Received: 13 May 2021 | Accepted: 21 February 2022

DOI: 10.1111/rssa.12841

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

Nearest neighbour ratio imputation with incomplete multinomial outcome in survey sampling

Chenyin Gao[1] | **Katherine Jenny Thompson**[2] | **Jae Kwang Kim**[3] | **Shu Yang**[1]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

[2]U.S. Census Bureau, Washington, District of Columbia, USA

[3]Department of Statistics, Iowa State University, Ames, Iowa, USA

Correspondence
Katherine Jenny Thompson, U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233, USA.
Email: Katherine.J.Thompson@census.gov

Abstract

Nonresponse is a common problem in survey sampling. Appropriate treatment can be challenging, especially when dealing with detailed breakdowns of totals. Often, the nearest neighbour imputation method is used to handle such incomplete multinomial data. In this article, we investigate the nearest neighbour ratio imputation (NNRI) estimator, in which auxiliary variables are used to identify the closest donor and the vector of proportions from the donor is applied to the total of the recipient to implement ratio imputation. To estimate the asymptotic variance, we first treat the NNRI as a special case of predictive matching imputation and build on earlier work to linearize the imputed estimate. To account for the non-negligible sampling fractions, parametric and generalized additive models are employed to incorporate the smoothness of the imputation estimator, which results in a valid variance estimator. We apply the proposed method to estimate expenditures detail items based on empirical data from the 2018 collection of the Service Annual Survey, conducted by the United States

Any views expressed here are those of the authors and not those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied (Approval ID: CBDRB-FY21-ESMD002-032).

Census Bureau. Our simulation results demonstrate the validity of our proposed estimators and also confirm that the derived variance estimators have good performance even when the sampling fraction is non-negligible.

KEYWORDS
generalized additive models, hot deck imputation, non-response

1 | INTRODUCTION

Sample surveys are often designed to estimate totals (e.g. revenue, earnings). However, in addition to collecting this information, many surveys request and produce sets of compositional variables (details) that sum to a total, such as a breakdown of total expenditures by type of expenditure or a breakdown of total income by source. Examples from the United States Census Bureau include the 2020 Current Population Survey's Annual Social and Economic (ASEC) Supplement which defines total household earnings as the sum of total wages and salaries, farm income, and self-employment income and the 2018 Services Annual Survey (SAS) which requests detailed breakdown of the sampled business' reported total expenditures by expenditures on annual payroll, fringe benefits for employees, and expenditures on software, among other categories.

This paper is concerned with the missing data treatment for these sets of detail items. In contrast to collected totals items, reliable auxiliary variable data for *all* detail items are rarely available for all requested compositional variables. Furthermore, there are often true zeros that differ across units. These limitations make it difficult to develop feasible parametric imputation models for each individual detail item and motivate the usage of hot deck imputation instead. Hot deck imputation—sometimes called 'donor imputation'—obtains replacement values for nonresponding (or missing) data items by matching a donor record containing valid data to a recipient record containing invalid or missing data, imputing the missing values from the donor record (Andridge & Little, 2010; Beaumont & Bocci, 2009). Kalton and Kaspryzk (1982) recommend using the same donor to impute sets of compositional details to preserve inter-item relationships in the multinomial data. To ensure additivity as well, it is sensible to use some form of hot deck imputation to impute the *vector* of proportions from a matched respondent (donor), then derive the imputed values for each imputed item by multiplying the (donated) proportion by the nonresponding recipient's total (Andridge & Thompson, 2015; Bankier et al., 2000; Little et al., 2008).

If the proportions associated with each detail item within a imputation class appear to be approximately the same for each unit, then the donor can be selected at random within imputation cells. However, it is possible that the multinomial distribution of the details could be related to unit size. In this case, the unit size should be incorporated into the hot deck matching procedure. The simplest version selects the nearest donor using a suitable distance function such as unit size or—as in this case—the total, assumed available for all donors and recipients. Since the total is univariate, the absolute value of the difference is a natural distance function, yielding estimates that are asymptotically unbiased (Yang & Kim, 2020). Hereafter, we refer to the imputation process that selects the nearest neighbour as donor and imputes the sets of donor proportions as the nearest neighbour ratio imputation (NNRI) method.

Consider the Service Annual Survey (SAS), the subject of the empirical application presented in Section 4. This program collects aggregate and detailed revenues and expenses from a stratified

sample of business firms with paid employees in selected industries in the services sector. Given a lack of auxiliary data, weak historic reporting, and a high reported zero rate, developing good predictive models for each individual detail item collected by the SAS is perhaps infeasible. That said, there are verifiable predictors of the *set* of detail items, that is, the multinomial distribution, specifically the industry in which the firm is classified, the tax-exempt status of the firm, and the size of the firm as measured by total revenue or total expenses. For example, a finance business will often report high proportions of total expenses in all three personnel cost categories, a scientific business is unlikely to report costs from temporary or leased employees, and a full-service restaurant would likely report most of its expenses from gross payroll and expensed equipment, materials, costs or supplies. Consequently, imputation classes in the SAS are defined by industry code and tax-exempt status. As regards unit size, larger businesses are more likely to expense costs and track depreciation than a smaller business in the same industry. Accordingly, one would expect zero or nearly zero proportions of costs and depreciation values from smaller businesses, and positively valued proportions for the same component items from the larger businesses. In turn, the proportion of total expenses represented by gross annual payroll tends to decrease as unit sizes increase.

In practice, it is often extremely difficult—if not impossible—to delineate exactly where the changes in multinomial distributions occur. However, using nearest neighbour donor selection procedure implicitly accounts for these subtle shifts.

The primary purpose of imputation is to 'fill in the blanks with plausible (i.e. realistic) and consistent values' (Sande, 1982). From a bias reduction perspective, there are advantages in preserving reported and previously imputed totals as is done with NNRI. Auxiliary data are often available for direct substitution of missing or invalid totals (Beaumont et al., 2011) or for modelling. Consequently, reported totals are generally validated (edited) and imputed when necessary early in the editing process, and missing values are imputed with very strong models. Not only are they made available for all units for subsequent hot deck imputation, such totals are usually 'goldplated' against further changes (Sigman & Wagner, 1997).

Fundamentally, totals are considered to be more reliably reported than sets of compositional details in the survey methodology literature, especially when the totals have accounting or financial record-keeping definitions (e.g. income, total sales, total expenses). In contrast, the queried sets of compositional details are generally collected for *statistical reporting purposes*, not accounting purposes, and are therefore not readily available (Willimack & Snijkers, 2013). Indeed, it is likely that the reported percentage distributions are more reliably reported than the values themselves. For example, when interviewing a non-probability sample of large businesses, Willimack and Nichols (2010) learned that 'company reporters resorted (sic) to estimation strategies rather than leaving items unreported (i.e. blank). Moreover, they only used estimation schemes when company data did not include the type of detail requested on the report'. This dovetails with the findings presented in Andridge and Thompson (2015) via a proxy pattern-mixture model analysis of selected items collected by the SAS: when annual payroll was used as the single predictor of total sales in a ratio imputation model, the fraction of missing information (FMI) values were close to zero, indicative of extremely accurate imputation models, whereas the FMI values for collected detail items using total sales as the sole predictor were near one (the maximum value). Finally, deriving hot deck imputed values by multiplying the recipient's total by the corresponding nearest-neighbour ratio is frequently used in business surveys (Beaumont & Bocci, 2009). In such positively skewed distributions, donating a ratio instead of a total guards against substitution of an overly large or small value from the respondent 'nearest neighbour'. Of course, this phenomena is not confined to business surveys. For example, the 2019 ASEC reports the U.S. median

income as \$68,703 ± (\$904); the 95th percentile is \$270,002 ± (\$4,831). See https://www.census.gov/content/dam/Census/library/publications/2020/demo/p60-270.pdf.

Estimation from nearest neighbour ratio imputed data is straightforward. Variance estimation is less so, in part because the donor selection procedure is deterministic. The variance estimation is further complicated in our setting due to the potential shift in multinomial distributions as described above. Andridge et al. (2021) investigates multiple imputations of proportions with nearest neighbour ratio hot deck imputation using the Approximate Bayesian Bootstrap, finding consistent underestimation of the variance. As in the cited reference, we assume that the *set* of details is reported or is missing; in practice, detail components that do not sum to their associated total and are not within a small raking tolerance are often treated as missing, and *all* detail items are imputed, regardless of their original reporting status.

Statistical inference under nearest neighbour imputation in survey sampling has been discussed by Chen and Shao (2001), Shao and Wang (2008), Kim et al. (2011), Yang and Kim (2019), among others. In this paper, we discuss statistical inference under NNRI in survey sampling. To make statistical inference, we derive the asymptotic linearization of the NNRI estimator, which allows us to decompose the asymptotic variance of the NNRI estimator into two components accounting for sampling and matching. Based on the asymptotic variance formula, we propose an alternative variance estimator by approximating these components by parametric or nonparametric approaches. We show the theoretical guarantees for the plug-in variance estimator. The proposed variance estimator is easily applied to a variety of probability sampling designs and accounts for non-negligible sampling fractions.

The rest of the paper is organized as follows. Section 2 introduces the basic setup and the NNRI procedure in detail, including asymptotic properties. Section 3 derives variance estimation for the NNRI estimator. In Section 4, we apply the proposed variance estimator to empirical expenditures data for reference year 2018 from a subset of industries surveyed in the SAS. The studied data are typical of many business surveys, in that many units report a total but may not provide the associated detail items, especially when the item definitions are complex or the number of requested detail items is large (Willimack et al., 2000). The SAS uses a stratified simple random sample without replacement (SRS-WOR) design with high sampling rates in several strata. The empirical application highlights the differences between our proposed variance estimator from a naive variance estimator but does not provide insight into its statistical properties. Consequently, Section 5 investigates the finite-sample performance of the proposed over repeated stratified SRS-WOR samples via a simulation study patterned off of Andridge et al. (2021). We close in Section 6 with some general observations along with directions for future research.

2 | BASIC SETUP

2.1 | Notation and assumptions

Let $y_i = (y_{i1}, \dots, y_{iT})$ be the study variable of interest and x_i be the auxiliary variable. We assume that x_i are observed throughout the sample but y_i are observed only for the subset of the sample. Let $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Let I_i be the sampling indicator where $I_i = 1$ if unit i is selected, and otherwise $I_i = 0$. Let $\mathcal{F}_N = \{(x_i, y_i, \delta_i) : i = 1, \dots, N\}$ be a finite random sample from a superpopulation model ζ with known N. We make the following assumption for the missing data process.

Assumption 1 (Missing at random and positivity). (i) The response indicator δ_i satisfies $P(\delta_i = 1|x_i, y_i) = P(\delta_i = 1|x_i)$, which can be denoted by $\pi(x_i)$, and (ii) $\pi(x_i) > \epsilon$ for a constant $\epsilon > 0$ w.p. 1.

Assumption 1 (i) states that the response indicator depends only on the observed x but not on the outcome value y. Essentially, it assumes that the covariates contain all the information for the outcome that affects the probability of response, that is, is missing at random in the population level (Berg et al., 2016; Rubin, 1976). Assumption 1 (ii) indicates that all sampled units have a positive probability of responding given any possible value of x, in turn implying that the support of the respondents and the nonrespondents is the same. This assumption guarantees that *all* donor values are plausible. If the Assumption 1 (ii) is violated, our proposed estimator in Section 2.2 would no longer be asymptotically unbiased. Throughout, we assume this strong ignorability condition in Assumption 1 holds.

