Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference

Jae-Kwang Kim\(^1\) and Siu-Ming Tam\(^2\)

\(^1\)Department of Statistics, Iowa State University, Ames, Iowa, USA  
E-mail: jkim@iastate.edu  

\(^2\)Methodology Division, Australian Bureau of Statistics, Canberra and School of Mathematics and Statistics, University of Wollongong, Wollongong, New South Wales, Australia

Summary

The statistical challenges in using big data for making valid statistical inference in the finite population have been well documented in literature. These challenges are due primarily to statistical bias arising from under-coverage in the big data source to represent the population of interest and measurement errors in the variables available in the data set. By stratifying the population into a big data stratum and a missing data stratum, we can estimate the missing data stratum by using a fully responding probability sample and hence the population as a whole by using a data integration estimator. By expressing the data integration estimator as a regression estimator, we can handle measurement errors in the variables in big data and also in the probability sample. We also propose a fully nonparametric classification method for identifying the overlapping units and develop a bias-corrected data integration estimator under misclassification errors. Finally, we develop a two-step regression data integration estimator to deal with measurement errors in the probability sample. An advantage of the approach advocated in this paper is that we do not have to make unrealistic missing-at-random assumptions for the methods to work. The proposed method is applied to the real data example using 2015–2016 Australian Agricultural Census data.

Key words: Calibration weighting; Measurement error; Non-response; Regression estimation; Selection bias.

1 Introduction

Suppose we are interested in estimating some finite population parameters, for example, the finite population mean, of a target population based on a data set. If the data set comes from a probability sample, parameter estimation is straightforward, and we can draw on the extensive literature on survey sampling over the past century, for example, Fuller (2009), Särndal et al. (1992) and Chambers and Clark 2012. However, if the data set comes from a non-probability sample, for example, from a big data source, the estimation is less
straightforward, and the theory for making inference with non-probability samples is not fully
developed. Tam & Clarke (2015) and Pfefffermann (2015) addressed methodological uses and
challenges of big data in the production of official statistics.

The perils and pitfalls in using big data are primarily under and over coverage and self-
selection. Bias from under coverage is akin to bias from non-random samples for inference,
and the bias from self-selection is akin to non-response bias in surveys. These biases have been
discussed extensively in the statistics literature (see, e.g. Elliott & Valliant, 2017; Groves, 2006;
Groves & Peytcheva, 2008; Hand, 2018; Kaplan et al., 2014; Keiding & Louis, 2016; Lohr &
Raghunathan, 2017; Sax et al., 2003; Tam & Kim, 2018). Specific discussion of these biases
can be found in Baeza-Yates (2018) for web data and Brodie et al. (2018) on data from smart
phones and wearable devices and Olteanu et al. (2019) for social media data. The weighting
methods considered in Valliant and Dever (2011) and Elliot and Valliant (2017) are based on
a missing-at-random assumption (MAR) of Rubin (1976). The MAR assumption is a strong
assumption, and there is no way to verify this assumption from the data only.

Survey data integration, which is developed to combine information for two independent
surveys from the same target population, can be used to handle the selection bias of non-
probability samples by incorporating a probability sample. Rivers (2007) proposed a mass
imputation approach for survey integration. In Rivers (2007), the nearest neighbour matching
imputation is used to identify the imputed value for each element in the probability sample.
Zhang (2012) developed a statistical theory for register-based statistics and data integration.
Bethlehem (2016) discussed practical issues in sample matching for solving the selection bias
in the non-probability sample. While matching-based imputation is promising and potentially
useful in practice, it is still based on the MAR. Chen et al. (2020) also considered a weight-
ing adjustment method based on parametric model assumptions on the selection mechanism
for the non-probability sample, but the MAR assumption is still required. Rao (2020) provided
comprehensive reviews of statistical methods of data integration for finite population inference.

In this paper, we propose a novel method of data integration for handling big data by incor-
porating survey sample data. The sampling mechanism for big data is not necessarily MAR.
That is, there can be some systematic difference between the big data sample and the survey
sample even after adjusting for the auxiliary variables. We assume that the survey variables
are observed in both samples, but allow them to be inaccurately measured in one sample. Our
approach is to treat the big data sample as a finite population of incomplete (or inaccurate)
observations. Furthermore, the incomplete observations in the population can be treated as
auxiliary information for calibration weighting (Deville & Särndal, 1992; Kim & Park, 2010).
Thus, standard techniques such as calibration weighting for incorporating auxiliary information
from the finite population can be used directly. To conduct calibration estimation in the survey
data, we need to identify the subset of the probability sample that also belongs to the big data
sample. This is somewhat similar in spirit to dual frame estimation (Hartley, 1962; Skinner &
Rao, 1996). In our application, the big data sample is subject to coverage errors, but the survey
sample is not. The proposed method is particularly useful for government statistical agencies
which can effectively apply such matching. When the accurate matching is not possible, we
propose a novel classification method to identify the overlapping units using the matching vari-
able observed from two data sources. Fully nonparametric propensity scores are obtained from
the proposed classification procedure, and they can be used to correct for the bias in applying
the data integration estimator with inaccurate matching.

The paper is organised as follows. In Section 2, basic setup is introduced. In Section 3,
the basic idea for data integration is introduced. In Section 4, a semi-supervised classification
method is introduced to identify the overlapping units with big data. In Section 5, an efficient
method for data integration is introduced. In Section 6, the proposed method is extended to

International Statistical Review (2021), 89, 2, 382–401
© 2020 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
the case of measurement errors in the sample observation. Two limited simulation studies are presented in Section 7, and an application of the proposed method to an official statistics is presented in Section 8. Some concluding remarks are made in Section 9.

2 Basic Setup

Consider a finite population $U = \{1, \ldots, N\}$ of size $N$. From the finite population, we have two samples, denoted by $A$ and $B$, where $A$ is a probability sample and $B$ is a big data sample obtained by an unknown selection mechanism. From both samples, we measure the study variable $Y$. Initially, we assume that $Y$ is measured without measurement error in sample $A$, but we shall relax this assumption in Section 6. However, in sample $B$, $Y$ is not necessarily measured accurately. Thus, instead of observing $y_i$, we observe $y_i^*$, which is a contaminated version of $y_i$, from sample $B$. For simplicity, we assume that

$$y_i^* = \beta_0 + \beta_1 y_i + e_i,$$

where $(\beta_0, \beta_1)$ is an unknown parameter and $e_i \sim (0, \sigma^2)$. Model (1) implies that $y_i^*$ can be systematically different from $y_i$. In the special case of $(\beta_0, \beta_1) = (0, 1)$, there is no measurement bias in $y_i^*$. In addition, because the selection mechanism for the big data sample is unknown, it is subject to selection bias. Generally speaking, the selection bias of big data cannot be ignored, and adjusting for the selection bias is critical (Meng, 2018).

To correct for the selection bias and measurement errors in the big data, we assume that we have a gold standard survey sample. Obtaining survey sample data is often expensive, but the gold standard can be used to improve the quality of the big data sample. Furthermore, optimal allocation of the resources can make the final analysis more cost effective.

