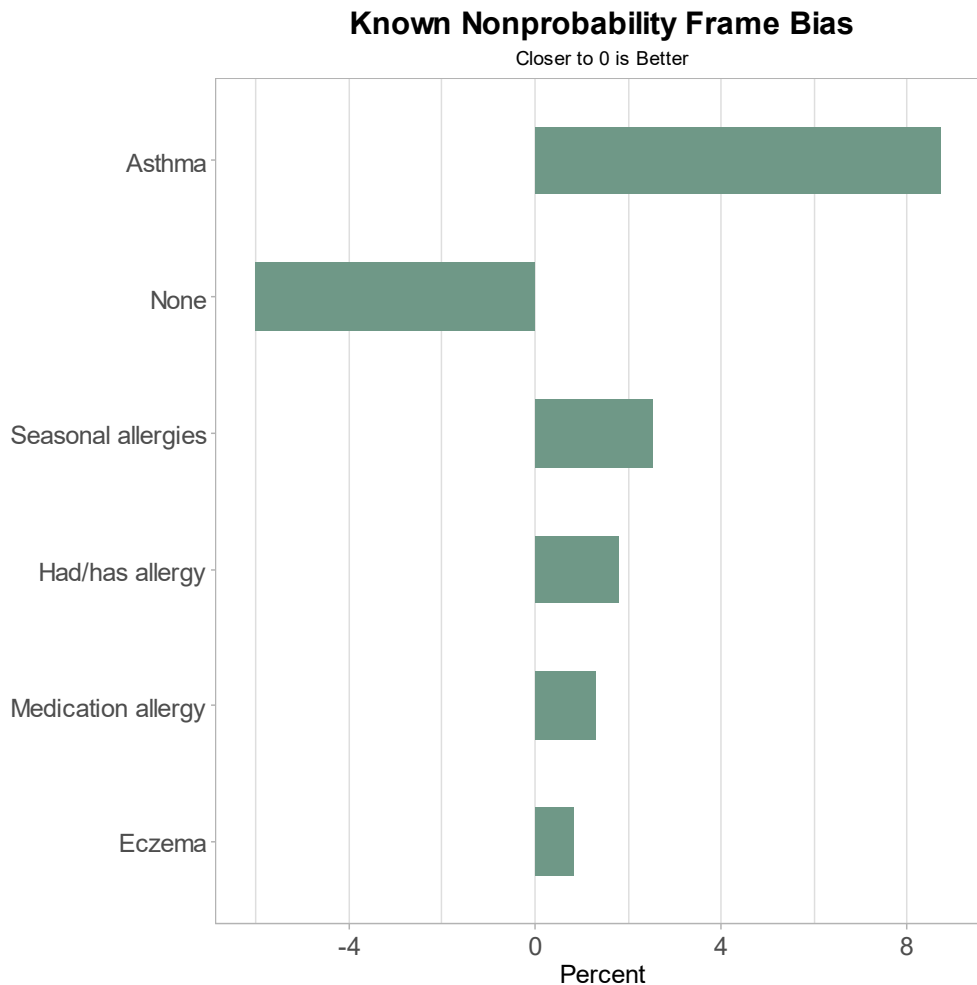
Known Nonprobability Frame Bias

Closer to 0 is Better

Figure 1: Six Response Variables with Largest Absolute Bias

---

[3] The six response variables with the largest bias are: (1) Asthma—Doctor diagnosed chronic conditions for Asthma, (2) None—No doctor diagnosed chronic conditions, (3) Seasonal Allergies—Doctor diagnosed chronic conditions for Hay fever/allergic rhinitis/seasonal allergies, (4) Had/has Allergy—Ever has doctor diagnosed allergies, (5) Medication Allergy—D Doctor diagnosed chronic conditions for Medication allergy, and (6) Eczema—Doctor diagnosed chronic conditions for Eczema. Note that the bias reported here was created for this simulation study only. The actual food allergy data do not exhibit these biases.

The key concern with nonprobability samples is potential estimation bias due to frame coverage bias and/or sample selection bias. We first compare the relative ability of each estimation method in reducing estimation bias.

