Using Statistical Matching to Account for Coverage Bias When Combining Probability and Nonprobability Samples

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

Abstract
Many methods have been developed to combine probability and nonprobability samples via quasi-randomization, superpopulation modeling, or doubly robust estimation (Valliant, 2020). Yang, et al (2018) observed that when using statistical matching (a quasi-randomization approach) there may be a proportion of probability sample units that do not match to nonprobability sample units. Given this observation, a reasonable conjecture is that this unmatched portion of the probability sample provides a means to assess the coverage bias of the nonprobability sample. Ma and Mulrow (2019) developed an approach that used statistical matching to produce estimates from combined probability and nonprobability samples, and observed its behavior via a case study. We explore this approach further using the simulation approach in Yang, et al (2019) to assess the bias reduction and confidence interval coverage of the matching approach compared to other methods.

Key Words: Nonprobability sample, statistical matching

1. Introduction

Probability sampling is the gold standard for survey research; however, demand for methods that utilize nonprobability samples, alone or combined with a probability sample, has grown in order to lower survey costs. Nonprobability samples may provide a lower cost alternative to probability samples; yet, 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 for 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 of this paper have evaluated methods for utilizing nonprobability samples in conjunction with a probability sample (Ganesh et al., 2017; Yang, et al. 2018, 2019). 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. A third quasi-randomization method that uses statistical matching (Matching) to match probability sample records to nonprobability sample records in order to assign weights to the nonprobability sample records was considered in the Yang et al 2018 case study, but was not part of the subsequent simulation study. An observation from the case study was that a number of probability sample records did not match to any nonprobability sample records. Given this observation, a reasonable conjecture is that this unmatched portion of the probability sample provides a means to assess the coverage bias of the nonprobability sample.

Ma and Mulrow (2019) developed an approach to combining probability and nonprobability samples that used matching along with propensity modeling to produce weights for the nonprobability sample, and observed properties of estimates via a case study. In this paper, we investigate properties of a matching approach using a simulation. While we take a naïve approach to imputing weights from the probability sample onto the nonprobability sample, the method performs well in terms of bias reduction and confidence interval coverage.

2. The Idea

Our approach is to determine appropriate weights for each record in a nonprobability sample by imputing weights from a companion probability sample. The circumstances are that the probability and nonprobability sample respondents are surveyed with the same instrument during a similar time period. The key difference is that the probability sample records each have a weight determined by the design of the sample, whereas the nonprobability sample was not designed and has missing weights. Using the probability sample records as donors, we match a probability sample record to each nonprobability record, and assign the nonprobability record the weight from the matching probability sample record. When a probability sample record is used as a donor more than once, the weights for all the nonprobability recipients are divided by the number of times the probability sample record was used as a donor.

2.1 Statistical Matching of Probability Sample Records to Nonprobability Sample Records

Statistical matching (Matching) is carried out using a nearest neighbor hot deck algorithm based on a distance measure. The matching process resembles imputation in the sense that a donor from the probability sample is matched to a recipient from the nonprobability sample based on a set of matching variables (Bethlehem, 2015). We used the R StatMatch package NND.hotdeck function (D'Orazio, 2017). Each nonprobability sample record is matched to one and only one probability sample record under the following conditions:

- A match to a nonprobability sample record is done by finding the closest probability sample record according to the Gower distance function. When for a given nonprobability sample record there are several probability sample records at the minimum distance, one of them is picked at random.
- Distances are measured using Gower's dissimilarity measure, which can use both categorical and continuous variables in the dissimilarity calculation.
- The nonprobability record assumes the weight of the matched probability record. However, when a probability record is matched to multiple nonprobability records, each matched nonprobability record's weight is the probability record weight divided by the number of matches.

