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

Survey Methodology

Integration of data from probability surveys and big found data for finite population inference using mass imputation

by Shu Yang, Jae Kwang Kim and Youngdeok Hwang

Release date: June 24, 2021

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 STATCAN.infostats-infostats.STATCAN@canada.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

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

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, June 2021
Vol. 47, No. 1, pp. 29-58
Statistics Canada, Catalogue No. 12-001-X

Integration of data from probability surveys and big found data for finite population inference using mass imputation

Shu Yang, Jae Kwang Kim and Youngdeok Hwang[1]

Abstract

Multiple data sources are becoming increasingly available for statistical analyses in the era of big data. As an important example in finite-population inference, we consider an imputation approach to combining data from a probability survey and big found data. We focus on the case when the study variable is observed in the big data only, but the other auxiliary variables are commonly observed in both data. Unlike the usual imputation for missing data analysis, we create imputed values for all units in the probability sample. Such mass imputation is attractive in the context of survey data integration (Kim and Rao, 2012). We extend mass imputation as a tool for data integration of survey data and big non-survey data. The mass imputation methods and their statistical properties are presented. The matching estimator of Rivers (2007) is also covered as a special case. Variance estimation with mass-imputed data is discussed. The simulation results demonstrate the proposed estimators outperform existing competitors in terms of robustness and efficiency.

Key Words: Calibration weighting; Data fusion; Generalized additive model; Matching; Nearest neighbor imputation; Post stratification.

1. Introduction

In finite population inference, probability sampling is the gold standard for obtaining a representative sample from the target population. Because the selection probability is known, the subsequent inference from a probability sample is often design-based and respect the way in which the data were collected; see Särndal, Swensson and Wretman (2003), Cochran (2007), Fuller (2009) for textbook discussions. However, large-scale survey programs continually face heightened demands coupled with reduced resources. Demands include requests for estimates for domains with small sample sizes and desires for more timely estimates. Simultaneously, program budget cuts force reductions in sample sizes, and decreasing response rates make nonresponse bias an important concern. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013) address the current challenges in using probability samples for finite population inferences.

To meet the new challenges, statistical offices face the increasing pressure to utilize convenient but often uncontrolled big data sources (also called big found data), such as satellite information (McRoberts, Tomppo and Næsset, 2010), mobile sensor data (Palmer, Espenshade, Bartumeus, Chung, Ozgencil and Li, 2013), and web survey panels (Tourangeau, Conrad and Couper, 2013). Couper (2013), Citro (2014), Tam and Clarke (2015), and Pfeffermann, Eltinge and Brown (2015) articulate the promise of harnessing big data for official and survey statistics but also raise many issues regarding big data sources. While such data sources provide timely data for a large number of variables and population elements, they are non-probability samples and often fail to represent the target population of interest because of inherent selection biases. Tam and Kim (2018) also cover some ethical challenges of big data for official

1. Shu Yang, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695, U.S.A. E-mail: syang24@ncsu.edu; Jae Kwang Kim, Department of Statistics, 1208 Snedecor Hall, Iowa State University, Ames, IA 50011, U.S.A. Youngdeok Hwang, Paul H. Chook Department of Information Systems and Statistics, Baruch College, City University of New York, New York, NY 10010, U.S.A.

statisticians and discuss some preliminary methods of correcting for selection bias in big data. See Keiding and Louis (2016), Elliott and Valliant (2017), Buelens, Burger and van den Brakel (2018), and Beaumont (2020) for recent reviews of the challenges in using non-probability samples for inferences.

To utilize modern data sources in statistically defensible ways, it is important to develop statistical tools for data integration for combining a probability sample with big non-probability data. Data integration for finite population inference is similar to the problem of combining randomized clinical trial studies and non-randomized epidemiological studies for causal inference of treatment effects (Keiding and Louis, 2016). We are particularly interested in developing data integration under the setup where the study variable is observed in the big data only, but some other variables are commonly observed in both data. In this case, survey statisticians and biostatisticians have provided different methods for combining information from multiple data sources. Lohr and Raghunathan (2017), Yang and Kim (2020), and Rao (2020) provide a review of statistical methods of data integration for finite population inference. Existing methods for data integration can be categorized into three types as follows.

The first type is the so-called propensity score adjustment (Rosenbaum and Rubin, 1983). In this approach, the probability of a unit being selected into the big sample, which is referred to as the propensity score, is modeled and estimated for all units in the big data sample. The subsequent adjustments, such as propensity score weighting or stratification, can then be used to adjust for selection biases; see, e.g., Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017). Stuart, Bradshaw and Leaf (2015), Stuart, Cole, Bradshaw and Leaf (2011), Buchanan, Hudgens, Cole, Mollan, Sax, Daar, Adimora, Eron and Mugavero (2018) use propensity score weighting to generalize results from randomized trials to a target population. O'Muircheartaigh and Hedges (2014) propose propensity score stratification for analyzing a nonrandomized social experiment. One of the notable disadvantages of the propensity score methods is that they rely on an explicit propensity score model and are biased if the model is mis-specified (Kang and Schafer, 2007).

The second type uses calibration (Deville and Särndal, 1992; Kott, 2006; Dong, Yang, Wang, Zeng and Cai, 2020). This technique can be used to calibrate auxiliary information in the big data sample with that in the probability sample, so that after calibration the big data sample is similar to the target population (DiSogra, Cobb, Chan and Dennis, 2011). Because calibration does not require parametric modeling, it is attractive to survey practitioners. However, this approach requires the information (such as the moments) of the auxiliary variables for the population is known or at least can be estimated from a probability sample.

The third type is mass imputation, where the imputed values are created for the whole elements in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample provide a training dataset for developing an imputation model. In the mass imputation, an independent big data sample is used as a training dataset, and imputation is applied to all units in the probability sample. While the mass imputation idea for incorporating information from big data is very natural, the literature on mass imputation itself is sparse. Breidt, McVey and Fuller (1996) discuss mass imputation for two-

phase sampling. Rivers (2007) proposes a mass imputation approach using nearest neighbor imputation but the theory is not fully developed. Kim and Rao (2012) develop a rigorous theory for mass imputation using two independent probability samples. Chipperfield, Chessman and Lim (2012) discuss composite estimation when one of the surveys is mass imputed. Bethlehem (2016) discuss practical issues in sample matching. Recently, Kim and Wang (2019) develop a theory for mass imputation for big data using a parametric model approach. However, the parametric model assumptions do not necessarily hold in practice. In order for mass imputation to be more useful and practical, the assumptions should be as weak as possible.

We summarize our contributions in this paper below:

1. We first develop a formal framework for mass imputation incorporating information from big data into a probability sample and present rigorous asymptotic results for the mass imputation estimators. Our framework covers the nearest neighbor imputation estimator of Rivers (2007). Unlike Kim and Wang (2019), we do not make strong parametric model assumptions for mass imputation. Thus, the proposed method is appealing to survey practitioners.
2. We also investigate two strategies for improving the nearest neighbor imputation estimator, one using k nearest neighbor imputation (Mack and Rosenblatt, 1979) and the other using generalized additive models (Wood, 2006). In k nearest neighbor imputation, instead of using one nearest neighbor, we identify multiple nearest neighbors in the big data sample and use the average response as the imputed value. This method is popular in the international forest inventory community for combining ground-based observations with imagines from remote sensors (McRoberts et al., 2010). In this paper, we establish asymptotic results for the k nearest neighbor estimator. In the second strategy, we investigate modern techniques of prediction for mass imputation with flexible models. We use generalized additive models (Wood, 2006) to learn the relationship of the outcome and covariates from the big data and create predictions for the probability samples. We note that this strategy can apply to a wider class of semi- and non-parametric estimators such as single index models, Lasso estimators (Belloni, Chernozhukov, Chetverikov and Kato, 2015), and machine learning methods such as random forests (Breiman, 2001).
3. Using a novel calibration weighting idea, we propose an efficient mass imputation estimator and develop its asymptotic results. The efficiency gain is justified under a purely design-based framework and no model assumptions are used. We consider the case when additionally the membership to the big data can be determined throughout the probability sample. The key insight is that the subsample of units in Sample A with the big data membership constitutes a second-phase sample from the big data sample, which acts as a new population. We calibrate the information in the second-phase sample to be the same as the new acting population. The calibration process in turn improves the accuracy of the mass imputation estimator without specifying any model assumptions.

The structure of the paper is as follows. In Section 2, we introduce the basic setup. In Section 3, we present the methodology for the nearest neighbor imputation and establish its asymptotic properties. In Section 4, we investigate two strategies for improving the nearest neighbor imputation estimator, one using k nearest neighbor imputation and the other using generalized additive models. In Section 5, we propose a regression calibration technique to improve the efficiency of the mass imputation estimators when additionally the big data membership is observed throughout the probability sample. In Section 6, we demonstrate that the proposed estimators are robust and efficient by simulation studies based on artificial data and real-life data from U.S. Census Bureau's Monthly Retail Trade Survey. In Section 7, we present a case study applying the proposed method to integrate national health survey data and national health insurance records. Section 8 concludes with a discussion.

2. Basic setup

2.1 Notation: Two data sources

Let $\mathcal{F}_N = \{(\mathbf{X}_i, Y_i): i \in U\}$ with $U = \{1, \ldots, N\}$ denote a finite population, where $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)$ is a p-dimensional vector of covariates, and Y_i is the study variable. We assume that $\mathcal{F}_N$ is a random sample from a superpopulation model ζ, and N is known. Our objective is to estimate the general finite population parameter $\mu_g = N^{-1}\sum_{i=1}^N g(Y_i)$ for some known $g(\cdot)$. For example, if $g(Y) = Y$, $\mu_g = N^{-1}\sum_{i=1}^N Y_i$ is the population mean of Y. If $g(Y) = \mathbf{1}(Y < c)$ for some constant c, $\mu_g = N^{-1}\sum_{i=1}^N \mathbf{1}(Y_i < c)$ is the population proportion of Y less than c.