To motivate the NNRI estimator, we first consider the full response case using the Horvitz-Thompson (HT) estimator. Under full response, we can use

$$\widehat{T}_y = \sum_{i \in S} w_i y_i,$$

to estimate $T_y = \sum_{i=1}^{N} y_i$, the population total of y_i, where w_i is the sampling (design) weight computed as the inverse of the sampling inclusion probability and S is the index set of the sample with $|S| = n$.

Let E_p and var_p be the expectation and variance with respect to the sampling mechanism; that is, $E_p(\cdot) = E(\cdot|\mathcal{F}_N)$ and $\text{var}_p(\cdot) = \text{var}(\cdot|\mathcal{F}_N)$. We assume a sequence of finite populations and samples in order to investigate the asymptotic properties as defined in Fuller (2009).

Assumption 2 The HT estimator of a population total given by $\hat{T}_y = \sum_{i \in S} w_i y_i$ satisfies (i) $C_1 \leq w_i n N^{-1} \leq C_2$; (ii) $\text{var}_p(N^{-1}\widehat{T}_y) = O_p(n^{-1})$ and $\{\text{var}_p(\widehat{T}_y)\}^{-1/2}(\widehat{T}_y - T_y)|\mathcal{F}_N \to \mathcal{N}(0, 1)$ in distribution, as $n \to \infty$.

Assumption 2 is widely accepted in survey sampling to allow for valid inferential conclusion via asymptotic normality.

2.2 | Nearest neighbour ratio imputation estimator

Nearest neighbour ratio imputation (NNRI) matches a donor to a recipient (nonrespondent), then multiplies the recipient's (available) total by the donated function $m(x_i)$ under the following assumption, assumed true for all $i \in S$

$$y_i = m(x_i) + e_i, \tag{1}$$

where $E_\zeta(e_i|x_i) = 0$ and $m(x_i) = x_i R(x_i)$ for some smooth function $R(\cdot)$. Let y_i^* be the imputed value of y_i using NNRI as

$$y_i^* = x_i R_{i(1)},$$

where the ratio $R_i = (y_{i1}, \ldots, y_{iT})/x_i$ is only available from the responding units (donors), and $i(1)$ is the index of the nearest neighbour of unit i within the same imputation cell where the nearest neighbour of i satisfies

$$\mathcal{D}(x_i, x_{i(1)}) \leq \mathcal{D}(x_i, x_j),$$

for all j in the subsample of respondents, where $\mathcal{D}(\cdot,\cdot)$ is a suitable distance function (in this application, the absolute value of the distance).

Then, the imputed estimator of T_y is given by

$$\widehat{T}_{y,I} = \sum_{i\in S} w_i \left\{\delta_i y_i + (1-\delta_i) y_i^*\right\}. \tag{2}$$

The goal is to estimate the variance of the imputed estimator in Equation (2). If we define

$$d_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is used as a donor for unit } j, \\ 0 & \text{otherwise,} \end{cases}$$

then we can express

$$y_i^* = x_i R_{i(1)} = x_i \sum_{j\in S} \delta_j d_{ji} R_j = \sum_{j\in S} \delta_j d_{ji} \left(x_i/x_j\right) y_j.$$

Thus, the imputed estimator in Equation (2) can be written as

$$\begin{aligned} \widehat{T}_{y,I} &= \sum_{i\in S} w_i \left\{ \delta_i y_i + (1-\delta_i) \sum_{j\in S} \delta_j d_{ji} \left(x_i/x_j\right) y_j \right\} \\ &= \sum_{i\in S} \delta_i w_i (1+\kappa_i) y_i, \end{aligned} \tag{3}$$

where

$$\kappa_i = \sum_{j\in S} \frac{w_j x_j}{w_i x_i} (1-\delta_j) d_{ij}.$$

Note that κ_i satisfies

$$\sum_{i\in S} \delta_i w_i (1+\kappa_i) x_i = \sum_{i\in S} w_i x_i. \tag{4}$$

To study the asymptotic properties of the NNRI estimator in Equation (3), we assume the following regularity condition holds.

Assumption 3 (i) Let the matching variable X be a random variable on a compact and convex support $\mathbb{X}$, with its density f_X bounded and bounded away from zero. Suppose that f_X is also differentiable in the interior of $\mathbb{X}$ with bounded derivatives; (ii) Let $R(x)$ be Lipschitz continuous in x, which means that $\exists C_3$ s.t. $|R(x_i) - R(x_j)| \leq C_3 |x_i - x_j|$, for any i, j.

Assumption 3 (i) imposes a compact and convex support for the random variable X, which will be essential for studying the asymptotic properties of the NNRI estimator; Assumption 3 (ii) restricts the ratio function R_i to be smooth in x_i (Abadie & Imbens, 2006; Yang & Kim, 2020). Note that the underlying ratio model in Equation (1) is a special case of the general models $m(\cdot)$ that clearly satisfies the Lipschitz condition since the Lipshitz condition on $R(x)$ implies that on $m(x)$.

The following lemma describes the key asymptotic property with the proof deferred to the Supplementary Material.

Lemma 1 *Under Assumptions 1–3, we have*

$$\sum_{i\in S}\delta_i w_i(1+\kappa_i)x_i R(x_i) = \sum_{i\in S} w_i x_i R(x_i) + O_p(n^{-1}N). \tag{5}$$

Combining with (5), *we have*

$$n^{1/2}N^{-1}\widehat{T}_{y,I} = n^{1/2}N^{-1}\sum_{i\in S} w_i\left[x_i R(x_i) + \delta_i(1+\kappa_i)\{y_i - x_i R(x_i)\}\right] + o_p(1). \tag{6}$$

Considering that $E(y_i|x_i) = x_i R(x_i)$ in Equation (1), we can express (6) as

$$n^{1/2}N^{-1}\widehat{T}_{y,I} = n^{1/2}N^{-1}\sum_{i\in S} w_i\left\{m_i + \delta_i(1+\kappa_i)e_i\right\} + o_p(1), \tag{7}$$

where $m_i = x_i R(x_i)$ and $e_i = y_i - m_i$. This decomposition assures that the first component m_i is uncorrelated with the second component e_i under the conditional argument of x_i. The decomposition leads to the asymptotic distribution of the NNRI estimator as follows.

Theorem 1 *Under Assumptions 1–3, suppose the ratio model in Equation (1) holds true and define* $\sigma_e^2(x_i) = \mathrm{var}(e_i \mid x_i) = E(e_i^2 \mid x_i)$. *Then,* $n^{1/2}N^{-1}(\widehat{T}_{y,I} - T_y) \to \mathcal{N}(0, V_y)$ *in distribution as* $n \to \infty$, *where*

$$V_y = V^m + V^e,$$

with

$$V^m = \lim_{n\to\infty}\frac{n}{N^2}E\left\{\mathrm{var}_p\left(\sum_{i\in S} w_i m_i - \sum_{i=1}^{N} m_i\right)\right\},$$

and

$$V^e = \lim_{n\to\infty}\frac{n}{N^2}E\left[\sum_{i=1}^{N}\{I_i w_i\delta_i(1+\kappa_i) - 1\}^2\sigma_e^2(x_i)\right]. \tag{8}$$

In particular, if $n/N = o(1)$, *then* V^e *reduces to*

$$V^e = \lim_{n\to\infty}\frac{n}{N^2}\sum_{i=1}^{N} E\left[I_i\delta_i\{w_i(1+\kappa_i)\}^2\sigma_e^2(x_i)\right]. \tag{9}$$

The detailed proof of Theorem 1 is presented in the Supplementary Material. The results in Theorem 1 are obtained by taking the reverse sampling arguments following Shao and Steel (1999) and Kim et al. (2006), so that the sample-response path begin with a census with nonrespondents from which a sample is selected. In the reverse sampling framework, the outer expectation (denoted E) is taken with respect to the superpopulation model for the census with nonrespondents, with inner expectations and variances with respect to the sampling

design conditional on $(\delta_1, \ldots, \delta_N)$. The component V^m is the variance due to the sampling design. The other component, V^e, constitutes the error variance added due to the deterministic ratio imputation model via the NNRI. Variance formula (8) requires access to population x_i's whereas variance formula (9) does not provided that the overall sampling fraction is negligible.

3 | VARIANCE ESTIMATION

Yang and Kim (2019, 2020) describe asymptotically unbiased variance estimators that fully account for the error term (V^m and V^e) presented in Section 2.2, under predictive mean matching imputation. Since NNRI is a special case of predictive mean matching, we modify their variance estimator to obtain asymptotically unbiased estimators i.e., $E(\widehat{V}^m + \widehat{V}^e) = \text{var}\{n^{1/2}N^{-1}(\widehat{T}_{y,I} - T_y)\}$.

Both the empirical application presented in Section 4 and the simulation study presented in Section 5 employ stratified SRS-WOR designs that include a certainty stratum (all units sampled with probability = 1) and at least one sampling strata with a large sampling rate (greater than 0.50). The sample design is highly characteristic of business surveys, which are sampled from highly skewed populations. The presented applications use approximate sampling variance (design based) estimators for $\widehat{V}_m$ to easily incorporate the non-negligible sampling fractions, although we also introduce a general replication variance estimator for use when sampling fractions are negligible.

3.1 | Estimation of m_i

Assumption 3 requires a smooth estimator of $R(x)$ that can be used for *all* sampled units to estimate m_i as $x_i R(x_i)$. However, NNRI is not a smooth imputation procedure. We approximate a smooth ratio function $R(x)$ in Equation (6) with a plug-in estimator $(\widehat{R}(x_i))$, considering

(a) PARAM1 Parametric ratio estimator with $\widehat{R}(x_i) = \widehat{\beta} = \sum_{i\in S} w_i\delta_i y_i / \sum_{i\in S} w_i\delta_i x_i$, the B.L.U.E. of the weighted simple linear no-intercept regression model $y_i w_i^{-1/2} = \beta x_i w_i^{-1/2} + e_i w_i^{-1/2}$, $e_i \sim (0, x_i\sigma^2)$. This is the PAR1 estimator utilized in Beaumont and Bocci (2009);
(b) PARAM2 Parametric ratio estimator with $\widehat{R}(x_i) = \widehat{\beta}_h$, where $\widehat{\beta}_h$ is estimated separately within each sampling stratum h by fitting the regression model from (a); and
(c) NONPARAM Generalized Additive Model (GAM) estimator (Hastie & Tibshirani, 1990) approximating the unknown smooth function of $R(x_i)$ with multinomial link functions.

