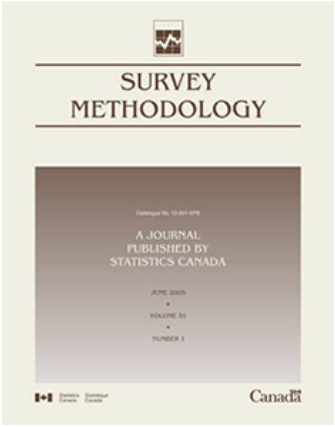
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Statistical inference with non-probability survey samples”



by Michael R. Elliott

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 319-329
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical inference with non-probability survey samples"

Michael R. Elliott[1]

Abstract

This discussion attempts to add to Wu's review of inference from non-probability samples, as well as to highlighting aspects that are likely avenues for useful additional work. It concludes with a call for an organized stable of high-quality probability surveys that will be focused on providing adjustment information for non-probability surveys.

Key Words: Pseudo-weighting; Propensity score; Doubly-robust estimation; Sensitivity analysis.

1. Introduction

Thanks to Dr. Changbao Wu for an excellent review of the previous work and open issues for statistical inference from non-probability samples. Given the large and rapidly developing work in this area, Dr. Wu was understandably unable to cover all of it; my own understanding has blinders as well but I will touch on a few additional approaches that relate to topics he considered. I will also discuss the issue of modeling versus weighting for different inferential targets, and use his discussion and conclusions to highlight the critical importance of probability samples – in particular high-quality studies that focus on estimation of relevant covariates – to improve inference for the profusion of non-probability samples used as replacements for traditional probability samples in many research and official statistics settings. To avoid notation confusion, all notation will follow that of Wu, except where new notation is required.

Section 2 reviews additional approaches to combining data from probability and non-probability surveys. Section 3 briefly reviews the issue of weighting versus modeling when adjusting non-probability survey data. Section 4 reviews some recent developments in sensitivity analyses of standard assumptions for adjusting non-probability survey data using probability survey data. Section 5 concludes with call to systematically design a set of probability surveys with the explicit purpose of adjusting non-probability surveys.

2. Additional approaches to combining data from probability and non-probability surveys

Dr. Wu's paper follows the general prescription of 1) using model estimation and subsequent calibration to probability-sample-estimated covariate distributions, 2) developing propensity score estimates based on discrepancies between the probability- and non-probability sample data, and 3) doubly-

1. Michael R. Elliott, Department of Biostatistics, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu.

robust methods that combine 1) and 2) in a manner such that only one of the two underlying models needs to be correct.

2.1 Propensity score estimators

Rivers (2007) appears to have been the first to suggest estimating propensity score using logistic regression with membership in the non-probability sample as the outcome and taking the reciprocal of the resulting propensity scores to use as inclusion weights. This approach was formalized further in Valliant and Dever (2011). Separately, using simple results from Bayes' theorem and discriminant analysis first described in Elliott and Davis (2005), Elliott, Resler, Flannagan and Rupp (2010) and Elliott (2013) developed a somewhat different estimator of the form

$$\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha}) = \hat{P}(i \in S_A) \propto P(i \in S_B) \frac{\hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})}{\hat{P}(i \in S_B \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})}. \quad (2.1)$$

$\hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$ can be obtained using logistic regression, or using one of the suite of machine learning-type approaches such as support vector machines (Soentpiet, 1999), targeted maximum likelihood estimation (Van Der Laan and Rubin, 2006), or Bayesian Additive Regression Trees (BART) (Chipman, George and McCulloch, 2010), and $\hat{P}(i \in S_A \mid i \in S_B \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$ obtained as $1 - \hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$. In principle $P(i \in S_B) = 1/d_i^B$ is known since sampling probabilities are known for all elements of the population, including those in the non-probability sample, but in practice analysts with access only to public use data may have to estimate this as well. (In addition, d_i^B may include calibration and non-response adjustments that are not known for the non-probability sample elements.) This last point is critical as use of the probability sample to develop propensity scores using only the discrepancies between the non-probability sample and the probability sample will be biased unless the probability sample used an equal probability (epsem) design, as noted by Wu.

In contrast, Chen, Li and Wu (2020) shows that using a pseudo-likelihood approach to estimating $\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ directly from the population likelihood for the indicators $I(i \in S_A)$ as a function of $\mathbf{x}_i$ yields an estimator that does not require $P(i \in S_B)$ for elements in the non-probability sample under the restriction that $\pi_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ follows a generalized linear model with a canonical link, i.e., logistic regression.

