Biometrika (2012), **99**, 1, *pp.* 85–100

Printed in Great Britain

doi: 10.1093/biomet/asr063
Advance Access publication 8 December 2011

Combining data from two independent surveys: a model-assisted approach

By JAE KWANG KIM

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
jkim@iastate.edu

and J. N. K. RAO

School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6
jrao@math.carleton.ca

Summary

Combining information from two or more independent surveys is a problem frequently encountered in survey sampling. We consider the case of two independent surveys, where a large sample from survey 1 collects only auxiliary information and a much smaller sample from survey 2 provides information on both the variables of interest and the auxiliary variables. We propose a model-assisted projection method of estimation based on a working model, but the reference distribution is design-based. We generate synthetic or proxy values of a variable of interest by first fitting the working model, relating the variable of interest to the auxiliary variables, to the data from survey 2 and then predicting the variable of interest associated with the auxiliary variables observed in survey 1. The projection estimator of a total is simply obtained from the survey 1 weights and associated synthetic values. We identify the conditions for the projection estimator to be asymptotically unbiased. Domain estimation using the projection method is also considered. Replication variance estimators are obtained by augmenting the synthetic data file for survey 1 with additional synthetic columns associated with the columns of replicate weights. Results from a simulation study are presented.

Some key words: Double sampling; Mass imputation; Synthetic data; Two-phase sampling.

1. Introduction

Traditional two-phase sampling, or double sampling, is based on selecting a large first phase sample and then collecting information on inexpensive auxiliary variables x, followed by a much smaller second phase subsample that collects information on both x and relatively expensive variables of interest y. This two-phase design, also called nested two-phase sampling, has a long history and is widely used in survey sampling (Rao, 1973; Cochran, 1977; Hidiroglou & Särndal, 1998), biostatistics (Reilly & Pepe, 1995; Clayton et al., 1998) and other applications. It is a cost-effective method for obtaining efficient estimators of finite population totals or means and other parameters of interest by utilizing the first phase data on x for stratification, ratio or regression estimation and other uses.

Nonnested two-phase sampling differs from traditional double sampling, as it involves two independent surveys from the same target population consisting of N elements. A large sample A_1 from survey 1 collects information only on x and a much smaller sample A_2 from survey 2

provides information on both y and x. It is assumed that the observed variable x is comparable in the two surveys. The two samples A_1 and A_2 may be selected from possibly different frames; for example a stratified random sample A_1 from a list frame and a stratified two-stage sample A_2 from an area frame. Although nonnested double sampling designs were studied long ago (Bose, 1943), such designs have received considerable attention recently in the context of combining data from two or more independent surveys. For example, Statistics Canada used two independent samples drawn from two different frames representing the same population for the Canadian Survey of Employment, Payrolls and Hours. A large sample A_1 was selected from a Canada Customs and Revenue Agency administrative data file and auxiliary variables x, which include the number of employees and the total amount of payroll, were collected. A much smaller sample A_2 was independently selected from the Statistics Canada Business Register and variables of interest y, number of hours worked by employees and summarized earnings, were collected in addition to the variables x (Hidiroglou, 2001).

Renssen & Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004, 2010), Wu (2004) and Ybarra & Lohr (2008) considered the problem of combining data from two independent surveys to estimate totals at the population and domain levels; our primary interest is in creating a single synthetic dataset of proxy values $\tilde{y}_i$ for the unobserved y_i in survey 1 and then using the proxy data together with the associated survey weights, w_{i1}, of survey 1 to produce projection estimators of the population total of y and domain totals of y. The proxy values $\tilde{y}_i$ $(i \in A_1)$ are generated by first fitting a working model relating y to x to the data $\{(y_i, x_i) : i \in A_2\}$ from survey 2 and then predicting y_i associated with x_i $(i \in A_1)$. This facilitates the use of only the synthetic data and associated survey weights reported in the survey 1 data file to produce estimates of population and domain totals of variables y not observed in survey 1. Schenker & Raghunathan (2007) reported several applications of the synthetic data approach, using a model-based method to estimate totals and other parameters associated with variables not observed in survey 1 but observed in a much smaller survey 2. In one application, survey 2 observed both self-reported health measurements, x_i, and clinical measurements from physical examinations, y_i, for a small sample A_2 of individuals, and the much larger survey 1 had only self-reported measurements, x_i. Only the imputed, or synthetic, data from survey 1 and associated survey weights are released to the public (Reiter, 2008). Reporting only synthetic values, $\tilde{y}_i$, on the data file might also minimize disclosure risk in some survey situations. Unlike the model-based approach of Schenker & Raghunathan (2007), our approach is robust against failure of the working model used to generate the synthetic values.

To describe the setup for two independent surveys, let w_{i1} be the sampling weight associated with unit i for survey 1 such that, if y_i were available in survey 1, $\hat{Y}_1 = \sum_{i \in A_1} w_{i1} y_i$ would be an unbiased estimator of the population total $Y = \sum_{i=1}^{N} y_i$, where y is a variable of interest. The estimator $\hat{Y}_1$ cannot be implemented from survey 1, unlike the following estimator of Y based on the synthetic values $\tilde{y}_i = m(x_i; \hat{\beta})$ reported in the survey 1 data file,

$$\hat{Y}_p = \sum_{i \in A_1} w_{i1} \tilde{y}_i, \tag{1}$$

for some known function $m(x_i; \hat{\beta}) \equiv \hat{m}_i$ of $\hat{\beta}$, where the estimator $\hat{\beta}$ is obtained from survey 2 using the data $\{(y_i, x_i) : i \in A_2\}$ and it is used in constructing the synthetic values $\tilde{y}_i$. In the case of a binary variable y, the data file in survey 1 will report binary synthetic values $\tilde{y}_i = 1$ and 0 with associated fractions $\hat{m}_i$ and $1 - \hat{m}_i$ for $i \in A_1$, and the projection estimator of Y, based on the reported synthetic values in the survey 1 data file, would be identical to (1).