We use extreme gradient boosting to determine the set of matching variables for use in the Gower dissimilarity calculation. This is an ensemble learning algorithm that constructs and combines weak learners in iterative fashion and eventually results in a strong learner. At each iteration, the algorithm tailors a learner for the local bias not successfully accounted for by the former learners. From a global perspective, each learner is weak (i.e., high bias) by itself, but the combined learner usually has low bias after a sufficient number of iterations (D'Orazio, Di, and Scanu, 2006). We use the R *xgboost* package to predict membership in the nonprobability sample using extreme gradient boosting, and select the top 20 influential features to form the set of matching variables.

2.2 Use a Composite Estimator based on Matching Results

We have found that it is almost always the case that there are unmatched records from the probability sample, which may be evidence of coverage bias of the nonprobability sample. If so, we should use this information in constructing an estimator that uses the combined set of probability and nonprobability sample records.

Notation:

- S_P^U, the set of unmatched probability sample records
- S_P^M, the set of matched probability sample records
- S_{NP}, the set of nonprobability sample records
- w_{Pi}, the weight of r_i (record i) in the probability sample
- $\mathcal{w}_{NPi}$, the imputed weight of r_i in the nonprobability sample
- X_i, an indicator variable for the attribute of interest for respondent i

We consider an estimator for the total number of units in the population with an attribute of interest. An estimator for the total number of people with the attribute based on the probability sample is

$$\hat{X}_P = \sum w_{Pi}X_{Pi} = \sum_{i \in S_P^U} w_{Pi}X_{Pi} + \sum_{i \in S_P^M} w_{Pi}X_{Pi} = \hat{X}_P^U + \hat{X}_P^M,$$

and an estimator for the total number of people with the attribute based on the nonprobability sample is

$$\hat{X}_{NP} = \sum \mathcal{w}_{NPi}X_{NPi}$$

To develop an estimator for the combined probability and nonprobability samples, we propose a composite estimator that blends the matched probability sample with the nonprobability sample, and leaves the unmatched probability sample "as-is."

$$\hat{X}_{comb} = \hat{X}_P^U + \lambda\hat{X}_P^M + (1-\lambda)\hat{X}_{NP}, \text{ where } 0 \le \lambda \le 1.$$

Note that this implies that the weights for the combined sample are

$$w^*{}_i = \begin{cases} w_{Pi} & r_i \in S_P^U \\ \lambda w_{Pi} & r_i \in S_P^M \\ (1-\lambda)\mathcal{w}_{NPi} & r_i \in S_{NP} \end{cases}$$

A proportion (mean) estimate for an attribute of interest is

$$\hat{p}_{comb} = \frac{\hat{X}_{comb}}{\sum w^{*}{}_{i}}$$

In what follows, we investigate this estimator via a simulation, and compare it to estimators based on propensity weighting and small area modeling. The latter two were investigated via simulation in Yang et al 2019.

3. Monte Carlo Simulation Setup

We use the same simulation process as in Yang et al 2019, in which the type of coverage bias typically exhibited in online opt-in nonprobability samples is mimicked by creating two sampling frames. One frame is a subset of the other, and both frames consist of adult survey completes from a large-scale national study about food allergies.

- 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, we sorted 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 the potential range of the generated values was the same as the range of row numbers—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 raking is used with gender, education, income, race, and age to derive a weights for each sample record;
- A nonprobability sample of size 800 is selected using SRSWOR from Frame 2, and a pseudo weight is derived for each record using the statistical matching method described in Section 2.1;
    i. Pseudo weights were also developed for the propensity and small area modeling methods as described in the Appendix;
- Combined sample estimates of the six response variables with the largest known bias are calculated for the combined sample using the method

described in Section 2.2. The value of λ is the proportion of the matched probability sample's effective sample size relative to the total of the effective samples for each sample;

  i. Combined sample estimates were also developed for the propensity and small area modeling methods as described in Yang et al 2019;

- Bias, mean squared error (MSE), and true confidence interval coverage associated with the composite estimates are derived for each of the six response variables.

Final summary statistics for each of the six response variable–weighting method combinations are computed by averaging over the 2,500 iterations.

