International Statistical Review

International Statistical Review (2016), 0, 0, 1–18 doi:10.1111/insr.12197

Variance Estimation of Imputed Estimators of Change for Repeated Rotating Surveys

Yves G. Berger[1] and Emilio L. Escobar[2]

[1] *University of Southampton, Southampton, UK*
E-mail: Y.G.Berger@soton.ac.uk
[2] *Numerika, Mexico City, Mexico*
E-mail: Emilio@numerika.mx

Summary

A common problem in survey sampling is to compare two cross-sectional estimates for the same study variable taken from two different waves or occasions. These cross-sectional estimates often include imputed values to compensate for item non-response. The estimation of the sampling variance of the estimator of change is useful to judge whether the observed change is statistically significant. Estimating the variance of a change is not straightforward because of the rotation in repeated surveys and imputation. We propose using a multivariate linear regression approach and show how it can be used to accommodate the effect of rotation and imputation. The regression approach gives a design-consistent estimation of the variance of change when the sampling fraction is small. We illustrate the proposed approach using random hot-deck imputation, although the proposed estimator can be implemented with other imputation techniques.

Key words: Longitudinal surveys; missing data; non-response; overlapping samples; rotation; unequal inclusion probabilities.

1 Introduction

Measuring change over time is a central problem for many users of social, economic and demographic data. The primary interest of many users is often in changes or trends from one time period to another. Smith *et al.* (2003) recognised that assessing change is one of the most important challenge in survey statistics. A common problem is to compare two cross-sectional estimates for the same study variable taken from two different waves or occasions. These cross-sectional estimates often include imputed values to compensate for item non-response (e.g. Lohr, 2009, ch. 8). The estimation of the variance of an estimator of change is useful to judge whether the observed change is statistically significant. Covariances play an important role in estimating the variance of an estimated change, and they are not straightforward to estimate with repeated surveys because of rotation.

We propose to use a multivariate linear regression approach to estimate these covariances. The proposed estimator is not a model-based estimator, as it is valid even if the underlying model does not fit the data (Berger & Priam, 2010; 2016). We show how this approach can be used to accommodate the effect of imputation. The regression approach gives design-consistent

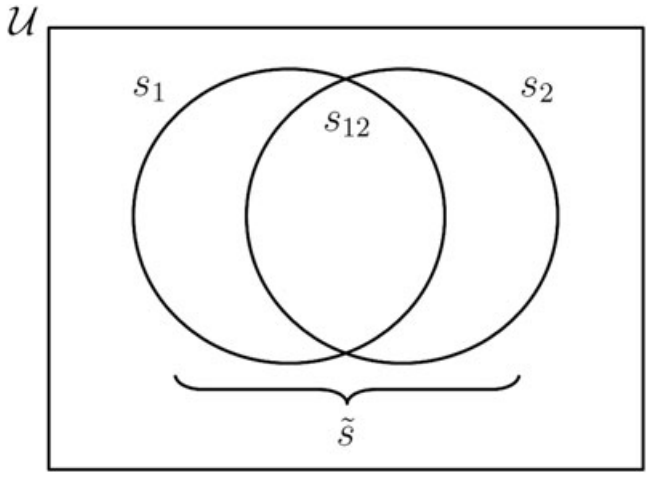

Figure 1. *The overall sample* $\tilde{s} = s_1 \cup s_2$.

estimation of the variance of change when the sampling fraction is small. We illustrate the proposed approach using random hot-deck imputation, although the proposed estimator can be implemented with any other imputation techniques.

2 Rotating Surveys

The estimation of variance of change would be relatively straightforward if cross-sectional estimates were based on the same sample. Furthermore, because of rotations that is used in repeated surveys, cross-sectional estimates are not independent. Let s_1 and s_2 denote respectively the first and the second wave samples. The samples s_1 and s_2 are usually not completely overlapping sets of units, because repeated surveys use rotation designs, which consist in selecting new units ($k \in s_2 \setminus s_1$) to replace old units ($k \in s_1 \setminus s_2$) that have been in the survey for a specified number of waves. Without lost of generality, we assume that s_1 and s_2 have the same sample size n. Let n_{12} denote the sample size of the common sample, $s_{12} = s_1 \cap s_2$. The units sampled on s_{12} represent usually a large fraction of s_1; that is, n_{12}/n is usually large. We denote the overall sample by $\tilde{s}$, where $\tilde{s} = s_1 \cup s_2$. The size of the overall sample is denoted by $\tilde{n} = \#(\tilde{s})$. Figure 1 gives a visual representation of the samples considered.

It is important to clarify that units in $s_1 \setminus s_2$ are not available at the second wave, and units in $s_2 \setminus s_1$ are not available at the first wave. This non-availability is not due to non-response (introduced in Section 3), but because of the rotation. The imputation is used to impute missing values because of non-response, not to impute the units, which are not available because of the rotation. In other words, those units in $s_2 \setminus s_1$ are not imputed at the first wave, and units in $s_1 \setminus s_2$ are not imputed at the second wave.

We assume that the rotation sampling design is such that n and n_{12} are fixed quantities. This class contains standard rotating sampling designs such as the rotating systematic sampling design (Holmes & Skinner, 2000), the rotation groups sampling design (e.g. Kalton 2009, Gambino & Silva, 2009), the rotating design that was proposed by Tam (1984) and the permanent random numbers rotating design (e.g. Ohlsson, 1995; Nordberg, 2000). Rotating sampling designs are widely used in practice for labour force surveys (e.g. Holmes & Skinner, 2000). Rotation design is also used for the European Union Statistics on Income and Living Conditions social survey (Eurostat, 2012).