The asymptotic design bias of $\hat{Y}_p$ with respect to both surveys 1 and 2 is

$$\text{Bias}(\hat{Y}_p) = E(\hat{Y}_p \mid \mathcal{F}_N) - Y \cong \sum_{i=1}^{N} \{m(x_i; \beta_0) - y_i\},$$

where the expectation $E(\cdot \mid \mathcal{F}_N)$, conditional on the finite population, refers to the design expectation with respect to both surveys 1 and 2 and β_0 is the probability limit of $\hat{\beta}$ with respect to the survey 2 design. Using the sample data from survey 2, we can estimate this bias by $\sum_{i \in A_2} w_{i2}(\tilde{y}_i - y_i)$, where w_{i2} is the sampling weight associated with unit i for survey 2 such that $\hat{Y}_2 = \sum_{i \in A_2} w_{i2} y_i$ is an unbiased estimator of the total Y. Thus, the bias-corrected estimator

$$\hat{Y}_{p,\text{bc}} = \sum_{i \in A_2} w_{i2} y_i + \sum_{i \in A_1} w_{i1} \tilde{y}_i - \sum_{i \in A_2} w_{i2} \tilde{y}_i$$

is asymptotically unbiased for Y. The estimator $\hat{Y}_{p,\text{bc}}$ in general depends on the information from both surveys, but it reduces to (1), if $\hat{\beta}$, estimated from survey 2, satisfies

$$\sum_{i \in A_2} w_{i2} \{y_i - m(x_i; \hat{\beta})\} = 0. \tag{2}$$

Thus, $\hat{Y}_p$, based only on the synthetic values $\tilde{y}_i$, is asymptotically design unbiased if (2) holds. The estimator $\hat{Y}_p$ is called a projection estimator, or synthetic estimator, because $\tilde{y}_i = m(x_i; \hat{\beta})$ can be viewed as a projection of y_i using the auxiliary variable x_i or as a synthetic value of y_i. The projection estimator (1) is derived from a working model $E(y_i \mid x_i) = m(x_i; \beta)$ but the results do not depend on the validity of the working model, although this affects the efficiency of estimators. To make the projection estimator asymptotically unbiased, we add a restriction, such as (2), when estimating the parameter β from survey 2. Projection estimators, using projected values instead of true values, have been widely used either implicitly or explicitly in survey sampling. The synthetic values, $\tilde{y}_i$, in the projection estimator are often referred to as imputed values. An advantage of the projection estimator is that a single weight, w_{i1}, is used across variables y_i even though different working models, depending on the choice of x_i may be used to generate the corresponding synthetic values $\tilde{y}_i$.

In spite of their wide use, projection estimators have not been fully investigated in the literature. An advantage of the projection approach is that it enables us to construct a data file of synthetic values associated with survey 1 which can be used for valid estimation of a total Y using only survey 1 weights, w_{i1}, without requiring access to survey 2 data $\{(y_i, x_i); i \in A_2\}$ as long as it satisfies (2). Furthermore, as can be seen in § 3, synthetic data can provide a useful tool for efficient domain estimation when the size of the survey 1 sample is much larger than the size of the survey 2 sample.

2. Projection estimator

Suppose that the working model is $E(y_i \mid x_i) = m(x_i, \beta) \equiv m_i$, that $\text{var}(y_i \mid x_i) = \sigma^2 a(m_i)$ for some known function $a(m_i)$ and that $\text{cov}(y_i, y_j \mid x_i, x_j) = 0$ for $i \neq j$. Then, appealing to estimation function theory (Godambe & Thompson, 1986), $\hat{\beta}$ is obtained as a solution to

$$\sum_{i \in A_2} w_{i2} \{(\partial m_i / \partial \beta)/a(m_i)\}(y_i - m_i) \equiv \sum_{i \in A_2} w_{i2}(y_i - m_i) h_i = 0. \tag{3}$$

To satisfy (2), we assume that the first element of h_i is equal to unity. For a continuous variable y and linear regression with $m(x_i;\beta)=x_i'\beta$ and $a(m_i)=1$, we have $h(x_i)=x_i$ and (2) is satisfied when the first element of x_i is equal to one. Similarly, for a binary variable y and logistic regression with $\text{logit}\{m(x_i;\beta)\}=x_i'\beta$ and $a(m_i)=m_i(1-m_i)$, we have $h(x_i)=x_i$ and (2) is satisfied when the first element of x_i is equal to one. For the commonly used ratio working model $E(y_i \mid x_i)=\beta x_i$ and $\text{var}(y_i \mid x_i)=\sigma^2 x_i$, we have $h_i=1$ and (2) is satisfied. We assume that the solution $\hat{\beta}$ to (3) is unique with probability one, to avoid unnecessary technical details.

To study the asymptotic properties of the projection estimator, we assume a sequence of finite populations and samples, as defined in Isaki & Fuller (1982), with bounded fourth moments of (x_i, y_i). We assume that

$$N^{-1}\sum_{i\in A_1} w_{i1}x_i x_i' - N^{-1}\sum_{i=1}^{N} x_i x_i' = O_p(n_1^{-1/2}) \tag{4}$$

and

$$N^{-1}\sum_{i\in A_2} w_{i2}x_i (x_i', y_i) - N^{-1}\sum_{i=1}^{N} x_i (x_i', y_i) = O_p(n_2^{-1/2}), \tag{5}$$

where n_1 is the size of A_1 and n_2 is the size of A_2.

Theorem 1 presents some asymptotic properties of the projection estimator (1) with $\hat{\beta}$ obtained from (3).

THEOREM 1. *Assume a sequence of finite populations and two independent samples with bounded fourth moments such that* (4) *and* (5) *hold. Assume further that:*

Condition 1. *under survey* 2,

$$\hat{\beta}-\beta_0 = o_p(1); \tag{6}$$

Condition 2. *for each* i, $m(x_i,\beta)$ *is differentiable with continuous partial derivatives* $\dot{m}(x_i;\beta)=\partial m(x_i;\beta)/\partial\beta$ *with respect to* β *in a compact set* $\mathcal{B}$ *containing* β_0*;*

Condition 3. *under survey* 1,

$$N^{-1}\left\{\sum_{i\in A_1} w_{i1}h(x_i)\dot{m}(x_i;\beta)' - \sum_{i=1}^{N} h(x_i)\dot{m}(x_i;\beta)'\right\} = O_p(n_1^{-1/2})$$

and, under survey 2,

$$N^{-1}\left\{\sum_{i\in A_2} w_{i2}h(x_i)\dot{m}(x_i;\beta)' - \sum_{i=1}^{N} h(x_i)\dot{m}(x_i;\beta)'\right\} = O_p(n_2^{-1/2})$$

uniformly in $\beta\in\mathcal{B}$. *Then,*

$$N^{-1}n_2^{1/2}(\hat{Y}_p-\tilde{Y}) = o_p(1), \tag{7}$$

for the projection estimator $\hat{Y}_p$ *satisfying* (2)*, where* $o_p(1)$ *refers to both survey designs, and*

$$\tilde{Y}=\hat{P}_1+\hat{Q}_2, \tag{8}$$

where

$$\hat{P}_1 = \sum_{i \in A_1} w_{i1} m(x_i, \beta_0), \quad \hat{Q}_2 = \sum_{i \in A_2} w_{i2}\{y_i - m(x_i, \beta_0)\},$$

and the subscripts 1 *and* 2 *in* (8) *denote that* $\hat{P}_1$ *is a quantity from survey* 1 *and* $\hat{Q}_2$ *is a quantity from survey* 2.