To make sample $A$ a gold standard sample, a probability sampling design for selecting sample $A$ is employed, and $y_i$ are accurately observed from the sample. From sample $A$, we can compute $\hat{T}_A = \sum_{i \in A} d_i y_i$, a design-unbiased estimator of $T = \sum_{i=1}^N y_i$, where $d_i = \pi_i^{-1}$ is the design weight of unit $i$ and $\pi_i$ is the first-order inclusion probability of unit $i$ in sample $A$. Table 1 presents the data structure of our setup. We also assume that it is possible to identify elements in sample $A$ also belonging to sample $B$. That is, we can create $\delta_i$ for $i \in A$, where

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise} \end{cases}.$$

Thus, we can observe $\delta_i$ in sample $A$ if the individual-level matching is possible. We shall relax this assumption in Section 5.

Our goal is to combine the observations in the two data sets to find an improved estimator of $T$. By making a proper use of big data through weighting, we can obtain an improved estimator of $T$ over $\hat{T}_A$, which completely ignores the information in the big data sample. Combining two data sources is called data integration, and we will consider data integration as a general tool for making a proper use of big data for finite population inference. Challenges in data

| Data | $Y^*$ | $Y$ | Representative? |
|------|-------|-----|----------------|
| A    | ✓     | Yes |               |
| B    | ✓     | No  |               |
integration are outlined in Lohr and Raghunathan (2017) and Hand (2018). Tam and Kim (2018) provided methods for adjusting such bias by using data integration. This paper extends the work of Tam and Kim (2018) to non-binary variables and also addresses situations when there are measurement errors or matching errors in the data sets.

3 Data Integration for Handling Selection Bias

We first consider the simple case of no measurement errors in $Y$, that is, $y_i^* = y_i$. Now, we can conceptually define $\delta_i$ in (2) throughout the finite population. Thus, the set of elements with $\delta_i = 1$ is the big data sample. We can decompose

$$ T = \sum_{i=1}^{N} y_i = T_b + T_c, $$

where $T_b = \sum_{i=1}^{N} \delta_i y_i$ and $T_c = \sum_{i=1}^{N} (1 - \delta_i) y_i$. Because $T_b$ can be obtained from sample $B$, we only have to estimate $T_c$ from sample $A$. Thus, we can use

$$ \widehat{T}_{DI} = T_b + \sum_{i \in A} d_i (1 - \delta_i) y_i $$

as a design-based estimator of $T$ obtained from two samples. If the population size $N$ is known, a better estimator is

$$ \widehat{T}_{PDI} = T_b + (N - N_b) \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i}{\sum_{i \in A} d_i (1 - \delta_i)}, $$

(3)

where $N_b = \sum_{i=1}^{N} \delta_i$ is the size of sample $B$. Estimator $\widehat{T}_{PDI}$ in (3) is essentially a post-stratified estimator with the two post-strata defined by $\delta_i = 1$ and $\delta_i = 0$, respectively.

The design variance of $\widehat{T}_{PDI}$ in (3) is

$$ \text{Var}(\widehat{T}_{PDI}) = (N - N_b)^2 \text{Var} \left( \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i}{\sum_{i \in A} d_i (1 - \delta_i)} \right) \approx \text{Var} \left( \sum_{i \in A} d_i (1 - \delta_i)(y_i - \overline{Y}_c) \right), $$

where $\overline{Y}_c = \sum_{i=1}^{N} (1 - \delta_i) y_i / (N - N_b)$. Here, the approximate equality follows from Taylor linearisation applied to the ratio component in (3). If the sampling design for sample $A$ is simple random sampling of size $n$ with $n/N \approx 0$, we have

$$ \text{Var}(\widehat{T}_{PDI}) \approx (1 - W_b) \frac{N^2}{n} S_c^2, $$

(4)

where $W_b = N_b/N$ and $S_c^2 = (N - N_b)^{-1} \sum_{i=1}^{N} (1 - \delta_i)(y_i - \overline{Y}_c)^2$. Thus, the variance reduction of $\widehat{T}_{PDI}$ compared with $\widehat{T}_a = \sum_{i \in A} d_i y_i$ is

$$ \frac{\text{Var}(\widehat{T}_{PDI})}{\text{Var}(\widehat{T}_a)} = (1 - W_b) \frac{S_c^2}{S^2}. $$

International Statistical Review (2021, 89, 2, 382–401) © 2020 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
If $S^2_c \approx S^2$, the data integration estimator is always more efficient than the design-based estimator using sample $A$ only. In fact, from (4), the effective sample size using the post-stratified data integration estimator is

$$n^* = n \frac{1}{1 - W_b} \frac{S^2}{S^2_c}.$$ 

Thus, if we define $c_a$ and $c_b$ to be the per-unit cost of observing $y_i$ in sample $A$ and sample $B$, respectively, the total cost function using post-stratified data integration estimation is $C_{DI} = c_a n + c_b N_b$, while the total cost required to obtain the same efficiency of $\hat{T}_a$ is $C_a = c_a n^*$. If $S^2_c \approx S^2$, we have

$$C_{DI} - C_a = c_b N W_b - c_a n \frac{W_b}{1 - W_b}.$$ 

Therefore, given the same efficiency, the cost for using post-stratified data integration estimator is lower than using sample $A$ only if

$$\frac{c_b}{c_a} \leq \frac{n}{N} \frac{1}{1 - W_b}. \quad (5)$$

Thus, if the under-coverage rate of $B$ is less than $(c_a/c_b) \cdot (n/N)$, the proposed data integration estimation is cost effective by (5).

## 4 Efficient Estimation

We now discuss how to further improve the efficiency of the data integration estimator. One approach is to use the idea of ratio estimation for $T$ by treating $x_i = \delta_i y_i$ as the auxiliary variable, which is observed throughout the finite population. Thus,

$$\hat{R} = \frac{\sum_{i=1}^N x_i}{\sum_{i \in A} d_i x_i}$$

can be multiplied to direct estimator to reduce the variance, that is, to improve efficiency. The resulting ratio estimator is

$$\hat{T}_{Rat DI} = \hat{T}_a \hat{R} = \hat{T}_a \frac{\hat{T}_a}{\hat{T}_b}, \quad (6)$$

where $\hat{T}_b = \sum_{i \in A} d_i \delta_i y_i$ and $\hat{T}_a = \sum_{i \in A} d_i y_i$. Thus, $\hat{T}_{Rat DI}$ in (6) is called the ratio data integration estimator. Note that we can express $\hat{T}_{Rat DI}$ as

$$\hat{T}_{Rat DI} = \sum_{i \in A} d_i \left( \frac{T_b}{T_b} \right) y_i = \sum_{i \in A} w_i y_i,$$

where $w_i$ satisfies

$$\sum_{i \in A} w_i x_i = \sum_{i \in A} d_i \left( \frac{T_b}{T_b} \right) \delta_i y_i = \sum_{i=1}^N \delta_i y_i = \sum_{i=1}^N x_i. \quad (7)$$

International Statistical Review (2021), 89, 2, 382–401
© 2020 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
Thus, equality (7) implies that the ratio data integration estimator satisfies the calibration property of the auxiliary variable in the sense that the estimator applied to $x_i$ matches the known population total of $x_i$.