Let $y_{\ell;k}$ denote the value of the variable of interest y_ℓ for the wave ($\ell = 1, 2$). Suppose, we wish to estimate the absolute change

$$\Delta = \tau_2 - \tau_1, \tag{1}$$

between two population totals τ_1 and τ_2 from waves 1 and 2, where $\tau_\ell = \sum_{k \in \mathcal{U}} y_{\ell;k}$. Here, $\mathcal{U}$ denotes the population of size N, assumed to be the same at both waves. It is possible to extend the approach we proposed for other measures of changes, such as relative change or change between means.

Suppose that the population is split into H strata: $U_1, \ldots, U_h, \ldots, U_H$. Suppose that two samples are selected from each strata according to the rotating design described in Figure 1. Let $z_{\ell;i}^{(h)}$ be the wave-strata sample indicator variable (Berger & Priam, 2016), defined by

$$z_{\ell;k}^{(h)} = \begin{cases} 1 \text{ when } k \in U_h \text{ and } k \in s_\ell, \\ 0 \text{ otherwise,} \end{cases} \qquad (h = 1, \ldots, H \quad \text{and } \ell = 1, 2)\cdot \tag{2}$$

The values of $z_{\ell;i}^{(h)}$ are assumed to be known for all $k \in s_\ell$. We also consider the wave indicator defined by

$$z_{\ell;k} = \sum_{h=1}^{H} z_{\ell;k}^{(h)} = \begin{cases} 1 \text{ when } k \in s_\ell \\ 0 \text{ otherwise} \end{cases}. \tag{3}$$

When we have a single stratum, $H = 1$ and $z_{\ell;i}^{(1)} = z_{\ell;k}$.

In Section 3, we introduce the uniform non-response mechanism for rotating sampling designs, and we show how random hot-deck imputed values can be used to compensate for item non-response. In Section 4, we propose to use a reverse approach (Fay, 1991) to estimate the variance of the imputed estimator of change. The proposed variance estimator depends on a covariance matrix, which shall be estimated using a multivariate (general) linear regression approach described in Section 6. In Section 7, we extend the variance estimator proposed to missing at random (MAR) response mechanism involving multiple imputation-classes. In Section 8.2, we consider a missing not at random (MNAR) response mechanism. In Section 8, a simulation study illustrates our findings.

3 Non-response

The main objective of this article is to address the problem of variance estimation under non-response rather than the non-response issue. Little has been performed on variance estimation of change. However, there are many design-based variance estimators of cross-sectional estimates (e.g. Wolter, 2007). The use of models to address non-response is also popular. A model-assisted approach can be found in Deville & Särndal (1994); Fay (1994); Steel & Fay (1995), Särndal & Lundström (2005). A Bayesian treatment of imputation can be found for example in Rubin (1987). See Brick & Montaquila (2009) for a wide discussion on non-response. A discussion on which inference-approach to use for non-response in surveys can be found in Haziza (2009). These approaches deal with cross-sectional estimators and cannot be directly implemented with estimators of changes.

We propose to use a design-based approach combined with random hot-deck imputation. A recent review on cross-sectional hot-deck imputation can be found in Andridge & Little (2010). The random hot-deck imputation has the advantage of guaranteeing unbiased estimation

of population distributions (Rao & Shao, 1992). The approach proposed is also valid under deterministic regression imputation.

Because of non-response, some of the values $y_{\ell;k}$ can be missing in each sample s_ℓ, ($\ell = 1, 2$). We propose to impute these missing values. Let

$$a_{\ell;k} \;=\; \begin{cases} 1 \text{ if } y_{\ell;k} \text{ is observed, and } k \in s_\ell, \\ 0 \text{ if } y_{\ell;k} \text{ is missing because of non-response, and } k \in s_\ell, \end{cases} \qquad (\ell = 1, 2).$$

The distribution of the random variables $a_{\ell;k}$ represents the response mechanism for the wave ℓ. We assume that the values $a_{\ell;k}$ are known for all $k \in s_\ell$. To simplify, we use the same notation for the random variables and their observed values.

As far as the response mechanism is concerned, we consider the usual cross-sectional design-based assumption in the succeeding texts (e.g. Fay 1991; Rao and Shao 1992; Rao and Sitter 1995; Shao and Steel 1999), but adapted for rotating sampling designs.

Assumption 1 (single imputation-class). *The response probability for the variable of interest in each wave is uniform on $\mathcal{U}$, and it is strictly positive (i.e. $P\{a_{\ell;k} = 1\} > 0$). $a_{\ell;k'}$ and $a_{\ell;k''}$ are independent for all $k' \neq k''$, where $k', k'' \in s_\ell$. The responses between waves can be dependent; that is, $a_{\ell';k'}$ and $a_{\ell'';k'}$ may be dependent, where $k' \in s_{\ell'}$ and $k' \in s_{\ell''}$. The imputation is conducted independently within waves.*

Assumption 1 implies a missing completely at random (MCAR) response mechanism. The MAR response mechanisms are covered in Section 7. In Section 8.2, we consider a simulation study with a MNAR response mechanism.

The setting of a single imputation class is the simplest case when handling non-response. However, the assumption may be considered unrealistic (e.g. Rao & Shao, 1992, p. 818). We therefore also consider, in Section 7, multiple imputation-classes where values are imputed within imputation classes (e.g. Haziza & Beaumont, 2007). First, we show how the proposed approach can be implemented when we have a single imputation class. In Section 7, we extend the approach proposed under a multiple imputation-classes setting (Assumption 2).