Suppose that there are two data sources, one from a probability sample, referred to as Sample A, and the other from a big data source, referred to as Sample B. Table 2.1 illustrates the observed data structure. Sample A contains observations $\mathcal{O}_A = \{(d_i = \pi_i^{-1}, \mathbf{X}_i): i \in A\}$ with sample size $n = |A|$, where $\pi_i = P(i \in A)$ is known throughout Sample A, and Sample B contains observations $\mathcal{O}_B = \{(\mathbf{X}_i, Y_i): i \in B\}$ with sample size $N_B = |B|$. Often the probability sample contains many other items but we only use those items overlapping with our big data. Although the big data source has a large sample size, the sampling mechanism is often unknown, and we cannot compute the first-order inclusion probability for Horvitz-Thompson estimation. The naive estimators without adjusting for the sampling process are subject to selection biases. On the other hand, although the probability sample with sampling weights represents the finite population, it does not observe the study variable.

Table 2.1
Two data sources. "√" and "?" indicate observed and unobserved data, respectively

		Sample weight $d = \pi^{-1}$	Covariate $\mathbf{X}$	Study Variable Y
Probability Sample	1	√	√	?
$\mathcal{O}_A$	⋮	⋮	⋮	⋮
	n	√	√	?
Big Data Sample	1	?	√	√
$\mathcal{O}_B$	⋮	⋮	⋮	⋮
	N_B	?	√	√

Sample A is a probability sample, and Sample B is a big data but may have selection biases.

2.2 Assumptions

Let $f(Y|\mathbf{X})$ be the conditional density function of Y given $\mathbf{X}$ in the superpopulation model ζ. Let $f(\mathbf{X})$ and $f(\mathbf{X}|\delta_B = 1)$ be the density function of $\mathbf{X}$ in the finite population and Sample B, respectively, where δ_B is the indicator of selection to Sample B. We first make the following assumptions.

Assumption 1 (Ignorability). *Conditional on* $\mathbf{X}$, *the density of* Y *in Sample B follows the superpopulation model; i.e.,* $f(Y|\mathbf{X}; \delta_B = 1) = f(Y|\mathbf{X})$.

Assumptions 1 and 2 constitute the strong ignorability condition (Rosenbaum and Rubin, 1983). This setup has previously been used by several authors; see, e.g., Rivers (2007), Vavreck and Rivers (2008). Assumption 1 states the ignorability of the selection mechanism to Sample B conditional upon the covariates. Assumption 1 also implies that $P(\delta_B = 1|\mathbf{X}, Y) = P(\delta_B = 1|\mathbf{X})$. This assumption holds if the set of covariates contains all predictors for the outcome that affect the possibility of being selected in Sample B. Under this assumption, the missing outcomes in Sample A are missing at random (Rubin, 1976).

Assumption 2 (Common support). *The vector of covariates* $\mathbf{X} \in \mathbb{R}^p$ *has a compact and convex support, with its density bounded and bounded away from zero. There exist constants* C_l *and* C_u *such that* $C_l \leq f(\mathbf{X})/f(\mathbf{X}|\delta_B = 1) \leq C_u$ *almost surely.*

Assumption 2 implies that the support of $\mathbf{X}$ in Sample B is the same as that in the finite population. This assumption can also be formulated as a positivity assumption that $P(\delta_B = 1|\mathbf{X}) > 0$ for all $\mathbf{X}$. Assumption 2 does not hold if certain units would never be included in the big data sample. The plausibility of this assumption can be judged by subject matter knowledge. For diagnosis purpose, we can examine the distribution of the estimated propensity scores or the distribution of the propensity score weights in Sample A. Values of propensity score close to zero or extreme large values of the propensity score weights indicate the possible positivity violation. We assume all covariates are continuous. Categorical variables can be handled by first defining imputation classes using the partition of the categories and then estimating the average of the outcome using the nearest neighbor imputation within imputation classes. In our context, Sample B is a big data sample and therefore the size of donors for each imputation class can be reasonable large.

3. Methodology

3.1 Nearest neighbor imputation

For simplicity, we will focus on the Horvitz-Thompson type estimator, although our discussion applies to other type of estimators. If Y_i were observed throughout Sample A, the Horvitz–Thompson estimator

$\hat{\mu}_{g,\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i)$ can be used. We consider the imputation estimator of μ_g, given by $\hat{\mu}_{g,I} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i^*)$, where Y_i^* is an imputed value for Y_i. Creating imputed values for the whole data is called mass imputation (Chipperfield et al., 2012; Kim and Rao, 2012).

To find suitable imputed values, we consider nearest neighbor imputation; that is, find the closest matching unit from Sample B based on the $\mathbf{X}$ values and use the corresponding Y value from this unit as the imputed value. This approach has been called *Sample Matching* by Rivers (2007). To investigate the theoretical properties, we first consider matching with replacement with single imputation; the discussion on k nearest neighbor imputation is presented in Section 4.

The nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the nearest neighbor from Sample B with the minimum distance between $\mathbf{X}_j$ and $\mathbf{X}_i$. Let $i(1)$ be the index of its nearest neighbor, which satisfies $d(\mathbf{X}_{i(1)}, \mathbf{X}_i) \le d(\mathbf{X}_j, \mathbf{X}_i)$, for $j \in B$, where $d(\mathbf{X}_i, \mathbf{X}_j)$ is a distance function between $\mathbf{X}_i$ and $\mathbf{X}_j$. If there are ties, randomly select one as the nearest neighbor. Without loss of generality, we use the Euclidean distance, $d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|$, where $\|\mathbf{X}\| = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}$, to determine neighbors.

Step 2. The nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\,\mathrm{nni}} = \frac{1}{N}\sum_{i\in A}\pi_i^{-1}g(Y_{i(1)}). \tag{3.1}$$

Remark 1. *Our theoretical development applies to a general class of distances* $\|\mathbf{X}\|_\Sigma = (\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X})^{1/2}$, *where* Σ *is a positive definite matrix (Abadie and Imbens, 2006). This class includes the standard Mahalanobis distance by taking* Σ *to be the empirical covariance matrix of* $\mathbf{X}$. *Write* $\Sigma = L^{\mathrm{T}}L$. *Notice that* $\|\mathbf{X}\|_\Sigma = \{(L\mathbf{X})\,L\mathbf{X}\}^{1/2} = \|L\mathbf{X}\|$. *Hence, using* $\|\cdot\|_\Sigma$ *and* $\mathbf{X}$ *is equivalent to using* $\|\cdot\|$ *and* $L\mathbf{X}$. *So, we can carry over the the theoretical result to the case with* $\|\mathbf{X}\|_\Sigma$.

Comparing to model-based imputation, nearest neighbor imputation has several advantages. First, it does not require strong parametric model assumptions and therefore is robust to model misspecification. Second, nearest neighbor imputation is donor-based, where the imputed value is a value that was actually measured and will always be within the bounds of observed values. Third, in contrast to regression imputation approaches, nearest neighbor imputation can retain the complex variance covariance structure of the data. Moreover, for the same imputed dataset, one can estimate different parameters by choosing reasonable $g(\cdot)$. Recall that p is the dimension of $\mathbf{X}$. The asymptotic bias of $\hat{\mu}_{g,\mathrm{nni}}$ is of order $O_{\mathrm{p}}(N_B^{-1/p})$ (Abadie and Imbens, 2006), which is negligible when the number of continuous covariates is fixed at a reasonable number and the size of the matching donor pool is huge as in our big data setup. In the presence of a large dimension of $\mathbf{X}$, variable selection is necessary for the nearest neighbor imputation estimator to have good statistical properties. In this case, we suggest selecting important variables that are associated with the outcome in order to ensure Assumption 1 holds and also to increase estimation precision (Brookhart, Schneeweiss, Rothman, Glynn, Avorn and Stürmer, 2006).

3.2 Asymptotic results

To study the asymptotic properties of $\hat{\mu}_{g,\text{nni}}$, we impose the following regularity conditions.

Assumption 3. *(i)* $f(\mathbf{X})$ *and* $\mu_g(\mathbf{X}) = E\{g(Y)|\mathbf{X}\}$ *are continuously differentiable for any continuous and bounded* $g(Y)$, *and (ii)* $E\{g(Y)^\beta|\mathbf{X}\}$ *is bounded for* $\beta = 1, 2$.

Assumption 4. *(i) There exist positive constants* C_1 *and* C_2 *such that* $C_1 \le Nn^{-1}\pi_i \le C_2$, *for* $i = 1, \ldots, N$; *(ii) the sampling fraction for Sample A is negligible,* $nN^{-1} = o(1)$; *and (iii) the sequence of the Horvitz-Thompson estimators* $\hat{\mu}_{g,\text{HT}}$ *satisfies* $\text{var}_p(\hat{\mu}_{g,\text{HT}}) = O(n^{-1})$ *and* $\{\text{var}_p(\hat{\mu}_{g,\text{HT}})\}^{-1/2}(\hat{\mu}_{g,\text{HT}} - \mu_g)|\mathcal{F}_N \to \mathcal{N}(0,1)$ *in distribution, as* $n \to \infty$, *where* $\text{var}_p(\cdot) = \text{var}(\cdot|\mathcal{F}_N)$ *is the variance under the sampling design for Sample A.*

For clarification, the probability distribution underpinning the notation $E(\cdot)$, $\text{var}(\cdot)$, $o_\text{p}(\cdot)$ and $O_\text{p}(\cdot)$ is the joint distribution of the superpopulation model and the sampling processes for Samples A and B. Assumption 3 is a technical condition imposed on the functional continuity and finite moments, which holds for common models; see, e.g., Mack (1981). Assumption 4 holds for standard sampling designs in survey practice (Fuller, 2009; Chapter 1). It requires the sampling weights to behave well in the sense that there do not exist extremely large weights that dominate other weights. This occurs when subjects when certain characteristics are largely underrepresented in the sample. Sufficient conditions for Assumption 4 (iii) can be found in Chapter 3 of Fuller (2009).