The first two estimators are frequently employed in the survey research methods literature (e.g. Magee, 1998, among others). Of course, these estimators require a very specific relationship between the independent and auxiliary variable, and this strong association is less likely for rarely-reported independent variables. Furthermore, the parametric approximations develop *separate* regression models for each detail item component in y_i. As a nonparametric alternative approach, we appeal to the GAM for a more flexible representation. In short, GAMs can be considered as fitting several spline smoothers to approximate the *unknown* smooth function, in this case the ratio function $R(x)$. By applying the order-n basis expansions of $R(x)$ under the multinomial link function, one could observe that for $i = 1, \ldots, n$

$$\eta^{(t)}(x_i) = \sum_{k=1}^{n} \beta_k^{(t)} b_k^{(t)}(x_i), t = 1, \ldots, T-1; \eta^{(T)}(x_i) = 1, \tag{10}$$

$$\widehat{R}(x_i) = \left\{ \frac{\exp\{\eta^{(1)}(x_i)\}}{\sum_{t=1}^{T} \exp\{\eta^{(t)}(x_i)\}}, \frac{\exp\{\eta^{(2)}(x_i)\}}{\sum_{t=1}^{T} \exp\{\eta^{(t)}(x_i)\}}, \ldots, \frac{\exp\{\eta^{(T)}(x_i)\}}{\sum_{t=1}^{T} \exp\{\eta^{(t)}(x_i)\}} \right\}. \tag{11}$$

The $\{\beta_k^{(t)}\}_{k=1}^{n}, t = 1, 2, \ldots, T-1$ are the regression coefficients for the $t-$th detail item and $\{b_k^{(t)}(x_i)\}_{k=1}^{n}, t = 1, 2, \ldots, T-1$ are the known n basis functions (e.g. splines, radial functions, etc.) associated with the n sample points, which are usually assumed to have good approximation theoretical properties. To alleviate the overfitting issue, a model penalty can be imposed during the model fitting to reduce the number of basis functions from n to K (Wood, 2003; Wood et al., 2016). The penalized objection function is specified as

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \delta_i \| x_i \widehat{R}(x_i) - y_i \|^2 + \lambda J\{\widehat{R}(x)\} \right\},$$

where the first term measures the closeness of our fitted functions while the second term $J\{\widehat{R}(x)\}$ penalizes the wiggliness of the function associated with the tuning parameter λ, which can be obtained by the cross validation technique. Here, we adopt the wiggliness penalty functional from Wood (2003) illustrated in Example 1.

Example 1 Let $\widehat{R}(x) \in \mathcal{H}$, where $\mathcal{H}$ is an arbitrary reproducing kernel Hilbert space. Begin with a simple objective function:

$$\min_{\beta} \sum_{i=1}^{n} \delta_i \| x_i \widehat{R}(x) - y \|^2 + \lambda \int_a^b \widehat{R}''(x)^{\intercal} \widehat{R}''(x) dx, \tag{12}$$

with arbitrary value a and b as long as they cover the variable in question (Hastie & Tibshirani, 1990). The resulting solution $\widehat{R}^*(x)$ can be considered as a type of the *basis spline* or B-spline (See Figure 1 for an illustrative example). However, its estimation requires $O(n^3)$ operations in the univariate case, which is computational infeasible for the large-scale of datasets. One approach to tackle this problem is to employ the *regression splines*, from which an optimal approximation of $\widehat{R}(x)$ can be produced via choosing the truncated bases with lower ranks $K^{(t)}$ for $t = 1, \ldots, T-1$ (Wood, 2003).

3.2 | Estimation of V^m

Using the pseudo observations $\widehat{m}_i$ with postulated parametric or nonparametric models of $R(x)$ in Section 3.1, a standard design-based estimator under complete response is given by

$$\widehat{V}^m = \sum_{i \in S} \sum_{j \in S} \Omega_{ij} \widehat{m}_i \widehat{m}_j, \tag{13}$$

where Ω_{ij} accounts for various sampling designs. For example, under simple random sampling, the variance expression of V_m simplifies to $(1 - n/N)(n-1)^{-1} \sum_{i \in S} (\widehat{m}_i - \overline{m})^2$ with

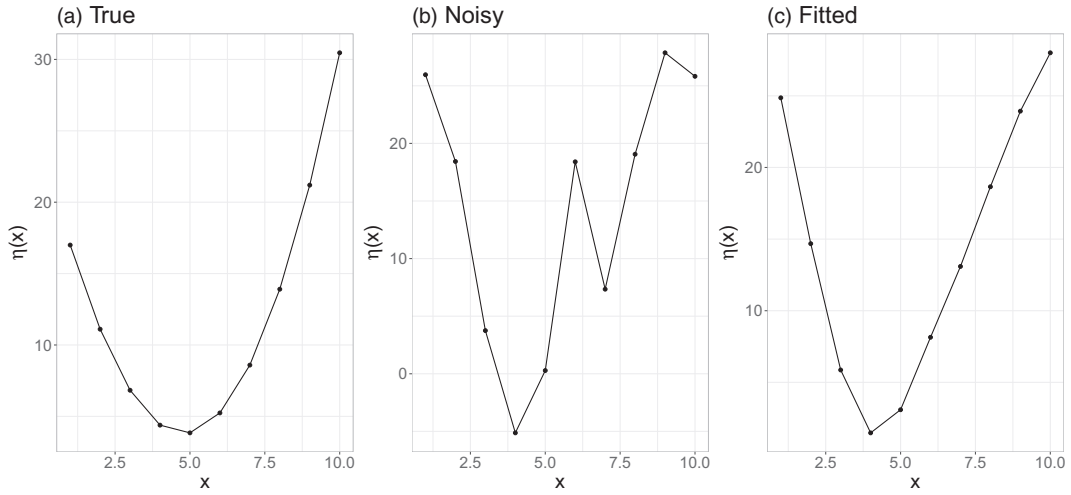

FIGURE 1 An illustrative example of the nonparametric B-spline, where x is a scalar variable ranging from 1 to 10, with 10 observed sample points as 1, 2, … , 10. Leftmost (a) is the underlying true values of $\eta(x) = \sqrt{|x|} + |x-5|^2 + \log(|x|)$; middle (b) is the observed values of $\eta(x)$ corrupted by the additive error term $e \sim \mathcal{N}(0, 10^2)$; rightmost (c) is the fitted values of $\eta(x)$ by implementing the B-spline approach based on (12)

$\overline{m} = 1/n \sum_{i \in S} \hat{m}_i$. With a stratified SRS-WOR design as in Sections 4 and 5, the same formula is used to obtain $\hat{V}^{mh}$ independently in each stratum, then aggregated ($\hat{V}^m = \sum_h \hat{V}^{mh}$).

For sampling designs with negligible sampling fractions, the ease of a replication variance estimator (Wolter, 2007) may be preferable for the NNRI estimator. The general replicate variance estimator when y_i is observed throughout the sample is

$$\hat{V}_{\text{rep}}(\hat{T}_y) = \frac{n}{N^2} \sum_{k=1}^{L} c_k \left(\hat{T}_y^{(k)} - \hat{T}_y \right)^2, \tag{14}$$

where c_k is the kth replication factor, and $\hat{T}_y^{(k)} = \sum_{i \in S} w_i^{(k)} y_i$ in which $w_i^{(k)}$ is the k-th replicate weight for unit i, using a replication method that appropriately accounts for the complex sampling design [Note: for a stratified SRS-WOR sample with non-negligible sampling fractions, the c_k should be modified to include the finite population correction factors $(1 - n_h/N_h)$]. The replicates are constructed such that $E\{\hat{V}_{\text{rep}}(\hat{T}_y)\} = \text{var}\{n^{1/2} N^{-1} (\hat{T}_y - T_y)\}\{1 + o(1)\}$. We illustrate replicate weights through the following example.

Example 2 Suppose that the probability sample is obtained by a single-stage design where each unit i has a sampling weight w_i, the delete-1-jackknife method yields an unbiased estimate of the sampling variance under complete response. Therefore, $L = n, c_k = (n-1)/n$, and $w_i^{(k)} = n w_i/(n-1)$ if $i \neq k$, and $w_k^{(k)} = 0$ if $i = k$.

3.3 | Estimation of V^e

We now discuss estimation of the V^e term. If an asymptotically unbiased estimator of $\sigma_e^2(x_i)$ is available, then we may use

$$\widehat{V}^e = \frac{n}{N^2}\sum_{i\in S}\{w_i^2\delta_i(1+\kappa_i)^2 + w_i - 2w_i\delta_i(1+\kappa_i)\}\widehat{\sigma}_e^2(x_i). \tag{15}$$

If additionally, we assume the Lipschitz continuity of $\sigma_e^2(x_i)$ in x, a similar result to Lemma 1 can be obtained

$$\sum_{i\in S}\delta_i w_i(1+\kappa_i)\sigma_e^2(x_i) = \sum_{i\in S} w_i\sigma_e^2(x_i) + O_p(n^{-1}N). \tag{16}$$

Substituting Equation (16) back into Equation (8) yields

$$V^e = \frac{n}{N^2}\sum_{i\in S}\{w_i^2\delta_i(1+\kappa_i)^2 - w_i\delta_i(1+\kappa_i)\}\sigma_e^2(x_i). \tag{17}$$

Now, we can directly use the residuals $\widehat{e}_i$ obtained from the modeled values i.e., $\widehat{e}_i = y_i - \widehat{m}_i$ to estimate $\sigma_e^2(x_i)$, where $\widehat{m}_i$ is obtained using the PARAM1, PARAM2, or NONPARAM models presented in Section 3.1. In particular, we have

$$\widehat{V}^e = \frac{n}{N^2}\sum_{i\in S}\{w_i^2\delta_i(1+\kappa_i)^2 - w_i\delta_i(1+\kappa_i)\}\widehat{e}_i^2.$$

Alternatively, we can model the variation of the residuals to estimate $\sigma_e^2(x_i)$ in Equation (15). We consider three approaches:

(a) PARAM1(M) Plug-in variance estimator as $\widehat{\sigma}_e^2(x_i) = x_i\widehat{\beta}(1-\widehat{\beta})$, specified by the multinomial distribution of y_i, where $\widehat{\beta}$ is the PARAM1 estimator from Section 3.1.
(b) PARAM2(M) Parametric linear estimator of $\widehat{\sigma}_e^2(x_i) = \widehat{\alpha}_{0,h} + \widehat{\alpha}_{1,h}x_i$ obtained by the OLS regression of $e_i^2 = (y_i - \widehat{\beta}_h x_i)^2 = \alpha_{0,h} + \alpha_{1,h}x_i$ for $\delta_i = 1$ within each strata, where $\widehat{\beta}_h$ is PARAM2 estimator from Section 3.1.
(c) NONPARAM(M) Nonparametric estimator of $\widehat{\sigma}_e^2(x_i) = \sum_{k=1}^K \widehat{\beta}_k b_k(x_i)$ obtained by fitting a GAM of $e_i^2 = \delta_i(y_i - \widehat{m}_i)^2 = \sum_{k=1}^K \beta_k b_k(x_i)$ for all strata, where $\widehat{m}_i$ is obtained using the NONPARAM estimator described in Section 3.1.

Combining the V^m components obtained with the proposed estimators of m_i described in Section 3.1 and the V^e components yields six candidate variance estimators.

4 | EMPIRICAL APPLICATION (SERVICE ANNUAL SURVEY)

In this section, we apply the proposed variance estimator to empirical data from the 2018 collection of the SAS. Conducted by the U.S. Census Bureau, the SAS is a mandatory survey of approximately 78,000 employer businesses (companies) having one or more establishments located in the U.S. that provide services to individuals, businesses, and governments. The SAS collects aggregate and detailed revenues and expenses, e-commerce, exports and inventories data from a stratified sample of business firms with paid employees in selected industries in the services sector. As mentioned in Section 1, the key items collected by SAS are total revenue and total expenses; detailed breakdowns of these two totals items are requested from all sampled firms. The revenue detail items vary by industry within sector. Expense detail items, however, are primarily

the same for all sectors, with an occasional additional expense detail or two collected for select industries. Complete information on the SAS methodology is available at https://www.census.gov/programs-surveys/sas/technical-documentation/methodology.html.