3.1 The Imputed Estimator of Change

Suppose that the change Δ in (1) is estimated by

$$\widehat{\Delta}^I \;=\; \widehat{\tau}_2^I \;-\; \widehat{\tau}_1^I, \tag{4}$$

where

$$\widehat{\tau}_\ell^I \;=\; \sum_{k \in \tilde{s}} \frac{y_{\ell;k}^I}{\pi_{\ell;k}} \qquad (\ell = 1, 2) \tag{5}$$

is the cross-sectional imputed Narain (1951); Horvitz & Thompson (1952) estimators at wave ℓ. Here, $y_{\ell;k}^I$ is defined by

$$y_{\ell;k}^I \;=\; z_{\ell;k}\,\{(1 - a_{\ell;k})\, y_{\ell;k}^* \;+\; a_{\ell;k}\, y_{\ell;k}\}, \tag{6}$$

where $y_{\ell;k}^*$ denotes an imputed value, which depends on the imputation technique. For example, in what follows, $y_{\ell;k}^*$ is defined by random hot-deck imputation, although the approach proposed can be generalised for other imputation techniques. The deterministic mean imputation is

a particular case of hot-deck imputation. Note that the imputation is only used for missing data because of non-response and not to impute the values $y_{2;k}$ of $k \in s_1 \setminus s_2$ and $y_{1;k}$ of $k \in s_2 \setminus s_1$, which are the non-available values of the units that rotate in and out.

3.2 Random Hot-deck Imputation

With random hot-deck imputation, the values $y^*_{\ell;k}$ used in Equation (6) are

$$\begin{aligned} y^*_{\ell;k} &= \widehat{\mu}^r_\ell \; + \; e_{\ell;k}, \\ e_{\ell;k} &= y_{\ell;j} \; - \; \widehat{\mu}^r_\ell, \end{aligned} \tag{7}$$

where j is a donor selected with-replacement with probabilities $p_{\ell;k} = \breve{a}_{\ell;k}/\widehat{N}^r_\ell$ from the sample of respondents $s^r_\ell = \{k : z_{\ell;k} = 1 \text{ and } a_{\ell;k} = 1\}$. Here, $\widehat{\mu}^r_\ell = \widehat{\tau}^r_\ell / \widehat{N}^r_\ell$ is the estimator of the respondents' mean, $\widehat{\tau}^r_\ell = \sum_{k \in \tilde{s}} \breve{y}_{\ell;k}$ is the estimator of the respondents' totals and $\widehat{N}^r_\ell = \sum_{k \in \tilde{s}} \breve{a}_{\ell;k}$ is the estimator of the number of respondents for waves $\ell = 1, 2$; with $\tilde{s} = s_1 \cup s_2$, and where

$$\breve{y}_{\ell;k} = \pi^{-1}_{\ell;k} \, z_{\ell;k} \, a_{\ell;k} \, y_{\ell;k}, \tag{8}$$

$$\breve{a}_{\ell;k} = \pi^{-1}_{\ell;k} \, z_{\ell;k} \, a_{\ell;k}. \tag{9}$$

$\pi_{\ell;k}$ denotes the first-order inclusion probability of the unit k at wave ℓ. If $e_{\ell;k} = 0$ in Equation (7), then $y^*_{\ell;k}$ from Equation (6) is the deterministic mean imputed value.

4 Population Variance of the Hot-deck Imputed Estimator of Change

We propose to estimate the variance of $\widehat{\Delta}^I$ defined by Equation (4) using a reverse approach for non-response (Fay, 1991; Shao & Steel, 1999). Let $\mathcal{U}^r_\ell$ be the population of respondents at wave ℓ, where $\mathcal{U}^r_\ell \subset \mathcal{U}$. In other words, at both waves, the population is randomly split into a population of respondents and a population of non-respondents according to an unknown response mechanism (see Figure 2). Note that the response mechanisms can be such that the set of respondents of wave 2 depends on the set of respondents at wave 1. That is, $a_{1;k}$ and $a_{2;k}$ can be dependent random variables (Assumption 1).

Let $E_r\{\cdot\}$ and $V_r\{\cdot\}$ denote respectively the expectation and variance operators with respect to the response mechanism. Rotation samples s_1 and s_2 are selected from the population $\mathcal{U}$ according to a rotation sampling design (Section 1). The samples of respondents are given by $s^r_\ell = \mathcal{U}^r_\ell \cap s_\ell, (\ell = 1, 2)$. Let $E_d\{\cdot\}$ and $V_d\{\cdot\}$ denote the expectation and the variance operators with respect to the sampling design. Furthermore, we suppose that the random hot-deck imputation described in Section 3.1 is used. Let $E_I\{\cdot\}$ and $V_I\{\cdot\}$ denote the expectation and the variance operators with respect to the random imputation.