We derive the asymptotic theory for $\hat{\mu}_{g,\text{nni}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 1. *Under Assumptions 1–3 and* $NN_B^{-1} = O(1)$, $\hat{\mu}_{g,\text{nni}}$ *has the same distribution as* $\hat{\mu}_{g,\text{HT}}$ *as* $N_B \to \infty$. *Furthermore, under Assumption 4,* $\hat{\mu}_{g,\text{nni}}$ *is consistent for* μ_g, *and*

$$n^{1/2}(\hat{\mu}_{g,\text{nni}} - \mu_g) \to \mathcal{N}(0, V_\text{nni}), \tag{3.2}$$

where

$$V_\text{nni} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\text{var}_p\left\{\sum_{i\in A} \pi_i^{-1} g(Y_i)\right\}\right].$$

Theorem 1 implies that the standard point estimator can be applied to the imputed data $\{(\mathbf{X}_i, Y_{i(1)}): i \in A\}$ as if the $Y_{i(1)}$'s were observed values. Let π_{ij} be the joint inclusion probability for units i and j. We show in the Supplementary Material that the direct variable estimator based on the imputed data

$$\hat{V}_\text{nni} = \frac{n}{N^2} \sum_{i\in A}\sum_{j\in A} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j}$$

is consistent for V_nni.

4. Other techniques for mass imputation

4.1 k-nearest neighbor imputation

Instead of using a single imputed value, we now consider fractional imputation with k imputed values for each missing outcome. Fractional imputation is designed to reduce the variance of the final estimator due to imputation (Kalton and Kish, 1984; Kim and Fuller, 2004).

Assume no matching ties, let $\mathcal{J}_k(i)$ be the set of k nearest neighbors for unit i

$$\mathcal{J}_k(i) = \left\{ l \in B : \sum_{j \in B} 1_{\{d(\mathbf{X}_j, \mathbf{X}_i) \le d(\mathbf{X}_l, \mathbf{X}_i)\}} \le k \right\} = \{i(1), \ldots, i(k)\}.$$

The k nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the k nearest neighbors from Sample B, $\mathcal{J}_k(i)$. Impute the Y value for unit i by $\hat{\mu}_g(\mathbf{X}_i) = k^{-1} \sum_{j=1}^{k} g(Y_{i(j)})$.

Step 2. The k nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_g(\mathbf{X}_i). \tag{4.1}$$

In the non-parametric estimation literature, researchers have investigated the asymptotic properties of the k nearest neighbor imputation estimators extensively. See, e.g., Mack and Rosenblatt (1979) and Mack (1981) for early references. Cheng (1994) establishes root-n consistency of the k nearest neighbor imputation estimator of the outcome mean when the outcome is subject to missingness. We derive the asymptotic theory for $\hat{\mu}_{g,\text{knn}}$ in the context of mass imputation combining a probability sample and a big data sample in the following theorem and defer its proof to the Supplementary Material.

Theorem 2. *Under Assumptions 1–4,* $n(k/N)^{4/p} \to 0$, $k/n \to 0$, *and* $k^2/n \to \infty$,

$$n^{1/2}(\hat{\mu}_{g,\text{knn}} - \mu_g) \to \mathcal{N}(0, V_{\text{knn}}), \tag{4.2}$$

where

$$V_{\text{knn}} = \lim_{n \to \infty} \frac{n}{N^2} \left(E\left[\text{var}_p \left\{ \sum_{i \in A} \pi_i^{-1} \mu_g(\mathbf{X}_i) \right\} \right] + E\left\{ \frac{1 - \pi_B(\mathbf{X})}{\pi_B(\mathbf{X})} \sigma_g^2(\mathbf{X}) \right\} \right),$$

and $\pi_B(\mathbf{X}) = P(\delta_B = 1 | \mathbf{X})$ *and* $\sigma_g^2(\mathbf{X}) = \text{var}\{g(Y) | \mathbf{X}\}$.

If $\pi_B(\mathbf{X})$ goes to 1, V_{knn} reduces to $\lim_{n \to \infty}(n/N^2) E\left[\text{var}_p \left\{\sum_{i \in A} \pi_i^{-1} \mu_g(\mathbf{X}_i)\right\}\right]$. It suggests that if the big sample is a large fraction of the target population, V_{knn} can be smaller than V_{nni}, suggesting that $\hat{\mu}_{g,\text{knn}}$ gains efficiency over $\hat{\mu}_{g,\text{nni}}$. In finite samples, Beretta and Santaniello (2016) conduct a simulation study to compare nearest neighbor imputation and k nearest neighbor imputation in the setting with independent and identically distributed data. They found that k nearest neighbor imputation with a small k outperforms nearest neighbor imputation in terms of mean squared error. On the one hand, a larger k can use more information in the big data sample and leads to more efficiency gain; on the other hand, k

cannot be too large, in order to control the bias of our estimator. In practice, we suggest using data-driven methods, such as cross-validation, to choose a reasonable k, and conducting sensitivity analysis varying the choice of k.

4.2 Generalized additive models

Nearest neighbor imputation methods are non-parametric. On the other hand, parametric models especially linear models are sensitive to model misspecification. We now consider semiparametric methods for mass imputation. Among semiparametric methods, generalized additive models (Hastie and Tibshirani, 1990) are flexible regarding model specification of the dependence of Y on $\mathbf{X}$ by specifying the model only through smooth functions rather than assuming a parametric relationship. As other non-parametric methods, the performance of generalized additive models will deteriorate as the dimension of $\mathbf{X}$ becomes large. For $\mathbf{X}$ with a moderate dimension, we apply generalized additive models to leverage the predictive power of the big data sample to produce a predictive model for Y given $\mathbf{X}$, so as to facilitate mass imputation for the probability sample.

We assume that $g(Y_i)$ given $\mathbf{X}_i$ follows some exponential family distribution, and

$$h^{-1}\{\mu_g(\mathbf{X}_i)\} = f_1(X_i^1) + f_2(X_i^2) + \ldots f_p(X_i^p), \tag{4.3}$$

where $h(\cdot)$ is an inverse link function, and each $f_k(\cdot)$ is a smooth function of X^k, for $k = 1, \ldots, p$. Model (4.3) allows for rather flexible specification of the dependence of Y on $\mathbf{X}$. The estimated function $f_k(X^k)$ can reveal possible nonlinearities of the relationship of Y and X^k.

There are several challenges in fitting model (4.3). First, $f_k(x)$ is an infinite-dimensional parameter, estimation of which often relies on some approximation. Second, we need to decide how smooth the $f_k(x)$ should be to balance the trade-off between model complexity and overfitting to the data at hand.

To solve the first issue, a common way to approximate $f_k(x)$ using splines. Let $B_m(x)$ be the basis spline functions for $m = 1, \ldots, M$ (Ruppert, Wand and Carroll, 2009). We approximate $f_k(x)$ by $f_k(x) = \sum_{m=1}^{M} \gamma_m^k B_m(x)$ with spline coefficients γ_m^k. This leads to an approximation of model (4.3):

$$h^{-1}\left[\hat{E}\{g(Y_i)\,|\,\mathbf{X}_i\}\right] = \sum_{k=1}^{p}\sum_{m=1}^{M} \gamma_m^k B_m(X_i^k). \tag{4.4}$$

In (4.4), a large M allows for increased model complexity and also an increased chance of overfitting; while a small M may result in an inadequate model. This trade-off is balanced by choosing a relatively large M and then penalizing the model complexity in the estimation stage (Eilers and Marx, 1996). Let the vector of spline coefficients be $\gamma_k^{\mathrm{T}} = (\gamma_1^k, \ldots, \gamma_m^k)$ and $\gamma^{\mathrm{T}} = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}})$. The estimate $\hat{\gamma}$ is obtained by maximizing the penalized likelihood:

$$-2l(\gamma) + \sum_{k=1}^{p} \lambda_k \gamma_k^{\mathrm{T}} S_k \gamma_k \tag{4.5}$$

where $l(\gamma)$ is the log likelihood function of γ, S_k is a matrix with the $(m, l)^{\text{th}}$ component $\int B_m''(x) B_l''(x)\,\mathrm{d}x$, $\gamma_k^{\mathrm{T}} S_k \gamma_k$ regularizes f_k to be smooth for which the degree of smoothness is controlled

by λ_k. Given the smoothing parameter $\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_p)$, the penalized likelihood function in (4.5) is optimized by a penalized version of the iteratively reweighted least squares algorithm (Nelder and Baker, 1972; McCullagh, 1984) to obtain $\hat{\gamma}$. Regarding the choice of λ, we note that λ controls the trade-off between model complexity and overfitting, which can be estimated separately from other model coefficients using generalized cross-validation or estimated simultaneously using restricted maximum likelihood estimation (Wood, 2006). In practice, the model performance is not sensitive to the choice of the number of basis functions as long as the number of basis functions is large relative to the sample size in the specification, but rather estimation of the smoothing parameter is critical to control the model complexity.

Once fitting the model, we can create an imputed value for each element i in Sample A as

$$\hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i) = h\left\{\hat{f}_1(X_i^1) + \hat{f}_2(X_i^2) + \ldots \hat{f}_p(X_i^p)\right\},$$

where $\hat{f}_k(x) = \sum_{m=1}^{M} \hat{\gamma}_m^k B_m(x)$ for $k = 1, \ldots, p$. The mass imputation estimator based on the generalized additive model is

$$\hat{\mu}_{g,\,\mathrm{GAM}} = \frac{1}{N}\sum_{i\in A} \pi_i^{-1} \hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i).$$

Because in our context, the sample size of Sample B is much larger than that of Sample A, the estimation error in the imputation model can be negligible compared to the sampling variability of $\hat{\mu}_{g,\mathrm{GAM}}$.

To close this subsection, it is worth commenting on the assumption of additive effects of $\mathbf{X}$ in model (4.3). This assumption may be fairly strong one. To relax the additivity assumption, we can extend model (4.3) to include interactions through using the tensor product basis. For example, we can include a bivariate interaction surface $f_{12}(X^1, X^2) = \sum_{m=1}^{M}\sum_{l=1}^{L} \gamma_{ml} B_m(X^1) B_l(X^2)$. When using the tensor product basis, care should be taken with respect to the penalty function in order to result in appropriate effective degrees of freedom for the smoother. This topic has been investigated extensively in the literature; see, e.g., Wood (2006).