Proof. See the Appendix. □

The asymptotic result in Theorem 1 shows that $\text{var}(\hat{Y}_p) \cong \text{var}(\tilde{Y})$ for large n_2. Furthermore, since the two surveys are independent, ignoring lower order terms, the variance of the sum of linearized terms in (8) is

$$\text{var}(\tilde{Y}) \cong \text{var}\left\{\sum_{i \in A_1} w_{i1} m(x_i, \beta_0)\right\} + \text{var}\left[\sum_{i \in A_2} w_{i2}\{y_i - m(x_i, \beta_0)\}\right]. \tag{9}$$

The first term in (9) is the variance due to sampling in survey 1 and the second term is the variance due to sampling in survey 2. The first term is of order $O(n_1^{-1}N^2)$ and the second term is $O(n_2^{-1}N^2)$.

It follows from (9) that a Taylor linearization variance estimator of $\hat{Y}_p$ is readily obtained from the formulae for the variance estimator of a total from survey 1 and survey 2. In particular, suppose that the estimators of a total $Z = \sum_{i=1}^N z_i$ from surveys 1 and 2 are denoted by $\hat{Z}_1 = \sum_{i \in A_1} w_{i1} z_i$ and $\hat{Z}_2 = \sum_{i \in A_2} w_{i2} z_i$ with corresponding variance estimators denoted by $v_1(z_i)$ and $v_2(z_i)$, using an operator notation. Then the linearization variance estimator of $\hat{Y}_p$ is

$$v(\hat{Y}_p) = v_1(\tilde{y}_i) + v_2(\hat{e}_i), \tag{10}$$

where $\hat{e}_i = y_i - \tilde{y}_i$. The variance estimator (10) is asymptotically unbiased with respect to both designs, but it requires access to data from both surveys. It follows from (10) that ignoring the second component $v_2(\hat{e}_i)$ based on the design and data for survey 2, and using only the first component $v_1(\tilde{y}_i)$, based only on the design for survey 1 and the synthetic data $\{\tilde{y}_i : i \in A_1\}$, can lead to serious underestimation, although the bias-correction term $\sum_{i \in A_2} w_{i2}\hat{e}_i$ is zero under (2). However, in § 4 we present a pseudo-replication method for correct variance estimation without requiring access to the data $\{(y_i, x_i) : i \in A_2\}$ from survey 2.

The above results can be extended to the case of known population totals $T = \sum_{i=1}^N t_i$ of auxiliary variables t observed in surveys 1 and 2: $\{(x_i, t_i) : i \in A_1\}$ and $\{(y_i, x_i, t_i) : i \in A_2\}$. We first obtain calibration weights (Deville & Särndal, 1992) $\tilde{w}_{i1}$ and $\tilde{w}_{i2}$ such that $\sum_{i \in A_1} \tilde{w}_{i1} t_i = T$ and $\sum_{i \in A_2} \tilde{w}_{i2} t_i = T$. In the second step, we replace w_{i2} by $\tilde{w}_{i2}$ in (3) to get $\hat{\beta}$ and $\tilde{y}_i$ and then use (1) with w_{i1} replaced by $\tilde{w}_{i1}$. The variance estimators $v_1(z_i)$ and $v_2(z_i)$ now refer to the calibration estimators of Z from surveys 1 and 2, respectively.

3. Domain estimation

In this section, we consider the estimation of a subpopulation total; subpopulations are called domains in the sampling literature. Creating an imputed dataset associated with the sample A_1, sometimes called mass imputation, is often used for domain estimation. For example, the World Bank has been using a simulated census method, developed by Elbers et al. (2003), to estimate

domain poverty measures in some developing countries. Breidt et al. (1996) discussed the possibility of using imputation to get improved estimates for domains.

Under the setup of § 2, a projection estimator of a domain total $Y_d = \sum_{i=1}^{N} \delta_i(d) y_i$, obtained from the imputed values $\{\tilde{y}_i : i \in A_1\}$ in the survey 1 data file, is given by

$$\hat{Y}_{d,p} = \sum_{i \in A_1} w_{i1} \delta_i(d) \tilde{y}_i, \tag{11}$$

where $\delta_i(d) = 1$ if unit i belongs to domain d and $\delta_i(d) = 0$ otherwise. Fuller (2003) proposed a similar projection domain estimator in the context of two-phase sampling and working model $m(x_i) = x_i'\beta$. The estimator (11), in general, is asymptotically design biased. A bias-corrected domain estimator is

$$\hat{Y}_{d,p,\mathrm{bc}} = \hat{Y}_{d,p} + \hat{Y}_{d,\mathrm{bc}}, \tag{12}$$

where $\hat{Y}_{d,\mathrm{bc}} = \sum_{i \in A_2} w_{i2} \delta_i(d)(y_i - \tilde{y}_i)$ is the bias-correction term. The bias-corrected domain estimator (12) was first proposed by Battese et al. (1988) in the context of small area estimation, when $m(x_i; \beta) = x_i'\beta$. It follows from (2) that the bias-corrected domain estimator (12), when summed over the domains d defining a partition of A_1, agrees with the projection estimator (1) of the grand total Y. This is a desirable internal consistency property. However, (12) cannot be implemented from the survey 1 data file containing the synthetic values $\tilde{y}_i$, unlike the projection estimator (11). It is therefore of interest to explore conditions that ensure either (i) $\hat{Y}_{d,\mathrm{bc}} = 0$ or (ii) the asymptotic bias of (11) relative to the domain total Y_d is negligible. The condition $\hat{Y}_{d,\mathrm{bc}} = 0$ is essentially the congeniality condition of Meng (1994) used in the context of multiple imputation.

We first study case (i) and note that $\hat{Y}_{d,\mathrm{bc}} = 0$ if and only if $\delta_i(d) = a'h(x_i)$ for some a, where $h(x_i)$ is used in (3). In the special case of linear or logistic augmented regression working models, $\hat{Y}_{d,\mathrm{bc}} = 0$ if the vector of domain indicators $\delta_i(d)$ is in the column space of the matrix X with rows x_i' $(i \in A_2)$. If the domains are planned or specified in advance, then X can be augmented by including the domain indicator $\delta_i(d)$ to ensure that $\hat{Y}_{d,\mathrm{bc}} = 0$ in the case of a linear or logistic regression working model. Information on the planned domain indicators should be supplied with the survey 1 data file to facilitate domain estimation.

Turning now to case (ii), the asymptotic bias of the projection domain estimator relative to the domain total is given by

$$\mathrm{RB}(\hat{Y}_{d,p}) = -\frac{\mathrm{cov}\{\delta_i(d), r_i\}}{\bar{\delta}(d)\bar{Y}_d}, \tag{13}$$

where $\bar{Y}_d$ is the domain mean, $\bar{\delta}(d)$ is the mean of the domain indicators $\delta_i(d)$ and $\mathrm{cov}\{\delta_i(d), r_i\}$ is the population covariance of $\delta_i(d)$ and $r_i = y_i - m(x_i, \beta_0)$. It follows from (13) that $\mathrm{RB}(\hat{Y}_{d,p})$ is negligible if $\delta_i(d)$ is approximately unrelated to r_i. This will be the case if the working model is correctly specified. In the simulation study of § 6, we generated $\delta_i(d)$ independent of (x_i, y_i) which ensured that $\delta_i(d)$ is unrelated to r_i. In the case of unplanned domains, d, and domain indicators $\delta_i(d)$ related to r_i, the domain projection estimator (11) can be significantly biased.