(None of these approaches actually has the correct intercept to obtain a true propensity score; however, as noted in Wu, weighted estimation usually uses Hájek-type estimators [using weights to estimate a population total for denominators; Hájek, 1971] so that propensity scores estimated up to a normalizing constant are sufficient.)

2.2 Doubly-robust estimators

If inference is focused on a particular variable Y available only in the non-probability sample, we can return to the model-assisted estimators that date back to Cassel, Särndal and Wretman (1976), which posit a model for the expectation $E(y_i \mid \mathbf{x}_i) = m_i$. Combining this with propensity score estimates of the

probability of being in the non-probability sample (which we are treating as an "unknown probability sample" – more about this under Assumptions below) yields estimators of the form

$$\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}+\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i \tag{2.2}$$

corresponding to $\hat{\mu}_{\mathrm{DR}2}$ of (4.11) in Wu. The intuition is that any bias due to the model misspecification in estimation of m_i in $\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i$ will be equal to and opposite in sign of $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ if the model for π_i^A is correctly specified. Conversely, if the model for π_i^A is misspecified but m_i is correctly specified, $y_i-\hat{m}_i$ will be iid with mean zero and consequently $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ will also have mean 0, yielding an unbiased estimator. Chen, Valliant and Elliott (2019) used LASSO for prediction in combination with generalized regression estimators (McConville, Breidt, Lee and Moisen, 2017) when $\mathbf{X}$ is of high dimension. As Wu notes, Wu and Sitter (2001) show the equivalence between GREG applied to predicted values and DR estimators of the form in (2.2), which indicates that the Chen et al. (2019) approach was equivalent to (2.2) with LASSO estimation for m_i and an assumption of simple random sampling for the non-probability sample.

A disadvantage of using (2.1) as opposed to Chen et al. (2020) as the estimator of π_i^A, and thus of d_i^A, is the requirement that the probability sample weights d_i^B be known or at least estimated for the non-probability sample. An advantage of using (2.1), is that non-linear models and machine learning methods can be used in estimation. Rafei, Flannagan and Elliott (2020) uses BART to estimate both m_i and π_i^A, reducing the impact of potential model misspecification. Simulations showed considerable improvement in bias and variance reduction over the method of Chen et al. (2020) when the linear models is misspecified. Variance estimation can proceed by adapting Rubin's multiple imputation rules: from M independent draws from BART, the mean of the variances computed treating the draw of d_i^A as known using standard complex sample design estimators and added to $\frac{M+1}{M}$ times the variance of the point estimates computed across the draws of d_i^A yield an approximately unbiased variance estimator.

An alternative approach to doubly-robust estimation uses the fact that the propensity score is the coarsest possible "balancing score" that contains all of the information about the association between the sampling indicator and the outcome of interest. This has led to the development of mean estimators that use smooth functions of weights to produce consistent estimators that can be more efficient when weights are highly variable or only weakly related to the outcome (Elliott and Little, 2000; Zheng and Little, 2005). Zhou, Elliott and Little (2019) extended this idea into the causal inference setting in non-randomized settings, in which probability of assignment to a treatment or exposure (propensity score) is estimated as a function of covariates $P_Z\left(\mathbf{x}_i,\boldsymbol{\alpha}\right)$ using logistic regression, and then non-observed potential outcomes Y^z under treatment arm $z_i'\neq z_i$ for observed treatment z_i are imputed from

$$Y_i^Z\sim N\left(s\left(\hat{P}_Z^*\left(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}\right)\middle|\boldsymbol{\theta}_Z\right)\right)+g_Z\left(\hat{P}^*\left(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}\right),\mathbf{x}_i\middle|\boldsymbol{\beta}_Z\right),\sigma^2 \tag{2.3}$$

where P^* is the logit transformation of P, $s(\hat{P}_Z^* \mid \boldsymbol{\theta}_Z)$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996) of propensity, and $g_Z(\hat{P}^*, \mathbf{x}_i \mid \boldsymbol{\beta}_Z)$ is a general function of covariates including the propensity scores. The resulting estimator is doubly robust in the sense that if either $P_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ or $E(Y^z) = g_Z(\hat{P}^*, \mathbf{x}_i \mid \boldsymbol{\beta}_Z)$ is correctly specified, $Y^{(z)}$ will be approximately unbiased; see Zhang and Little (2009). This can be implemented in the non-probability setting by replacing $\hat{P}_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ in the mean model for (2.3) with $\hat{\pi}_i^A$ estimated using (2.1) to obtain a draw of $Y_i^{(b)}$. (Note this requires obtaining $\hat{\pi}_i^A$ for the probability sample elements requiring prediction.) Inference can proceed by obtaining $b = 1, \ldots, B$ draws from the posterior distribution of the estimated population quantity of interest, e.g., for the population mean