For each response variable, let $\hat{p}_{P,m}$ denote the proportion estimate associated with Monte Carlo probability sample m (of size 400). Similarly, let $\hat{p}_{NP,m}$ denote the proportion estimate associated with Monte Carlo nonprobability sample m (of size 800) under a particular estimation method[4]. And, let $\hat{p}_{Comp,m} = \lambda\hat{p}_{P,m} + (1 - \lambda)\hat{p}_{NP,m}$, with $\lambda = \frac{n_P}{n_P + n_{NP}}$, denote the composite estimate for Monte Carlo combined sample m (of size 1,200). The composite estimate is a weighted probability and nonprobability sample estimates with the weights proportional to the relative sample size.

The estimated bias associated with the composite estimate for each iteration is defined as the difference between $\hat{p}_{Comp,m}$ and the true population proportion $P_{Frame\ 1}$,

$$\hat{b}_m = P_{Frame\ 1} - \hat{p}_{Comp,m}$$

For each response variable, the average bias associated with the composite estimate under each estimation method is computed as

$$\bar{\bar{b}} = \frac{1}{2500}\sum \hat{b}_m$$

Finally, for each response variable, *percent absolute bias reduction* is computed as

$$\frac{|B_{pop} - \bar{\bar{b}}|}{B_{pop}}\%$$

Figure 2 below compares the percent of bias reduction under each estimation approach for the six response variables. The vertical orange line represents the median percent of bias reduction for each method over the six variables. All estimation methods achieve some level of bias reduction, ranging from roughly one third to two thirds. However, with a median of about 56 percent, the Propensity method stands out as the clear winner with most bias reduction, while the other three methods lead to similar and much smaller bias reduction with a median of about 34 percent. The lone exception is for response variable "had/has allergy" (whether the respondent currently has or ever had any diagnosed food allergy) which has the largest percent of bias reduction at 65% under Small Area Modeling. This variable happens to be the variable that is modeled as the dependent variable under the Small Area Modeling method. In general, our simulations have shown consistently that, for the modeled response variables, the Small Area method leads to greater bias reduction than the other methods, with the caveat that the small area models are correctly specified. We also observe from Figure 2 that, under each method, the smaller the known bias for a variable, the higher the percent of bias reduction for the variable,

---

[4] For simplicity, we suppressed the subscript that identifies each specific estimation method.

although there are occasional exceptions. This pattern implies that it is generally challenging to remove large biases associated with nonprobability samples.
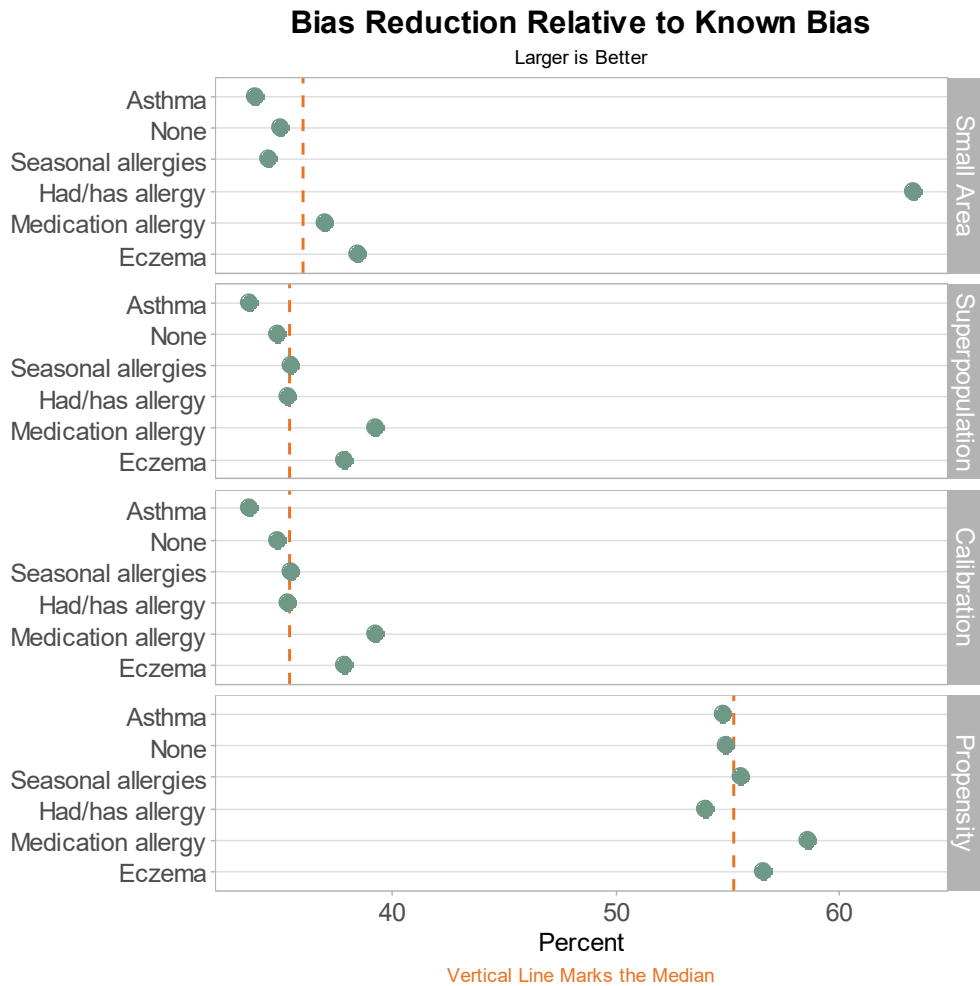