Turning to variance estimation, the decomposition $\hat{Y}_{d,p} \doteq \hat{P}_{d1} + \hat{Q}_{d2}$ with

$$\hat{P}_{d1} = \sum_{i \in A_1} w_{i1} \delta_i(d) m_i(x_i, \beta_0), \quad \hat{Q}_{d2} = \sum_{i \in A_2} w_{i2} \delta_i(d)\{y_i - m(x_i, \beta_0)\}$$

holds under case (i) or case (ii). Since the two surveys are independent, the total variance is given by

$$\mathrm{var}(\hat{Y}_{d,p}) \doteq \mathrm{var}(\hat{P}_{d1}) + \mathrm{var}(\hat{Q}_{d2}), \tag{14}$$

where $\text{var}(\hat{Q}_{d2})$ is based on the design for survey 2 while $\text{var}(\hat{P}_{d1})$ depends on the design for survey 1.

Using the operator notation of § 2, it follows from (14) that a linearization variance estimator of $\hat{Y}_{d,p}$ is given by

$$v(\hat{Y}_{d,p}) = v_1\{\delta_i(d)\tilde{y}_i\} + v_2\{\delta_i(d)\hat{e}_i\}. \tag{15}$$

The linearization variance estimator (15) is obtained from (10) by simply replacing $\tilde{y}_i$ and $\hat{e}_i$ by $\delta_i(d)\tilde{y}_i$ and $\delta_i(d)\hat{e}_i$, respectively. Note that (15) requires access to full data from both the surveys, similar to the variance estimator (10) of $\hat{Y}_p$. However, in § 4 we present a pseudo-replication method of variance estimator without requiring access to data from survey 2.

4. Replication variance estimation

In this section, we propose a pseudo-replication method for variance estimation that requires the generation of synthetic data $\{\tilde{y}_i^{(k)}, i \in A_1\}$ corresponding to each set of replication weights $\{w_{i1}^{(k)}, i \in A_1\}$ associated with survey 1 only. This method enables the user to correctly estimate the variance of the projection estimator $\hat{Y}_p$ without access to the data $\{(y_i, x_i) : i \in A_2\}$ from survey 2. The data file will contain additional columns of $\{\tilde{y}_i^{(k)} : i \in A_1\}$ associated with the columns of replicate weights $\{w_{i1}^{(k)} : i \in A_1\}$ $(k = 1, \ldots, L_1)$, where L_1 is the number of replicates created from survey 1. Hence, the price one pays to use only survey 1 synthetic data is to increase the number of columns in the data file by L_1 for each variable y for which synthetic values are generated. Typically, the number of such variables may be small.

Suppose that some pseudo-replication weights $\{w_{i1}^{(k)} : i \in A_1\}$ $(k = 1, \ldots, L_1)$ are used to compute a variance estimator of $\hat{Z}_1 = \sum_{i \in A_1} w_{i1} z_i$ for a variable z_i from survey 1 as

$$v_{1,\text{rep}}(\hat{Z}_1) = \sum_{k=1}^{L_1} c_k (\hat{Z}_1^{(k)} - \hat{Z}_1)^2, \tag{16}$$

where $\hat{Z}_1^{(k)} = \sum_{i \in A_1} w_{i1}^{(k)} z_i$ and c_k is a factor associated with replicate k. For example, the pseudo-replication weights could be based on a resampling method, such as the jackknife method or the bootstrap method, see Rust & Rao (1996). Assume that the replication variance estimator (16) is design-consistent with respect to survey 1. That is,

$$\{\text{var}(\hat{Z}_1)\}^{-1} v_{1,\text{rep}}(\hat{Z}_1) = 1 + o_p(1). \tag{17}$$

To generate the $\tilde{y}_i^{(k)} (k = 1, \ldots, L_1)$, we use the following steps,

Step 1. From the sample of survey 2, obtain an estimator of the variance of $\hat{Y}_2 = \sum_{i \in A_2} w_{i2} y_i$, denoted by $v_2(\hat{Y}_2)$, using a linearization method or a resampling method.

Step 2. Create L_1 columns of replicate weights $\{w_{i2}^{(k)} : k = 1, \ldots, L_1, i \in A_2\}$ such that

$$\sum_{k=1}^{L_1} c_k (\hat{Y}_2^{(k)} - \hat{Y}_2)^2 = v_2(\hat{Y}_2) \tag{18}$$

holds for any y, where $v_2(\hat{Y}_2)$ is obtained from Step 1 and $\hat{Y}_2^{(k)} = \sum_{i \in A_2} w_{i2}^{(k)} y_i$.

Step 3. For each k, using the replication weights $w_{i2}^{(k)}$ obtained from Step 2, solve

$$\sum_{i\in A_2} w_{i2}^{(k)}\{y_i - m(x_i;\beta)\}h_i = 0 \tag{19}$$

to get $\hat{\beta}^{(k)}$ and the associated synthetic values $\tilde{y}_i^{(k)} = m(x_i, \hat{\beta}^{(k)})$.

This procedure is applicable for any replication method of variance estimation for survey 1 satisfying (17). We assume that the solution to (19) is uniquely determined. Instead of solving (19) iteratively, one can use a one-step Newton–Raphson method to get an approximate for $\hat{\beta}^{(k)}$ (Rao & Tausi, 2004). Using the $\tilde{y}_i^{(k)}$, we obtain the replicate estimator

$$\hat{Y}_p^{(k)} = \sum_{i\in A_1} w_{i1}^{(k)}\tilde{y}_i^{(k)}. \tag{20}$$

The resulting replication variance estimator of $\hat{Y}_p$ is then

$$v_{1,\mathrm{rep}}(\hat{Y}_p) = \sum_{k=1}^{L_1} c_k(\hat{Y}_p^{(k)} - \hat{Y}_p)^2 \tag{21}$$

The crucial part of the proposed method is the construction of a set of L_1 replicate weights $\{w_{i2}^{(k)} : k = 1, \ldots, L_1, i \in A_2\}$ that satisfy (18). A method of constructing such replicate weights in general sampling designs is given in Appendix A2. However, if variance estimation for survey 2 also uses L_1 replicate weights for each $i \in A_2$, then those weights can be used in (19) to get $\hat{\beta}^{(k)}$ $(k = 1, \ldots, L_1)$; for example, when both surveys use the same number of bootstrap weights.

Theorem 2 shows that the proposed variance estimator (21) is design-consistent.