4. Comparisons of Combined 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 absolute bias associated with the six variables that have the largest absolute bias.[1] 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 variables only.

[1] 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.

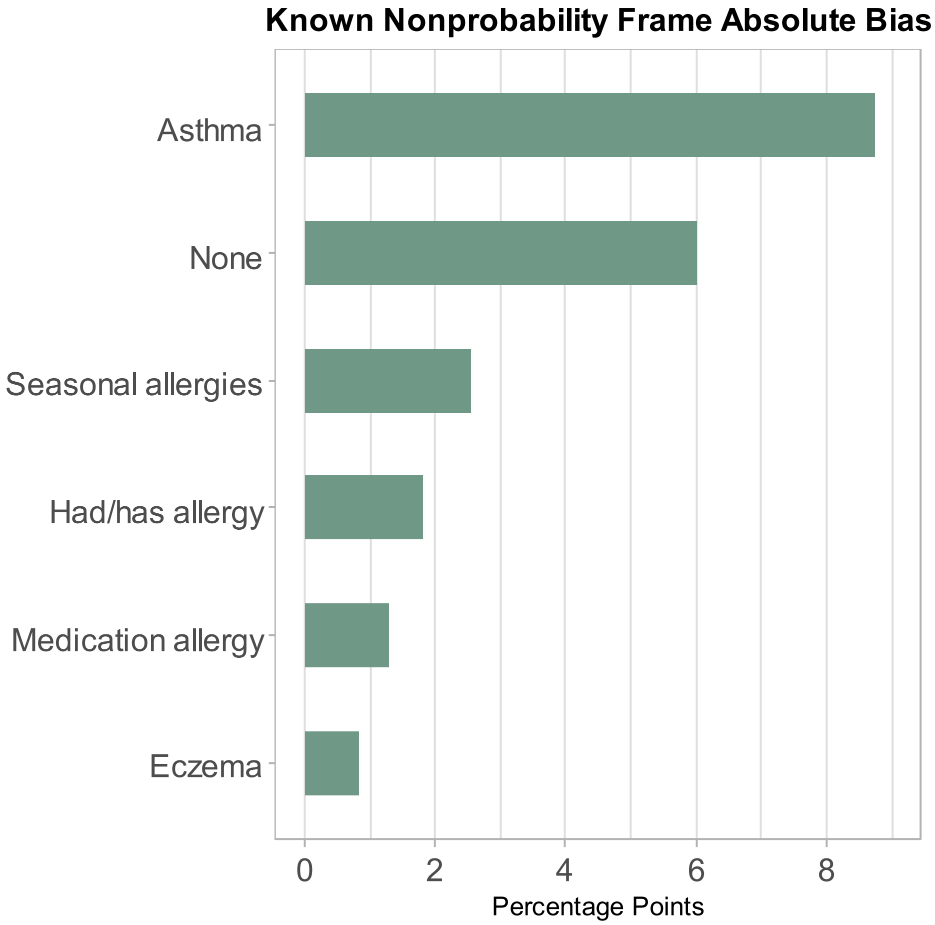

Figure 1: Six Response Variables with Largest Absolute Bias

The key concern with nonprobability samples is potential bias due to frame coverage bias and/or sample selection bias. We first compare the relative ability of each estimation method in reducing such bias. The estimated bias associated with a combined estimate for each iteration is defined as the difference between $\hat{p}_{Comb,m}$ and the true population proportion $P_{Frame\ 1}$,

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

For each response variable, the average bias across the 2,500 iterations associated with the combined estimate under each estimation method is computed as

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

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

$$\frac{\left|B_{pop} - \bar{\hat{b}}\right|}{\left|B_{pop}\right|} \times 100\%$$

Figure 2 below compares the percent of bias reduction under each weighting approach for the six response variables with the largest known bias. 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, with the medians ranging from approximately 55% to 90%. However, with a median of about 90 percent, the Matching method stands out as the method with the most bias reduction, while the Propensity method has the lowest, but consistent, bias reduction with a median of 55 percent. Even though the Small Area method has a reasonably good median absolute bias reduction of 75%, the bias reduction across all six response variables is inconsistent. However, we note that method has close to 100% absolute bias reduction for two variables—one of which has the second highest known absolute bias—and when the method did not perform well, the known absolute bias is small (under 1.5 %). Small Area also has close to 90% bias reduction for asthma, which is the variable with the largest know bias.