5. Regression calibration

In practice, especially for government agencies, one nearest neighbor may be preferred because of its simplicity in implementation and data storage. We now consider another strategy to improve the efficiency for $\hat{\mu}_{g,\mathrm{nni}}$ when additionally the membership to Sample B can be determined throughout Sample A with the indicator δ_B. In some situation, we can obtain δ_B by matching the membership to Sample B (i.e., data linkage). We focus on the ideal setting without linkage errors. The key insight is that the subsample of units in Sample A with $\delta_B = 1$ constitutes a second-phase sample from Sample B, where Sample B acts as a new population. Standard regression calibration requires all calibration variables to be observed in Sample A and Sample B, and thus rules out the possibility of using Y as the calibration variable due to lack of the outcome data from Sample B. One of the advantages of mass imputations is that we can leverage the imputed outcomes to facilitate calibration of Y.

Let $\mathbf{h}(\delta_B, \mathbf{X}, Y)$ be a multi-dimensional function of δ_B, $\delta_B\mathbf{X}$ and $\delta_B Y$, e.g., $\mathbf{h}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1-\delta_B, \delta_B\mathbf{X}, \delta_B Y)^{\mathrm{T}}$. For simplicity of notation, we use $\mathbf{h}_i$ to denote $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_i)$ and $\mathbf{h}_i^*$ to denote $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_{i(1)})$. We can calculate the population quantity $\mathbf{H} = N^{-1}\sum_{i=1}^{N}\mathbf{h}_i$ from Sample B. This insight enables the typical calibration weighting in survey sampling with known marginal totals. In Sample A, we treat the imputed values as observed values, and the design weighted estimator of $\mathbf{H}$ is $\hat{\mathbf{H}}_A = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^*$. In general, $\hat{\mathbf{H}}_A$ is not equal to $\mathbf{H}$. We can use the known information $\mathbf{H}$ to improve the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$.

This suggests the following calibration strategy. We modify the original design weights $\{d_i: i \in A\}$ in $\hat{\mu}_{g,\mathrm{nni}}$ to a new set of weights $\{\omega_i: i \in A\}$ by minimizing a distance function

$$\sum_{i\in A} G(\omega_i, d_i) = \sum_{i\in A} d_i\left(\frac{\omega_i}{d_i} - 1\right)^2, \tag{5.1}$$

subject to the calibration constraints $N^{-1}\sum_{i\in A}\omega_i\mathbf{h}_i^* = \mathbf{H}$. By Lagrange multiplier, the solution to the constraint minimization problem is

$$\omega_i = d_i + \left(N\times\mathbf{H} - \sum_{k\in A} d_k\mathbf{h}_k^*\right)^{\mathrm{T}}\left(\sum_{k\in A} d_k\mathbf{h}_k^*\mathbf{h}_k^{*\mathrm{T}}\right)^{-1} d_i\mathbf{h}_i^*,$$

for $i \in A$. The resulting weights $\{\omega_i: i \in A\}$ can be called generalized regression weights.

The proposed estimator utilizing the new set of weights is

$$\hat{\mu}_{g,\mathrm{RC}} = \frac{1}{N}\sum_{i\in A}\omega_i g\left(Y_{i(1)}\right), \tag{5.2}$$

which is asymptotically equivalent to a generalized regression estimator (Park and Fuller, 2012). Following Yang and Ding (2020), one can show that $\hat{\mu}_{g,\mathrm{RC}}$ is the optimal estimator among the class of $\left\{\hat{\mu}_{g,\mathrm{nni}} + \left(N\times\mathbf{H} - \sum_{k\in A} d_k\mathbf{h}_k^*\right)^{\mathrm{T}}\boldsymbol{\gamma}: \boldsymbol{\gamma}\in\mathbb{R}^{\dim(\mathbf{h})}\right\}$.

We derive the asymptotic theory for $\hat{\mu}_{g,\mathrm{RC}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 3. *Under Assumptions 1-4,*

$$n^{1/2}\left(\hat{\mu}_{g,\mathrm{RC}} - \mu_g\right) \to \mathcal{N}(0, V_{\mathrm{RC}}), \tag{5.3}$$

in distribution, as $n \to \infty$, *where*

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\frac{n}{N^2} E\left(\mathrm{var}_p\left[\sum_{i\in A}\pi_i^{-1}\left\{g(Y_i) - \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\beta}_N\right\}\right]\right),$$

and $\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{N}\mathbf{h}_i g(Y_i)$.

The calibrated estimator $\hat{\mu}_{g,\mathrm{RC}}$ improves the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$ in the sense that V_{RC} is at most as large as V_{nni} given in Theorem 1. If $\mathbf{h}_i$ explains a proportion of the variability of $g(Y_i)$, V_{RC} is strictly less than V_{nni} and the efficiency gain does not require any parametric model assumption.

Remark 2 (Choice of distance functions). *Different distance functions in (5.1) can be considered. If we choose* $G(\omega_i, d_i) = -d_i\log(\omega_i/d_i)$, *it leads to empirical likelihood estimation (Newey and Smith,*

2004). If we choose the Kullback-Leibler distance function $G(\omega_i, d_i) = \omega_i \log(d_i/\omega_i)$, *it leads to exponential tilting estimation (Kitamura and Stutzer, 1997; Imbens, Johnson and Spady, 1998; Schennach, 2007; Dong et al., 2020). Under mild conditions, these procedures provide a set of weights that is asymptotically equivalent to the set of regression weights (Deville and Särndal, 1992; Breidt and Opsomer, 2017).*

For variance estimation, by Theorem (3), we construct a consistent variance estimator for $\hat{\mu}_{g,\mathrm{RC}}$ as $\hat{V}_{\mathrm{RC}}/n$, where

$$\hat{V}_{\mathrm{RC}} = \frac{n}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j},$$

with $\hat{e}_i = g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\hat{\boldsymbol{\beta}}$, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{h}_i^*\mathbf{h}_i^{*\mathrm{T}}\right)^{-1}\begin{pmatrix}\sum_{i=1}^{N}\delta_{Bi}g(Y_i)\\ \sum_{i\in A}\pi_i^{-1}(1-\delta_{Bi})g\left(Y_{i(1)}\right)\\ \sum_{i=1}^{N}\delta_{Bi}\mathbf{X}_i g(Y_i)\\ \sum_{i=1}^{N}\delta_{Bi}Y_i g(Y_i)\end{pmatrix}.$$

6. Empirical experiments

In this section, we evaluate the finite sample performance of the proposed estimator using simulation studies, one based on artificial data using simple random sampling and the other based on a synthetic population file from a single month sample of the U.S. Census Bureau's Monthly Retail Trade Survey using stratified sampling.

6.1 Kim-Wang example

We use the simulation example in Kim and Wang (2019) to compare various estimators. We generate the data according to the following mechanism. We first generate a finite population $\mathcal{F}_N = \{\mathbf{X}_i = (X_{1i}, X_{2i}), \mathbf{Y}_i = (Y_{1i}, Y_{2i}): i = 1, \ldots, N\}$ with size N = 1,000,000, where Y_{1i} is a continuous outcome and Y_{2i} is a binary outcome. From the finite population, we select a big data Sample B where the inclusion indicator $\delta_{Bi} \sim \mathrm{Ber}(p_i)$ with p_i the inclusion probability for unit i with the sample size around 700,000. We obtain a representative Sample A of size n = 1,000 using simple random sampling. The parameters of interest are the population mean $N^{-1}\sum_{i=1}^{N}\mathbf{Y}_i$ and the conditional population mean of Y_1 given $Y_2 = 1$.

For generating the finite population, we consider linear models

$$Y_{1i} = 1 + X_{1i} + X_{2i} + \alpha_i + \varepsilon_i, \tag{6.1}$$
$$P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) = \mathrm{logit}(1 + X_{1i} + X_{2i} + \alpha_i),$$

and nonlinear models

$$Y_{1i} = 0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i + \varepsilon_i, \tag{6.2}$$

$$P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) = \text{logit}\left\{0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i\right\},$$

where $X_{1i} \sim \mathcal{N}(1, 1)$, $X_{2i} \sim \text{Exp}(1)$, $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, and X_{1i}, X_{2i}, α_i and ε_i are mutually independent. The variables α_i induce the dependence of Y_{1i} and Y_{2i} even adjusting for X_{1i} and X_{2i}. For the big-data inclusion probability, we also consider a logistic linear model

$$\text{logit}(p_i) = X_{2i}, \tag{6.3}$$

and a nonlinear logistic model

$$\text{logit}(p_i) = -3 + (X_{1i} - 1.5)^2 + (X_{2i} - 2)^2. \tag{6.4}$$

We consider the following combinations: I. (6.1) and (6.3); II. (6.1) and (6.4); II. (6.2) and (6.3); and IV. (6.2) and (6.4) for data generating mechanisms. Therefore, the simulation setup is a 2×2 factorial design with two levels in each factor.