More generally, we can apply the calibration estimation method to $x_i = (1, \delta_i, \delta_i y_i)^T$, because $\sum_{i=1}^{N}(1, \delta_i, \delta_i y_i) = (N, N_b, T_b)$ is known. Specifically, we can find $\{w_i : i \in A\}$ that minimises an objective function $Q(d, w)$ subject to the calibration equation $\sum_{i \in A} w_i x_i = \sum_{i=1}^{N} x_i$. The regression estimator is based on

$$Q(d, w) = \sum_{i \in A} d_i \left( \frac{w_i}{d_i} - 1 \right)^2.$$  

The solution to the optimisation problem is

$$w_i = d_i X_N^T \left( \sum_{i \in A} d_i x_i x_i^T \right)^{-1} x_i,$$  

where $X_N = \sum_{i=1}^{N} x_i$.

To understand the solution in (8), if we write $x_i = (1 - \delta_i, x_{1i}^T)^T$ with $x_{1i} = \delta_i (1, y_i)^T$, the regression weight in (8) reduces to

$$w_i = \begin{cases} d_i X_{1i}^T \hat{\Sigma}_{xx11}^{-1} x_{1i} & \text{if } \delta_i = 1 \\ d_i (N_c / N_c) & \text{if } \delta_i = 0, \end{cases}$$

where $X_1 = \sum_{i=1}^{N} x_{1i}$, $\hat{\Sigma}_{xx11} = \sum_{i \in A} d_i x_{1i} x_{1i}^T$, $N_c = N - N_b$ and $\hat{N}_c = \sum_{i \in A} d_i (1 - \delta_i)$.

The weights in (9) satisfy

$$\sum_{i \in A} w_i (\delta_i, \delta_i y_i) = (N_b, T_b), \quad \sum_{i \in A} w_i (1 - \delta_i) = N_c.$$  

The regression data integration estimator is then defined as

$$\hat{T}_{RegDI} = \sum_{i \in A} w_i y_i,$$  

where $w_i$ is defined in (9). Inserting (9) into (10), we can write

$$\hat{T}_{RegDI} = \sum_{i=1}^{N} \delta_i (1, y_i)^T \hat{\beta}_1 + N_c \frac{\hat{\beta}_c}{\hat{N}_c},$$  

where $\hat{\beta}_c = \sum_{i \in A} d_i (1 - \delta_i) y_i$ and

$$\hat{\beta}_1 = \left( \sum_{i \in A} d_i \delta_i (1, y_i)(1, y_i)^T \right)^{-1} \sum_{i \in A} d_i \delta_i (1, y_i)^T y_i = (0, 1)^T.$$  

Therefore, the regression data integration estimator in (11) is algebraically equivalent to the post-stratified data integration estimator in (3). However, we can include other auxiliary variables observed throughout the finite population in the calibration equation; refer to Remark 1 for details.
For variance estimation, standard linearisation methods or replication methods for regression estimator can be applied. For example, a linearisation variance estimator for (10) can be written as

\[
\hat{V}(\hat{\alpha}_{\text{RegDI}}) = \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \hat{e}_i \hat{e}_j, \tag{12}
\]

where \(\pi_{ij}\) is the joint inclusion probability of unit \(i\) and \(j\), \(\hat{e}_i = y_i - x_i^T \hat{\beta}\) and \(\hat{\beta} = (\sum_{i \in A} d_i x_i x_i^T)^{-1} \sum_{i \in A} d_i x_i y_i\).

Remark 1. In addition to \(y_i\), if there is another variable \(z_i\) observed in both samples, we can incorporate this information into calibration estimation. That is, we use \(x_i = (1 - \delta_i, \delta_i y_i, \delta_i z_i)^T\) in the calibration estimation. If \(z_i\) is observed throughout the finite population, we can use \(x_i = (1 - \delta_i, \delta_i, \delta_i y_i, z_i)^T\).

Remark 2. In some cases, the big data may have duplication and lead to over-coverage problems. In this case, we can still apply the idea of calibration estimation by modifying the definition of \(\delta_i\) to be the number of times that the unit appears in sample B. In this case, we can use

\[
\sum_{i \in A} w_i (1, \delta_i, \delta_i y_i) = \sum_{i \in U} (1, \delta_i, \delta_i y_i) \tag{13}
\]
as the calibration equation.

Remark 3. The proposed method is also applicable when measurement errors exist in addition to selection bias in big data sample. That is, instead of observing \(y_i\), we observe \(y_i^*\), an inaccurate measurement for \(y_i\), in sample B. In sample A, in addition to observing \((y_i, \delta_i)\), we assume that it is possible to obtain \(y_i^*\) for units with \(\delta_i = 1\) by matching. Thus, we observe \((y_i^*, \delta_i, \delta_i y_i^*)\) in sample A. In this case, we can still use \(\delta_i y_i^*\) as a control for the calibration equation. Thus, instead of using \(x_i = (1 - \delta_i, \delta_i y_i)^T\), we can use \(x_i^* = (1 - \delta_i, \delta_i y_i^*)^T\) in (9) to get the calibration weights satisfying \(\sum_{i \in A} w_i (1 - \delta_i) = N_c, \sum_{i \in A} w_i \delta_i = N_b\) and \(\sum_{i \in A} w_i \delta_i y_i^* = \sum_{i \in B} y_i^*\).

5 Semi-supervised Classification

The proposed method in Section 3 is based on the assumption that the big-data indicator function \(\delta_i\) is observed for every element in sample A. If we have an access to the unique identifiers, then it is possible to match the records accurately and obtain \(\delta_i\). In other cases, we only have matching variables such as name, zip code and date of birth. In this case, we use these matching variables to obtain the best guess of \(\delta_i\), denoted by \(\hat{\delta}_i\), based on the observed value of the matching variables \(z_i\). Obtaining \(\hat{\delta}_i\) from the matching variables is a challenging classification problem. Furthermore, finding a bias-corrected estimator under misclassification error is not considered in the literature. In the context of multiple frame surveys, Lohr (2011) developed a bias-adjustment method assuming that the misclassification probabilities are known.

In our setup, note that \(\delta_i\) is observed for sample B, as \(\delta_i = 1\) if \(i \in B\) by definition. We do not observe \(\delta_i\) if \(i \in A\). Thus, this is a semi-supervised classification problem because the true label (\(\delta_i\)) for classification is available only for sample B. Here, we shall propose a maximum
likelihood method of semi-supervised classification under the setup of data integration. Note that unlike the probabilistic record linkage, we do not have to identify the pairs of matches and non-matches as in Fellegi and Holt (1976). We have only to determine whether each unit in sample \( A \) belongs to the particular subpopulation \( B \) or not.

To formally describe the idea of the proposed method, recall that the finite population \( U \) is decomposed into two groups, \( U = B \cup B^c \). We assume that \( \pi = P(\delta = 1) \) is known and given by \( \pi = N_b / N \). We have a probability sample \( A \) selected from \( U \) and observe \( z_i \) instead of observing \( \delta_i \) for all \( i \in A \). If the densities for two groups, \( p(z \mid \delta = 1) \) and \( p(z \mid \delta = 0) \), are known or estimated from the samples, then we can use

\[
P(\delta_i = 1 \mid z_i) = \frac{\pi p(z_i \mid \delta_i = 1)}{(1 - \pi) p(z_i \mid \delta_i = 0) + \pi p(z_i \mid \delta_i = 1)}
\]

to make classification for unit \( i \in A \). We use \( \hat{\delta}_i = 1 \) if we classify unit \( i \) as \( i \in B \). Otherwise, we use \( \hat{\delta}_i = 0 \). The decision rule is

\[
\hat{\delta}_i = 1 \iff \hat{P}(\delta_i = 1 \mid z_i) > \frac{1}{2},
\]

where

\[
\hat{P}(\delta_i = 1 \mid z_i) = \frac{\pi \hat{p}(z_i \mid \delta_i = 1)}{(1 - \pi) \hat{p}(z_i \mid \delta_i = 0) + \pi \hat{p}(z_i \mid \delta_i = 1)}.
\]

Note that \( p(z \mid \delta = 1) \) means the marginal density function of \( z \) among big data. Estimation of \( p(z \mid \delta = 1) \) is straightforward as long as we have access to the big data. Thus, we have only to estimate parameters in \( p(z \mid \delta = 0) \).