The overall variance of the imputed estimator of change $\widehat{\Delta}^I$ is given by

$$V(\widehat{\Delta}^I) \; = \; A \; + \; B \; + \; C, \tag{10}$$

which is an overall three stage variance, where

$$A = E_r\{V_d\{E_I\{\widehat{\Delta}^I | S, R\} | R\}\}, \tag{11}$$

$$B = E_r\{E_d\{V_I\{\widehat{\Delta}^I | S, R\} | R\}\}, \tag{12}$$

$$C = V_r\{E_d\{E_I\{\widehat{\Delta}^I | S, R\} | R\}\}, \tag{13}$$

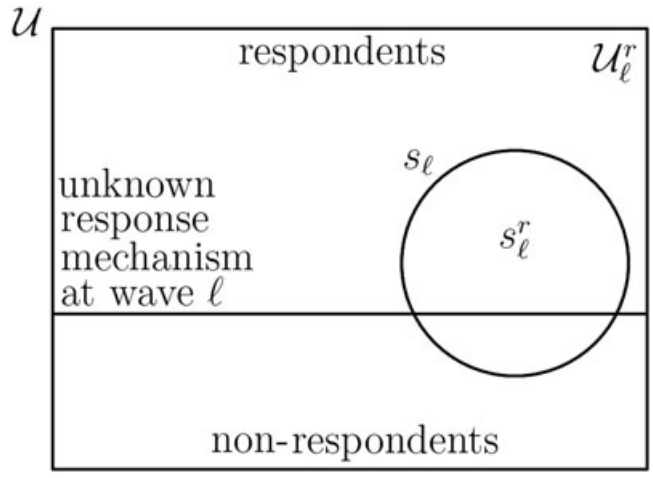

Figure 2. *Non-response at wave* $\ell = 1, 2$.

with $S = \{s_1, s_2\}$, $R = \{s_1^r, s_2^r\}$. The variance (10) includes the effect of the response mechanism, the sampling design and the imputation. We now focus on each term.

4.0.1 The term A

Turning to the term A given by Equation (11), as $E_I\{e_{\ell;k}|S, R\} = 0$, from Equation (7), we have that $E_I\{y^*_{\ell;k}|S, R\} = \widehat{\mu}_\ell^r$. Hence, from Equations (5) and (6), it can be shown that $E_I\{\widehat{\tau}_\ell^I|S, R\} = \widehat{N}_\ell \widehat{\tau}_\ell^r / \widehat{N}_\ell^r$. Thus,

$$E_I\{\widehat{\Delta}^I|S, R\} = \widehat{N}_2 \frac{\widehat{\tau}_2^r}{\widehat{N}_2^r} - \widehat{N}_1 \frac{\widehat{\tau}_1^r}{\widehat{N}_1^r}, \tag{14}$$

where $\widehat{N}_\ell = \sum_{k\in\tilde{s}} \breve{z}_{\ell;k}, (\ell = 1, 2)$ is an estimator of N. Here, $\breve{z}_{\ell;k} = \pi^{-1}_{\ell;k}\, z_{\ell;k}$. Note that $E_I\{\widehat{\Delta}^I|S, R\} = f(\widehat{\boldsymbol{\tau}})$, where $f(\cdot)$ is a function of estimated totals $\widehat{\boldsymbol{\tau}} = (\widehat{\boldsymbol{\tau}}_1^\top, \widehat{\boldsymbol{\tau}}_2^\top)^\top$, with

$$\widehat{\boldsymbol{\tau}}_\ell = \left(\widehat{N}_\ell, \widehat{N}_\ell^r, \widehat{\tau}_\ell^r\right)^\top, \tag{15}$$

is a vector of Narain–Horvitz–Thompson totals (Narain, 1951; Horvitz & Thompson, 1952). Using a Taylor linearisation (e.g. Särndal, Swensson & Wretman, 1992, §5.5, 5.7), we have that

$$E_I\{\widehat{\Delta}^I|S, R\} - \Delta \;\doteq\; \boldsymbol{\nabla}(\boldsymbol{\tau})^\top (\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}), \tag{16}$$

where

$$\boldsymbol{\nabla}(\boldsymbol{\tau}) = \left(\frac{-\tau_1^r}{N_1^r}, \frac{N\tau_1^r}{(N_1^r)^2}, \frac{-N}{N_1^r}, \frac{\tau_2^r}{N_2^r}, \frac{-N\tau_2^r}{(N_2^r)^2}, \frac{N}{N_2^r}\right)^\top, \tag{17}$$

is the gradient of $f(\boldsymbol{\tau})$ at $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^\top, \boldsymbol{\tau}_2^\top)^\top$, with

$$\boldsymbol{\tau}_\ell = \left(N, N_\ell^r, \tau_\ell^r\right)^\top. \tag{18}$$

Here, τ_ℓ^r is the population total of the variable of interest over the respondents at wave ℓ; and N_ℓ^r is the total number of respondents in the population at wave ℓ, ($\ell = 1, 2$).

The expression (16) implies the following approximation

$$A = E_r\{V_d\{E_I\{\widehat{\Delta}^I|S, R\}|R\}\} \;\simeq\; \boldsymbol{\nabla}(\boldsymbol{\tau})^\top\, E_r\{\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)\}\, \boldsymbol{\nabla}(\boldsymbol{\tau}), \tag{19}$$

where $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$ is the covariance matrix of the vector $\widehat{\boldsymbol{\tau}}$ with respect to the design. Thus, an approximately design-unbiased estimator for (19) is given by

$$\widehat{A} = \widehat{V}_d\{E_I\{\widehat{\Delta}^I|S, R\}|R\} \;=\; \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}})^\top\, \widehat{\boldsymbol{V}}_d(\widehat{\boldsymbol{\tau}}|R)\, \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}}), \tag{20}$$

where $\widehat{\boldsymbol{V}}_d(\widehat{\boldsymbol{\tau}}|R)$ is the approximately design-unbiased estimator of the covariance $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$. The estimator $\widehat{\boldsymbol{V}}_d(\widehat{\boldsymbol{\tau}}|R)$ is defined in Equation (28). Note that in Equation (20), the $a_{\ell;k}$s are treated as fixed quantities, as $\widehat{V}_d\{E_I\{\widehat{\Delta}^I|S, R\}|R)\}$ is a conditional variance given R.