Chen, Li and Wu (2020) propose the inverse propensity score weighting estimator using the estimated probability of selection into Sample B and the doubly robust estimator which further incorporates an outcome regression model. To evaluate the robustness and efficiency, we compare the following estimators:

1. $\hat{\mu}_{\text{HT}}$, the Horvitz–Thompson estimator assuming the Y_i's were observed in Sample A for the purpose of benchmark comparison;
2. $\hat{\mu}_{\text{ipw}}$, the inverse propensity score weighting estimator,

   $$\hat{\mu}_{\text{ipw}} = \frac{1}{N}\sum_{i \in B} \frac{1}{p_i(\hat{\eta})} Y_i,$$

   where $p_i(\eta) = P(\delta_{Bi} = 1 \mid X_{2i}; \eta)$ is a logistic regression model with the linear predictor X_{2i} with an unknown parameter η, and $\hat{\eta}$ is an estimator of η obtained by maximizing the modified likelihood function of η (Chen et al., 2019) based on Samples A and B;
3. $\hat{\mu}_{\text{dr}}$, the doubly robust estimator of Chen et al. (2019),

   $$\hat{\mu}_{\text{dr}} = \frac{1}{N}\sum_{i \in B} \frac{1}{p_i(\hat{\eta})}\left(Y_i - \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\beta}}\right) + \frac{1}{n}\sum_{i \in A} \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\beta}},$$

   where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficients using (6.1) as the working outcome regression model based on Sample B;
4. $\hat{\mu}_{\text{nni}}$, the nearest neighbor imputation estimator;
5. $\hat{\mu}_{\text{knn}}$, the k nearest neighbor imputation estimator with $k = 5$;
6. $\hat{\mu}_{\text{GAM}}$, the generalized additive model imputation estimator;
7. $\hat{\mu}_{\text{RC}}$, the regression calibration estimator based on $\hat{\mu}_{\text{nni}}$ with calibration variables $\mathbf{H}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1 - \delta_B, \delta_B\mathbf{X}, \delta_B Y)^{\text{T}}$.

All simulation results are based on 1,000 Monte Carlo runs. Table 6.1 summarizes the simulation results with biases, standard errors, and coverage rates of 95% confidence intervals using asymptotic

normality of the point estimators. The following observations can be made from Table 6.1. $\hat{\mu}_{\text{ipw}}$ has large biases when the propensity score is misspecified. $\hat{\mu}_{\text{dr}}$ gains robustness over $\hat{\mu}_{\text{ipw}}$ if one of the outcome regression model or the propensity score is correctly specified. However, if both models are misspecified, $\hat{\mu}_{\text{dr}}$ has a larger bias. $\hat{\mu}_{\text{nni}}$ has small biases across four scenarios, which shows its robustness. Importantly, the performance of $\hat{\mu}_{\text{nni}}$ is close to that of $\hat{\mu}_{\text{HT}}$ in terms of standard errors and coverage rates, which is consistent with our theory in Theorem 1. Moreover, as predicted by our theoretical results, $\hat{\mu}_{\text{knn}}$ improves $\hat{\mu}_{\text{nni}}$ in terms of efficiency. Also, $\hat{\mu}_{\text{GAM}}$ shows robustness because of the flexibility of the model specification. The regression calibration estimator $\hat{\mu}_{\text{RC}}$ has small biases across all scenarios and therefore shows robustness against model specifications for sampling score and outcome. Moreover, it has smaller standard errors than both $\hat{\mu}_{\text{nni}}$ and $\hat{\mu}_{\text{knn}}$. The coverage rates are all close to the nominal level.

Table 6.1
Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on 1,000 Monte Carlo samples. OM: outcome model; PS: propensity score model (all numbers in the table are the numerical results multiplied by 100)

	Scenario I			Scenario II			Scenario III			Scenario IV		
OM	linear			linear			nonlinear			nonlinear		
PS	linear			nonlinear			linear			nonlinear		
	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.
						Population Mean of Y_1						
$\hat{\mu}_{\text{HT}}$	0.2	6.5	96.0	-0.2	6.4	94.5	0.61	15.2	95.7	-0.5	15.6	93.5
$\hat{\mu}_{\text{ipw}}$	-0.1	1.6	95.3	22.2	35.8	97.5	-0.1	4.2	95.3	432.7	284.5	75.6
$\hat{\mu}_{\text{dr}}$	0.0	4.6	94.5	0.0	4.3	96.5	0.5	14.2	95.2	229.8	168.8	35.8
$\hat{\mu}_{\text{nni}}$	0.2	6.5	95.1	-0.3	6.4	94.7	0.7	15.2	94.6	-0.6	15.6	93.7
$\hat{\mu}_{\text{knn}}$	0.2	4.9	96.1	-0.3	4.9	95.6	0.5	14.5	94.6	-0.6	14.9	93.8
$\hat{\mu}_{\text{GAM}}$	0.1	4.5	95.7	-0.2	4.5	96.0	0.5	14.3	94.9	-0.6	14.8	93.4
$\hat{\mu}_{\text{RC}}$	0.0	3.2	95.5	-0.2	4.1	95.3	-0.1	4.8	95.0	0.1	6.7	95.5
						Population Mean of Y_2						
$\hat{\mu}_{\text{HT}}$	-0.0	1.5	96.2	-0.0	1.6	95.1	-0.1	1.6	95.2	0.1	1.6	94.4
$\hat{\mu}_{\text{ipw}}$	0.0	0.2	95.0	-12.1	3.1	0.0	-0.0	0.3	95.4	3.0	1.8	94.7
$\hat{\mu}_{\text{dr}}$	-0.0	0.9	95.0	-1.1	1.8	68.6	0.0	0.4	94.9	-2.9	2.2	59.8
$\hat{\mu}_{\text{nni}}$	0.0	1.4	95.3	-0.0	1.6	95.3	-0.1	1.6	94.6	0.1	1.6	95.3
$\hat{\mu}_{\text{knn}}$	0.0	1.0	95.8	-0.0	1.1	95.8	-0.0	1.0	95.2	0.0	0.9	96.1
$\hat{\mu}_{\text{GAM}}$	-0.0	0.9	95.3	-0.0	0.9	94.8	-0.0	0.8	96.2	0.0	0.8	94.5
$\hat{\mu}_{\text{RC}}$	0.0	1.2	95.5	-0.1	1.4	94.2	-0.0	1.4	94.1	0.1	1.5	95.6
						Conditional Mean of Y_1 given $Y_2 = 1$						
$\hat{\mu}_{\text{HT}}$	0.0	7.3	95.1	-0.3	7.2	95.2	0.2	9.3	95.3	-0.1	9.8	94.1
$\hat{\mu}_{\text{ipw}}$	-0.1	1.6	95.2	-9.1	10.3	69.8	-0.1	4.3	95.0	534.2	329.8	65.3
$\hat{\mu}_{\text{dr}}$	0.1	4.7	95.6	2.5	4.6	93.2	9.8	18.0	93.1	452.0	465.4	65.6
$\hat{\mu}_{\text{nni}}$	-0.0	7.3	95.0	-0.3	7.3	95.3	0.1	9.2	95.4	-2.2	9.5	95.2
$\hat{\mu}_{\text{knn}}$	-0.1	4.7	96.8	-0.3	4.6	96.5	0.1	6.0	94.8	0.0	6.4	93.6
$\hat{\mu}_{\text{GAM}}$	0.0	4.8	94.2	-0.3	4.5	96.0	-0.1	6.5	95.5	-0.6	6.8	94.8
$\hat{\mu}_{\text{RC}}$	-0.0	3.9	94.8	-0.2	5.0	96.0	-0.2	5.4	95.1	-0.1	5.4	96.7

6.2 Monthly retail trade survey

To demonstrate the practical relevance, we consider the U.S. Census Bureau's 2014 Monthly Retail Trade Survey (Mulry, Oliver and Kaputa, 2014). The Monthly Retail Trade Survey is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. This survey selects a sample of about 12,000 retail businesses each month with paid employees to collect data on sales and inventories. It employs an one-stage stratified sample with stratification based on major industry, further substratified by the estimated annual sales referred to as the size variable.

For simulation purpose, we use the simulated data from the 2014 Monthly Retail Trade Survey to suggest the data generating model and the true parameter values (https://ww2.amstat.org/meetings/ices/2016/contests.cfm). We generate a finite population of $N = 812{,}765$ retail businesses with 16 strata with a stratum identifier h, sales Y, inventories $\mathbf{X}$, and a size variable Z on the log scale. Table 6.2 reports some summary statistics. We generate the inventory data from $X_{hi} \sim N\left(\mu_{X,h}, \sigma^2_{X,h}\right)$ for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 16$, and the sales data from a linear model

$$Y_{hi} = \beta_0 + X_{hi} + \varepsilon_{hi}, \tag{6.5}$$

and a nonlinear model

$$Y_{hi} = \beta_0 + 0.5X^2_{hi} + \varepsilon_{hi}, \tag{6.6}$$

where $\varepsilon_{hi} \sim \mathcal{N}(0, 0.25)$. In (6.5) and (6.6), we specify different values for β_0 so that the parameter of interest, $\mu = N^{-1}\sum_{h=1}^{16}\sum_{i=1}^{N_h} Y_{hi}$, matches with the true population mean 12.73.

Table 6.2
The stratum size, sample allocation, mean and standard error of the inventory data on the log scale extracted from the 2014 Monthly Retail Trade simulated dataset

Stratum h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N_h	366	20	2,015	4,646	7,402	700	12,837	17,080	29,808	2,400	41,343	57,518	83,465	95,244	115,028	342,893
n_h	37	5	34	57	74	7	103	115	116	12	184	196	218	200	220	336
$\mu_{X,h}$	16.8	16.7	16.6	16.4	16.1	15.6	16.0	15.7	15.6	15.5	15.4	15.1	14.8	14.5	13.9	11.5
$\sigma_{X,h}$	1.1	0.8	0.4	0.3	0.4	0.6	0.4	0.4	0.4	0.3	0.4	0.4	0.3	0.7	0.5	1.1
$\mu_{Z,h}$	5.9	2.3	5.8	6.3	6.6	4.2	6.9	7.0	7.4	4.8	7.5	7.6	7.7	7.6	7.7	8.1

We also generate a big data sample $\mathcal{S}_B$ where the inclusion indicator $\delta_{hi} \sim \text{Ber}(p_{hi})$ with the inclusion probability p_{hi} for unit i in stratum h. The big data sample in practice is often available from E-commercial companies who monitor inventories and sales for retail businesses. For the big data inclusion probability, let $Z_{hi} \sim N\left(\mu_{Z,h}, \sigma^2_{Z,h}\right)$, for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 16$. We consider a logistic linear model

$$\text{logit}(p_{hi}) = \alpha_0 + Z_{hi}, \tag{6.7}$$

and a nonlinear logistic model

$$\text{logit}(p_{hi}) = \alpha_0 + X_{hi} + Z_{hi}^2, \tag{6.8}$$

where we specify different values for α_0 so that the mean inclusion probability is about 30%. Lastly, we generate a representative sample $\mathcal{S}_A$ by stratified sampling with simple random sampling within strata without replacement; see Table 6.2 for the sample allocation.