$$Y^{(b)} = \frac{\sum_{i \in S_R} N_i^{(b)} Y_i^{(b)} + \sum_{i \in S_A} \left(y_i - Y_i^{(b)} \right)}{N}$$

where now $N_i^{(b)}$ is a estimate of the population represented by the weight d_i^R obtained from a finite population Bayesian bootstrap (Little and Zheng, 2007); more complete FBPP extensions to complex sample designs that include clustering and stratification are available in Dong, Elliott and Raghunathan (2014).

As in the estimation of (2.1), the non-parametric (spline) component of (2.3) can be replaced with other machine-learning estimators; see Chapter 4 of Rafei (2021) for implementation using Gaussian processes. Also, extensions to non-normal models are direct, although not necessarily computational easy.

2.3 Poststratified estimators

Wu also describes the use of poststratified estimators in the context of quota sampling, which is not only a very old form of non-probability sampling but indeed the standard before Neyman made the case for stratified random sampling (Neyman, 1934). Wu's Section 5 suggests a robust alternative to the propensity score estimates obtained by ordering observations in the probability sample by $\hat{\pi}_i$, stratifying into K strata based on this ordering, and computing the predicted proportion of the population belonging to the k^{th} stratum as proportion of the sample weights W_k in this stratum using the probability sample, with

$$\hat{\mu}_{\text{PST}} = \sum_k \hat{W}_k \bar{y}_k \tag{2.4}$$

where $\bar{y}_k$ is the mean within the k^{th} stratum in the non-probability sample. Wu notes the tradeoff between choosing K to be large enough to retain homogeneity within units but small enough to obtain stable estimates of $\bar{y}_k$, suggesting 30 as the old "rule of thumb" for "large [enough] sample sizes". I would add that a more formal approach discussed in Little (1986) suggests a method to generate strata (there in the context of non-response adjustment) that minimizes mean square error by maximizing the

between-stratum-to-within-stratum variance. It would seem such an approach would be appropriate to consider in the non-probability post-stratified estimator as well.

A more direct approach to obtain estimates using a post-stratified type estimator is multilevel regression and poststratification (Wang, Rothschild, Goel and Gelman, 2015; Downes and Carlin, 2020). Here only data from the non-probability sample is used in the outcome model:

$$E\left(Y_{k[i]}\right) = \beta_0 + \mathbf{x}_k^T\boldsymbol{\beta} + \sum_j a_{l[k]}^j \tag{2.5}$$

where $k = 1, \ldots, K$ indexes the poststratum developed from $j = 1, \ldots, J$ variables, $a_{l[k]}^j \sim N(0, \sigma_j^2)$ for $l = 1, \ldots, L_j$ and $l[k]$ maps the postratum cell k to the appropriate category l of variable j. The poststratifed estimator is still given by (2.4) with $\hat{W}_k$ now replaced with known population totals W_k; posterior inference is obtained though posterior draws of β_0, $\boldsymbol{\beta}$, and $a_{l[k]}^j$ to obtain a draw of

$$\hat{\mu}_{\text{PST}}^{(b)} = \sum_k W_k \left[\frac{1}{n_k} \sum_{i \in k} \left(\beta_0^{(b)} + \mathbf{x}_k^T\boldsymbol{\beta}^{(b)} + \sum_j a_{l[k]}^{j(b)} \right) \right].$$

Though not technically doubly-robust, it has been shown to work well in some applications where J is large enough to capture all of the important discrepancies between the probability and non-probability sample, and the non-probability sample is sufficiently large to allow reasonably accurate estimation of $a_{l[k]}^j$. In the absence of known joint distributions of a high dimensional $\mathbf{X}$, this approach has the weakness of relying on estimated distributions, which are unstable. A possible alternative might be replace the simple $\bar{y}_k$ with (2.5) in Wu's poststratified estimator (2.4), using the fact that the sampling weights d_i^R summarize the information about $\mathbf{X}$ in the probability sample similar to that of the propensity score for non-probability sample.