4.0.2 The term B

We now turn to the term B given by expression (12). Under Assumption 1, we have

$$\begin{aligned} B &= E_r\left\{E_d\left\{\sum_{\ell=1}^{2} V_I\left\{\widehat{\tau}_\ell^I|S, R\right\}\right\}\right\} \\ &= E_r\left\{E_d\left\{\sum_{h=1}^{H}\sum_{\ell=1}^{2} V_I\{y_{\ell;k}^*|S, R\}\sum_{k\in\tilde{s}}\frac{z_{\ell;k}^{(h)}}{\pi_{\ell;k}^2}(1-a_{\ell;k})\right\}\right\}, \end{aligned} \tag{21}$$

with $V_I\{y_{\ell;k}^*|S, R\} = V_I\{e_{\ell;k}|S, R\} = \sum_{k\in\tilde{s}} a_{\ell;k}\; p_{\ell;k}\, e_{\ell;k}^2$ as $E_I\{e_{\ell;k}|S, R\} = 0$. Note that, under deterministic mean imputation, we have $V_I\{\widehat{\Delta}^I|S, R\} = 0$. This is also the case for regression imputation.

The expression (21) implies that an unbiased estimator of expression (12) is given by

$$\widehat{B} = \widehat{V}_I\{\widehat{\Delta}^I|S, R\} = \sum_{h=1}^{H}\sum_{\ell=1}^{2} V_I\{y_{\ell;k}^*|S, R\}\sum_{k\in\tilde{s}}\frac{z_{\ell;k}^{(h)}}{\pi_{\ell;k}^2}(1-a_{\ell;k})\cdot \tag{22}$$

Note that we use the same notation for the random variables $e_{\ell;k}$ and their observed values.

4.0.3 The term C

We now turn to the term C given by Equation (13). Let

$$\Upsilon_\ell \;=\; E_d\{E_I\{\widehat{t}_\ell^I|S, R\}|R\} \;=\; N_\ell\,\frac{\tau_\ell^r}{N_\ell^r}.$$

We have from Equation (4) that $E_d\{E_I\{\widehat{\Delta}^I|S, R\}|R\} = \Upsilon_2 - \Upsilon_1$. Hence, from Equation (13),

$$C \;=\; V_r\,(\Upsilon_1) \;+\; V_r\,(\Upsilon_2) \;-\; 2\,Corr_r(\Upsilon_1, \Upsilon_2)\,\sqrt{V_r\,(\Upsilon_1)\;V_r\,(\Upsilon_2)}, \tag{23}$$

where $Corr_r(\cdot)$ denotes the correlation operator with respect to the response mechanism. Here, $V_r(\Upsilon_\ell) = V_r\{E_d\{E_I\{\widehat{t}_\ell^I|S,R\}|R\}\}$ is the cross-sectional variance for wave ℓ under the response mechanism given the random imputation and the sampling design.

We recall from Shao & Steel (1999, pp. 256, 257) that the cross-sectional variances $V_r(\Upsilon_\ell)$ are of order $\mathcal{O}(N_\ell)$ implying $C = \mathcal{O}(N_\ell)$, because the correlation $Corr_r(\Upsilon_1, \Upsilon_2)$ is between -1 and 1 in Equation (23). Given standard assumptions for linearised variances of functions of totals (e.g. Robinson & Särndal, 1983; Särndal, *et al.*, 1992, secs. 5.5, 5.7), the linearised version of the term A from Equation (19) is of order $\mathcal{O}(N_\ell^2/n)$, which is the dominant term of the overall variance $V(\widehat{\Delta}^I)$ from Equation (10). Furthermore, $C/A = \mathcal{O}(n/N_\ell)$. Thus, for negligible n/N_ℓ, the contribution of C to Equation (10) is negligible (e.g. Haziza, 2009, pp. 238-240). Thus,

$$V(\widehat{\Delta}^I) \doteq A + B. \tag{24}$$

From Equation (23), we note that the response mechanisms can be dependent between waves. In other words, even if $a_{1;k}$ and $a_{2;k}$ are dependent, Equation (24) still holds.

5 The Proposed Variance Estimator

We proposed to estimate the variance of the imputed estimator (4) by

$$\widehat{V}\{\widehat{\Delta}^I\} = \widehat{A} + \widehat{B} = \widehat{V}_d\{E_I\{\widehat{\Delta}^I|S,R\}|R\} + \widehat{V}_I\{\widehat{\Delta}^I|S,R\}, \tag{25}$$

where $\widehat{V}_d\{E_I\{\widehat{\Delta}^I|S,R\}|R\}$ and $\widehat{V}_I\{\widehat{\Delta}^I|S,R\}$ are defined by expressions (20) and (22). In Section 6, we propose a multivariate (or general) linear regression model to estimate the covariance matrix $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$ involved in the estimator (20) (Berger & Priam, 2010; 2016). The proposed estimator (25) is an approximately unbiased estimator of the variance $V(\widehat{\Delta}^I)$ given by Equation (10), as the overall expectation of the estimator (25) is given by

$$\begin{aligned} E_r\{E_d\{E_I\{\widehat{V}(\widehat{\Delta}^I)|S,R\}|R\}\} &= E_r\{E_d\{E_I\{\widehat{V}_d(E_I\{\widehat{\Delta}^I|S,R\}|R)|S,R\}|R\}\} \\ &\quad + E_r\{E_d\{E_I\{V_I\{\widehat{\Delta}^I|S,R\}\}|R\}\} \\ &\doteq E_r\{V_d\{E_I\{\widehat{\Delta}^I|S,R\}|R\}\} \\ &\quad + E_r\{E_d\{V_I\{\widehat{\Delta}^I|S,R\}|R\}\} \\ &\doteq V(\widehat{\Delta}^I), \end{aligned}$$

by using the result (24) and the fact that the estimator (20) does not depend on the $e_{\ell;k}$ for $\ell = 1, 2$.