We consider the seven estimators in Section 6.1 adopted for stratified sampling. In each mass imputed dataset, we apply the following point estimator and variance estimator: $\hat{\mu} = N^{-1}\sum_{h=1}^{H} N_h \bar{y}_{n_h}$ with $\bar{y}_{n_h}$ is the sample mean of y in the h^{th} stratum, $\hat{V}(\hat{\mu}) = N^{-2}\sum_{h=1}^{H} N_h^2 (1 - n_h/N_h) s_{n_h}^2 / n_h$ with $s_{n_h}^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{n_h})^2$.

Table 6.3 summarizes the simulation results. A similar discussion to Section 6.1 applies. $\hat{\mu}_{\text{ipw}}$ is sensitive to misspecification of the selection model; while $\hat{\mu}_{\text{dr}}$ has double robustness feature, which still relies on at least one model to be correctly specified. Mass imputation based on nearest neighbor imputation, k nearest neighbor imputation and generalized additive model shows good performances by leveraging the representativeness of the survey sample and the predictive power of the big data sample. In addition, if the big data membership is known throughout the survey data, the regression calibration estimator gains efficiency while maintaining the robustness against model misspecification.

Table 6.3
Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on 1,000 Monte Carlo runs for the 2014 Monthly Retail Trade Survey. OM: outcome model; PS: propensity score model (all numbers in the table are the numerical results multiplied by 100)

OM	Scenario I linear			Scenario II linear			Scenario III nonlinear			Scenario IV nonlinear		
PS	linear			nonlinear			linear			nonlinear		
	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.
$\hat{\mu}_{\text{HT}}$	0.0	3.0	95.0	0.0	3.0	95.0	1.1	31.5	95.0	1.1	31.5	95.0
$\hat{\mu}_{\text{ipw}}$	-0.6	5.8	96.6	-55.5	1.7	0.0	-7.3	76.2	96.6	-735.8	22.3	0.0
$\hat{\mu}_{\text{dr}}$	-0.3	2.7	94.4	-0.2	2.7	94.0	-3.3	34.6	93.8	-52.3	33.2	65.0
$\hat{\mu}_{\text{nni}}$	0.1	3.1	94.5	-0.1	3.1	94.6	1.1	31.5	95.3	-0.3	31.7	94.6
$\hat{\mu}_{\text{knn}}$	0.1	2.7	94.4	-0.2	2.7	94.3	1.0	31.4	94.9	-2.3	31.4	94.1
$\hat{\mu}_{\text{GAM}}$	0.1	2.7	94.9	0.1	2.7	94.9	1.1	31.6	94.9	-2.5	31.4	94.2
$\hat{\mu}_{\text{RC}}$	0.1	2.9	94.1	-0.1	2.6	95.1	0.6	30.7	94.6	-0.5	26.9	95.0

7. Real-data application

7.1 Data description

To demonstrate the practical use, we apply the proposed method to the survey data from the Korea National Health and Nutrition Examination Survey (KNHANES) and the big data from National Health Insurance Sharing Service (NHISS). The KNHANES is an annual national survey that studies the health and nutritional status of Koreans since 1998. The surveys have been conducted by the Korea Centers for

Disease Control and Prevention. This nationally representative cross-sectional survey includes approximately 10,000 individuals each year as a survey sample and collects information on socioeconomic status, health-related behaviours, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for non-communicable diseases and dietary intakes with three component surveys: health interview, health examination, and nutrition survey. More details of the KNHANES can be found in Kweon, Kim, jin Jang, Kim, Kim, Choi, Chun, Khang and Oh (2014). The data set used in this study has 4,929 samples.

On the other hand, the big data from NHISS provides health-related information collected from National Health Screening Program (NHSP) in South Korea. The NHSP was launched with the goal of improving the overall health of the South Korean citizens and preventing the costly chronic diseases. All beneficiaries are eligible for screening once every year or two depending on their demographic or occupational status. The specific screening items are stipulated by the implementation standards, which include, but not limited to, various blood tests and cancer screening. The total number of eligible beneficiaries is about 16 million, where approximately 75% of them participated the screening. The data that we have used in this study is the subset corresponding to the blood test results that are associated with metabolic syndrome from the 2014 program. The variables in this data set are demographics as sex and age, and clinical measurements such as total glycerides (mg/dL), total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL, mg/dL), and medical diagnosis on whether having anemia. The data set is made publicly available after anonymization and randomly selecting 1 million observations (National Health Insurance Data Sharing Service, 2014). Note that more thorough data can be purchased with a paid subscription and expert panel review.

7.2 Analysis and results

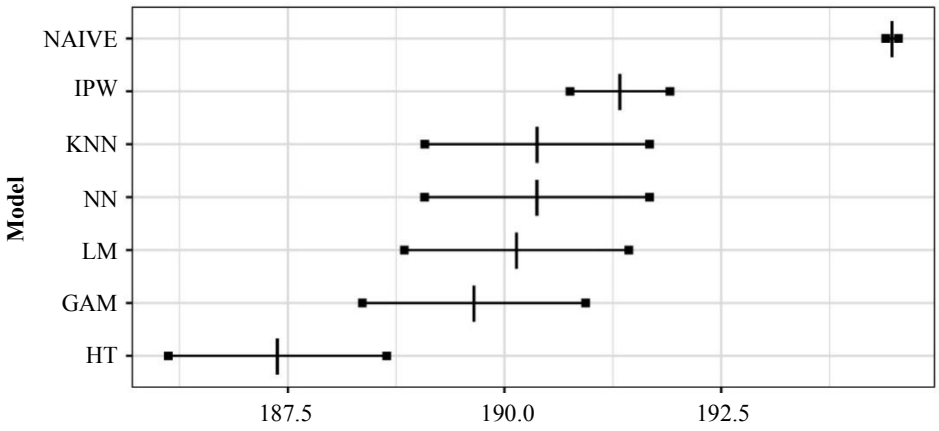
To apply the proposed method of mass imputation, we assume that total cholesterol is not available in KNHANES data, and use the big data from NHSP to perform mass imputation for total cholesterol variable. The actual survey values from KNHANES are used to compute a benchmark so that we can validate the efficacy of our proposed method. We consider the six different estimators:

- HT: the Horvitz-Thompson estimator based on the Sample A data. This is used for a benchmark comparison;
- NN: the nearest neighbor imputation estimator;
- kNN: the k nearest neighbor imputation estimator with $k = 5$;
- GAM: the generalized additive model imputation estimator;
- LM: the linear regression model imputation estimator using sex, age group, HDL cholesterol and total glycerides as the covariates;
- IPW: the inverse propensity score weighting estimator;
- NAIVE: the naive estimator using the Sample B without any treatment.

Total cholesterol is affected by the amount of HDL, because HDL is one of the components that constitute the total cholesterol, and is known to be also affected by sex and age. Unless Sample B is from

a particular sub-population such as cardiovascular stenosis patients group, we may assume that the relationship between the total cholesterol and other variables remain the same. Hence ignorability holds. Also the covariates are all medical/biological measurements, meaning they should stay within the similar range both for Samples A and B. The variance estimator for each estimator is calculated, and 95% Wald confidence interval for μ_g is obtained using asymptotic normality. Figure 7.1 depicts the intervals, where the population mean estimate from each method is presented as a vertical bar. The interval obtained from HT can be viewed as a reference. It can be seen that all estimators produce intervals that are slightly overestimated compared to the one from HT. It is because of the inherent bias in total cholesterol level in NHSP data; the sample mean values of the total cholesterol from NHSP data is 7 point or 3.7 per cent higher than HT estimator calculated with KNHANES data, as seen from the naive result. One can see that all the proposed methods substantially reduce such bias to make the estimator close to the HT estimator, which shows the benefit of the proposed methods. IPW estimator produced a relatively poor estimate compared to other methods, which is probably because of either misspecification in logistic regression model, or considerable discrepancy in sample sizes between KNHANES (4,929) and NHSP data (1 million). We also tried the DR estimator but not included here, because the effect from the IPW is very marginal due to the limited auxiliary variables available.

Figure 7.1 Estimated 95% confidence interval for the total cholesterol level.