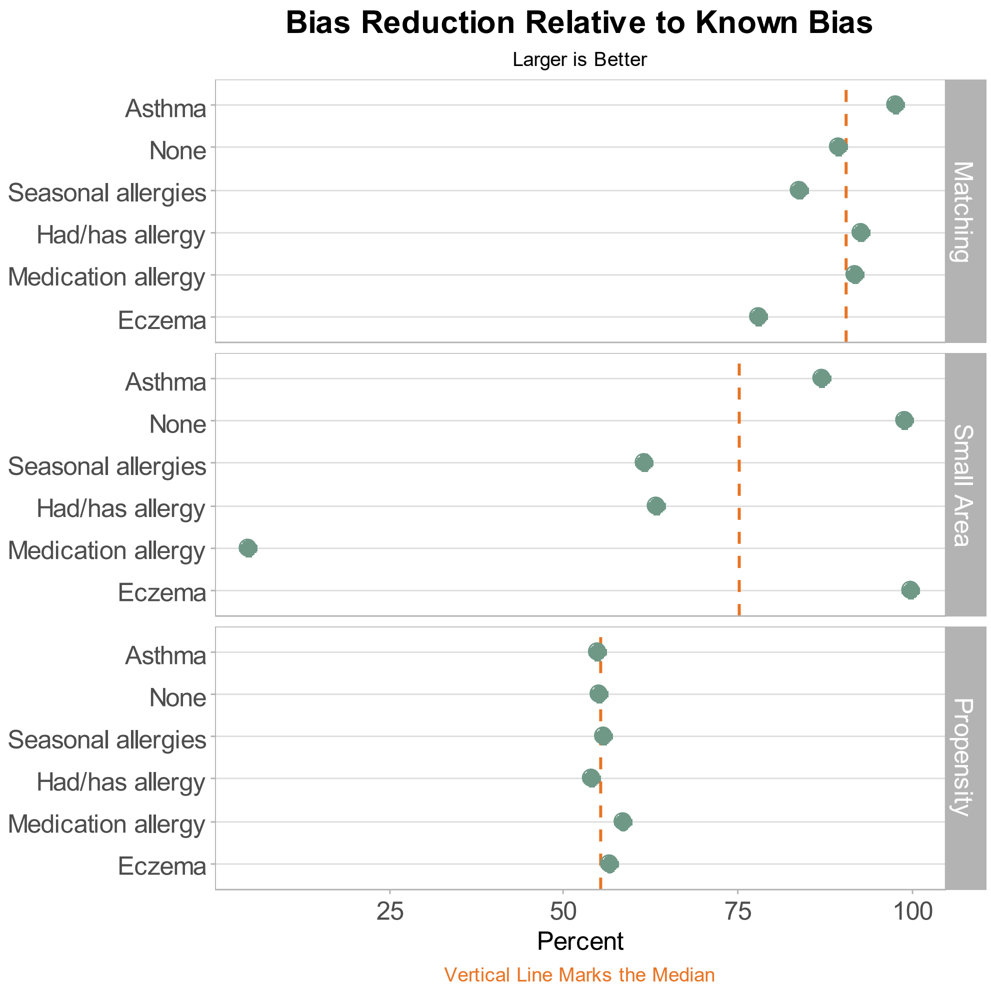

Figure 2: Percent Bias Reduction for the Response Variables under Each Method

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

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

To evaluate confidence interval coverage, we construct a 95 percent confidence interval around each combined estimate $\hat{p}_{Comb,m}$ using as standard error the square root of $\widehat{MSE}_{Comb}$. Next, we calculate the percentage of the 2,500 intervals that contain the population true value. The results are reported in Figure 3.