3. Weighting vs. modeling for the general user

Wu's paper and the above addendums tend to follow the long-trodden path regarding weighting versus modeling in the finite population inference setting, dating back at least to Hansen, Madow and Tepping (1983). In thinking about this choice I believe it is important to distinguish between models used to derive so-called descriptive parameters – in the sense of Kalton (1983) – and models that are of interest in and of themselves, so-called analytic parameters in regression models, latent classes analysis, etc. For the former distinguishing a descriptive target of interest Y from potential modeling covariates $\mathbf{X}$ has the advantage of creating doubly-robust estimators that are targeted to a single descriptive parameter. This also requires assumptions such as A1 in Section 2.1 (propensity score does not depend on Y conditional on $\mathbf{X}$). When models themselves are the targets of interest, it may be that developing weights via propensity scores to account for selection bias and, as Wu notes, employing standard weighted estimating equations may be the most sensible choice, since typically a wide number of models may be considered. This comes at the cost

of double robustness, since there is usually no attempt to model the analytic parameter directly. Developing ways to extend double-robustness into a broader class of model parameter estimates may be a fruitful exercise.

4. Unverifiable assumptions: Recent developments in sensitivity analysis

Wu provides four key assumptions required to correct for selection bias in non-probability surveys using data from probability surveys: they can be roughly summarized as "selection at random" or SAR (covariates in the non-probability sample explain the probability of selection in the non-probability sample); "positivity" (all elements in the population have a non-zero probability of selection into the non-probability sample); "independence" (elements are selected independently into the non-probability sample); and "common covariates" (there exists a probability survey with covariates whose subset matched the covariates required for the MAR assumption to hold). It might be worth noting that the first two assumptions basically require the non-probability survey to be a probability survey "in disguise" – that is, there really are non-zero probabilities of selection into the non-probability survey for all elements in the population, but we as analysts just do not know what they are.

In practice neither of these assumptions probably hold precisely. Some recent work has focused on the failure of the first, the SAR assumption. Some existing measures borrowed from the non-response literature have been repurposed here: for example, the R-indicator measure (Schouten, Cobben and Bethlehem, 2009), which in this context is the measure of the variability in the probabilities of selection in the non-probability sample:

$$\hat{R} = 1 - 2\sqrt{\frac{1}{n_a - 1}\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A - \sum_{j=1}^{n_A}\hat{\pi}_j^A / n_a\right)^2}$$

$\hat{R}$ can range between 0 and 1, where 1 is achieved when probabilities of selection are constant – suggesting something akin to a simple random sample, with less chance for selection bias – and 0 – suggesting all elements are either included with probability 1 or 0, maximizing the risk of selection bias.

Of course, in the absence of the outcome Y in the probability sample, there is no way to directly assess selection bias. Hence recent work has extended Andridge and Little (2011), which develops a sensitivity analysis using a pattern-mixture model, wherein selection into non-probability sample is allowed to depend entirely on a scalar reduction to the covariates $\mathbf{X}$, entirely on the outcome Y, or some convex combination thereof. Little, West, Boonstra and Hu (2020), Andridge, West, Little, Boonstra and Alvarado-Leiton (2019), and West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton (2021) consider sensitivity to this assumption in the estimation of the mean of a normally distributed variable, the mean of a binary outcome, and the regression parameters in a linear regression model, respectively, in non-probability samples. By varying the convex mixing parameter ϕ, sensitivity to the SAR assumption

can be assessed. Boonstra, Little, West, Andridge and Alvarado-Leiton (2021) finds that these "standard measures of bias" (SMB) compare favorably with alternatives such as $\hat{R}$ in a simulation study. An important point to note is that the methods that extend Andridge and Little (2011) do not depend on assumption of common covariates in a probability sample. This suggests that methods that use information available in the probability sample to assess SAR are an open area for development.

The second assumption – positivity – is also unlikely to exist precisely in many practical settings. My own work in this area has focused on naturalistic driving studies, which typically involve convenience samples in a limited geographical area: for example, the Second Strategic Highways Research Program (SHRP2) recruited drivers in six specific geographic regions across the United States (Transportation Research Board (TRB) of the National Academy of Sciences, 2013). This corresponds to the second scenario given by Wu in Section 7.2, where only a subpopulation has any chance of being selected into the non-probability sample, which as he notes has "no simple fix". Following his notation of D providing an indicator of membership in the subpopulation, it would seem that if $D_i \perp \mathbf{X}_i, Y_i \mid \pi_i^A$ – that is, if the distribution of $\mathbf{X}, Y$ is the same for $D = 0$ and $D = 1$ after weighting for π_i^A within the $D = 1$ stratum – then lack of positivity would have no impact on inference. This is likely a tall order in the most general settings but might be reasonably well approximated if the analysis of interest involves a subset of $\mathbf{X}, Y$ that is only weakly associated with D even before adjustment.