The advantages of the proposed variance estimator (25) are that it is approximately unbiased under the unknown response mechanisms and that it does not involve the estimation of the response probabilities. Moreover, note that the estimator (25) can be generalised for other types of imputation, as long as $E_I\{\widehat{\Delta}^I|S,R\}$ is a function of Narain–Horvitz–Thompson estimators of totals. In this situation, the gradient (17) would have a different expression depending on the imputation considered.

6 Estimation of the Covariance Using the Multivariate Regression Approach

We derive an estimator for the covariance matrix $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$ in Equation (19) under the rotation sampling design. Note that this covariance is not straightforward to estimate because it involves

a covariance between all the components of $\widehat{\boldsymbol{\tau}}$ defined from different overlapping samples, s_1 and s_2. Several estimators can be used (e.g. Kish, 1965; Tam, 1984; Holmes & Skinner, 2000; Nordberg, 2000; Berger, 2004; Qualité and Tillé, 2008; Wood, 2008; Goga, Deville & Ruiz-Gazen, 2009; Münnich and Zins, 2011; Knottnerus and van Delden, 2012). We propose to use a multivariate (or general) linear regression model to estimate this covariance matrix (Berger & Priam, 2010; 2016).

Consider the following $\tilde{n} \times 6$ matrix $\breve{\boldsymbol{Y}}_{(\tilde{n}\times 6)} = (\breve{\boldsymbol{y}}_1, \ldots, \breve{\boldsymbol{y}}_k, \ldots, \breve{\boldsymbol{y}}_{\tilde{n}})^\top$, where $\tilde{n} = \#(s_1 \cup s_2)$, $\breve{\boldsymbol{y}}_k = (\breve{\boldsymbol{y}}_{1k}, \breve{\boldsymbol{y}}_{2k})$ and

$$\breve{\boldsymbol{y}}_{\ell k} \;=\; (\breve{z}_{\ell;k}, \breve{a}_{\ell;k}, \breve{y}_{\ell;k})^\top \tag{26}$$

with $\breve{z}_{\ell;k} = \pi^{-1}_{\ell;k}\, z_{\ell;k}$. Here, $z_{\ell;k}$, $\breve{a}_{\ell;k}$ and $\breve{y}_{\ell;k}$ are defined by the expressions (3), (9) and (8), with $\ell = 1, 2$. Consider the following multivariate (general) regression model

$$\breve{\boldsymbol{Y}} \;=\; \boldsymbol{Z}_s\, \boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{27}$$

where $\boldsymbol{\alpha}$ is a 3×6 matrix of regression parameters, the residuals $\boldsymbol{\varepsilon}$ have a 6×6 covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{Z}_s$ is a $\tilde{n} \times 3$ design matrix, which specifies the fixed-size constraints of the rotation design. The matrix $\boldsymbol{Z}_s$ is defined by $\boldsymbol{Z}_s = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k, \ldots, \boldsymbol{z}_{\tilde{n}})^\top$ with

$$\boldsymbol{z}_k = \left(z^{(1)}_{1;k}, \ldots, z^{(H)}_{1;k}, z^{(1)}_{2;k}, \ldots, z^{(H)}_{2;k}, z^{(1)}_{1;k} \times z^{(1)}_{2;k}, \ldots, z^{(H)}_{1;k} \times z^{(H)}_{2;k}\right)^\top.$$

With a single stratum, $\boldsymbol{z}_k = (z_{1;k}, z_{2;k}, z_{1;k} \times z_{2;k})^\top$. The model (27) belongs to the class of general linear model. In fact, the model (27) is also a multivariate analysis of variance model, as the covariates are all dummy variables. Note that we have $\sum_{k\in\tilde{s}} z^{(h)}_{\ell;k} = n_h$, $\sum_{k\in\tilde{s}} z^{(h)}_{1;k}\, z^{(h)}_{2;k} = n_{h;12}$, where n_h is the sample size of the h-th stratum and $n_{h;12}$ is the number of waves 1 and 2 units, which belong to the h-th stratum. These sums are the sample size restriction imposed on the samples. Thus, by using the design variables as covariates in the model (27), we implicitly condition on them. This takes into account the fixed size constraints in the estimation of the covariance (see Berger and Priam, 2010, 2016). Note that the model (27) includes the within strata interactions between the variable $z^{(h)}_{1;k}$ and $z^{(h)}_{2;k}$. These interactions capture the rotation of the sampling design, which is represented by the constraint $\sum_{k\in\tilde{s}} z^{(h)}_{1;k}\, z^{(h)}_{2;k} = n_{h;12}$.

To estimate $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$, Berger & Priam (2010, 2016) proposed the estimator

$$\widehat{\boldsymbol{V}}_d(\widehat{\boldsymbol{\tau}}|R) \;=\; \widehat{\boldsymbol{D}}^\top\, \widehat{\boldsymbol{\Sigma}}\, \widehat{\boldsymbol{D}}, \tag{28}$$

where the matrix $\widehat{\boldsymbol{\Sigma}}$ is the *Ordinary Least Squares* residual covariance matrix estimated from the model (27) and $\widehat{\boldsymbol{D}}$ is a diagonal matrix with the diagonal elements: $\{\widehat{V}(\widehat{\tau}_q|R)/\widehat{\Sigma}_{qq}\}^{1/2}$, where $\widehat{V}(\widehat{\tau}_q|R)$ is a design-based variance estimator of the q-th component of $\widehat{\boldsymbol{\tau}}$ and $\widehat{\Sigma}_{qq}$ is the q-th diagonal component of $\widehat{\boldsymbol{\Sigma}}$. Any approximately unbiased variance estimator can be used to calculate $\widehat{V}(\widehat{\tau}_q|R)$. Note that the variance matrix (28) is positive definite, as $\widehat{\boldsymbol{\Sigma}}$ is always positive definite. Hence, the proposed variance estimator (25) is always positive.