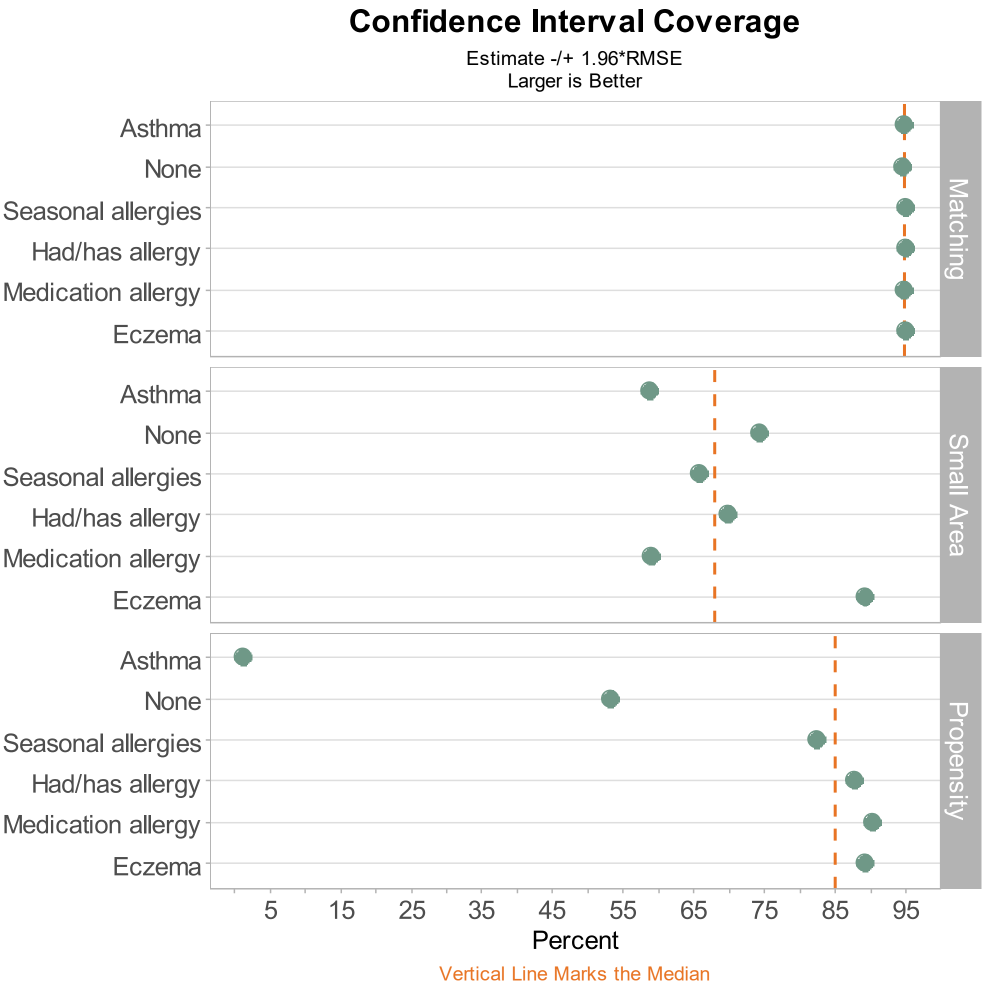

Figure 3: True 95 Percent Confidence Interval Coverage

Matching is the clear winner in this comparison as it is the only method that consistently has 95% confidence interval coverage. The Propensity method sometimes achieves coverage above 90%, but performs poorly for the two variables with the highest known bias. The Small Area method has the lowest median confidence interval coverage, but its minimum across the six response variables is higher than the minimum for the Propensity method.

5. Summary

We have taken a very simple, and perhaps naïve, approach to deriving weights for nonprobability sample records using statistical matching. Our simulations show that the Matching method consistently outperforms the other methods in terms of bias reduction and confidence interval coverage. It is somewhat surprising that the method of imputing weights from a companion probability sample performs well compared to other more sophisticated methods.