To discuss parameter estimation, suppose that \( z = (z_1, \ldots, z_K) \) and each \( z_k \) can take one of \( D \) values among the set \( Z_k = \{z_k^{(1)}, \ldots, z_k^{(D)}\} \) with unknown probabilities. We assume that

\[
p(z \mid \delta = 1) = \prod_{k=1}^K p_k(z_k \mid \delta = 1),
\]

where \( p_k(z_k \mid \delta = 1) = m_{kd} \) if \( z_k = z_k^{(d)} \) and \( \sum_{d=1}^D m_{kd} = 1 \). Because we can observe \( z_i \) among \( \delta_i = 1 \), we can estimate \( m_{kd} \) using

\[
\hat{m}_{kd} = \frac{1}{N_B} \sum_{i \in B} I(z_{ik} = z_k^{(d)}).
\]

Now, for the model for \( p(z \mid \delta = 0) \), we assume that

\[
p(z \mid \delta = 0) = \prod_{k=1}^K p_k(z_k \mid \delta = 0),
\]

where \( p_k(z_k \mid \delta = 0) = u_{kd} \) if \( z_k = z_k^{(d)} \) and \( \sum_{d=1}^D u_{kd} = 1 \). If we define

\[
\gamma_{ik}^{(d)} = \begin{cases} 1 & \text{if } z_{ik} = z_k^{(d)} \\ 0 & \text{otherwise}, \end{cases}
\]

then we can express \( m_{kd} = P(\gamma_{ik}^{(d)} = 1 \mid \delta_i = 1) \) and \( u_{kd} = P(\gamma_{ik}^{(d)} = 1 \mid \delta_i = 0) \).
To estimate $u_{kd}$, we use the following EM algorithm:

1. First note that if $\delta_i$ were observed, the complete-sample pseudo log-likelihood would be

$$l_{com}(u | \delta, \gamma) = \sum_{i \in A} d_i \delta_i \log \left\{ \prod_{k=1}^{K} m_{ik} \right\} + \sum_{i \in A} d_i (1 - \delta_i) \log \left\{ (1 - \pi) \prod_{k=1}^{K} u_{ik} \right\}$$

where

$$(m_{ik}, u_{ik}) = \sum_{d=1}^{D} Y_{ik}^{(d)} (m_{kd}, u_{kd}).$$

Note that there is no need to estimate $m_{ld}$ again, because we have access to big data directly. Only $u_{kd}$ are the parameters of interest.

2. In the E-step, we need to evaluate the conditional expectation of $l_{com}(u | \delta, \gamma)$ given the observed data. Thus, given the current parameters, we have only to compute

$$Q(u | u^{(t)}) = E \left\{ l_{com}(u | \delta, \gamma) | u^{(t)} \right\}$$

$$= \sum_{i \in A} d_i \bar{p}_{i}^{(t)} \log \left\{ \prod_{k=1}^{K} m_{ik} \right\} + \sum_{i \in A} d_i (1 - \bar{p}_{i}^{(t)}) \log \left\{ (1 - \pi) \prod_{k=1}^{K} u_{ik} \right\} ,$$

where

$$\bar{p}_{i}^{(t)} = E(\delta_i | \gamma_i; \bar{u}^{(t)})$$

$$= \frac{\pi \prod_{k=1}^{K} m_{ik}}{\pi \prod_{k=1}^{K} m_{ik} + (1 - \pi) \prod_{k=1}^{K} \bar{u}_{ik}^{(t)}}$$

and $(m_{ik}, \bar{u}_{ik}^{(t)}) = \sum_{d=1}^{D} Y_{ik}^{(d)} (m_{kd}, \bar{u}_{kd}^{(t)})$.

3. The M-step is to maximise the $Q$ over $u$ to update the parameters. The updating formula is

$$\bar{u}_{kd}^{(t+1)} = \frac{\sum_{i \in A} d_i (1 - \bar{p}_{i}^{(t)}) Y_{ik}^{(d)}}{\sum_{i \in A} d_i (1 - \bar{p}_{i}^{(t)})}.$$

4. Set $t = t + 1$ and go to Step 2. Continue until convergence.

Once $\hat{\delta}_i$ are computed, we may want to use, instead of (13),

$$\sum_{i \in A} w_i (1, \hat{\delta}_i, \hat{\delta}_i y_i) = \sum_{i \in U} (1, \hat{\delta}_i, \hat{\delta}_i y_i)$$

as the calibration equation. The calibration estimator using (17) is equivalent to

$$\hat{T}_{PID_2} = T_{b2} + (N - N_{b2}) \frac{\sum_{i \in A} d_i (1 - \hat{\delta}_i) y_i}{\sum_{i \in A} d_i (1 - \hat{\delta}_i)}.$$
where \((N_{b2}, T_{b2}) = \sum_{i \in U} \delta_i (1, y_i)\). However, unless \(\hat{\delta}_i = \delta_i\), we do not observe \(N_{b2}\) and \(T_{b2}\) and cannot compute \(T_{PDI2}\).

To overcome this difficulty, note that \(\hat{\rho}_i\) in (16) is a consistent estimator of \(E(\delta_i | z_i)\). Thus, as long as

\[
P(\delta = 1 | z, y) = P(\delta = 1 | z)
\]

holds, we can estimate \(N_{b2}\) and \(T_{b2}\) consistently by applying the standard propensity score method using \(\hat{\rho}_i\). That is, use

\[
\left(\hat{N}_{b2}, \hat{T}_{b2}\right) = \sum_{i \in U} \frac{\delta_i}{\hat{\rho}_i} \hat{\delta}_i (1, y_i) = \sum_{i \in B} \frac{\delta_i}{\hat{\rho}_i} (1, y_i)
\]

(19)

as a propensity score estimator of \((N_{b2}, T_{b2}) = \sum_{i \in U} \delta_i (1, y_i)\). Unlike Chen et al. (2020), the estimated propensity scores \(\hat{\rho}_i\) are fully nonparametric. Ignoring estimation errors in \(\hat{\rho}_i\), we have

\[
E_\delta\{\left(\hat{N}_{b2}, \hat{T}_{b2}\right)\} \approx E_\delta \left\{ \sum_{i \in U} \frac{\delta_i}{\hat{\rho}_i} \hat{\delta}_i (1, y_i) \right\} = \sum_{i \in U} \hat{\delta}_i (1, y_i) = (N_{b2}, T_{b2}),
\]

where \(E_\delta(\cdot)\) denotes the expectation with respect to \(\delta\) and the first equality holds because \(E(\delta_i | z_i, y_i) = p_i\). Thus, the resulting data integration estimator is

\[
\hat{T}_{PDI2} = \hat{T}_{b2} + (N - \hat{N}_{b2}) \frac{\sum_{i \in A} d_i (1 - \hat{\delta}_i) y_i}{\sum_{i \in A} d_i (1 - \hat{\delta}_i)}.\]