Using Berger & Priam's (2010, 2016) results, we have that the estimator (28) is an approximately design unbiased estimator for $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}}|R)$ when the finite population corrections are negligible, even when model (27) does not fit the data (Berger & Priam, 2010; 2016).

In a series of simulations based on the Swedish Labour Force Survey, Andersson *et al.* (2011a) & Andersson *et al.* (2011b) showed that under full response, the estimator that was

proposed by Berger (2004) gives more accurate estimates than the standard variance estimators (e.g. Tam, 1984; Qualité & Tillé, 2008) when we are interested in change between strata domains. Berger & Priam (2010, 2016) showed that the estimator that was proposed by Berger (2004) reduces to the variance (28) when the sampling fractions are negligible.

7 Multiple Imputation-classes

We now consider the situation when the response mechanism is not uniform. Hence, instead of Assumption 1, we have the following assumption:

Assumption 2 (**multiple imputation-classes**). *The population $\mathcal{U}$ can be divided into C_1 imputation classes for wave 1 and C_2 imputation classes for wave 2. The response probability for the variable of interest is uniform within wave-class combinations, and it is strictly positive. The units' responses within and across classes are independent; and responses between waves can be dependent.*

Assumption 2 holds under MAR response mechanism (given the set of classes).

Let $\mathcal{U}_{\ell;1},\ldots,\mathcal{U}_{\ell;c},\ldots,\mathcal{U}_{\ell;C_\ell}$ be the C_ℓ class of wave ℓ. Let $b_{\ell;k}^{(c)}$ be the following imputation classes indicator for wave ℓ.

$$b_{\ell;k}^{(c)} \;=\; \begin{cases} 1 \text{ if } k \in \mathcal{U}_{\ell;c}, \\ 0 \text{ otherwise,} \end{cases}$$

where $c = 1,\ldots,C_\ell$. The random hot-deck imputed values $y_{\ell;k}^{I}$ from Equation (6) are now given by $y_{\ell;k}^{*} = \sum_{c=1}^{C} b_{\ell;k}^{(c)}\,(\widehat{\mu}_\ell^{r(c)} + e_{\ell;k}^{(c)})$, where $e_{\ell;k}^{(c)} = y_{\ell;j} - \widehat{\mu}_\ell^{r(c)}$ instead of expression (7), where j is a donor selected with-replacement with probabilities $p_{\ell;k} = b_{\ell;k}^{(c)}\breve{a}_{\ell;k}/\widehat{N}_\ell^{r(c)}$ from the sample $s_\ell^{r(c)}$ of respondents of class c. Here, $s_\ell^{r(c)} = \{k : z_{\ell;k} = 1, a_{\ell;k} = 1, b_{\ell;k}^{(c)} = 1\}$ and

$$\begin{aligned} \widehat{\mu}_\ell^{r(c)} &= \frac{\widehat{\tau}_\ell^{r(c)}}{\widehat{N}_\ell^{r(c)}}, \\ \widehat{\tau}_\ell^{r(c)} &= \sum_{k\in\tilde{s}} b_{\ell;k}^{(c)}\,\breve{y}_{\ell;k}, \\ \widehat{N}_\ell^{r(c)} &= \sum_{k\in\tilde{s}} b_{\ell;k}^{(c)}\,\breve{a}_{\ell;k}, \end{aligned}$$

are respectively the estimates of the respondents' means, totals and number of respondents for each wave-class combination ($c = 1,2,\ldots,C_\ell; \ell = 1,2$). Note that under multiple-classes deterministic mean imputation, we have $e_{\ell;k}^{(c)} = 0$.

With multiple imputation-classes, the population variance $V(\widehat{\Delta}^I)$ from Equation (10) is different. In term A from expression (11), Equations (14), (15), (17) and (18) are replaced by

$$E_I\{\widehat{\Delta}^I|S,R\} = \widehat{N}_2\,\frac{\sum_{c=1}^{C_2}\widehat{\tau}_2^{r(c)}}{\sum_{c=1}^{C_2}\widehat{N}_2^{r(c)}} \;-\; \widehat{N}_1\,\frac{\sum_{c=1}^{C_1}\widehat{\tau}_1^{r(c)}}{\sum_{c=1}^{C_1}\widehat{N}_1^{r(c)}}, \tag{29}$$

$$\widehat{\boldsymbol{\tau}}_\ell = \left(\widehat{N}_\ell,\, \widehat{N}_\ell^{r(1)},\,\ldots,\,\widehat{N}_\ell^{r(C_1)},\,\widehat{\tau}_\ell^{(1)},\,\ldots,\,\widehat{\tau}_\ell^{r(C_2)}\right)^\top, \tag{30}$$