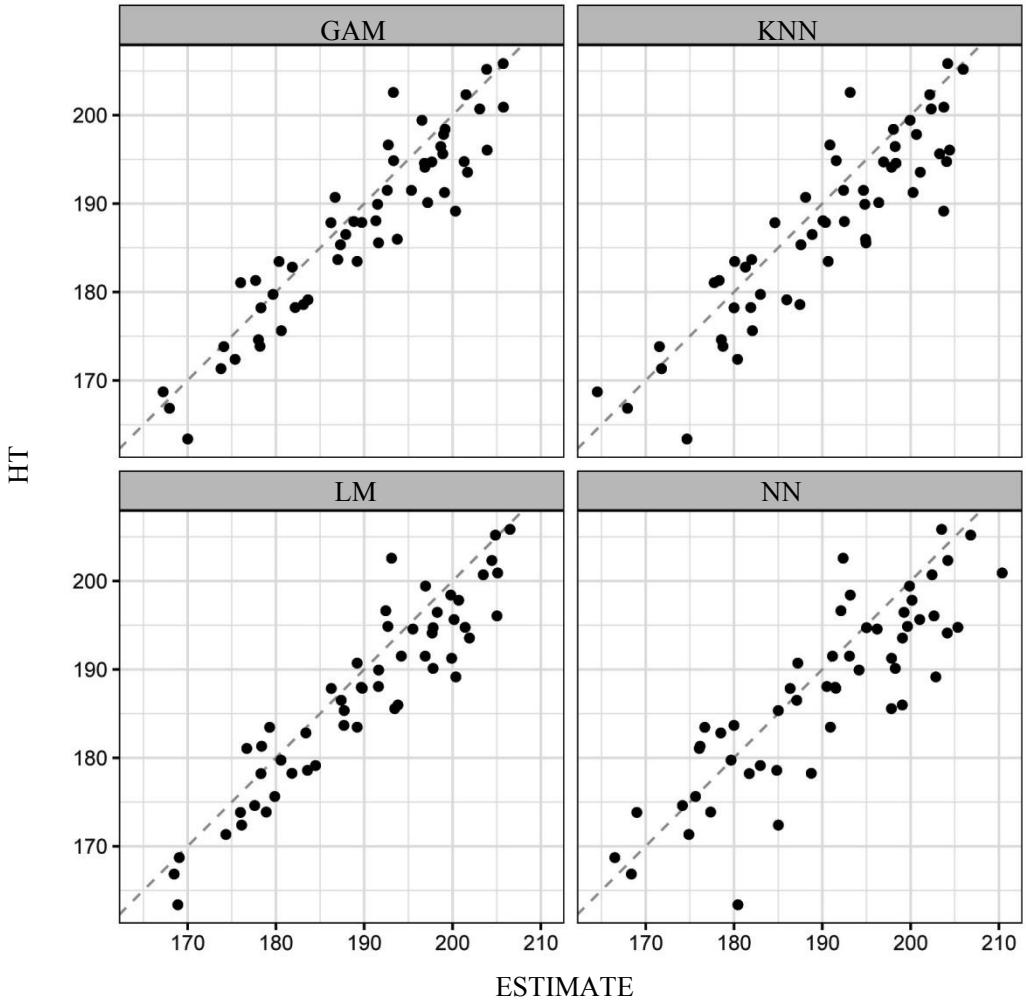
To better understand the prediction performance of the mass imputation methods, we calculated RMSE, mean bias, and correlation of imputed values by comparing the imputed values and actual survey values. Because we can observe the actual survey values from KNHANES, we can compute the prediction quality measures. Table 7.1 presents the summarized table, where we compared the results at individual levels and subgroup mean levels divided by age group and sex. For subgroup levels, we first obtain the subgroup mean estimates and then calculate the statistics aggregated over different groups. It can be seen that GAM performs better than the other methods in terms of RMSE and correlation. Overall, mass imputation method provides reasonable results for subgroup level as can be seen in Figure 7.2.

These quality measures need a predicted value for Sample A, hence IPW estimators are excluded in the comparison. Estimating the population and subgroup means using Sample B can give a very biased result – in the case of NHSP data, the difference between the mean of NHSP data and the HT estimator from KNHANES is about 7.09, or 3.7 per cent.

Table 7.1
Comparison of the imputation methods

	Method	RMSE	Bias	Corr.
Individual	NN	43.94	2.87	0.26
	KNN	32.62	2.86	0.42
	GAM	29.15	2.13	0.54
	LM	30.35	2.59	0.48
Group Means	NN	6.33	2.68	0.85
	KNN	5.44	2.70	0.90
	GAM	4.33	2.03	0.93
	LM	4.57	2.52	0.93

Figure 7.2 Comparison of the HT estimates and estimates using mass imputation for subgroup average.



8. Discussion

Mass imputation is an important technique for survey data integration. When the training dataset for imputation is obtained from a probability sample, the theory of Kim and Rao (2012) can be directly applied. If the training dataset is a non-probability sample and its size is huge, we have shown in this paper that various non-parametric methods can be used for mass imputation, and the estimation error in the imputation model can be safely ignored, under the assumption that the sampling mechanism for training data is missing at random in the sense of Rubin (1976). If the sampling mechanism is believed to be missing not at random, imputation techniques can be developed under the strong structural assumptions for the sampling mechanism (e.g., Riddles, Kim and Im, 2016; Morikawa and Kim, 2018) or the outcome model (e.g., Yang, Zeng and Wang, 2020). Also, when the training dataset has a hierarchical structure, multi-level models can be used to develop mass imputation. This is closely related to unit-level small area estimation in survey sampling (Rao and Molina, 2015).

The mass imputation estimator is not necessarily efficient. In Section 5, we have described a method of using calibration weighting as a tool for efficient data integration with big data. The calibration weighting requires correct matching between two data sources, as investigated by Kim and Tam (2020). Also, if the fraction of big data in the finite population is not substantial, the efficiency gain will be limited. Instead, one could improve the efficiency by combining the mass imputation estimator with the inverse propensity weighting estimator in the big data (Yang, Kim and Song, 2020). However, the correct specification of the propensity score model will be challenging. These are topics for future research.

Acknowledgements

We thank two anonymous referees and the associated editor for very constructive comments. Dr. Yang is partially supported by NSF grant DMS 1811245 and NIA grant 1R01AG066883. Dr. Kim is partially supported by NSF grant MMS 1733572 and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

Appendix

A.1 Proof for Theorem 1

For a given $\mathbf{X}_i = \mathbf{x}$ in Sample A, we show that $\mathbf{X}_{i(1)}$ converges to $\mathbf{x}$ in probability as $N_B \to \infty$. Consider for any $\varepsilon > 0$, we show that

$$P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = P\left\{d\left(\mathbf{X}_j, \mathbf{x}\right) > \varepsilon, \forall j \in B\right\} \tag{A.1}$$

converges to zero, and therefore $\mathbf{X}_{i(1)}$ converges to $\mathbf{x}$ in probability as $N_B \to \infty$, where the probability is induced by the sampling process of Sample B of size N_B. We show this fact by contradiction. Assume that for some $\varepsilon > 0$, $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\}$ does not coverage to zero as $N_B \to \infty$. Define the region $\mathcal{R}_\varepsilon = \{\mathbf{X}: d(\mathbf{X}, \mathbf{x}) \le \varepsilon\}$. Then, we must have $f(\mathbf{X} \mid \delta_B = 1) = 0$ for $\mathbf{X} \in \mathcal{R}_\varepsilon$; otherwise, there exists $\tilde{\mathbf{X}} \in \mathcal{R}_\varepsilon$ with a positive probability in Sample B as $N_B \to \infty$, and therefore $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = 0$ as $N_B \to \infty$. But the claim that $f(\mathbf{X} \mid \delta_B = 1) = 0$ for $\mathbf{X} \in \mathcal{R}_\varepsilon$ implies that $\mathcal{R}_\varepsilon$ is a non-overlap region of the distribution of $\mathbf{X}$ between Sample A (and also the population) and Sample B, violating Assumption 2.

Given $\mathbf{X}_i = \mathbf{x}$ in Sample A, for any continuous and bounded $g(y)$,

$$\begin{aligned} E\{g(Y_{i(1)}) | \mathbf{X}_i = \mathbf{x}, i \in A\} &= E\left[E\{g(Y_{i(1)}) | \mathbf{X}_{i(1)}, \mathbf{X}_i = \mathbf{x}, i \in A\} | \mathbf{X}_i = \mathbf{x}, i \in A\right] \\ &= E\left[E\{g(Y_{i(1)}) | \mathbf{X}_{i(1)}\} | \mathbf{X}_i = \mathbf{x}, i \in A\right] \\ &= E\{\mu_g(\mathbf{X}_{i(1)}) | \mathbf{X}_i = \mathbf{x}, i \in A\} \to E\{\mu_g(\mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}, i \in A\} \\ &= E\{g(Y_i) | \mathbf{X}_i = \mathbf{x}, i \in A\}, \end{aligned}$$

in probability as $N_B \to \infty$, where $\to$ follows from the fact that $\mu_g(\mathbf{x})$ is bounded and continuous. Then, by Portmanteau Lemma (Klenke, 2006), $Y_{i(1)} \to Y_i | (\mathbf{X}_i = \mathbf{x}, i \in A)$ in distribution as $N_B \to \infty$. By Assumption 1, $g(Y_{i(1)}) | (\mathbf{X}_i, i \in A) \to \mu_g(\mathbf{X}_i) + e_g^*(\mathbf{X}_i)$ in distribution as $N_B \to \infty$, where $e_g^*(\mathbf{X}_i)$ has the same distribution as $\{g(Y_i) | (\mathbf{X}_i, i \in A)\} - \mu_g(\mathbf{X}_i)$.

We now show that for $i \neq j \in A$, $e_g^*(\mathbf{X}_i)$ and $e_g^*(\mathbf{X}_j)$ are conditionally independent, given data $\mathcal{O}_A$. It is sufficient to show that $P\{i(1) = j(1)\} \to 0$ as $N_B \to \infty$; in other words, the same unit can not be matched for unit i and unit j with probability 1. This can be shown using (A.1) with $\varepsilon = \min_{i \neq j \in A} \| \mathbf{X}_i - \mathbf{X}_j \|$.

Therefore, conditional on data $\mathcal{O}_A$, we have

$$\hat{\mu}_{g,\text{nni}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_{i(1)}) \to \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_i) = \hat{\mu}_{g,\text{HT}}$$

in distribution as $N_B \to \infty$. This completes the proof for Theorem 1.

Let

$$\tilde{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_i)}{\pi_i} \frac{g(Y_j)}{\pi_j}. \tag{A.2}$$

Then, $\tilde{V}_{\text{nni}}$ is consistent for V_{nni}.

Similar to the above argument, for $i, j \in A$, conditional on data $\mathcal{O}_A$, $g(Y_{i(1)}) g(Y_{j(1)}) \to g(Y_i) g(Y_j)$ as $N_B \to \infty$. Therefore, conditional on data $\mathcal{O}_A$,

$$\hat{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j} \to \tilde{V}_{\text{nni}}, \tag{A.3}$$

in distribution as $N_B \to \infty$. Combining (A.2) and (A.3), $\hat{V}_{\text{nni}}$ is consistent for V_{nni}.