Finally, regarding the fourth assumption – existence of a probability sample with available $\mathbf{X}$ – I very much second Wu's observation that methods to take advantage of multiple probability surveys need more development. However, it remains more likely that a researcher will struggle to find a single probability sample with sufficient covariates than struggle with a surfeit of options (Wu's "rich person's problem"). To this end I will conclude with a call to action by the survey community.

5. Probability sampling in the 21st century: Now more than ever

I learned statistics, and particularly survey statistics, near the end of the 20th century, when probability sampling was the unchallenged touchstone of survey design. I was first introduced to the problem of making inference from non-probability samples in the late 00's in the context of injury analysis using Crash Injury Research (CIREN) data, where analysts were treating a highly-restricted sample of individuals in passenger vehicle crashes as if they were a random sample of crash victims and consequently finding non-sensible results (Elliott et al., 2010). About the same time web surveys were exploding in popularity and survey statisticians were somewhat at a loss as to how to make inference from such data. I will admit to a rather paternalistic attitude at the time – I almost avoided trying to do research in this area because I thought it would only encourage "bad behavior" regarding sample design. I did not think I could single-handedly stop it, but I did not want to participate in what I perceived as the downgrading of science. I came to recognize, however, that many of these new data sources have advantages beyond what can be achieved through the traditional probability sample, certainly within

limited budgets. This is above and beyond the increasing challenges to implementing probability surveys, especially in general populations, due to non-response, lack of adequate sampling frames, etc.

However, I remain concerned that the idea that we have developed methods to deal with the limitations of non-probability surveys means that probability sampling is passe is becoming entrenched among scientists and policy makers with limited statistical training, despite efforts like those of Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Marek, Tervo-Clemmens, Calabro et al. (2022). However, as Wu's review notes, the absence of probability samples unmoors the non-probability sample from the possibility of even partial calibration or other adjustment approaches (although sensitivity analyses such as those SMB approaches noted above do not require benchmarking probability samples). Hence I believe it is increasingly critical for an organized and ideally government funded stable of high-quality probability surveys to be put into place for routine data collection. Some of these obviously already exist – the US Census' American Community Survey and the National Center for Health Statistics National Health Interview Survey premier among them – but going forward I believe it would be valuable for statistical agencies to explicitly coordinate around the need for high quality probability surveys to serve a role as analytic partners to the non-probability survey world rather than just as stand-alone products. This means thinking carefully about important covariates across a variety of public health and social science roles in which survey data play a role. Choices will have to be made given limited budget constraints, and at the same time provisions should be made for sufficient funding to retain the quality needed for adjustment. Finally, while some methods do not require microdata and thus can use summary measures such as those avaiable in the American Communities Survey, other will require such data, which likely means new areas of research to be explored in the fields of privacy and confidentiality research as applied to the combining of data from probability and non-probability surveys.

References

Andridge, R.R., and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society*, C68, 1465-1483.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society*, 68, 657-681.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266-298.

Dong, Q., Elliott, M.R. and Raghunathan, T.E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40, 1, 29-46. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14003-eng.pdf.

Downes, M., and Carlin, J.B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, 62, 479-491.

Eilers, P.H., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Elliott, M.R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. *Journal of the Royal Statistical Society*, C54, 595-609.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Elliott, M.R., Resler, A., Flannagan, C.A. and Rupp, J.D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference*, 236.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175-188.

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

Little, R.J.A., and Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8, 1-20.

Little, R.J.A., West, B.T., Boonstra, P.S. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.

Marek, S., Tervo-Clemmens, B., Calabro, F.J. et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, in press.

McConville, K.S., Breidt, F.J., Lee, T. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5, 131-158.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. University of Michigan Thesis. Accessible at https://www.overleaf.com/project/6228db145a47be05f8da3777.

Rafei, A, Flannagan, C.A.C. and Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8, 148-180.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*. Available at https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Soentpiet, R. (1999). *Advances in Kernel Methods: Support Vector Learning*. Boston: MIT Press.

Transportation Research Board of the National Academy of Sciences (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.*

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40, 105-137.

Van Der Laan, M.J., and Rubin, D.R. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 980-991.

West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zhang, G., and Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 911-918.

Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Zhou, T., Elliott, M.R., and Little, R.J.A. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114, 1-19.