Figure 2: Percent Bias Reduction for the Response Variables under Each Method

For each response variable, the mean squared error (MSE) of the composite estimate is defined as

$$\widehat{MSE}_{Comp} = \frac{1}{2500} \sum (\hat{p}_{Comp,m} - P_{Frame\ 1})^2$$

To compare MSEs, which vary greatly depending on the underlying proportion distribution, we consider the ratio of MSE to Known Variance. The latter is defined as follows for each response variable,

$$V_{pop} = P_{Frame\ 1} * (1 - P_{Frame\ 1})/1200$$

Figure 3 shows the comparisons of the ratios for the response variables under each estimation method. For each variable, the ratios are very close under Small Area, Superpopulation, and

Calibration, following a pattern similar to that observed in Figure 2. Note that the ratio for the variable "had/has allergy" is not particularly small under Small Area method in spite of the much greater bias reduction as shown in Figure 2. This implies that the composite estimate under Small Area has relatively large variance which dominates the MSE given that the known bias for this variable is quite small. Although the differences are not large, Figure 3 still shows clearly that the Propensity method performs better than the other methods based on the MSE to Known Variance ratio.
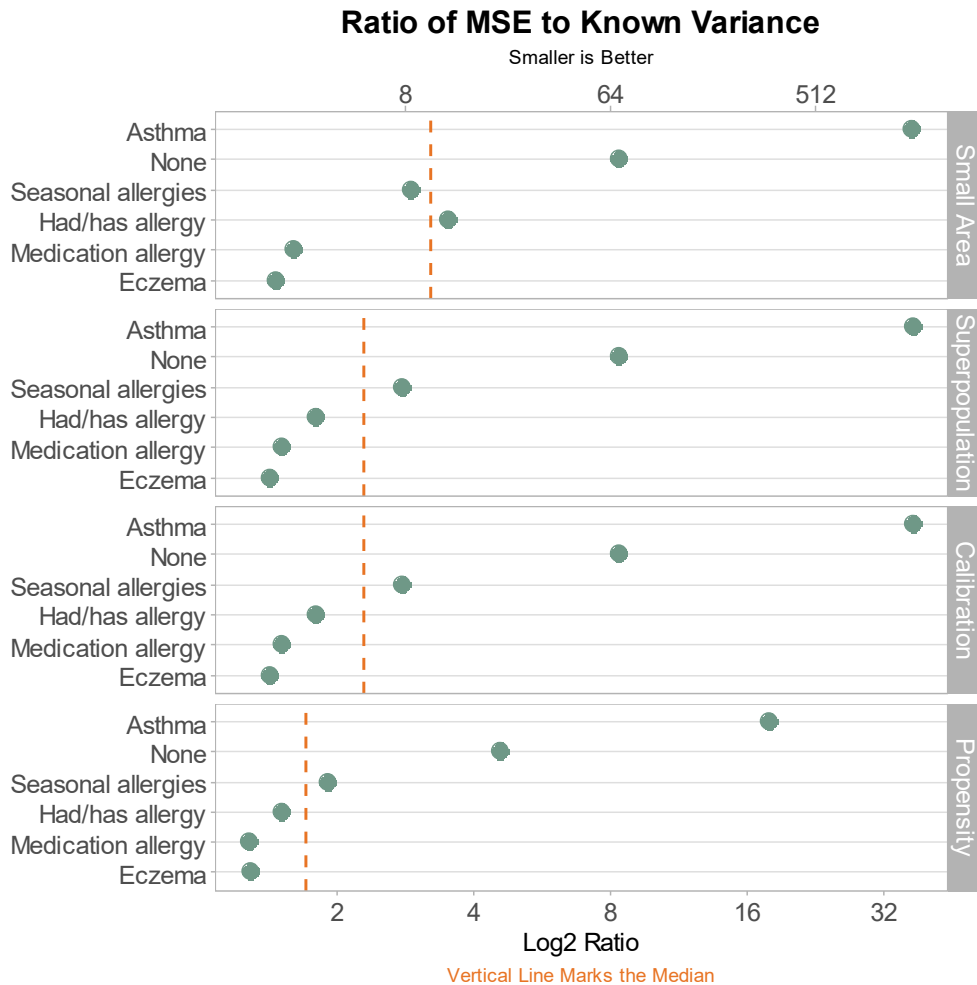