SAS uses imputation to account for unit and item nonresponse, relying heavily on the ratio imputation models presented in Section 3.1 (specifically, PARAM1 and PARAM2). For total receipts and total expenses, the independent variables are either a highly correlated data item from the same reference period or historic data for the same item. The model for detail items use the corresponding totals item (reference period only). Thompson and Washington (2013) evaluated these imputation models in two of the sectors included in the SAS, explicitly fitting weighted no-intercept linear regression models within industry using respondent data to assess model fit (i.e. the PARAM1 method). For the totals, the ratio imputation models were appropriate, with adjusted-R^2 consistently above 95%. Given such strong predictors, the nonresponse adjustment is robust to the assumed response mechanism and should decrease the variance (Little & Vartivarian, 2005). However, the results for the details items were far less convincing, with adjusted-R^2 values often well-below 75-percent and non-significant slopes ($\alpha = 0.10$). The weak predictive power of this ratio imputation approach is exacerbated in the 2020 data collection year, as many businesses were closed or had business limited due to the COVID-19 pandemic. Not surprisingly, the SAS program managers were interested in exploring other imputation methods for these detail items, which are historically less frequently and reliably reported than their associated totals and whose imputed values are generally more difficult to independently validate.

The empirical application is restricted to five industries, collectively representing a cross-section of the survey's *expenditures* data collections. We chose a subset of industries from a candidate list provided by subject matter experts, requiring a minimum of three sampled units per strata in addition to validating that the NNRI model assumptions appeared to hold. Consequently, these industries are *not* representative of the larger survey. The input data for this application study consists of the sampled companies in the selected industries that tabulated a *positive* (non-zero) total expenses value; companies reporting zero-valued expenses were dropped. Furthermore, the SAS uses industry-average ratio imputation (not NNRI) for missing and invalid expenses items and implements a naïve random group variance estimator for all item. For these reasons, our estimates and variance estimates differ from the official published values.

Tables 1 and 2 describes selected features of the study industries. The sample sizes and response rates are rounded to comply with internal regulations. The unweighted response rates presented in Table 1 represent the proportion of sampled units that provided *a complete donor record* and do not correspond to the official response rates. On inspection, the response patterns displayed in Table 1 appear to be atypical of business surveys, in that generally the larger businesses (e.g. the certainty companies) respond at higher rates. However for NNRI, *all* detail items are imputed if either (1) the company was a nonrespondent or (2) the reported set of expenditures details did not add to the total, thus reducing the total number of respondents.

Figure 2 presents histograms of total expenses within study industry. Data values are suppressed for confidentiality protection. Typical of many business surveys, all five distributions are highly positively skewed, and the largest businesses are sampled with probability = 1 (certainty). These distributions of the total expenses variable are well-approximated by lognormal distributions in all industries. Notice that the variance of the approximate lognormal distribution is small in three industries (517210, 541210, and 713110), so that the compact support requirement in Assumption 3 is approximately true. Unfortunately, conformance to this requirement is unlikely for the other two study industries, although the convex support assumption should hold for all five industries.

TABLE 1 Number of expenditures detail items (details) and unweighted response rates of expenditures details (in percentages) by certainty and noncertainty status for Services Annual Survey (SAS) study industries. Businesses that are included with certainty are sampled with probability = 1; noncertainty businesses are sampled with probability less than 1

		Details	Response rates (in Percentages)		
Industry	**Description**	y_i	**Total**	**Certainty**	**Noncertainty**
221122	Electric power distribution	7	75	65	95
517210	Wireless telecommunications carriers	9	30	60	30
541211	Offices of certified public accountants	7	80	75	85
621410	Family planning centers	9	85	80	90
713110	Amusement and theme parks	7	60	60	60

TABLE 2 Sample design characteristics of Services Annual Survey (SAS) study industries. The range of strata sampling rates excludes the certainty stratum

		Sample size (n)		Number of	Strata sampling rates (n_h/N_h)	
Industry	**Overall n/N**	**Certainty**	**Noncertainty**	**Strata**	**Minimum**	**Maximum**
221122	0.0785	80	30	7	0.0061	0.3448
517210	0.0620	60	350	15	0.0140	0.5000
541211	0.0035	40	150	13	0.0020	0.0894
621410	0.1055	50	50	7	0.0122	0.4468
713110	0.0709	40	30	4	0.0160	0.1125

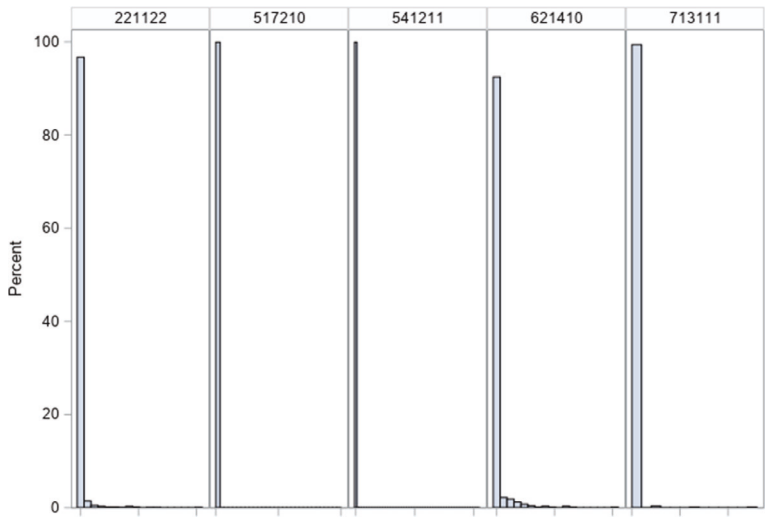

FIGURE 2 Histograms of total expenses within industry for the studied industries. Data source: Service Annual Survey (2018, U.S. Census Bureau)

Table 2 provides sample design characteristics of the study industries. The overall sampling rate (n/N) is non-negligible in four of the five study industries. All industries include a certainty strata (all units included with probability = 1), which is excluded from the range provided in Table 2 but is included in the computation of the overall sampling rate. The sampling rates for the noncertainty industry strata vary greatly, with at least one stratum in each industry having a non-negligible sampling rate.

Table 3 provides ratios of the NNRI detail item totals to their corresponding total expenses values (denoted $\hat{R}_y$) and ratios of the NNRI detail items variance estimates computed with our proposed variance estimators to their corresponding naïve variance counterparts (denoted R_{METHOD}). The $\hat{R}_y$ estimates are presented without error bounds and are provided to illustrate the prevalence of a reported detail item in an industry. Since the naïve estimates do not account for variance caused by nonresponse and imputation, we expect the naïve variance estimates to be smaller than those obtained by the proposed variance estimates in general, although ratios close to 1 could indicate an extremely effective hot deck prediction. Variance ratios greater than 1000 are indicated by an 'XXX'.

The parametric ratio estimator employed by the $\hat{V}_{\text{PARAM1}}$ is clearly problematic. The variance estimates obtained with the $\hat{V}_{\text{PARAM1}}$ for the rarely reported detail items (values less than 10%) are often considerably larger than the counterparts obtained with non-naïve variance estimates. However, using modeled residuals for σ_e^2 appears to underestimate the variances, as evidenced by frequency of variance ratios with values less than one. Similarly the variance estimates obtained using the modeled residuals with the PARAM2 model ($\hat{V}_{\text{PARAM2(M)}}$) tend to be much larger than those obtained with any other considered variance estimator, providing evidence of a poor model fit for σ_e^2.

Table 4 presents the coefficients of variation (c.v's) of each item for each considered variance estimator in percentages. At the 95% confidence level, a total with an associated c.v. greater than 51% (=1/1.96) is not significantly different from zero. Thus, the c.v.'s provide a measure of the practical impact of the variance estimators on inference.

As one might expect, given the results in Table 3, the c.v.'s obtained using $\hat{V}_{\text{PARAM1}}$ or $\hat{V}_{\text{PARAM2(M)}}$ estimates are much larger than those obtained with any other considered variance estimator. Recall that the PARAM1 model utlizes an *industry level* ratio estimator, likely inappropriate. Table 3 provides further indications that the modeled residuals in ($\hat{V}_{\text{PARAM1(M)}}$) do not improve this ratio estimator's performance, as the associated c.v.'s tend to be smaller than their naïve variance counterparts. In contrast, the c.v.'s obtained with the $\hat{V}_{\text{PARAM2}}$, $\hat{V}_{\text{NONPARAM}}$, and $\hat{V}_{\text{NONPARAM(M)}}$ estimators are generally similar, with a few visible exceptions. These large differences could be due to model misspecifications, but could also be confounded with small sample size effects for many of the detail items.

Despite the similarity of their c.v.'s, the variance estimates of corresponding items obtained with $\hat{V}_{\text{PARAM2}}$, $\hat{V}_{\text{NONPARAM}}$, and $\hat{V}_{\text{NONPARAM(M)}}$ are quite different. This affects confidence interval width, and expected coverage by extension. Without a gold standard against which to measure these variances, however, we have no viable recommendation. Consequently, we conducted the Monte Carlo simulation study described in Section 5.

5 | SIMULATION STUDY

To evaluate the finite-sample performance of the proposed method over repeated samples, we conducted a simulation study. The simulation varies in four factors: parametric distribution of the

TABLE 3 Ratios of nearest neighbour ratio imputation (NNRI) detail item totals to total expenses (line 1 in each set of industry results) and ratios of proposed variance estimators to corresponding naïve variance estimates within industry (lines 2–7 in each set of industry results). The detail item proportions will not necessarily add to 1 due to rounding. Entries with 'XXX' indicate ratio values greater than 1000. For all industries, the detail items are Y1 = gross annual payroll; Y2 = fringe benefits; Y3 = temporary staff payroll; Y4 = expensed software; Y5 = depreciation costs; Y6 = expensed equipment; and Y7 = other expenses. In industry 517210, Y8 = access charges and Y9 = universal service and similar charges. In industry 621410, Y8 = professional liability insurance and Y9 = medical supply costs. Data Source: Service Annual Survey (2018), U.S. Census Bureau