THEOREM 2. *Assume that conditions* (*C*1)–(*C*3) *of Theorem* 1 *hold. Assume that the replication method based on survey* 1 *satisfies* (17)–(19) *and that the variance estimator* $v_2(\hat{Y}_2)$ *obtained from survey* 2 *satisfies*

$$\{\mathrm{var}(\hat{Y}_2)\}^{-1}v_2(\hat{Y}_2) = 1 + o_p(1). \tag{22}$$

Further assume that the projection estimator satisfies condition (2). *Then,* (21) *satisfies* $v_{1,\mathrm{rep}}(\hat{Y}_p) = \mathrm{var}(\hat{Y}_p) + o_p(n_2^{-1}N^2)$, *where* $o_p(\cdot)$ *refers to both survey designs.*

Proof. See the Appendix. □

Theorem 2 also holds for the domain projection estimator satisfying case (i) or case (ii). That is, the variance estimator

$$v_{1,\mathrm{rep}}(\hat{Y}_{d,p}) = \sum_{k=1}^{L_1} c_k(\hat{Y}_{d,p}^{(k)} - \hat{Y}_{d,p})^2,$$

with $\hat{Y}_{d,p}^{(k)} = \sum_{i\in A_1} w_{i1}^{(k)}\delta_i(d)\tilde{y}_i^{(k)}$, is consistent for the variance of $\hat{Y}_{d,p}$.

5. EFFICIENT ESTIMATOR: FULL INFORMATION

If the full data from both surveys were available, then it would be possible to obtain an efficient estimator of the total Y based on the full data. We now derive such an estimator, which is used

in § 6 to study the relative efficiency of the projection estimator. For simplicity, we consider the case of a scalar variable x.

The full data are $\{(y_i, x_i, w_{i2}) : i \in A_2\}$ and $\{(x_i, w_{i1}) : i \in A_1\}$. Based on survey 2 data, we obtain an unbiased estimator $(\hat{X}_2, \hat{Y}_2) = \sum_{i \in A_2} w_{i2}(x_i, y_i)$ of (X, Y), where $X = \sum_{i=1}^{N} x_i$. Similarly, based on survey 1 data, an unbiased estimator of X is given by $\hat{X}_1 = \sum_{i \in A_1} w_{i1} x_i$ which is independent of $(\hat{X}_2, \hat{Y}_2)$. Now it is easy to show that the best linear unbiased estimator of Y, based on the estimators $\hat{X}_2$, $\hat{Y}_2$ and $\hat{X}_1$, is given by

$$\tilde{Y}_{\text{opt}} = \hat{Y}_2 + B_{y.x2}(\tilde{X}_{\text{opt}} - \hat{X}_2), \tag{23}$$

where

$$\tilde{X}_{\text{opt}} = \frac{V_{xx2}\hat{X}_1 + V_{xx1}\hat{X}_2}{V_{xx1} + V_{xx2}} \tag{24}$$

is the best linear unbiased estimator of the total X, $B_{y.x2} = V_{yx2}/V_{xx2}$, $V_{xx1} = \text{var}(\hat{X}_1)$, $V_{xx2} = \text{var}(\hat{X}_2)$ and $V_{yx2} = \text{cov}(\hat{Y}_2, \hat{X}_2)$. In practice, we replace V_{xx1}, V_{xx2} and V_{yx2} in (23) and (24) by corresponding design-consistent estimators, leading to asymptotically optimal estimates $\hat{Y}_{\text{opt}}$ and $\hat{X}_{\text{opt}}$. The estimator $\hat{Y}_{\text{opt}}$ agrees with the estimator (3.27) of Hidiroglou (2001), but he did not formally establish its asymptotic optimality. Note that $\hat{Y}_{\text{opt}}$ is purely design-based, unlike the model-assisted projection estimators based on a working model.

Optimal estimators based on domain-specific variances $V_{yx2,d} = \text{cov}(\hat{Y}_{2d}, \hat{X}_{2d})$, $V_{xx2,d} = \text{var}(\hat{X}_{2d})$ and $V_{xx1,d} = \text{var}(\hat{X}_{1d})$ can also be constructed, where $\hat{X}_{2d} = \sum_{i \in A_2} w_{i2}\delta_i(d)x_i$ and $\hat{X}_{1d} = \sum_{i \in A_1} w_{i1}\delta_i(d)x_i$ are unbiased estimators of the domain total X_d, and $\hat{Y}_{2d} = \sum_{i \in A_2} w_{i2}\delta_i(d)y_i$ is an unbiased estimator of Y_d. Best linear unbiased estimators of Y_d and X_d, based on $\hat{X}_{2d}$, $\hat{Y}_{2d}$ and $\hat{X}_{1d}$, are given by

$$\tilde{Y}_{\text{opt},d} = \hat{Y}_{2d} + B_{y.x2,d}(\tilde{X}_{\text{opt},d} - \hat{X}_{2d}), \quad \tilde{X}_{\text{opt},d} = \frac{V_{xx2,d}\hat{X}_{1d} + V_{xx1,d}\hat{X}_{2d}}{V_{xx1,d} + V_{xx2}, d}, \tag{25}$$

where $B_{y.x2,d} = V_{yx2,d}/V_{xx2,d}$. In practice, we replace the domain-specific variances in (25) by design-consistent estimators, leading to asymptotically optimal estimators $\hat{Y}_{\text{opt},d}$ and $\hat{X}_{\text{opt},d}$. The estimator $\hat{Y}_{\text{opt},d}$ does not satisfy the internal consistency property, and it may not be stable for domains containing few A_2 sample units.

6. Simulation study

We conducted a small simulation study on the finite sample bias and efficiency of the projection estimator $\hat{Y}_p$, the optimal estimator $\hat{Y}_{\text{opt}}$, the domain projection estimator $\hat{Y}_{d,p}$ and the domain optimal estimator $\hat{Y}_{\text{opt},d}$. We assumed simple random sampling for both surveys with sample sizes n_1 and n_2. For this special case, $w_{i1} = N/n_1$, and $w_{i2} = N/n_2$.

Two artificial finite populations, denoted by A and B, each of size $N = 10\,000$, were generated from two different models. Values $\{(y_i, x_i, z_i) : i = 1, \ldots, N\}$ for population A were generated independently from $x_i \sim \chi^2(2)$, $y_i = 1 + 0{\cdot}7x_i + e_i$, $e_i \sim N(0, 2)$, and $z_i \sim \text{Un}(0, 1)$, where z_i is independent of (x_i, y_i). Population B used the same (x_i, z_i) of population A but y_i was generated from $y_i = 0{\cdot}7x_i + u_i$ with $u_i \sim N(0, x_i)$. In both populations, the population variance of y is about 4·1 and the correlation coefficient between x and y is about 0·71.