There is more to the method than weight derivation. The matching process identifies records in the probability sample that do not share similar characteristics with any of the nonprobability records, and this may be an indicator of nonprobability sample's coverage bias. So, in addition to determining weights for nonprobability sample records, the combined estimator attempts to correct for coverage bias by leaving the weights of the unmatched probability sample record alone.

In this paper, we have provided some basic measures to evaluate the Matching method, and compare it to two other methods. While Matching performs well based on the metrics we have used in our evaluation, there are additional metrics could be used. For example, we suspect that the design effect (DEFF) under matching may be higher compared to other methods because the weight variation under matching is much higher that other methods. We have begun to investigate this, and plan to report on it in the future. For now, we believe that Matching is a method for combining probability and nonprobability samples that should be included in one's toolbox. Careful analysis of a nonprobability sample should be conducted before any combining methods is applied.

Appendix: Background on the Propensity and Small Area Methods

Here we provide some details on how we implemented the Propensity and Small Area methods for determining nonprobability weights when there is a companion probability sample. These descriptions are taken from Yang et al, 2017.

A.1 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.

A.2 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 ($\mathrm{y}_\mathrm{d}^\mathrm{P}$) and the nonprobability sample ($\mathrm{y}_\mathrm{d}^\mathrm{NP}$):

$$\mathrm{y}_\mathrm{d}^\mathrm{P} = \mathbf{x}'_\mathrm{d}\boldsymbol{\beta} + \nu_\mathrm{d} + \varepsilon_\mathrm{d}^\mathrm{P}$$
$$\mathrm{y}_\mathrm{d}^\mathrm{NP} = \mathrm{b} + \alpha_\mathrm{d}^\mathrm{NP} + \mathbf{x}'_\mathrm{d}\boldsymbol{\beta} + \nu_\mathrm{d} + \varepsilon_\mathrm{d}^\mathrm{NP}$$

    - d is a demographic group (e.g. 18-34 year old, male, Hispanic);
    - $\mathbf{x}_\mathrm{d}$ 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;
    - $\varepsilon_\mathrm{d}^\mathrm{P}$, $\varepsilon_\mathrm{d}^\mathrm{NP}$are the sampling errors associated with $\mathrm{y}_\mathrm{d}^\mathrm{P}$, $\mathrm{y}_\mathrm{d}^\mathrm{NP}$.

- 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.

References

J. Bethlehem. 2015. “Solving the nonresponse problem with sample matching?” *Social Science Computer Review*, Vol. 34, No. 1, pp. 59–77.

M. D'Orazio. 2017. StatMatch: Statistical Matching. R package version 1.2.5. https://CRAN.R-project.org/package=StatMatch.

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

M. R. Elliot and R. Valliant. 2017. “Inference for Nonprobability Samples,” *Statistical Science* 2017, Vol. 32, No. 2, 249–264.

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.

N. Ganesh, V. Pineau, A. Chakraborty, J. M. Dennis. 2017. Combining Probability and Non-Probability Samples Using Small Area Estimation. Joint Statistical Meetings Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1657-1667.

Q. Ma and E. Mulrow. 2019. Exploring Hybrid Methods for Estimation with Combined Probability and Nonprobability Samples. JSM Presentation.

Y. M. Yang, N. Ganesh, E. Mulrow, and V. Pineau. 2018. Estimation Methods for Nonprobability Samples with a Companion Probability Sample. JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1715-1723.

Y. M. Yang, N. Ganesh, E. Mulrow, and V. Pineau. 2019. Evaluating Estimation Methods for Combining Probability and Nonprobability Samples through a Simulation Study. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1714-1727.

R. Valliant. 2020. “Comparing Alternatives for Estimation from Nonprobability Samples,” *Journal of Survey Statistics and Methodology*, Vol. 8, No. 2, pp. 231–263.