Industry	Variance	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9
221122	$\hat{R}_y$	0.1	0.0	0.0	0.0	0.2	0.0	0.6		
	R_{PARAM1}	6.2	2.9	XXX	14.4	25.9	7.7	1.6		
	$R_{\text{PARAM1(M)}}$	1.5	0.5	82.0	0.1	4.0	0.1	0.4		
	R_{PARAM2}	16.8	2.9	XXX	18.0	19.7	2.7	0.8		
	$R_{\text{PARAM2(M)}}$	16.7	6.1	XXX	91.4	47.4	494.4	40.2		
	R_{NONPARAM}	8.6	3.8	XXX	2.0	9.3	2.0	0.7		
	$R_{\text{NONPARAM(M)}}$	8.6	3.8	XXX	2.0	9.3	2.0	0.7		
517210	$\hat{R}_y$	0.1	0.0	0.0	0.1	0.2	0.3	0.3	0.0	0.0
	R_{PARAM1}	XXX	677.6	XXX	1.3	XXX	XXX	1.3	9.6	0.5
	$R_{\text{PARAM1(M)}}$	0.1	0.2	0.5	1.0	2.5	2.6	0.3	0.0	0.4
	R_{PARAM2}	4.3	2.5	4.5	6.3	13.7	18.8	1.4	2.9	5.1
	$R_{\text{PARAM2(M)}}$	3.2	171.4	292.1	3.3	7.6	612.1	243	15.9	787.0
	R_{NONPARAM}	2.5	2.7	2.5	1.2	10.1	4.2	2.2	2.3	2.8
	$R_{\text{NONPARAM(M)}}$	3.5	2.7	3.3	4.4	8.1	13	1.7	2.9	4.2
541211	$\hat{R}_y$	0.5	0.1	0.0	0.0	0.0	0.0	0.3		
	R_{PARAM1}	1.0	1.9	9.9	1.2	2.0	1.4	2.6		
	$R_{\text{PARAM1(M)}}$	0.7	0.5	0.3	0.2	0.2	0.0	1.0		
	R_{PARAM2}	0.9	1.5	2.2	1.1	1.7	0.9	2.1		
	$R_{\text{PARAM2(M)}}$	1.1	124.4	15.5	1.3	89.3	XXX	6.8		
	R_{NONPARAM}	1.0	1.2	1.3	1.1	1.2	0.8	1.9		
	$R_{\text{NONPARAM(M)}}$	1.0	1.4	1.7	1.0	1.6	0.8	1.3		
621410	$\hat{R}_y$	0.5	0.1	0.0	0.0	0.0	0.0	0.3	0.0	0.1
	R_{PARAM1}	0.9	2.0	4.7	3.5	3.8	1.4	1.5	1.7	1.5
	$R_{\text{PARAM1(M)}}$	0.5	1.5	0.4	0.1	0.8	0.1	0.3	0.1	0.7
	R_{PARAM2}	1.0	7.0	4.5	1.4	4.6	1.4	1.0	1.1	4.5
	$R_{\text{PARAM2(M)}}$	19.9	7.1	5.1	52.7	7.2	87.3	107.8	158.6	19.2
	R_{NONPARAM}	1.0	6.2	3.5	1.1	5.1	1.4	1.0	1.3	3.6
	$R_{\text{NONPARAM(M)}}$	0.8	2.5	3.6	1.2	1.9	1.5	1.2	1.0	2.1
713110	$\hat{R}_y$	0.3	0.1	0.0	0.0	0.2	0.0	0.4		
	R_{PARAM1}	XXX	371	38.7	269.3	100.2	XXX	85.9		
	$R_{\text{PARAM1(M)}}$	0.4	1.3	0.1	0.7	0.5	0.0	0.5		
	R_{PARAM2}	6.0	31.2	5.6	344.1	2.4	7.4	5.2		
	$R_{\text{PARAM2(M)}}$	1.0	14.1	3.4	XXX	7.6	51.7	2.1		
	R_{NONPARAM}	1.5	2.5	0.0	58.8	2.7	3.6	2.1		
	$R_{\text{NONPARAM(M)}}$	6.6	30.1	5.6	343.3	2.2	7.2	5.3		

TABLE 4 Coefficients of variation (c.v.) for nearest neighbour ratio imputation (NNRI) detail items (HT Totals) in percentages. Data Source: Service Annual Survey (2018), U.S. Census Bureau

Industry	Variance	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9
221122	$\hat{V}_{\text{Naïve}}$	2.2	3.9	0.3	7.1	1.3	10.9	4.2		
	$\hat{V}_{\text{PARAM1}}$	5.4	6.6	34.1	27.0	6.8	30.2	5.4		
	$\hat{V}_{\text{PARAM1(M)}}$	2.6	2.7	2.6	2.6	2.7	3.1	2.8		
	$\hat{V}_{\text{PARAM2}}$	8.9	6.6	27.2	30.2	6.0	17.9	3.7		
	$\hat{V}_{\text{PARAM2(M)}}$	8.9	9.6	106.3	68.0	9.3	242.0	26.9		
	$\hat{V}_{\text{NONPARAM}}$	6.4	7.6	14.7	10.1	4.1	15.6	3.6		
	$\hat{V}_{\text{NONPARAM(M)}}$	8.7	6.9	27.6	29.9	4.8	16.5	3.8		
517210	$\hat{V}_{\text{Naïve}}$	1.1	0.8	0.4	0.3	0.2	0.2	0.6	2.3	0.5
	$\hat{V}_{\text{PARAM1}}$	34.1	19.9	49.7	0.4	54.3	46.7	0.6	7.2	0.4
	$\hat{V}_{\text{PARAM1(M)}}$	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	$\hat{V}_{\text{PARAM2}}$	2.2	1.2	1.0	0.8	0.7	0.8	0.7	4.0	1.1
	$\hat{V}_{\text{PARAM2(M)}}$	1.9	10.0	7.8	0.6	0.5	4.8	8.6	9.3	14.0
	$\hat{V}_{\text{NONPARAM}}$	1.7	1.3	0.7	0.4	0.6	0.4	0.8	3.5	0.8
	$\hat{V}_{\text{NONPARAM(M)}}$	2.0	1.3	0.8	0.7	0.6	0.7	0.7	4.0	1.0
541211	$\hat{V}_{\text{Naïve}}$	3.4	4.0	5.9	6.8	6.5	23.9	3.0		
	$\hat{V}_{\text{PARAM1}}$	3.5	5.5	18.6	7.9	9.1	29.9	4.7		
	$\hat{V}_{\text{PARAM1(M)}}$	2.9	2.8	3.2	2.9	2.9	3.0	2.9		
	$\hat{V}_{\text{PARAM2}}$	3.2	4.8	8.7	7.3	8.4	23.4	4.2		
	$\hat{V}_{\text{PARAM2(M)}}$	3.6	44.1	23.3	8.0	60.3	1035.5	7.6		
	$\hat{V}_{\text{NONPARAM}}$	3.4	4.3	6.8	7.3	7.0	22.3	4.0		
	$\hat{V}_{\text{NONPARAM(M)}}$	3.4	4.6	7.8	7.3	8.1	21.9	3.4		
621410	$\hat{V}_{\text{Naïve}}$	4.3	2.6	5.4	8.5	3.7	11.4	5.7	9.5	3.6
	$\hat{V}_{\text{PARAM1}}$	4.1	3.8	11.8	16.0	7.3	13.4	7.1	12.4	4.5
	$\hat{V}_{\text{PARAM1(M)}}$	3.0	3.3	3.2	3.2	3.3	3.0	3.2	3.2	3.1
	$\hat{V}_{\text{PARAM2}}$	4.3	7.0	11.5	10.0	8.0	13.4	5.8	10.1	7.6
	$\hat{V}_{\text{PARAM2(M)}}$	19.1	7.1	12.2	62.3	10.0	106.6	60.0	121.3	15.7
	$\hat{V}_{\text{NONPARAM}}$	4.3	6.6	10.1	9.2	8.4	13.5	5.9	10.9	6.8
	$\hat{V}_{\text{NONPARAM(M)}}$	3.7	4.2	10.2	9.5	5.1	14.1	6.4	9.8	5.1
713110	$\hat{V}_{\text{Naïve}}$	1.4	0.8	0.5	1.0	1.3	4.2	1.3		
	$\hat{V}_{\text{PARAM1}}$	93.9	15.8	24.2	15.5	12.8	167.7	11.9		
	$\hat{V}_{\text{PARAM1(M)}}$	0.9	0.9	1.0	0.8	0.9	0.8	0.9		
	$\hat{V}_{\text{PARAM2}}$	3.3	4.6	9.2	17.5	2.0	11.5	2.9		
	$\hat{V}_{\text{PARAM2(M)}}$	1.4	3.1	7.1	219.3	3.5	30.3	1.8		
	$\hat{V}_{\text{NONPARAM}}$	1.6	1.3	0.5	7.2	2.1	8.0	1.9		
	$\hat{V}_{\text{NONPARAM(M)}}$	3.5	4.5	9.2	17.5	1.9	11.3	2.9		

size (auxiliary) variable x_i, the size of the finite population (N), relationship of auxiliary variable and detail items (x_i and y_i), and response propensity. The data generation is largely patterned after the realistic procedures described in Andridge et al. (2021), with each separate process outlined below.

5.1 | Create and stratify the finite population

We generated *three* different sets of $B = 2000$ finite populations of size N by drawing the size (auxiliary) variable x_i from

Population Scenario 1: $x_i \sim 100{,}000 * U(0, 1)$
Population Scenario 2: $x_i \sim Lognormal(4.1, 0.66)$
Population Scenario 3: $x_i \sim Lognormal(12, 1.72)$

The first population scenario ensures the compact and convex support requirements of Assumption 3. Thus, these data represent ideal conditions for the proposed NNRI variance estimators. However, business data population such as the SAS industry populations discussed in Section 4 are generally positively skewed. Consequently, we consider two lognormal distributions. The second population scenario exhibits mild deviations from the required compact support requirement, while respecting the convex support requirement (similar to the 517210, 541210, and 713110 industries' total expenses distributions discussed in Section 4). The third population scenario creates finite populations that do not exhibit the compact support requirement of Assumption 3, but resemble 221122 and 621410 industries' total expenses distributions presented in Section 4. Thus, the two lognormal population scenarios provide insights into the empirical results while testing the robustness of the proposed variance estimation approach.

Each finite population is stratified using the strata boundaries provided in Table 5.

5.2 | Generate sets of detail items in stratified finite populations

We used a two-step process to generate the sets of detail items values $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5})^\intercal$ associated with each unit, capturing important data features observed in several of the U.S. Census Bureau's economic programs. Specifically, the y_i for all units have non-zero values assigned to (y_{i1}, y_{i2}) but may have zero values in (y_{i3}, y_{i4}, y_{i5}).

To justify the use of the NNRI procedure, the *number of non-zero detail items* is directly related to unit size. Within each stratum S_i, we generate C_i (the number of non-zero detail items) for

TABLE 5 Strata boundaries for the simulation study (Section 5)

Stratum S	Population Scenario 1	Population Scenario 2	Population 3
1	<25,000	<55	<40,000
2	$25{,}000 \leq X < 50{,}000$	$55 \leq X < 85$	$40{,}000 \leq X < 150{,}000$
3	$50{,}000 \leq X < 75{,}000$	$85 \leq X < 150$	$150{,}000 \leq X < 500{,}000$
4	≥75,000	≥150	≥500,000

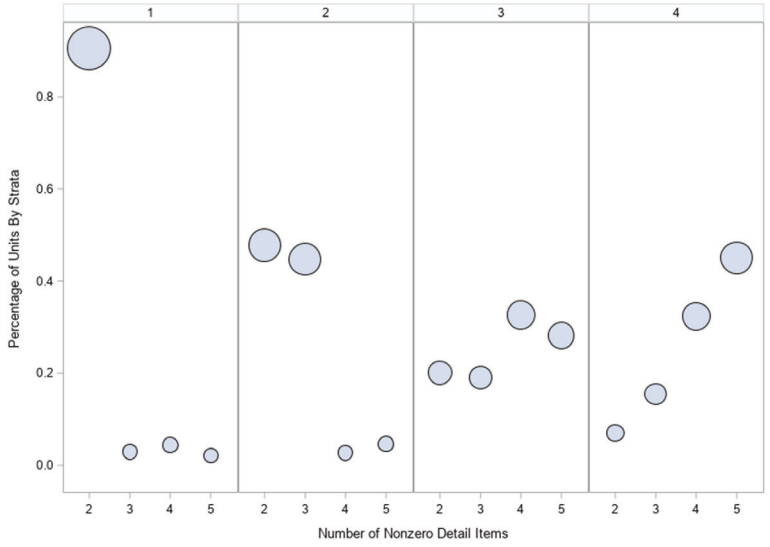

FIGURE 3 Bubble plots of realized nonzero detail items c in a single simulated finite population. Bubble plots are computed within sampling strata. Relative size of each bubble indicates stratum proportion. Strata are numbered in increasing order, with Stratum 1 containing the smallest units

unit i from a discrete distribution of {2,3,4,5} with selection probability $P(C_i = c|x_i) = p(x_i)$ where $p(x) = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ are given by

$$(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = \begin{cases} (0, .91, .03, .03, .03) & \text{in Stratum 1} \\ (0, .50, .40, .05, .05) & \text{in Stratum 2} \\ (0, .20, .20, .30, .30) & \text{in Stratum 3} \\ (0, .05, .15, .40, .40) & \text{in Stratum 4} \end{cases}.$$

Figure 3 presents bubble plots of the realized values of the number of nonzero details (c) from a single simulated population. As the unit size (x) increases, the number of nonzero detail items reported by each unit tends to likewise increase: for example, in the smallest unit size stratum (1), the majority of units provide two nonzero values, whereas in the largest unit size stratum (4), the majority of units provide four or five nonzero values.