(20)

The data integration estimator in (20) can be viewed as a calibration estimator with calibration equation

\[
\sum_{i \in A} w_i (1, \hat{\delta}_i, \hat{\delta}_i y_i) = \sum_{i \in U} (1, \hat{\delta}_i \delta_i / \hat{\rho}_i, \hat{\delta}_i \delta_i y_i / \hat{\rho}_i) = \left( N, \sum_{i \in B} \hat{\delta}_i / \hat{\rho}_i, \sum_{i \in B} \hat{\delta}_i y_i / \hat{\rho}_i \right),\]

(21)

which requires computing \(\hat{\rho}_i\) and \(\hat{\delta}_i\) for every unit in sample \(B\). Condition (18) can be understood as the ignorability condition of the sampling mechanism for sample \(B\). This condition is not as strong as it might look at first. If \(y\) is categorical, one can always include \(y\) into \(z\) and apply the proposed classification method. In this case, condition (18) is always satisfied. For continuous \(y\), we may categorise \(y\) first and include it into \(z\). Refer to the simulation study in Section 7.2 for an example.

6 Handling Measurement Errors in Survey Data

We now consider the case the measurement errors exist in the survey data. For example, survey data are collected annually, and the big data are available monthly. In this case, if we are interested in estimating parameters on a monthly basis, we can treat the observed values in the latest year from the survey data as an inaccurate measurement for \(y_i\). Thus, we observe \((\delta_i, y_i^*)\) from sample \(A\) and observe \(y_i\) from sample \(B\). In this case, we can use the measurement error model (1) to obtain a design-model-based estimator of \(T = \sum_{i=1}^{N} y_i\).

To estimate \(T\) under measurement errors in sample \(A\) and selection bias in sample \(B\), we consider the following two-step approach:
1. [Step 1] Using the measurement model, estimate the parameters in $E(y_i \mid y_i^*) = m(y_i^*; \beta)$, and obtain mass imputation for sample $A$. That is, create $\hat{y}_i = m(y_i^*; \hat{\beta})$ for all elements in sample $A$. If the measurement error model is (1), then we can use $\hat{y}_i = \hat{\beta}_1^{-1}(y_i^* - \hat{\beta}_0)$, where $(\hat{\beta}_0, \hat{\beta}_1)$ is the estimated parameter from the elements in $A \cap B$.

2. [Step 2] Apply calibration estimation using $x_i = (1 - \delta_i, \delta_i, \delta_i y_i)^T$. That is, the final estimator is

$$\hat{T}_{\text{RegDI}} = \sum_{i \in A} w_i \hat{y}_i,$$

where $w_i$ minimises $Q(d, w)$ subject to the calibration equation $\sum_{i \in A} w_i x_i = \sum_{i \in U} x_i$.

In Step 1, the bias-corrected predictor of $y_i$ is obtained from model (1). In principle, because we observe $(y_i, y_i^*)$ among those with $\delta_i = 1$ in sample $A$, we can treat this sample, $A \cap B$, as the validation sample in the calibration study. If the mechanism for $\delta_i = 1$ depends on $y$ only, then the measurement error model (1) is non-informative in the sense of Pfeffermann et al. (1998). In this case, we can estimate model parameters in (1) consistently by the complete-case analysis. That is, we can use

$$\sum_{i \in A} d_i \delta_i (y_i^* - \beta_0 - \beta_1 y_i)(1, y_i) = (0, 0)$$

as an estimating equation for $(\beta_0, \beta_1)$.

For variance estimation of $\hat{T}_{\text{RegDI}}$ in (22), we can use, similarly to (12),

$$\hat{V}(\hat{T}_{\text{RegDI}}) = \sum_{i \in A} \sum_{j \in A} \pi_{ij} - \pi_i \pi_j \hat{e}_i \hat{e}_j / \pi_i \pi_j,$$

(23)

where $\hat{e}_i = \hat{y}_i - x_i^T \hat{B}$ and $\hat{B} = (\sum_{i \in A} d_i x_i x_i^T)^{-1} \sum_{i \in A} d_i x_i \hat{y}_i$. Thus, we can safely ignore the effect of uncertainty of $\hat{B}$ in $\hat{y}_i = m(y_i^*; \hat{\beta})$ for variance estimation. See Appendix B1 for a sketched justification.

7 Simulation Study

7.1 Simulation study 1

In the first simulation, continuous $Y$ variable is considered from the following model:

$$y_i = 3 + 0.7(x_i - 2) + e_i,$$

where $x_i \sim N(2, 1)$, $e_i \sim N(0, 0.51)$ and $e_i$ is independent of $x_i$. We generate a finite population of size $N = 1000000$ from this model. Also, we generate

$$y_i^* = 2 + 0.9(y_i - 3) + u_i,$$

where $u_i \sim N(0, 0.5^2)$ and $u_i$ is independent of $y_i$.

In this simulation, we repeatedly obtain two samples, denoted by $A$ and $B$, by simple random sampling of size $n = 1000$ and by an unequal probability sampling of size $N_B = 500000$, respectively. In selecting sample $B$, we create two strata, where stratum 1 consists of elements with $x_i \leq 2$, and stratum 2 consists of those with $x_i > 2$. Within each stratum, we select $n_h$ elements by simple random sampling independently, where $n_1 = 300000$ and $n_2 = 200000$. 

Under this sampling mechanism, the sample mean of $B$ is smaller than the population mean. We assume that the stratum information is not available at the time of data analysis.

We consider the following three scenarios:

1. [Scenario I] No measurement errors in both samples. Thus, we observe $y_i$ in both samples.
2. [Scenario II] Measurement errors in sample $B$. Thus, we observe $y_i$ in sample $A$ and $y_i^{*}$ in sample $B$.
3. [Scenario III] Measurement errors in sample $A$. Thus, we observe $y_i^{*}$ in sample $A$ and $y_i$ in sample $B$.

In addition, assume that we observe the matching indicator $ı_i$ in sample $A$. If $ı_i = 1$ in sample $A$, we observe $(y_i, y_i^{*})$.

We consider the following four estimators for the population mean of $Y$:

1. Mean $A$. Mean of sample $A$ observations.
2. Mean $B$. Mean of sample $B$ observations.
3. Post-stratified data integration estimator of the form (3).
4. Regression data integration estimator of the form (10).

In Scenario II, the post-stratified data integration estimator is computed using

$$
\hat{\theta}_{PDI}^{II} = \frac{1}{N} \left\{ \sum_{i=1}^{N} \delta_i y_i^{*} + (N - N_b) \frac{\sum_{ı_i \in A} d_i (1 - \delta_i) y_i}{\sum_{ı_i \in A} d_i (1 - \delta_i)} \right\}.
$$

In Scenario III, the post-stratified data integration estimator is computed using

$$
\hat{\theta}_{PDI}^{III} = \frac{1}{N} \left\{ \sum_{i=1}^{N} \delta_i y_i + (N - N_b) \frac{\sum_{ı_i \in A} d_i (1 - \delta_i) y_i^{*}}{\sum_{ı_i \in A} d_i (1 - \delta_i)} \right\},
$$

and the regression data integration estimator is computed from the two-step approach in (22).