A.2 Proof for Theorem 2

To investigate the asymptotic properties of $\hat{\mu}_{g,\text{knn}}$, we re-express

$$\hat{\mu}_g(\mathbf{x}) = \frac{\sum_{j \in B} K_{R_\mathbf{x}}(\mathbf{x} - \mathbf{X}_j) g(Y_j)}{\sum_{j \in B} K_{R_\mathbf{x}}(\mathbf{x} - \mathbf{X}_j)},$$

where

$$K_h(u) = \frac{1}{h^p} K\left(\frac{u}{h}\right), \quad K(u) = 0.5 I(\| u \| \le 1),$$

and the bandwidth $h = R_{\mathbf{x}}$ is the random distance between $\mathbf{x}$ and its furthest among the k nearest neighbors. Therefore, $\hat{\mu}_{g,\text{knn}}$ can be viewed as a kernel estimator incorporating a data-driven bandwidth.

In the literature, asymptotic properties of the k nearest neighbor imputation estimator have been studied extensively. The result shown in the following lemma on k nearest neighbor imputation is extracted from Mack (1981).

Lemma 1. *Under Assumptions 1-3,*

$$N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j) g(Y_j) = f(\mathbf{x}) \pi_B(\mathbf{x}) \mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \tag{A.4}$$

We now express

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \mu_g(\mathbf{X}_i) + \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\}.$$

Let $T_N = N^{-1}\sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\}$. To study the properties for T_N, we first look at $\hat{\mu}_g(\mathbf{x})$, which can be expressed as

$$\hat{\mu}_g(\mathbf{x}) = \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})},$$

where $h_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j) g(Y_j)$ and $f_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j)$. By the result in Lemma 1, we obtain

$$h_N(\mathbf{x}) = f(\mathbf{x}) \pi_B(\mathbf{x}) \mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}$$

$$f_N(\mathbf{x}) = f(\mathbf{x}) \pi_B(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Now, by a Taylor expansion, we obtain

$$\begin{aligned}
\hat{\mu}_g(\mathbf{x}) - \mu_g(\mathbf{x}) &= \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})} - \mu_g(\mathbf{x}) \\
&= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})} \{h_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})\} \\
&\quad - \frac{f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})}{\{f(\mathbf{x})\pi_B(\mathbf{x})\}^2} \{f_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\} + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\} \\
&= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})} \{h_N(\mathbf{x}) - f_N(\mathbf{x})\mu_g(\mathbf{x})\} + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.
\end{aligned}$$

Therefore, we obtain

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Under the assumption in Theorem 2, it is easy to derive that $(k/N)^{2/p} + 1/k = o(n^{-1/2})$, and therefore,

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + o_{\mathrm{p}}(n^{-1/2}).$$

We then express T_N in a form of U-statistics (van der Vaart, 2000; Chapter 12):

$$T_N = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}h(\mathbf{Z}_i, \mathbf{Z}_j) + o_{\mathrm{p}}(n^{-1/2}),$$

where $\mathbf{Z}_i = (\mathbf{X}_i, Y_i, \delta_{A,i}, \delta_{B,i})$ and

$$\begin{aligned} h(\mathbf{Z}_i, \mathbf{Z}_j) &= \frac{1}{2}\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\right. \\ &\quad \left. + \frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\right] \\ &\equiv \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned}$$

Now, by Lemma 1, we obtain

$$\begin{aligned} E(\zeta_{ij}\,|\,\mathbf{Z}_i) &= E\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}E\left[\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}, \end{aligned}$$

and

$$\begin{aligned} E(\zeta_{ji}\,|\,\mathbf{Z}_i) &= E\left[\frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= \delta_{B,i}E\left(E\left[\frac{\delta_{A,j}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\,\middle|\,R_{\mathbf{X}_j}, \mathbf{Z}_i\right]\,\middle|\,\mathbf{Z}_i\right) \\ &= \frac{\delta_{B,i}}{\pi_B(\mathbf{X}_i)}\{g(Y_i) - \mu_g(\mathbf{X}_i)\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \end{aligned}$$

Therefore, by the theory of U-statistics, we obtain

$$\begin{aligned} T_N &= \frac{2}{N}\sum_{i=1}^{N} E\left\{h\left(\mathbf{Z}_i, \mathbf{Z}_j\right)\middle|\mathbf{Z}_i\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned}$$

Combining the above results leads to

$$\begin{aligned} \hat{\mu}_{g,\mathrm{knn}} - \mu_g &= \frac{1}{N}\sum_{i=1}^{N}\left\{\pi_i^{-1}\delta_{A,i}\mu_g\left(\mathbf{X}_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)} - 1\right\}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.5}$$

Then, the asymptotic results in Theorem 2 follow by Assumptions 1-4 and (A.5).

A.3 Proof for Theorem 3

The consistency and asymptotic normality of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ follow by the standard arguments under Assumptions 1-4. The remaining is to show that the asymptotic variance of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ is V_{nni}.

Using the distance function $G\left(\omega_i, d_i\right) = d_i\left(\omega_i/d_i - 1\right)^2$ in (5.1), the minimum distance estimation leads to generalized regression estimation (Park and Fuller, 2012). Therefore, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\omega_i g\left(Y_{i(1)}\right) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.6}$$

Similar to the argument in the proof for Theorem 1, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right\} + \frac{n^{1/2}}{N}\sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.7}$$

It is straightforward to show the variance of the second term in (A.7) is negligible given $nN^{-1} = o(1)$. Following the arguments in the proof for Theorems 1 and 2, $g\left(Y_{i(1)}\right)$ and $\mathbf{h}_i^*$ have the asymptotic distribution as $g\left(Y_i\right)$ and $\mathbf{h}_i$ given the data $\mathcal{O}_A$ from Sample A, respectively. Therefore, the asymptotic variance of $n^{1/2}\hat{\mu}_g$ is

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\mathrm{var}\left[\frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g\left(Y_i\right) - \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\beta}_N\right\}\right].$$

References

Abadie, A., and Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235-267.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-eng.pdf.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186, 345-366.

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16, 198-208.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, 34, 59-77.

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32, 190-205.

Breidt, F.J., McVey, A. and Fuller, W.A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, 49, 79-90.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149-1156.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., Adimora, A.A., Eron, J.J. and Mugavero, M.J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society, Series A, (Statistics in Society),* 181, 1193-1209.

Buelens, B., Burger, J. and van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *International Statistical Review*, 86, 322-343.

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

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of American Statistical Association*, 89, 81-87.

Chipperfield, J., Chessman, J. and Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, 54, 223-238.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, 2, 137-161. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf.

Cochran, W.G. (2007). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Couper, M.P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DiSogra, C., Cobb, C., Chan, E. and Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings (JSM), Survey Research Methods*, 4501-4515.

Dong, L., Yang, S., Wang, X., Zeng, D. and Cai, J. (2020). Integrative analysis of randomized clinical trials with real world evidence studies, *arXiv preprint arXiv:2003.01242*.

Eilers, P.H., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-102.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Fuller, W.A. (2009). *Sampling Statistics*, Wiley, Hoboken.

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*, NY: Chapman and Hall, Inc.

Imbens, G., Johnson, P. and Spady, R.H. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, 66, 333-357.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, 13, 1919-1939.

Kang, J.D., 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.

Keiding, N., and Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 179, 319-376.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., and Tam, S.M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*. Accepted (Available at https://doi.org/10.1111/insr.12434).

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87, 177-191.

Kitamura, Y., and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.

Klenke, A. (2006). *Probability Theory*, Springer-Verlag: Heidelberg.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9547-eng.pdf.

Kweon, S., Kim, Y., jin Jang, M., Kim, Y., Kim, K., Choi, S., Chun, C., Khang, Y.-H. and Oh, K. (2014). Data resource profile: The Korea National Health and Nutrition Examination Survey (KNHANES). *International Journal of Epidemiology*, 43, 69-77.

Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319-343.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Mack, Y.-P. (1981). Local properties of k-NN regression estimates. *SIAM Journal on Algebraic Discrete Methods*, 2, 311-323.

Mack, Y., and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9, 1-15.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16, 285-292.

McRoberts, R.E., Tomppo, E.O. and Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research*, 25, 368-381.

Morikawa, K., and Kim, J.K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Statistics & Probability Letters*, 140, 1-6.

Mulry, M.H., Oliver, B.E. and Kaputa, S.J. (2014). Detecting and treating verified influential values in a Monthly Retail Trade Survey. *Journal of Official Statistics*, 30, 721-747.

National Health Insurance Data Sharing Service (2014). National health screening data. https://nhiss.nhis.or.kr/bd/ab/bdabf006cv.do, [Accessed: 2019-07-11].

Nelder, J.A., and Baker, R.J. (1972). *Generalized Linear Models*, New York: John Wiley & Sons, Inc.

Newey, W.K., and Smith, R.J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219-255.

O'Muircheartaigh, C., and Hedges, L.V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 63, 195-210.

Palmer, J.R., Espenshade, T.J., Bartumeus, F., Chung, C.Y., Ozgencil, N.E. and Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50, 1105-1128.

Park, M., and Fuller, W.A. (2012). Generalized regression estimators. *Encyclopedia of Environmetrics*, 2, 1162-1166.

Pfeffermann, D., Eltinge, J.L. and Brown, L.D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3, 425-483.

Rao, J.N.K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*. pp. DOI 10.1007/s13571–020–00227–w.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, New York: John Wiley & sons, Inc.

Riddles, M.K., Kim, J.K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193-1256.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model-Assisted Survey Sampling*. Springer Science & Business Media, New York: Springer-Verlag.

Schennach, S.M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35, 634-672.

Stuart, E.A., Bradshaw, C.P. and Leaf, P.J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16, 475-485.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 174, 369-386.

Tam, S.-M., and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, 83, 436-448.

Tam, S.-M., and Kim, J.-K. (2018). Big data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, 34, 577-588.

Tourangeau, R., Conrad, F.G. and Couper, M.P. (2013). *The Science of Web Surveys*, New York: Oxford University Press.

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

van der Vaart, A.W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge, MA.

Vavreck, L., and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18, 355-366.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

Yang, S., and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115, 1540-1554.

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *Journal of the Royal Statistical Society, Series B, (Statistical Methodology)*, 82, 445-465.

Yang, S., Zeng, D. and Wang, X. (2020). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding, *arXiv preprint arXiv:2007.12922*.