Conditioning on the assigned c_i, we draw the R_i for each unit i from a multinomial distribution.

$$R_i|(x_i, c_i) \sim \text{Multinomial}\{x_i, (p_1, p_2, p_3, p_4, p_5)\},$$

with probabilities

$$(p_1, p_2, p_3, p_4, p_5) = \begin{cases} (.60, .40, .00, .00, .00) & \text{if } c = 2 \\ (.60, .30, .10, .00, .00) & \text{if } c = 3 \\ (.60, .25, .10, .05, .00) & \text{if } c = 4 \\ (.60, .20, .10, .05, .05) & \text{if } c = 5 \end{cases}.$$

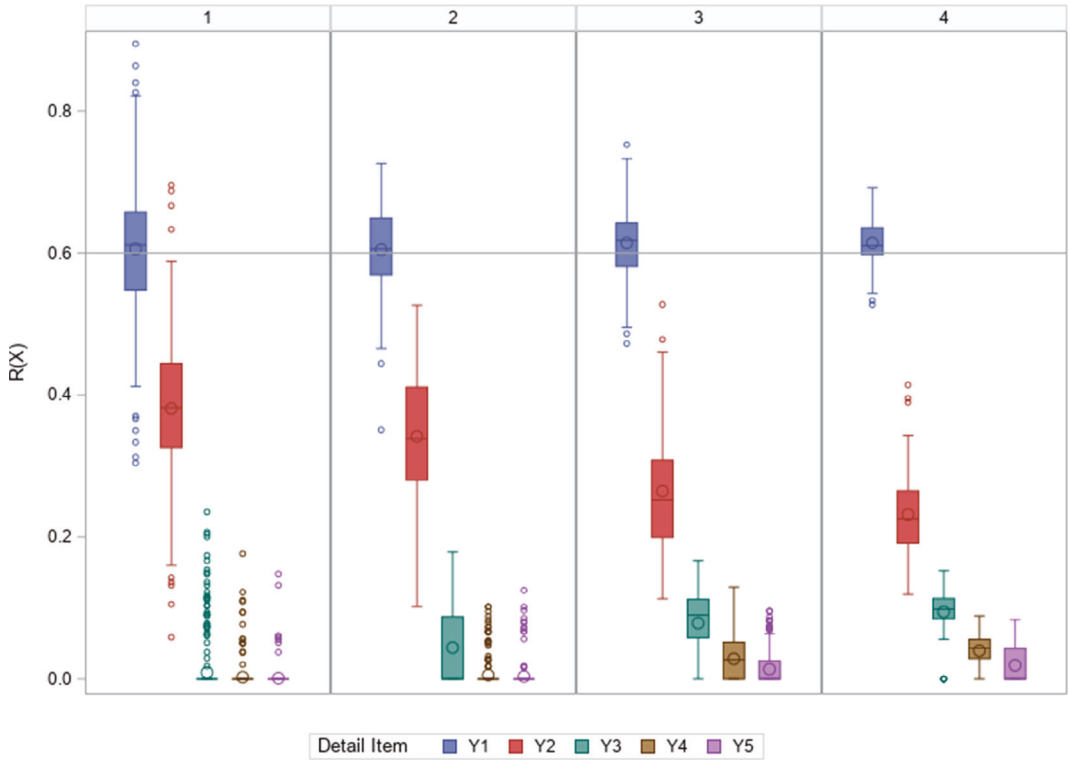

FIGURE 4 Distribution of detail item ratios $R(x)$ by sampling strata in a single simulated finite population

By design, $p_1 = 0.60$ for all units regardless its class. This detail item therefore represents about 60% of each unit's total (x_i). Figure 4 illustrates the subtle change in multinomial distributions as unit size (sampling strata) increases. The largest proportion of the total is always reported in item $Y1$, with item $Y2$ following. The remaining three items are more rarely reported, with the probability of a reported nonzero value being strongly related to unit size. This mimics patterns that were observed by Andridge et al. (2021) in several economic census datasets.

Lastly, using these R_i, we compute $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}) = x_i(R_{i1}, R_{i2}, R_{i3}, R_{i4}, R_{i5})$. This ensures that $x_i = \sum_{t=1}^{5} y_{it}$.

5.3 | Select stratified SRS-WOR samples

We select a single stratified SRS-WOR sample from each finite population. Table 6 provides the finite population size (N), the average stratum sizes N_h, the sampling fractions f_h, and average sample sizes n_h. Notice that the overall sampling fraction (f) is 302/1000 in the $N = 1000$ populations and is 152/500 in the $N = 500$ populations and are therefore both non-negligible. As typical of business surveys, the largest units are grouped into a certainty stratum, and a high proportion of the medium-sized units (grouped into strata 2 and 3) are sampled at a high (non-negligible) rates, requiring inclusion of the finite-population correction in $\hat{V}^m$.

TABLE 6 Averaged Stratum size and sample allocations. Sampling fractions are fixed in all simulations

Stratum S	1	2	3	4
$N = 1000$				
N_h	442	255	219	83
f_h	1/10	1/4	1/2	1
n_h	45	64	110	83
$N = 500$				
N_h	221	128	110	42
f_h	1/10	1/4	1/2	1
n_h	23	32	55	42

TABLE 7 Response propensities for the negative and positive MAR mechanisms

Stratum S	1	2	3	4
Negative MAR				
π_h	0.85	0.65	0.45	0.25
Positive MAR				
π_h	0.25	0.45	0.65	0.85

5.4 | Induce nonresponse and impute

Finally, within each sample, we generate missingness indicators for each unit i as $\delta_i \sim Bernoulli(\pi)$ under different response mechanisms: uniform (missing completely at random) with $\pi = 75\%$ and 50% (MCAR); uniform response within strata (missing at random) with smaller units being more likely to respond (Negative MAR); and uniform response within strata with larger units being more likely to respond (Positive MAR). Table 7 presents the strata response propensities for the negative and positive MAR response mechanisms. Notice that between-stratum differences in response propensities quite large; these discrepancies are exaggerated for illustration. The MCAR response propensities resemble the observed patterns for industry 713110 and to a lesser extent, for industries 541211 and 621410 presented in Section 4, although the latter two industries could equally be categorized with negative MAR response patterns. The negative MAR response mechanism mimics the observed pattern for industry 221122, whereas the positive MAR response mechanism mimics the observed pattern for industry 517210.

Imputation and estimation are performed separately within each strata, with x_i as the matching variable for the NNRI of $y_i = (y_{i1}, \ldots, y_{iT})$ as outlined in Section 2.2.

5.5 | Simulation results

To evaluate the performance of the proposed NNRI variance estimators over repeated samples selected from different population distributions (scenarios), finite population sizes, and response mechanisms, we compute the relative bias of each variance estimator $\widehat{V}_{yp}$ and the coverage rates of the approximate 95% confidence intervals constructed with $\widehat{T}_y$ and $\widehat{V}_{yp}$ for variance estimation method p.

Table 8 reports the relative biases using *directly obtained residuals* for all five detail items for each population scenario and response mechanism for the $N = 1000$ populations. Results for the $N = 500$ populations are similar and are consequently not presented here, but are available upon request to the authors. The relative bias of each variance estimator is computed as $\text{RB}\left(\hat{V}_{yp}^{(b)}\right) = \left[\left(\sum_{b=1}^{B} \hat{V}_{yp}^{(b)}/B\right)/V_{yp}\right] - 1$ where $\hat{V}_{yp}^{(b)}$ is the estimate from the b^{th} sample ($B = 2000$) and V_{yp} is the Monte Carlo empirical (true) variance. Recall that $\sum_{i=2}^{5} Y_i$ represents a *maximum* of 40% of the corresponding total, $\sum_{i=3}^{5} Y_i$ represents a *maximum* of 20% of the corresponding total, and $\sum_{i=4}^{5} Y_i$ represents 5% or less of the corresponding total. Consequently, the analysis of relative biases of the variance estimates will be confounded with small sample size effects for each detail item except Y_1, with small unit differences in totals potentially representing large percentage differences. The main results of Table 8 can be summarized as follows:

- The naïve variance estimator severely underestimates the true variance of detail items Y_2, Y_3, Y_4, and Y_5 for all populations and response mechanisms;
- In Population Scenarios 1 and 2, the PARAM1 variance estimator *overestimates* the true variance of all detail items except for Y_1, essentially confirming the conjecture of overestimation posited Section 4 for industries 517210, 541210 and 713110. This variance estimator generally *underestimates* the true variance of the same detail items in Population Scenario 3, mimicking the empirical results for industries 221122 and 631400.
- In Population Scenarios 1 and 2, the PARAM2 variance estimates are nearly unbiased, regardless of response mechanism. However, in Population Scenario 3, the PARAM2 variance estimator tends to underestimate the variance of all detail items except for Y_1, with the degree of underestimation being more severe than their PARAM1 variance estimate counterparts.
- In Population Scenarios 1 and 2, the NONPARAM variance have inconsistent relative bias performance, although generally improved over the corresponding naïve variance estimates. This is not true in Population Scenario 3, where the variances estimates for all detail items except Y_1 are severely underestimated, and the variance estimates for Y_1 are overestimates, with the level of overestimation related to the response mechanism.

The simulation conditions in Population 2 with a MCAR or negative MAR response mechanism resemble the 541210 and 713110 industries' conditions; Table 8 provides evidence that the $\hat{V}_{PARAM2}$ are likely the most accurate estimates in this situation, even for the rarely-reported detail items. The simulation conditions in Population 2 with a positive MAR response mechanism resemble the industry 517210 conditions; the simulation results are close to the same for the $\hat{V}_{PARAM2}$ and the $\hat{V}_{NONPARAM}$, without a clear-cut favourite. The simulation conditions in Population 3 with a negative MAR response mechanism resemble the 221122 and 621400 industries' data; the relative biases for the MCAR response mechanism are similar for $\hat{V}_{PARAM1}$ and $\hat{V}_{PARAM2}$, but the $\hat{V}_{PARAM2}$ are less biased under the negative MAR response mechanism (the $\hat{V}_{NONPARAM}$ are severe underestimates for all detail items).