Table 2 presents the result of the simulation study based on 1 000 Monte Carlo samples. From Table 2, mean $A$ estimator is unbiased except for scenario III, where systematic measurement errors exist in sample $A$. Mean $B$ estimator is always biased due to the selection bias in sample $B$. The bias is the largest in absolute values for scenario II, where measurement errors exist in addition to the selection bias. Variance of mean $B$ estimator is the smallest because of the large sample size of sample $B$ ($N_B = 500000$). The post-stratified data integration estimator is unbiased in scenario I, which is consistent with our theory in Section 3. The variance of the post-stratified estimator is about half of the variance of the mean $A$ estimator because $N_B / N = 0.5$. If the rate $W_B = N_B / N$ is larger, then the variance estimator post-stratified estimator will be smaller as equation (4) may suggest. However, in scenario II, the post-stratified data integration estimator is biased because $T_b = \sum_{i=1}^{N} \delta_i y_i$ is estimated without correcting for the measurement errors. In scenario III, it is biased because $T_c = \sum_{i=1}^{N} (1 - \delta_i) y_i$ is estimated from sample $A$ without correcting for the measurement errors. The regression data integration estimator is unbiased for all scenarios. It is the same as the post-stratified data integration estimator under scenario I, as discussed in (11).

In addition, we also compute variance estimators of the regression data integration estimator using formula (23). For example, in scenario II, we use

$$
\hat{\epsilon}_i = \begin{cases} 
    y_i - (\hat{b}_0 + \hat{b}_1 y_i^{*}) & \text{if } \delta_i = 1 \\
    y_i - \bar{y} & \text{if } \delta_i = 0,
\end{cases}
$$
Table 2. Results of the four estimators for simulation study 1 based on a Monte Carlo sample of size 1 000

| Scenario | Estimator | Bias  | SE    | RMSE  |
|----------|-----------|-------|-------|-------|
| I        | Mean A    | 0.00  | 0.031 | 0.031 |
|          | Mean B    | -0.11 | 0.001 | 0.113 |
|          | PDI       | 0.00  | 0.022 | 0.022 |
|          | RegDI     | 0.00  | 0.022 | 0.022 |
| II       | Mean A    | 0.00  | 0.031 | 0.031 |
|          | Mean B    | -1.10 | 0.001 | 1.101 |
|          | PDI       | -0.49 | 0.022 | 0.495 |
|          | RegDI     | 0.00  | 0.024 | 0.024 |
| III      | Mean A    | 1.00  | 0.033 | 1.001 |
|          | Mean B    | -0.11 | 0.001 | 0.113 |
|          | PDI       | -0.51 | 0.023 | 0.507 |
|          | RegDI     | 0.00  | 0.028 | 0.028 |

PDI, post-stratified data integration estimator; RegDI, regression data integration estimator; RMSE, root mean squared error; SE, standard error.

where \((\hat{b}_0, \hat{b}_1)\) is the solution to \(\sum_{i \in A} d_i \delta_i (y_i - b_0 - b_1 y_i^*) (1, y_i^*) = (0, 0)\). Based on 1 000 Monte Carlo samples, we compute the relative biases of the variance estimators. The relative biases are -0.0037, 0.028 and 0.019 for scenarios I, II and III, respectively. Thus, we conclude that the proposed variance estimators are nearly unbiased.

7.2 Simulation study 2

In the second simulation study, we study the performance of the data integration estimator using matching variables. In the simulation study, we first generate a finite population with \((z_{1i}, z_{2i}, \delta_i, y_i)\) as follows. First generate

\[ z_{1i} \sim \text{Unif}\{1, \ldots, 20\} \]

independently. Given \(z_{1i}\), we generate \(\delta_i\) from Bernoulli distribution with the probability

\[ P(\delta_i = 1 | z_{1i}) = \begin{cases} c & \text{if} \ z_{i1} \leq 10 \\ 2c & \text{if} \ z_{i1} > 10 \end{cases} \]

where \(c\) is chosen such that the sum of the probabilities over the finite population is equal to \(N_B\). We set \(N = 10000\) and \(N_B = 5000\) in this simulation. We also generate

\[ y_i = \begin{cases} 4 + 0.5(z_{i2} + e_i) & \text{if} \ z_{i1} \leq 10 \\ 6 + 0.3(z_{i2} + e_i) & \text{if} \ z_{i1} > 10 \end{cases} \]

where \(z_{2i} \sim \text{Unif}\{1, \ldots, 10\}, e_i \sim \text{Unif}(0, 1)\) and \(z_{2i}\) and \(e_i\) are mutually independent. Thus, we can treat \(z_{2i}\) as a categorisation of continuous variable \(y_i\).

From the finite population, we select sample \(A\) by simple random sampling of size \(n_A\). Two values of \(n_A = |A|\) are considered: \(n_A = 1000\) versus \(n_A = 2000\). From sample \(A\), we observe \((z_{1i}, z_{2i}, y_i)\) but not \(\delta_i\). Thus, we apply the semi-supervised classification method using \((z_{1i}, z_{2i})\) as the matching variable. Note that as \(z_{2i}\) is included in the matching to satisfy the ignorability condition (18) approximately.
From each sample, we consider five estimators of $\overline{Y} = N^{-1} \sum_{i=1}^{N} y_i$.

1. Mean $A$. Mean of sample $A$ observations.
2. Mean $B$. Mean of sample $B$ observations.
3. Naive data integration (DI) estimator: Treat $\hat{\delta}_i$ as if accurate and apply the data integration estimator using $\hat{\delta}_i$ to get
   \[ \hat{T}_{PDI} = T_B + (N - N_b) \frac{\sum_{i \in A} d_i (1 - \hat{\delta}_i)}{\sum_{i \in A} d_i (1 - \hat{\delta}_i)} y_i. \]
4. The proposed data integration estimator:
   \[ \hat{T}_{PDI2} = \hat{T}_{b_2} + (N - \hat{N}_{b_2}) \frac{\sum_{i \in A} d_i (1 - \hat{\delta}_i) y_i}{\sum_{i \in A} d_i (1 - \hat{\delta}_i)}, \]
   where $\hat{T}_{b_2}$ and $\hat{N}_{b_2}$ are defined in (19).
5. The original data integration estimator using the true indicator function $\delta_i$. This estimator is computed as a benchmark for comparison.

Table 3 presents the performance of the five estimators. Mean $B$ is seriously biased due to its selection bias. Naive DI estimator is also biased seriously due to the misclassification errors in $\hat{\delta}_i$. The proposed DI estimator is unbiased and is more efficient than the sample mean (Mean $A$) of the sample $A$, although the efficiency gain is not as significant as in the original DI estimator. The efficiency gain will increase with $W_b = N_b/N$.