Table 1. *Relative bias (%) and relative efficiency (%) of the projection estimators and the optimal estimator, based on 5000 samples*

		Population A		Population B	
Parameter	Estimator	RB	RE	RB	RE
Total	Regression projection	0	98	0	97
	Ratio projection	0	58	0	99
	Aug. reg. projection	0	97	0	97
	Aug. rat. projection	1	55	0	98
	Optimal	0	100	0	100
Domain	Regression projection	0	196	1	201
	Ratio projection	1	122	1	205
	Aug. reg. projection	0	105	0	98
	Aug. rat. projection	0	64	0	96
	Optimal	−1	100	−2	100

RB, relative bias; RE, relative efficiency; Aug. reg., Augmented regression; Aug. rat., Augmented ratio.

From each of the two simulated populations, two independent samples of size $n_1 = 500$ and $n_2 = 100$ were generated by simple random sampling. We generated 5000 pairs of independent samples. We considered the estimation of total Y and the domain total Y_d defined by $\delta_i(d) = 1$ if $z_i < 0{\cdot}3$ and $\delta_i(d) = 0$ otherwise. From each generated sample, we computed $\hat{Y}_{\text{opt}}$, a regression projection estimator $\hat{Y}_{p,1} = \sum_{i \in A_1} w_{i1} \tilde{y}_{i1}$ using synthetic data $\{\tilde{y}_{i1} : i \in A_1\}$ constructed from a working linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ $(i \in A_2)$ with $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$, and a ratio projection estimator $\hat{Y}_{p,2} = \sum_{i \in A_1} w_{i1} \tilde{y}_{i2}$ using synthetic data $\{\tilde{y}_{i2} : i \in A_1\}$ constructed from a working ratio model $y_i = \gamma x_i + e_i$ $(i \in A_2)$ with $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2 x_i$. We used $\hat{y}_{i1} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $\hat{y}_{i2} = \hat{\gamma} x_i$, where $(\hat{\beta}_0, \hat{\beta}_1)'$ is the least squares estimator of $(\beta_0, \beta_1)'$ and $\hat{\gamma} = (\sum_{i \in A_2} x_i)^{-1} (\sum_{i \in A_2} y_i)$ is the weighted least squares estimator of γ computed from survey 2 data. We also computed the domain projection estimators $\hat{Y}_{d,p1}$ and $\hat{Y}_{d,p2}$, and the optimal domain estimator $\hat{Y}_{\text{opt},d}$ in (25). Under the simulation set-up, $\delta_i(d)$ is unrelated to r_i because z_i is generated independent of (y_i, x_i). Hence, the domain projection estimators $\hat{Y}_{d,p1}$ and $\hat{Y}_{d,p2}$ are asymptotically unbiased.

Table 1 reports the relative bias and the relative efficiency of the estimators of the total Y. Relative bias of an estimator $\hat{Y}$ of Y is computed as $\text{RB}(\hat{Y}) = \{E(\hat{Y}) - Y\}/Y$ where the expectation is based on the 5000 simulation runs. Relative efficiency of an estimator $\hat{Y}$ is computed as $\text{RE}(\hat{Y}) = \text{mse}(\hat{Y}_{\text{opt}})/\text{mse}(\hat{Y})$ where the mean square error $\text{mse}(\hat{Y})$ is based on the 5000 simulation runs. Table 1 shows that the relative bias of all the estimators is negligible, less than 1%, confirming their asymptotic unbiasedness even when the working model is misspecified. Table 1 also shows that the regression projection estimator $\hat{Y}_{p1}$ is almost as efficient as $\hat{Y}_{\text{opt}}$ even when the true model is the ratio model. On the other hand, the ratio projection estimator $\hat{Y}_{p2}$ is considerably less efficient if the true model is the linear regression model with a substantial intercept term. This result suggests the need for model diagnostics to identify a good working model. We also computed the relative bias and relative efficiency of projection estimators based on working augmented linear regression and ratio models, denoted augmented projection estimators, where the augmentation is based on the domain indicators $\delta_i(d)$. Table 1 shows that the augmented projection estimators are similar to the corresponding projection estimators in terms of relative bias and relative efficiency.

Table 2. *Relative biases (%) of the jackknife variance estimator for the projection estimators, based on* 5000 *samples*

Point estimator	Parameter	Population A	Population B
Regression projection	Total	−1·3	2·4
	Domain	−3·0	0·6
Ratio projection	Total	3·2	0·0
	Domain	−0·1	−1·7
Aug. reg. projection	Total	3·3	4·0
	Domain	2·2	5·0
Aug. rat. projection	Total	5·9	3·0
	Domain	6·4	6·1

Aug. reg., Augmented regression; Aug. rat., Augmented ratio.

Results for domain estimation are also shown in Table 1. The relative bias of all the domain estimators is less than 3%, confirming their asymptotic unbiasedness under the simulation setup. In terms of efficiency, the regression projection estimator, $\hat{Y}_{d,p1}$, is considerably more efficient than the optimal estimator, $\hat{Y}_{\mathrm{opt},d}$, unlike in the case of estimating the total. This is because the projection estimator is based on domain units belonging to much larger sample A_1, unlike $\hat{Y}_{\mathrm{opt},d}$. Again, the ratio projection estimator is considerably less efficient than the regression projection estimator if the true model is the regression model. Table 1 also shows that the augmented domain projection estimators are considerably less efficient than the corresponding domain projection estimators, but the augmented domain regression projection estimator is slightly more efficient than the optimal domain estimator. Augmenting the working model by domain indicators affects efficiency of domain estimators under congeniality, but the estimators remain asymptotically unbiased regardless of congeniality. Hence, augmented domain projection estimators can be more efficient than the corresponding domain projection estimators if the congeniality condition does not hold.

We also computed jackknife variance estimators of the projection estimators. In the jackknife method, $L_1 = n_2 = 100$ replicates were created. To do this, we first partitioned the sample A_1 at random into $n_2 = 100$ disjoint subsets of equal size. Let $A_1^{(k)}$ be the kth random group obtained from the partitioning. We have $A_1 = \cup_{k=1}^{n_2} A_1^{(k)}$. Following the method in § 4, the kth replicate of $\hat{Y}_{p1} = \sum_{i \in A_1} w_{i1}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ was computed as

$$\hat{Y}_{p1}^{(k)} = \sum_{i \in A_1} w_{i1}^{(k)}(\hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)} x_i) = \sum_{i \in A_1} w_{i1}^{(k)} \tilde{y}_i^{(k)},$$

where $w_{i1} = N/n_1$,

$$w_{i1}^{(k)} = \begin{cases} (w_{i1} n_2/n_2 - 1) & (i \notin A_1^{(k)}), \\ 0 & (i \in A_1^{(k)}), \end{cases}$$

and $(\hat{\beta}_0^{(k)}, \hat{\beta}_1^{(k)})$ is computed by deleting the kth element in the survey 2 sample. The jackknife variance estimate is computed as

$$v_{1,\mathrm{JK}} = \sum_{k=1}^{n_2} \frac{n_2 - 1}{n_2} (\hat{Y}_{p1}^{(k)} - \hat{Y}_{p1})^2.$$

The jackknife variance estimator of the other projection estimators were also computed. The jackknife variance estimators of the projection estimators for the domain mean were computed similarly. Table 2 shows that the relative biases of the jackknife variance estimators are all small; $<6\%$ in absolute value. Thus, the proposed replication method in § 4 works well for estimating the variance of projection estimators in moderate sample sizes.