Figure 3: Ratio of MSE to Known Variance for the Response Variables under Each Method

We finally compare the true 95 percent confidence interval coverage associated with the composite estimates for each response variable. First, we construct a 95 percent confidence interval around each composite estimate $\hat{p}_{Comp,m}$ using as standard error the square root of $\widehat{MSE}_{Comp}$. Next, we calculate the percentage of the 2,500 intervals that contain the population true value. The results are reported in Figure 4.
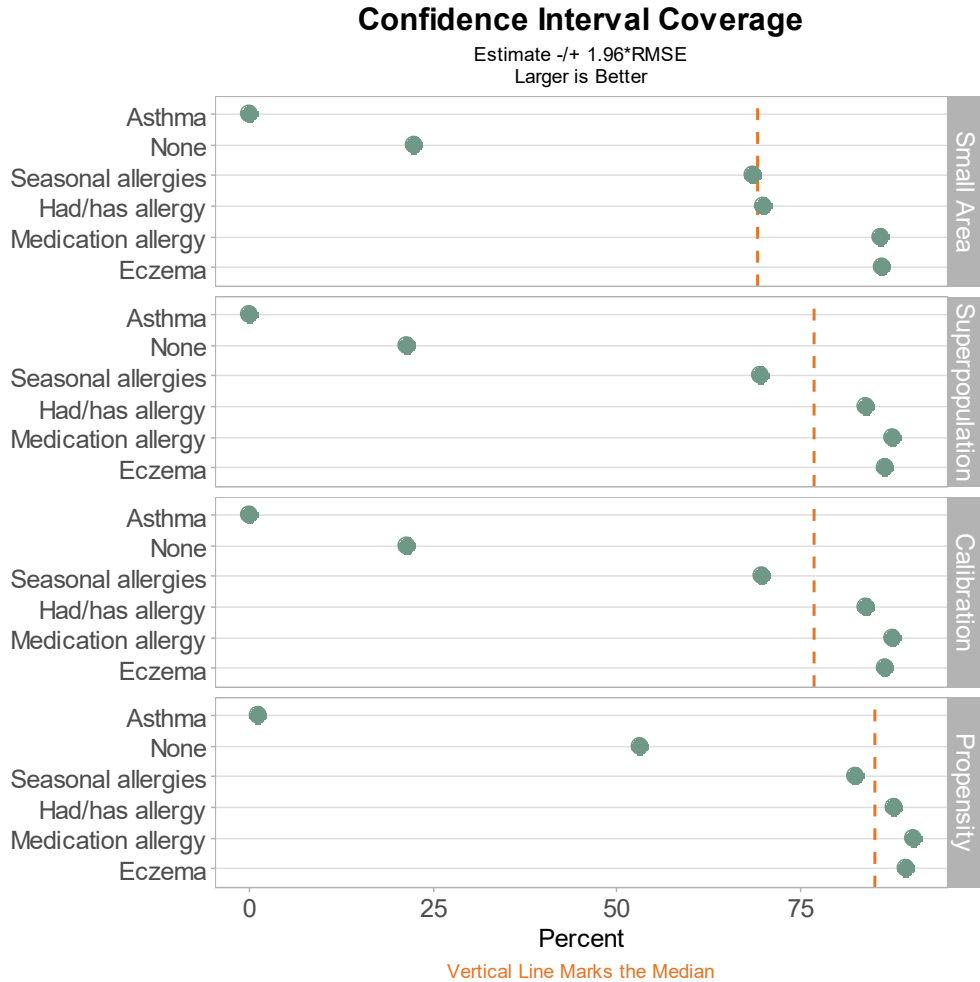
Confidence Interval Coverage