8 An Application in Official Statistics

We now consider an application of the proposed method to a real data problem using 2015–2016 Australian Agricultural Census as the big data, which has 85% response rate. In addition, we use the 2014–2015 Rural Environment and Agricultural Commodities Survey (REACS) as the probability sample (sample $A$) for calibration. Our interest is to combine the Agricultural Census data with the REACS data to estimate the total area of holdings (AOH), the total number of dairy cattle (DAIRY), the number of beef cattle (BEEF), and the number

| $n_a$ | Estimator | Bias  | SE   | RMSE  |
|-------|-----------|-------|------|-------|
| 1000  | Mean $A$  | 0.00  | 0.037| 0.037 |
|       | Mean $B$  | -0.14 | 0.011| 0.135 |
|       | Naive DI  | 0.12  | 0.036| 0.130 |
|       | Proposed DI | 0.00 | 0.035| 0.035 |
|       | Original DI | 0.00 | 0.024| 0.024 |
| 2000  | Mean $A$  | 0.00  | 0.025| 0.025 |
|       | Mean $B$  | -0.14 | 0.011| 0.135 |
|       | Naive DI  | 0.14  | 0.015| 0.136 |
|       | Proposed DI | 0.00 | 0.023| 0.023 |
|       | Original DI | 0.00 | 0.016| 0.016 |

RMSE, root mean squared error; SE, standard error.
of tonnes of wheat for grain or seed produced (WHEET) for 2015–2016. Thus, we observe \( y_i \) from the Agricultural Census data and observe \( y_i^* \) from REACS.

To apply the proposed method, define \( \delta_i = 1 \) if unit \( i \) participated at the census and \( \delta_i = 0 \) otherwise. Thus, in REACS sample, we observe \( y_i \) in addition to \( y_i^* \) for \( \delta_i = 1 \). Using the matched sample in sample \( A \), we can fit a measurement error model

\[
y_i^* = \beta_0 + \beta_1 y_i + u_i
\]

and obtain \( \hat{y}_i = \hat{\beta}_1^{-1}(y_i^* - \hat{\beta}_0) \) for all \( i \in A \). Here, \( y_i \) is the true value of the study variable from 2015–2016 census and \( y_i^* \) is its proxy value obtained from 2014–2015 REAC survey data.

For each parameter, we compute the following three estimators:

1. Survey estimate (from REACS sample): \( \hat{\theta}_H = \sum_{i \in A} w_i \hat{y}_i \)
2. Big data estimate (from census): \( \hat{\theta}_B = \sum_{i \in B} y_i \)
3. The two-step data integration estimate using calibration weighting:

\[
\hat{\theta}_{DI} = \sum_{i \in A} w_{i, cal} \hat{y}_i
\]

Figure 1. Three estimates for area of holding (AOH) for 2015–2016 by states. ACT, Australian Capital Territory; NSW, New South Wales; NT, Northern Territory; QLD, Queensland; SA, South Australia; TAS, Tasmania; VIC, Victoria; WA, Western Australia.
where $w_{i, \text{cal}}$ satisfies $\sum_{i \in A} w_{i, \text{cal}} (1 - \delta_i, \delta_i, \delta_i x_i) = \sum_{i \in U} (1 - \delta_i, \delta_i, \delta_i x_i)$ and $x_i$ includes major study variables.

The estimates are compared with the official numbers of the Australian Bureau of Statistics, which is obtained by applying imputation for item non-response in the census. Figures 1 and 2 present the estimation results for AOH and DAIRY, respectively, by eight states in Australia. We do not report the results for other commodities to save space. The confidence intervals are constructed using the asymptotic normality with 90% nominal coverage rates. The results in Figures 1 and 2 can be summarised as follows: (1) the big data estimates show serious negative biases due to the under-coverage of the big data (non-response in the census); (2) the proposed data integration estimator shows narrower confidence intervals than the survey estimate; (3) the effect of calibration weighting is reduced because of the measurement errors in sample $A$ observations. Overall, the confidence intervals obtained from the proposed data integration estimators cover the official Australian Bureau of Statistics estimates.

9 Discussion

The proposed data integration methods feature an independent probability sample for estimating the missing data stratum of the finite population, which can correct for the
under-coverage bias of the big data sample. By treating big data as an incomplete sampling frame for the finite population, we can apply the calibration weighting method. In addition, these methods are extended in this paper to handle measurement errors in either the big data source or the probability sample source. Also, a fully non-parametric approach to propensity score estimation for big data sample participation is developed using a new semi-supervised classification method.

In practice, our methods are useful provided the following conditions apply:

1. Existence of a probability sample \(A\), which also measures \(y\) or provides a proxy measure \(\hat{y}_i\). While the coincidental existence of such a sample is rare, where one, for example, a national statistical offices, determines the benefits in using big data for inference outweighs the costs, one can design, develop and implement such a random sample to collect the measure of interest. Where this occurs, the population count of the sample units, \(N\), is by definition known.

2. The calibration method is useful only if the coverage of \(B\) is substantial, which is not an unreasonable assumption if \(B\) is a big data set. Also, when \(B\) is big, it can be assumed that \(A \cap B\) is not empty for measurement error adjustment, where warranted.

The nonparametric propensity scores obtained from the semi-supervised classification method can be used to correct for the coverage bias in big data sample. How to make valid statistical inference, including variance estimation, under the nonparametric propensity score adjustment is not pursued here and will be covered elsewhere. Extensions to small domain estimation (Rao & Molina, 2015) and analytic inferences using big data will also be future research topics.

Appendix A: Justification for (23)

Let \(\theta = N^{-1}T\), the finite population mean of \(Y\), be the parameter of interest. We first consider variance estimation of the mass imputation estimator of the form

\[
\hat{\theta}_{DI} = \frac{1}{N} \sum_{i \in A} d_i \hat{y}_i,
\]

where \(\hat{y}_i\) is a predictor of \(y_i\) using \(y_i^*\). We use \(\hat{y}_i = \hat{m}^{-1}(y_i^*)\) where \(\hat{m}(y_i) = m(y_i; \hat{\beta}) = E(y_i^* | y_i; \hat{\beta})\). The estimating equation for \(\hat{\beta}\) can be written as

\[
\hat{U}_\beta(\beta) = N^{-1} \sum_{i \in A} d_i \delta_i \{y_i^* - m(y_i; \beta)\} h(y_i; \beta) = 0 \quad (A1)
\]

for some \(h(y; \beta)\) such that \(\hat{U}_\beta(\beta)\) is linearly independent. Writing \(\hat{\theta}_{DI} = \hat{\theta}_{DI}(\hat{\beta})\), we can use Taylor linearisation to estimate the variance of \(\hat{\theta}_{DI}\). Using the standard argument (Kim & Rao, 2009), we can obtain

\[
\hat{\theta}_{DI} = \hat{\theta}_{DI}(\beta_N) - E \left\{ \frac{\partial}{\partial \beta} \hat{\theta}_{DI}(\beta_N) \right\} \left[ E \left\{ \frac{\partial}{\partial \beta} \hat{U}_\beta(\beta_N) \right\} \right]^{-1} \hat{U}_\beta(\beta_N) + o_p(n^{-1/2}), \quad (A2)
\]

where \(\beta_N\) is the probability limit of \(\hat{\beta}\).