7. FURTHER WORK

We have only considered the case of synthetic values obtained by deterministic imputation. This method will not work for estimating population quantiles of a continuous variable. We propose to study other approaches based on stochastic imputation, including fractional imputation of Kim & Fuller (2004) to handle quantiles and other nonsmooth functions. We also plan to study other descriptive population parameters such as regression coefficients. Extension to small area estimation, using predicted, or synthetic, values based on small area models, will also be studied.

ACKNOWLEDGEMENT

We thank three referees for their very helpful comments. The research of the first author was partially supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The research of the second author was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Proof of Theorem 1. By (2), we can express the projection estimator (1) as

$$\hat{Y}_p(\hat{\beta}) = \sum_{i \in A_1} w_{i1} m(x_i; \hat{\beta}) + \sum_{i \in A_2} w_{i2}\{y_i - m(x_i; \hat{\beta})\}.$$

Using (6), we make a Taylor series expansion of $\hat{Y}_p(\hat{\beta})$ around $\hat{\beta} = \beta_0$ to get

$$\hat{Y}_p = \hat{P}_1 + \hat{Q}_2 + \left\{\sum_{i \in A_1} w_{i1} \dot{m}(x_i; \beta^*) - \sum_{i \in A_2} w_{i2} \dot{m}(x_i; \beta^*)\right\}' (\hat{\beta} - \beta_0),$$

where $\hat{P}_1$ and $\hat{Q}_2$ are defined after (8), β^* is a point on the line segment between $\hat{\beta}$ and β_0, and $\dot{m}(x_i; \beta) = \partial m(x_i; \beta)/\partial \beta$. By Condition 3,

$$\sum_{i \in A_1} w_{i1} \dot{m}(x_i; \beta^*) - \sum_{i \in A_2} w_{i2} \dot{m}(x_i; \beta^*) = O_p(n_2^{-1/2} N)$$

and, using Condition 1, the result (7) follows. □

Construction of replicates of $\hat{\beta}$

In principle, we can always construct a replication variance estimator that is algebraically equivalent to the traditional variance estimator in the linear case. We assume that the traditional variance estimator of

$\hat{Y}_2 = \sum_{i \in A_2} w_{i2} y_i$ is a quadratic function of the sample values $\{y_i : i \in A_2\}$. That is,

$$v_2 = \sum_{i \in A_2} \sum_{j \in A_2} \Delta_{ij2} y_i y_j \tag{A1}$$

for some Δ_{ij2}. Let Δ_2 be an $n_2 \times n_2$ matrix whose (i, j)th element is Δ_{ij2}. Assume that the matrix Δ_2 is nonnegative definite so that it can be decomposed into

$$\Delta_2 = \sum_{j=1}^{m} \eta_j b_j b_j' \tag{A2}$$

for some scalar η_j and n_2-dimensional vector b_j for $j = 1, \ldots, m$.

Suppose that we want to express the variance estimator v_2 in (A1) as a replication variance estimator of the form

$$v_1(\hat{Y}_2) = \sum_{k=1}^{L_1} c_k (\hat{Y}_2^{(k)} - \hat{Y}_2)^2, \tag{A3}$$

where $L_1 > m$ is the number of replications and $\hat{Y}_2^{(k)} = \sum_{i \in A_2} w_{i2}^{(k)} y_i$ is the kth replicate of $\hat{Y}_2 = \sum_{i \in A_2} w_{i2} y_i$. A natural question in the replication method is to find a condition for the replication variance estimator (A3) to be algebraically equivalent to v_2 in (A1). The following lemma gives a general result for such replication variance estimation.

LEMMA A1. *If*

$$w_{i2}^{(k)} = w_{i2} + \sum_{j=1}^{m} a_{kj} b_{ij} \tag{A4}$$

where b_{ij} is the ith element of b_j satisfying (A2) and the coefficients a_{kj} satisfy

$$\sum_{k=1}^{L_1} c_k a_{ki} a_{kj} = \begin{cases} \eta_i & (i = j), \\ 0 & (i \neq j), \end{cases} \tag{A5}$$

then the replication variance estimator in (A3) is equal to v_2 in (A1).

Proof. Let $w_2^{(k)}$ and w_2 be n_2-dimensional vectors with elements $w_{i2}^{(k)}$ and w_{i2}, respectively. Then,

$$\begin{aligned} \sum_{k=1}^{L_1} c_k (w_2^{(k)} - w_2)(w_2^{(k)} - w_2)' &= \sum_{k=1}^{L_1} c_k \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ki} a_{kj} b_i b_j' \\ &= \sum_{k=1}^{m} \eta_k b_k b_k' = \Delta_2. \end{aligned} \tag{A6}$$

Now expressing v_2 and $v_1(\hat{Y}_2)$ as $v_2 = y_2' \Delta_2 y_2$ and

$$v_1(\hat{Y}_2) = y_2' \left\{ \sum_{k=1}^{L_1} c_k (w_2^{(k)} - w_2)(w_2^{(k)} - w_2)' \right\} y_2,$$

the equivalence result follows from (A6). □

The condition (A5) is the minimum condition for the replication weights in (A4) to be valid for variance estimation. There are many coefficients $w_{2i}^{(k)}$ satisfying (A5). For example, the choice

$$w_{i2}^{(k)} = \begin{cases} w_{i2} + (\eta_k/c_k)^{1/2} b_{ik} & (k \leqslant m), \\ w_{i2} & (k > m), \end{cases}$$

satisfies (A6).

Proof of Theorem 2. Write

$$\tilde{Y}^{(k)}(\beta) = \hat{P}_1^{(k)}(\beta) + \hat{Q}_2^{(k)}(\beta) = \sum_{i \in A_1} w_{i1}^{(k)} m(x_i; \beta) + \sum_{i \in A_2} w_{i2}^{(k)} \{y_i - m(x_i; \beta)\}$$

as a function of β, where $w_{i1}^{(k)}$ is defined in (16) and $w_{i2}^{(k)}$ is computed by (A4). Because $\hat{\beta}^{(k)}$ is constructed to satisfy (19), the replicate estimator (20) can be written as $\hat{Y}_p^{(k)} = \tilde{Y}^{(k)}(\hat{\beta}^{(k)})$. Also, write

$$\tilde{Y}(\beta) = \hat{P}_1(\beta) + \hat{Q}_2(\beta) = \sum_{i \in A_1} w_{i1} m(x_i; \beta) + \sum_{i \in A_2} w_{i2} \{y_i - m(x_i; \beta)\}.$$