Estimate -/+ 1.96*RMSE
Larger is Better

Vertical Line Marks the Median

Figure 4: True 95 Percent Confidence Interval Coverage

Despite significant bias reduction under all methods and for all response variables, confidence interval coverage is low for variables with large bias. For the three variables with the smallest bias, the true coverage is not far from the nominal coverage of 95 percent. For variables with larger bias, the true coverage rate is much lower. The variable Asthma has the largest bias and lowest confidence interval coverage, followed by variable None which has the second largest bias and second lowest coverage. The median coverage is lowest for Small Area; Superpopulation and Calibration have the same median at about 75%; and Propensity has the highest median coverage at about 88%.

5. Summary and Discussions

Our simulations show that the Propensity Weighting method consistently outperforms the other methods in terms of bias reduction, MSE, and true confidence interval coverage. In particular, the Propensity Weighting method achieves much greater bias reduction than its competitors, which should be considered a tremendous advantage given that our main concern with nonprobability samples is potential bias due to coverage and selection bias.

The Small Area Modeling method achieves the most bias reduction for the modeled response variable(s), which is another consistent finding in our research based on real survey data or/and simulations[5]. For the other response variables, Small Area Modeling provides results similar to Calibration and Superpopulation in terms of bias, MSE, and confidence interval coverage.

Calibration and Superpopulation methods consistently lead to modeled weights and composite estimates with very similar properties, which should be expected as both methods rely on the same set of covariates and corresponding population benchmarks.

In general, all the estimation methods investigated in this study are model-based, but the underlying model and the covariates differ across the methods. The covariates incorporated into Calibration and Superpopulation Modeling are limited to basic demographics because population totals are needed. These demographic variables are: gender, age, race/ethnicity, income, and education. Small Area Modeling includes all these demographic variables in the small area models, plus additional domain level variables from external data sources such as the American Community Survey (ACS) and the Current Population Survey (CPS). The logistic regression model under the Propensity method incorporates a larger number of predictor variables, including response variables collected from the underlying survey. It is apparent that the Propensity Method's advantage derives primarily from the fact that it has the flexibility to incorporate a more comprehensive set of covariates, including response variables, into the logistic regression model. Based on our test results, this advantage shrinks or disappears when we restricted the predictor variables to the same set of covariates as used by Calibration and Superpopulation Modeling.

All nonprobability samples are not the same as they are likely to be biased in different ways. In the absence of any overarching theoretical guidance, we need to rely on empirical research and simulation studies to increase our knowledge about the properties of estimates based on nonprobability samples. We hope that the accumulation of such knowledge will help us to develop fit-for-purpose nonprobability estimation methods for different types of nonprobability samples and response variables. For example, opt-in web panels may differ from the general population in predictable ways and such knowledge should be taken into account in developing estimation methods for opt-in web panels. Empirical research may also help to identify the key covariates for different types of response variables. For example, basic demographic and socioeconomic variables, the most accessible covariates in practice, may be highly predictive of some response variables, but they may be quite inadequate for predicting other response variables.

In this study, we evaluated the properties of composite estimates based on four estimation methods for combining probability and nonprobability samples. While our earlier research did not identify any preferred methods, this simulation study shows clearly that Small Area Modeling and Propensity Weighting tend to perform better than Calibration and Superpopulation Modeling. Given their limited ability in incorporating covariates under Calibration and Superpopulation Modeling, we intend to devote most of our future research to the more promising Small Area Modeling and Propensity Weighting methods.

---

[5] Our other studies show that the Small Area method also produces better subgroup estimates, an important result that will be reported separately.

Small Area Modeling is primarily designed to enable better small domain estimation. As such, we would like to expand our current investigation to compare small domain composite estimates under the different estimation approaches. Preliminary findings indicate that Small Area Modeling achieves better confidence interval coverage than the other methods at the domain level. Given this and our observation that Propensity Weighting consistently outperform the other methods, we will explore adapting the Small Domain Model to work with Propensity Weighting.

References

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.

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.

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

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., Dorfman, A. H. and Royall, R. M. (2000). Finite Population Sampling and Inference: A Prediction Approach. Wiley, New York.

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

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.