After some algebra, we can express (A2) as

\[
\hat{\theta}_{DI} = \frac{1}{N} \sum_{i \in A} d_i \{q_i + \delta_i (y_i^* - m(y_i; \beta)) \kappa^T h_i\} + o_p(n^{-1/2}). \quad (A3)
\]
where \( q_i = q_i(\beta_N) \) is the solution to \( y_i^* = m(q_i; \beta_N), h_i = h(y_i; \beta_N) \) and \( \kappa \) satisfies

\[
\sum_{i=1}^N \delta_i \hat{m}_i h_i^T \kappa = \sum_{i=1}^N \hat{q}_i
\]

with \( \hat{m}_i = \partial m(y_i; \beta) / \partial \beta \) and \( \hat{q}_i = \partial q_i(\beta) / \partial \beta \). Using (A3), we can express

\[
\hat{\theta}_{DI} - \theta = (\bar{q}_N - \theta) + (\bar{u}_{HT} - \bar{u}_N) + o_p(n^{-1/2}),
\]

(A4)

where \( \bar{q}_N = N^{-1} \sum_{i=1}^N q_i, u_i = q_i + \delta_i \left( y_i^* - m(y_i; \beta_N) \right) \left( \kappa^T h_i \right), \bar{u}_{HT} = N^{-1} \sum_{i \in A} d_i u_i \) and \( \bar{u}_N = N^{-1} \sum_{i=1}^N u_i \).

From (A4), we can obtain

\[
\text{Var}\left( \hat{\theta}_{DI} - \theta \right) = \text{Var}(\bar{q}_N - \theta) + \text{Var}(\bar{u}_{HT} - \bar{u}_N) = V_1 + V_2.
\]

(A5)

The first term is of order \( O(N^{-1}) \), and the second term is \( O(n^{-1}) \). The first term is negligible if \( n/N = o(1) \). To estimate the second term of (A5), we can use

\[
\hat{V}_2 = \frac{1}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{u}_i \hat{u}_j}{\pi_i \pi_j},
\]

(A6)

where

\[
\hat{u}_i = \hat{q}_i + \delta_i \left( y_i^* - m(y_i; \beta) \right) \left( \hat{\kappa}^T \hat{h}_i \right)
\]

and

\[
\hat{\kappa} = \left\{ \sum_{i \in A} d_i \delta_i \hat{m}_i h_i^T \right\}^{-1} \sum_{i \in A} d_i \hat{q}_i.
\]

Next, we consider variance estimation for the calibration estimator

\[
\hat{\theta}_{RegDI} = \frac{1}{N} \sum_{i \in A} w_i \hat{y}_i,
\]

where \( w_i \) are the calibration weights satisfying the calibration equation \( \sum_{i \in A} w_i x_i = \sum_{i=1}^N x_i \).

In this case, the linearisation in (A3) changes to

\[
\hat{\theta}_{RegDI} = \frac{1}{N} \sum_{i \in A} d_i \left\{ e_i + \delta_i \left( y_i^* - m(y_i; \beta_N) \right) \kappa_2^T h_i \right\} + o_p(n^{-1/2}),
\]

(A7)

where \( e_i = q_i - x_i^T \hat{B}, B = \left( \sum_{i=1}^N x_i x_i^T \right)^{-1} \sum_{i=1}^N x_i q_i \) and \( \kappa_2 \) satisfies

\[
\sum_{i=1}^N \delta_i \hat{m}_i h_i^T \kappa_2 = \sum_{i=1}^N e_i.
\]

Because \( x_i \) includes an intercept term, we have \( \sum_{i=1}^N e_i = 0 \), which implies \( \kappa_2 = 0 \). Therefore, for variance estimation of \( \hat{\theta}_{RegDI} \), we can use (23), where \( \hat{e}_i = \hat{q}_i - x_i^T \hat{B} \) and \( \hat{B} = \left( \sum_{i \in A} d_i x_i x_i^T \right)^{-1} \sum_{i \in A} d_i x_i \hat{q}_i \).
Acknowledgements

The authors are grateful to two anonymous referees and the co-editor for the very constructive comments. The research of the first author was partially supported by a grant from US National Science Foundation (MMS-1733572).

References

Baeza-Yates, R. (2018). Bias on the web. *Commun. ACM*, 61, 54–61.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Soc. Sci. Comput. Rev.*, 34, 59–77.

Brodie, M.A., Pliner, E.M., Ho, A., Li, K., Chen, Z., Gandevia, S.C. & Lord, S.R. (2018). Big data vs accurate data in health research: large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Med. Hypotheses*, 119, 32–36.

Chambers, R.L. & Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press: London.

Chen, Y., Li, P. & Wu, C. (2020). Doubly robust inference with non-probability survey samples. *J. Am. Stat. Assoc.* Accepted (available at https://doi.org/10.1080/01621459.2019.1677241).

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, 87, 376–382.

Elliott, M. & Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249–264.

Fellegi, I.P. & Holt, D. (1976). A systematic approach to automatic data editing. *J. Am. Stat. Assoc.*, 71, 17–35.

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

Groves, R. (2006). Non response rates and nonresponse bias in household surveys. *Public Opin. Q.*, 72, 167–189.

Groves, R. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opin. Q.*, 72, 167–189.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data. *J. R. Stat. Soc. Ser. A*, 181, 203–206.

Hartley, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social statistics Section*. American Statistical Association. 1–24.

Kaplan, R.M., Chambers, D.A. & Glasgow, R.E. (2014). Big data and large sample size: a cautionary note on the potential of bias. *American Society for Clinical Pharmacology and Therapeutics*, 7, 342–346.

Keiding, N. & Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussions). *J. R. Stat. Soc. Ser. A*, 179, 1–28.

Kim, J.K. & Park, M. (2010). Calibration estimation in survey sampling. *Int. Stat. Rev.*, 78, 21–39.

Kim, J.K. & Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917–932.

Lohr, S.L. (2011). Alternative survey sample designs: sampling with multiple overlapping frames. *Surv. Methodol.*, 37, 197–213.

Lohr, S. & Raghunathan, T. (2017). Combining survey data with other data sources. *Stat. Sci.*, 32, 293–312.

Meng, X.L. (2018). Statistical paradoxes and paradoxes in big data (i): law of large populations, big data paradox, and 2016 US Presidential Election. *Ann. Appl. Stat.*, 12, 685–726.

Olteanu, A., Castillo, C., Daiz, F. & Kiciman, E. (2019). Social data: biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 1–33.

Pfeffermann, D., Krieger, A.M. & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Stat. Sin.*, 8, 1087–1114.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *J. Surv. Stat. Methodol.*, 3, 425–483.

Rao, J.N.K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*. Accepted (available at https://doi.org/10.1007/s13571-020-00227-w).

Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation*, 2. Hoboken, New Jersey: Wiley.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Method Section*, American Statistical Association. 1–26.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581–590.

Särndal, C.E., Cassel, C.M. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag: New York.

Sax, L.I., Gilmer, S. & Bryant, A. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Res. High Educ.*, 44, 409–432.

Skinner, C.J. & Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *J. Am. Stat. Assoc.*, 91, 349–356.

Tam, S.-M. & Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *Int. Stat. Rev.*, 83, 436–448.
Tam, S.-M. & Kim, J.K. (2018). Big data, selection bias and ethics—an official statistician's perspective. *Stat. J. IAOS*, **34**, 577–588.

Valliant, R. & Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Social. Methods Res.*, **40**, 105–137.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Stat. Neerl.*, **66**, 41–63.

[Received August 2019, Revised June 2020, Accepted October 2020]