By Condition 2 and (22),

$$\sum_{k=1}^{L_1} c_k \{\hat{Q}_2^{(k)}(\beta) - \hat{Q}_2(\beta)\}^2 = O_p(n_2^{-1} N^2)$$

uniformly in $\beta \in \mathcal{B}$. Using this result, it can be shown that

$$\max_{1 \leqslant k \leqslant L_1} N^{-1} c_k^{1/2} |\hat{Q}_2^{(k)}(\beta) - \hat{Q}_2(\beta)| \to 0$$

in probability uniformly in $\beta \in \mathcal{B}$. Because $\hat{\beta}^{(k)}$ is a unique solution to $\hat{Q}_2^{(k)}(\beta) = 0$, we can apply Lemma 3 of Kim & Park (2010) to get

$$\hat{\beta}^{(k)} - \hat{\beta} = o_p(1). \tag{A7}$$

By a Taylor expansion of $\tilde{Y}^{(k)}(\hat{\beta}^{(k)})$ around $\hat{\beta}$, we have

$$\begin{aligned} \hat{Y}_p^{(k)} = \tilde{Y}^{(k)}(\hat{\beta}^{(k)}) = \hat{P}_1^{(k)}(\hat{\beta}) + \hat{Q}_2^{(k)}(\hat{\beta}) + \{\hat{U}(\beta_k^*)\}'(\hat{\beta}^{(k)} - \hat{\beta}) \\ + \{\hat{U}^{(k)}(\beta_k^*) - \hat{U}(\beta_k^*)\}'(\hat{\beta}^{(k)} - \hat{\beta}), \end{aligned}$$

for some β_k^* in the line segment between $\hat{\beta}^{(k)}$ and β_0, where

$$\hat{U}^{(k)}(\beta) = \sum_{i \in A_1} w_{i1}^{(k)} \dot{m}(x_i; \beta) - \sum_{i \in A_2} w_{i2}^{(k)} \dot{m}(x_i; \beta), \quad \hat{U}(\beta) = \sum_{i \in A_1} w_{i1} \dot{m}(x_i; \beta) - \sum_{i \in A_2} w_{i2} \dot{m}(x_i; \beta).$$

By Condition 3,

$$N^{-1} \hat{U}(\beta) = O_p(n_2^{-1/2}) \tag{A8}$$

uniformly in $\beta \in \mathcal{B}$. Thus, we have

$$\begin{aligned} \tilde{Y}^{(k)}(\hat{\beta}^{(k)}) - \tilde{Y}(\hat{\beta}) = \hat{P}_1^{(k)}(\hat{\beta}) - \hat{P}_1(\hat{\beta}) + \hat{Q}_2^{(k)}(\hat{\beta}) - \hat{Q}_2(\hat{\beta}) \\ + \{\hat{U}^{(k)}(\beta_k^*) - \hat{U}(\beta_k^*)\}'(\hat{\beta}^{(k)} - \hat{\beta}) + \{\hat{U}(\beta_k^*)\}'(\hat{\beta}^{(k)} - \hat{\beta}). \end{aligned} \tag{A9}$$

By Condition 2, (17) and (22), we have

$$\sum_{k=1}^{L_1} c_k \{\hat{U}^{(k)}(\beta) - \hat{U}(\beta)\}^2 = O_p(n_2^{-1} N^2) \tag{A10}$$

uniformly in $\beta \in \mathcal{B}$. Also, it can be shown that

$$\sum_{k=1}^{L_1} c_k(\hat{\beta}^{(k)} - \hat{\beta})(\hat{\beta}^{(k)} - \hat{\beta})' = o_p(1). \tag{A11}$$

Hence, combining (A9), (A10) and (A11) and using (A7) and (A8), we have

$$\begin{aligned}\sum_{k=1}^{L_1} c_k\{\hat{Y}_p^{(k)} - \hat{Y}_p\}^2 = &\sum_{k=1}^{L_1} c_k\{\hat{P}_1^{(k)}(\hat{\beta}) - \hat{P}_1(\hat{\beta})\}^2 + \sum_{k=1}^{L_1} c_k\{\hat{Q}_2^{(k)}(\hat{\beta}) - \hat{Q}_2(\hat{\beta})\}^2 \\ &+ 2\sum_{k=1}^{L_1} c_k\{\hat{P}_1^{(k)}(\hat{\beta}) - \hat{P}_1(\hat{\beta})\}\{\hat{Q}_2^{(k)}(\hat{\beta}) - \hat{Q}_2(\hat{\beta})\} + o_p(n_2^{-1}N^2). \end{aligned} \tag{A12}$$

The first term in (A12) estimates the variance of $\hat{P}_1(\beta_0)$, the second term estimates the variance of $\hat{Q}_2(\beta_0)$, and the third term estimates two times the covariance between $\hat{P}_1(\beta_0)$ and $\hat{Q}_2(\beta_0)$, which is equal to zero because of the independence of the two surveys. Hence, the variance estimator (21) is consistent with respect to both survey designs. □

References

Battese, G., Harter, R. & Fuller, W. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Am. Statist. Assoc.* **83**, 28–36.

Bose, C. (1943). Note on the sampling error in the method of double sampling. *Sankhya* **6**, 330.

Breidt, F. J., McVey, A. & Fuller, W. A. (1996). Two-phase sampling by imputation. *J. Indian Soc. Agric. Statist.* **49**, 79–90.

Clayton, D., Spiegelhalter, D., Dunn, G. & Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *J. R. Statist. Soc.* B **60**, 71–87.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica* **71**, 355–364.

Fuller, W. A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, Ed. R. L. Chambers and C. J. Skinner. Chichester: Wiley.

Godambe, V. & Thompson, M. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *Int. Statist. Rev.* **54**, 127–38.

Hidiroglou, M. (2001). Double sampling. *Survey Methodol.* **27**, 143–54.

Hidiroglou, M. & Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodol.* **24**, 11–20.

Isaki, C. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Assoc.* **77**, 89–96.

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

Kim, J. K. & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika* **91**, 559–78.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **9**, 538–73.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *J. Am. Statist. Assoc.* **99**, 1131–9.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *J. R. Statist. Soc.* B **72**, 27–48.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* **60**, 125–33.

Rao, J. N. K. & Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Commun. Statist.* **33**, 2087–95.

Reilly, M. & Pepe, M. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

Reiter, J. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* **95**, 933–46.

Renssen, R. H. & Nieuwenbroek, N. (1997). Aligning estimates for common variables in two or more sample surveys. *J. Am. Statist. Assoc.* **92**, 368–75.

RUST, K. F. & RAO, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statist. Meth. Med. Res.* **5**, 283.
SCHENKER, N. & RAGHUNATHAN, T. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statist. Med.* **26**, 1802.
WU, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Can. J. Statist.* **32**, 15–26.
YBARRA, L. & LOHR, S. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* **95**, 919–31.

[*Received January* 2010. *Revised September* 2011]