Table 9 reports the relative biases using the *modeled residuals* (see Section 3.3) for all five detail items for each population scenario and response mechanism for the $N = 1000$ populations. With the 'high' uniform response rate (i.e. MCAR with $\pi = 0.75$), as well as the negative and positive MAR response mechanism, modelling the residuals for $\sigma_e^2(x_i)$ generally reduces the relative bias of the PARAM2 variance estimates. Notice that the PARAM2(M) variance

TABLE 8 Relative biases of variance estimates using *directly-obtained residuals* for all detail items by population scenario and response mechanism computed from 2000 independent stratified SRS-WOR samples from the $N = 1000$ populations. *Negative* relative biases are in parenthesis. Population Scenario 1 = Uniform; Population Scenario 2 = Lognormal ($\mu = 4.1$, $\sigma = 0.66$); Population Scenario 3 = Lognormal ($\mu = 12.0$, $\sigma = 1.7$)

Population Scenario	Method ($\hat{V}_{yp}$)	MCAR ($\pi = 0.75$)					MCAR ($\pi = 0.50$)				
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_1	Y_2	Y_3	Y_4	Y_5
Population Scenario 1	NAÏVE	(0.00)	(0.77)	(0.89)	(0.93)	(0.95)	(0.01)	(0.54)	(0.75)	(0.82)	(0.85)
	PARAM1	(0.00)	0.45	0.47	0.56	0.12	(0.00)	0.31	0.51	0.64	0.23
	PARAM2	(0.00)	(0.04)	(0.07)	(0.07)	(0.05)	(0.00)	(0.02)	(0.05)	(0.03)	0.03
	NONPARAM	0.01	(0.02)	0.02	(0.01)	(0.04)	0.00	(0.05)	0.04	0.03	0.05
Population Scenario 2	NAÏVE	(0.13)	(0.36)	(0.49)	(0.57)	(0.62)	(0.29)	(0.57)	(0.73)	(0.78)	(0.82)
	PARAM1	(0.00)	0.14	0.70	0.54	0.18	0.02	0.33	0.67	0.58	0.18
	PARAM2	(0.01)	(0.07)	(0.02)	(0.04)	(0.03)	0.00	(0.01)	(0.02)	(0.00)	(0.02)
	NONPARAM	0.02	(0.12)	0.11	0.00	(0.04)	0.03	(0.03)	0.09	0.01	(0.07)
Population Scenario 3	NAÏVE	0.04	(0.96)	(0.96)	(0.99)	(1.00)	0.04	(0.98)	(0.98)	(1.00)	(1.00)
	PARAM1	0.04	(0.22)	(0.08)	(0.35)	(0.28)	0.04	(0.08)	(0.02)	(0.15)	(0.20)
	PARAM2	0.04	(0.28)	(0.17)	(0.39)	(0.30)	0.04	(0.14)	(0.10)	(0.19)	(0.21)
	NONPARAM	0.09	(0.73)	(0.57)	(0.78)	(0.82)	0.21	(0.76)	(0.69)	(0.77)	(0.81)
		Negative MAR					**Positive MAR**				
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_1	Y_2	Y_3	Y_4	Y_5
Population Scenario 1	NAÏVE	(0.01)	(0.60)	(0.75)	(0.88)	(0.92)	(0.00)	(0.49)	(0.72)	(0.74)	(0.75)
	PARAM1	(0.01)	0.45	0.64	0.57	0.11	(0.00)	0.68	1.08	1.49	0.46
	PARAM2	(0.01)	0.02	(0.00)	(0.02)	(0.03)	(0.00)	(0.06)	(0.02)	(0.07)	(0.06)
	NONPARAM	0.00	(0.02)	0.06	(0.00)	(0.04)	0.00	(0.07)	0.16	0.06	0.02
Population Scenario 2	NAÏVE	(0.22)	(0.52)	(0.72)	(0.83)	(0.88)	(0.44)	(0.68)	(0.77)	(0.77)	(0.75)
	PARAM1	(0.01)	0.53	0.81	0.67	0.16	(0.01)	0.59	1.32	1.23	0.56
	PARAM2	(0.02)	(0.01)	(0.04)	(0.03)	(0.05)	(0.05)	(0.11)	(0.09)	(0.09)	(0.02)
	NONPARAM	(0.01)	(0.10)	(0.03)	(0.15)	(0.26)	(0.01)	(0.04)	0.32	0.06	0.07
Population Scenario 3	NAÏVE	0.04	(0.99)	(0.99)	(1.00)	(1.00)	0.04	(0.94)	(0.93)	(0.99)	(0.99)
	PARAM1	0.04	(0.11)	(0.04)	(0.14)	(0.24)	0.04	(0.26)	(0.07)	(0.40)	(0.30)
	PARAM2	0.04	(0.21)	(0.17)	(0.21)	(0.28)	0.04	(0.33)	(0.19)	(0.45)	(0.33)
	NONPARAM	0.46	(0.86)	(0.83)	(0.84)	(0.88)	0.08	(0.69)	(0.47)	(0.76)	(0.82)

TABLE 9 Relative biases of variance estimates using *modeled residuals* for all detail items by population scenario and response mechanism computed from 2000 independent stratified SRS-WOR samples from the $N = 1000$ populations. *Negative* relative biases are in parenthesis. Population Scenario 1 = Uniform; Population Scenario 2 = Lognormal ($\mu = 4.1$, $\sigma = 0.66$); Population Scenario 3 = Lognormal ($\mu = 12.0$, $\sigma = 1.7$)

Population Scenario	Method ($\widehat{V}_{yp}$)	MCAR ($\pi = 0.75$)					MCAR ($\pi = 0.50$)				
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_1	Y_2	Y_3	Y_4	Y_5
Population Scenario 1	PARAM1(M)	(0.00)	(0.90)	(0.98)	(0.99)	(1.00)	0.00	(0.81)	(0.95)	(0.97)	(0.99)
	PARAM2(M)	(0.00)	(0.04)	(0.06)	(0.06)	(0.05)	(0.00)	(0.01)	(0.04)	(0.02)	0.04
	NONPARAM(M)	0.01	2.42	16.03	0.14	0.06	0.00	0.07	0.48	0.18	0.16
Population Scenario 2	PARAM1(M)	0.00	(0.47)	(0.46)	(0.23)	(0.47)	0.02	(0.47)	(0.50)	(0.30)	(0.55)
	PARAM2(M)	(0.01)	(0.06)	(0.01)	(0.02)	(0.02)	0.00	(0.01)	(0.02)	0.00	(0.01)
	NONPARAM(M)	0.03	(0.08)	0.42	0.15	0.16	0.05	0.04	0.41	0.16	0.12
Population Scenario 3	PARAM1(M)	0.04	(0.98)	(0.98)	(1.00)	(1.00)	0.04	(0.99)	(0.99)	(1.00)	(1.00)
	PARAM2(M)	0.04	0.10	0.34	0.00	0.05	0.04	0.22	0.29	0.22	0.10
	NONPARAM(M)	0.20	(0.40)	0.26	(0.51)	(0.65)	0.41	(0.55)	(0.22)	(0.57)	(0.69)
		Negative MAR					**Positive MAR**				
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_1	Y_2	Y_3	Y_4	Y_5
Population Scenario 1	PARAM1(M)	(0.00)	(0.81)	(0.96)	(0.99)	(1.00)	(0.00)	(0.81)	(0.93)	(0.95)	(0.98)
	PARAM2(M)	(0.00)	0.01	(0.00)	(0.02)	(0.02)	(0.00)	(0.05)	(0.01)	(0.04)	(0.02)
	NONPARAM(M)	0.02	0.00	0.15	0.04	(0.01)	0.04	0.54	10.88	0.40	0.18
	PARAM1(M)	(0.01)	(0.49)	(0.68)	(0.71)	(0.84)	0.01	(0.46)	(0.24)	0.48	0.25
	PARAM2(M)	(0.02)	(0.00)	(0.03)	(0.02)	(0.04)	(0.05)	(0.09)	(0.08)	(0.06)	(0.00)
	NONPARAM(M)	0.00	(0.07)	0.06	(0.09)	(0.19)	0.06	0.17	1.25	0.30	0.43
Population Scenario 3	PARAM1(M)	0.04	(1.00)	(1.00)	(1.00)	(1.00)	0.04	(0.97)	(0.97)	(1.00)	(1.00)
	PARAM2(M)	0.04	(0.05)	(0.02)	(0.02)	(0.07)	0.04	(0.03)	0.17	(0.10)	(0.00)
	NONPARAM(M)	0.55	(0.82)	(0.77)	(0.80)	(0.85)	0.72	8.17	16.58	5.78	0.81

estimates are nearly unbiased in Population Scenario 3 for all detail items, except the lowest uniform response rate mechanism (MCAR with $\pi = 0.50$); in this situation, the bias effects are likely overstated for the most rarely reported detail items (e.g. Y_4 and Y_5). This results contrast with those presented in the empirical case study in Section 4, and we suspect this could be an artifact of the data simulation process. The other approaches (PARAM1(M) and NONPARAM(M)) do not yield consistent improvements over their counterparts obtained with directly-obtained residuals. Given that the 2nd parametric method of estimating m_i (PARAM2) generally yields accurate variance estimates and the associated procedures for obtaining $\widehat{\sigma}_e^2$ appear tractable, we dropped the alternative parametric approach (PARAM1) from further consideration.

Regardless of population scenario, population size, and considered response mechanism, all NNRI *totals* were unbiased across the repeated samples. Again, we conjecture that this is an artifact of the simulation design, which ensures appropriate conditions for the distance function used to select the nearest neighbour for imputation. Nevertheless, coverage rates provide a

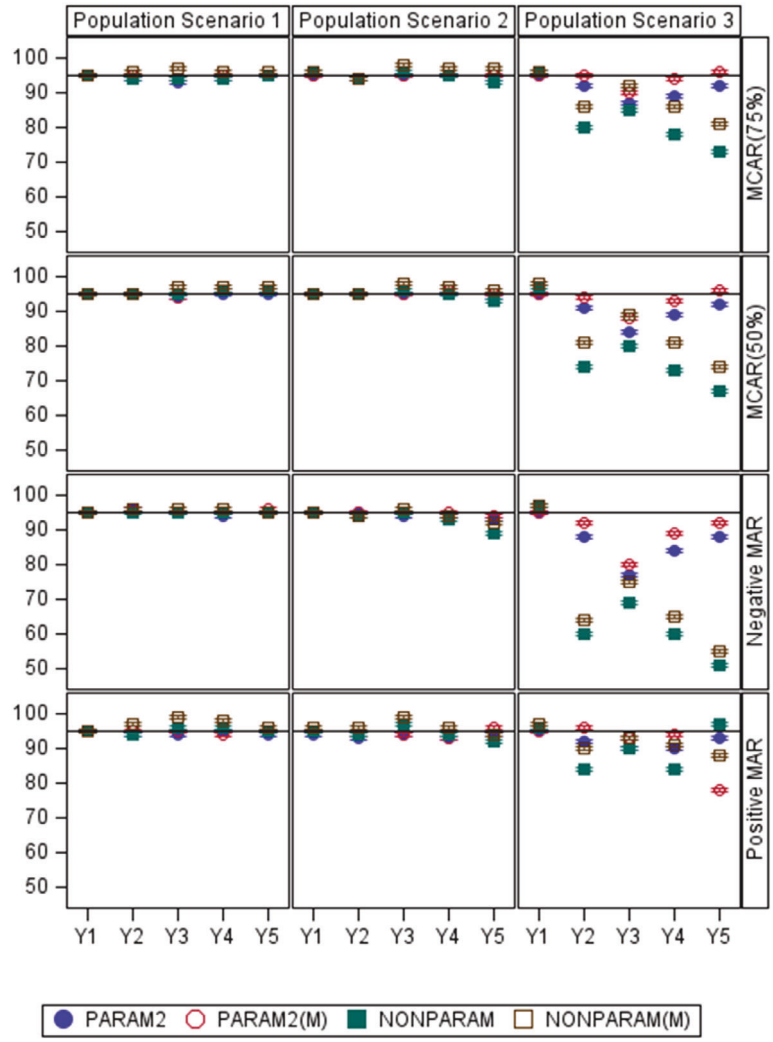

FIGURE 5 Coverage rates with Monte Carlo confidence bounds (in percentages) by population scenario and response mechanism using directly-obtained and modeled residuals with the PARAM2 and NONPARAM methods for the $N = 1000$ populations. Nominal coverage indicated by horizontal asymptote

measure of the practical impact of the bias of the considered variance estimators, given the unbiased estimates. Figure 5 presents the coverage rates using confidence intervals constructed from the PARAM2, PARAM2(M), NONPARAM, and NONPARAM(M) variance estimates.

Figure 5 provides evidence of

- Nominal or slight undercoverage with the PARAM2 variance estimates regardless of response mechanism when Assumption 3 is fully met (Population Scenario 1) or is mildly violated (Population Scenario 2). Moderate undercoverage under strong violations of Assumption 3, regardless of response mechanism. However, the undercoverage is nearly abated with the PARAM2(M) variances, except for the most infrequently reported item (Y_5) with the

positive MAR response mechanism. In this case, the modeled residuals are inadequate, probably because small units are less likely to respond under this response mechanism and Y_5 is (by design) rarely reported by smaller units.

- Inconsistent but rarely nominal coverage for the NONPARAM variance estimates for all response mechanisms in Population Scenarios 1 and 2 and consistent overcoverage with the NONPARAM(M) variance estimates in the same scenarios. Severe undercoverage with NONPARAM variance estimates for all response mechanisms in Population Scenario 3, with no improvements offered by the NONPARAM(M) approach.

Taken collectively, the simulation results support the generally poor performance of the PARAM1 and PARAM1(M) methods demonstrated in Section 4. Ultimately, the results obtained with PARAM2 are promising, especially given that the data generation models were not congenial to the *WLS* regression used to obtain $\hat{m}_i$ and the data violations in the third data generation scenario (found in two of the five studied empirical distributions). Despite its poor performance in this simulation study, the nonparametric method remains appealing due to its flexbility. It is possible that the model fit and estimation might be improved with a different choice of basis functions, although we would not recommend utilizing either variation with the studied industries in Section 4. That said, some caution should be exercised in over-generalizing these results, as the simulation utilizes a very specific multinomial distribution and a single sampling design.

6 | CONCLUDING REMARKS

Nearest neighbour ratio imputation (NNRI) is a useful approach for imputing an *entire set* of component detail items. Instead of directly imputing the set of detail item values (y_i) from the donor, NNRI imputes the proportions of donor ratios (R_i), which are in turn multiplied by the recipient's available total to derive imputed values for all items. This imputation method has several appealing properties, especially from a bias reduction perspective as discussed in Section 1.

NNRI guarantees additivity, as the summed details always equal the associated total. It accommodates subtle changes in unit level multinomial distributions that are associated with unit size. It yields realistic microdata, preserving multivariate relationships. In contrast to other frequently used imputation methods, the NNRI avoids inadvertently imposing possibly outdated historical patterns in the imputed data, as it is entirely restricted to current data (Andridge et al., 2021).

The numerous advantages of the NNRI procedure can be offset by the difficulty of obtaining a valid variance estimator. However, Yang and Kim (2019) show that by identifying the nearest donor using a single scalar with a suitable distance function, the NNRI estimator is asymptotically consistent. Following the frameworks of Shao and Steel (1999) and Yang and Kim (2019), we decompose our asymptotic variance into two parts and extend the variance estimation further to include the case of non-negligible sampling fractions employing both parametric and nonparametric models to obtain smooth estimators for set of ratios. Selecting an appropriate model for the data set at hand is essential, as demonstrated by the empirical application and the simulation study. Model determination is not completely straightforward: in both the empirical application and our simulation, the frequently reported detail items

tend to be strongly associated with the total, whereas the relationship between rarely-reported detail items and the total is less obvious. Nevertheless, our simulation studies provide fairly promising results in terms of bias and coverage, even with some model misspecification. In practice, one would expect that methodologists would develop and validate any implemented models after careful data analysis before implementing the proposed variance estimator.

Although promising, the empirical and simulation study results collectively suggest several areas of future research. First, we limited ourselves to uniform and lognormally distributed size variables in our study and restricted the response mechanisms to tractable MCAR or MAR models. Thus, the sensitivity of our variance estimator to more complex MAR models or even to non-random response mechanism bears study. Second, if the proportion of recipients to donors is large, then NNRI may repeatedly use the same donor, yielding insufficient variation within each imputation cell. Andridge et al. (2021) propose a modification of the NNRI method that addresses this issue in a multiple imputation framework; it would be useful to develop a single imputation analogue. Third, in practice, many auxiliary variables can be used to determine nearest neighbours, in which case, dimension reduction is necessary to mitigate matching discrepancies. Several techniques such as propensity score (Rosenbaum & Rubin, 1983), prognostic score (Hansen, 2008), or their combination (Yang & Zhang, 2020) can be potentially adopted in the NNRI framework. Finally, this paper considers only population totals. However, extending the current framework to general parameter estimation is also feasible (Yang & Kim, 2020). Given that hot deck imputation is used to create realistic microdata (as well as macrodata), such extensions are especially compelling topics for our future research.

ACKNOWLEDGEMENTS

The authors thank Carol Caldwell, William Davie Jr., Matthew Thompson, Katrina Washington, two anonymous referees, the associate editor, and the editor for their helpful comments on earlier versions of the manuscript, and Stephen Kaputa for his contributions to the empirical analysis. Yang is partially supported by the NSF grant DMS 1811245, IA grant 1R01AG06688, and NIEHS grant 1R01ES031651. Kim is partially supported by the NSF grant MMS 1733572.

ORCID

Chenyin Gao https://orcid.org/0000-0002-3850-5587
Katherine Jenny Thompson https://orcid.org/0000-0002-5564-2161

REFERENCES

Abadie, A. & Imbens, G.W. (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235–267.

Andridge, R.R. & Little, R.J. (2010) A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40–64.

Andridge, R. & Thompson, K.J. (2015) Assessing nonresponse bias in a business survey: proxy pattern-mixture analysis for skewed data. *The Annals of Applied Statistics*, 9(4), 2237–2265.

Andridge, R., Bechtel, L. & Thompson, K.J. (2021) Finding a flexible hot-deck imputation method for multinomial data. *Journal of Survey Statistics and Methodology*, 9(4), 789–809.

Bankier, M., Poirier, P., Lachance, M. & Mason, P. (2000) A generic implementation of the nearest-neighbour imputation methodology (nim). In: *Proceedings of the Joint Statistical Meetings*. American Statistical Association Alexandria, VA, pp. 571–578.

Beaumont, J.-F. & Bocci, C. (2009) Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37(3), 400–416.

Beaumont, J., Haziza, D. & Bocci, C. (2011) On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515–537.

Berg, E., Kim, J.K. & Skinner, C. (2016) Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4(4), 436–462.

Chen, J. & Shao, J. (2001) Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453), 260–269.

Fuller, W.A. (2009) *Sampling statistics*. Hoboken: John Wiley & Sons.

Hansen, B.B. (2008) The prognostic analogue of the propensity score. *Biometrika*, 95(2), 481–488.

Hastie, T.J. & Tibshirani, R.J. (1990) *Generalized additive models*, Volume 43. Boca Rotan: CRC Press.

Kalton, G. & Kaspryzk, D. (1982) The treatment of missing survey data. *Survey Methodology*, 12(1), 1–16.

Kim, J.K., Navarro, A. & Fuller, W.A. (2006) Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101(473), 312–320.

Kim, J.K., Fuller, W.A. & Bell, W.R. (2011) Variance estimation for nearest neighbor imputation for us census long form data. *The Annals of Applied Statistics*, 5(2A), 824–842.

Little, R. & Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2), 161–168.

Little, R.J., Yosef, M., Cain, K.C., Nan, B. & Harlow, S.D. (2008) A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events. *Statistics in Medicine*, 27(1), 103–120.

Magee, L. (1998) Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 60(1), 115–126.

Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63(3), 581–592.

Sande, I.G. (1982) Imputation in surveys: coping with reality. *The American Statistician*, 36(3a), 145–152.

Shao, J. & Steel, P. (1999) Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445), 254–265.

Shao, J. & Wang, H. (2008) Confidence intervals based on survey data with nearest neighbor imputation. *Statistica Sinica*, 18, 281–297.

Sigman, R. & Wagner, D. (1997) Algorithms for adjusting survey data that fail balance edits. In: *Proceedings of the Section on Survey Research Methods*.

Thompson, K.J. & Washington, K.T. (2013) Challenges in the treatment of unit nonresponse for selected business surveys: a case study. *Survey Methods: Insights from the Field*. Available from: https://surveyinsights.org/?p=2991 [Accessed 18th April 2022].

Willimack, D. & Nichols, E. (2010) A hybrid response process model for business surveys. *Journal of Official Statistics*, 26(1), 3–24.

Willimack, D.K. & Snijkers, G. (2013) The business context and its implications for the survey response process. In: Snijkers, G., Haraldsen, G., Jones, J. & Willimack, D.K. (Eds.) *Designing and conducting business surveys*, Chapter 2. Hoboken: John Wiley and Sons, pp. 39–82.

Willimack, D.K., Anderson, A. & Thompson, K.J. (2000) Using focus groups to identify analysts' editing strategies in an economic survey. In: *Proceedings of the Second International Conference on Establishment Surveys*. American Statistical Association Alexandria, VA.

Wolter, K. (2007) *Introduction to variance estimation*. Berlin: Springer Science & Business Media.

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114.

Wood, S.N., Pya, N. & Säfken, B. (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

Yang, S. & Kim, J.K. (2019) Nearest neighbor imputation for general parameter estimation in survey sampling. In: Huynh, K.P., Jacho-Chávez, D.T. & Tripathi, G. (Eds.) *The econometrics of complex survey data*. Bingley: Emerald Publishing Limited.

Yang, S. & Kim, J.K. (2020) Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics*, 47(3), 839–861.

Yang, S. & Zhang, Y. (2020) Multiply robust matching estimators of average and quantile treatment effects. *arXiv preprint arXiv:2001.06049*.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Gao, C., Thompson, K.J., Kim, J.K. & Yang, S. (2022) Nearest neighbour ratio imputation with incomplete multinomial outcome in survey sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–28. Available from: https://doi.org/10.1111/rssa.12841