$$\nabla(\boldsymbol{\tau}) = \left(\frac{-\tau_1^r}{N_1^r}, \underbrace{\frac{N\tau_1^r}{(N_1^r)^2}, \ldots,}_{C_1 \text{ times}} \underbrace{\frac{-N}{N_1^r}, \ldots,}_{C_1 \text{ times}} \frac{\tau_2^r}{N_2^r}, \underbrace{\frac{-N\tau_2^r}{(N_2^r)^2}, \ldots,}_{C_2 \text{ times}} \underbrace{\frac{N}{N_2^r}, \ldots}_{C_2 \text{ times}} \right)^\top, \tag{31}$$

$$\boldsymbol{\tau}_\ell = \left(N,\, N_\ell^{r(1)}, \ldots,\, N_\ell^{r(C_\ell)},\, \tau_\ell^{r(1)}, \ldots,\, \tau_\ell^{r(C_\ell)} \right)^\top, \tag{32}$$

with $\tau_\ell^r = \sum_{c=1}^{C_\ell} \tau_\ell^{r(c)}$ and $N_\ell^r = \sum_{c=1}^{C_\ell} N_\ell^{r(c)}$, where $\tau_\ell^{r(c)}$ and $N_\ell^{r(c)}$ are, respectively, the respondents' population totals of the variable y_k and the number of respondents at each wave-class combination, $(c = 1, 2, \ldots, C_\ell; \ell = 1, 2)$.

Under Assumption 2, we have that

$$\widehat{V}_I\{\widehat{\Delta}^I | S, R\} = \sum_{\ell=1}^{2} \sum_{c=1}^{C_\ell} \sum_{k \in \tilde{s}} a_{\ell;k}\, p_{\ell;k}\, \{e_{\ell;k}^{(c)}\}^2 \sum_{k' \in \tilde{s}} \frac{b_{\ell;k'}^{(c)}\, z_{\ell;k'}}{\pi_{\ell;k'}^2} (1 - a_{\ell;k'}) \cdot \tag{33}$$

Thus, the estimator proposed is given by expression (25), where $\widehat{V}_d\{E_I\{\widehat{\Delta}^I | S, R\} | R\}$ and $V_I\{\widehat{\Delta}^I | S, R\}$ are now given by Equations (20) and (33). The quantities $E_I\{\widehat{\Delta}^I | S, R\}$, $\widehat{\boldsymbol{\tau}}_\ell$, $\nabla(\boldsymbol{\tau})$ and $\boldsymbol{\tau}_\ell$ are now given by Equations (29), (30), (31) and (32).

As in Section 6, the covariance matrix $\boldsymbol{V}_d(\widehat{\boldsymbol{\tau}} | R)$ in Equation (19) can be estimated using a multivariate (or general) linear regression model. With multiple imputation-classes, the model (27) now uses a $\tilde{n} \times (2 + 2C_1 + 2C_2)$ matrix $\breve{\boldsymbol{Y}} = (\breve{\boldsymbol{y}}_1, \ldots, \breve{\boldsymbol{y}}_k, \ldots, \breve{\boldsymbol{y}}_{\tilde{n}})^\top$, where $\breve{\boldsymbol{y}}_k = (\breve{\boldsymbol{y}}_{1k}, \breve{\boldsymbol{y}}_{2k})$, with

$$\breve{\boldsymbol{y}}_{\ell k} = \left(\breve{z}_{\ell;k}, b_{\ell;k}^{(1)} \breve{a}_{\ell;k}, \ldots, b_{\ell;k}^{(C_\ell)} \breve{a}_{\ell;k}, b_{\ell;k}^{(1)} \breve{y}_{\ell;k}, \ldots, b_{\ell;k}^{(C_\ell)} \breve{y}_{\ell;k} \right)^\top$$

replacing the expression (26). Now, $\boldsymbol{\alpha}$ is a $3 \times (2 + 2C_1 + 2C_2)$ matrix, and $\boldsymbol{\Sigma}$ is a $(2 + 2C_1 + 2C_2) \times (2 + 2C_1 + 2C_2)$ matrix.

The variance estimator (28) is based on the implicit assumption that the class indicator $b_{\ell;k}^{(c)}$ are constants defined at population level. In other words, $b_{\ell;k}^{(c)}$ do not depend on the samples selected. For example, this is the case when we use the strata as imputation classes; that is, the $b_{\ell;k}^{(c)}$ are the strata indicators $z_{\ell;i}^{(h)}$ defined by (2). In Section 8.2, we will consider the situation when the $b_{\ell;k}^{(c)}$ are defined form the sampled data; that is, $b_{\ell;k}^{(c)}$ are random variable because they depend on the samples selected. In this case, the variance estimator (28) does not take the randomness of $b_{\ell;k}^{(c)}$ into account. In the simulation study of Section 8.2, we did not observe major impacts of this randomness.

8 Simulation Study

8.1 Labour Force Population

We use the *Labour Force Population* dataset from Valliant, Dorfman & Royall (2000, Appendix B.5) available at the John Wiley worldwide website. The dataset is duplicated 50 times to obtain a large population suitable for different levels of rotation and small sampling fractions in the sampling design. We consider the following two variables: the weekly wages and the hours worked per week (HW). The units with the value 99 for the weekly wage and

Table 1. *RB, RRMSE and Coverage of 95% confidence interval of the variance estimators. Hot-deck imputed point estimator* $\hat{\Delta}^I$. $\pi_{1;k} = n/N$.

				RB		RRMSE		Coverage	
		g	f	Prop.	Naïve	Prop.	Naïve	Prop.	Naïve
$q_{1;k}$	$q_{2;k}$	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
0.70	0.86	40	0.5	-2.8	-33.8	15.5	35.3	95.0	88.7
			1.0	-0.7	-32.3	11.2	33.2	94.8	89.3
			1.5	-0.4	-32.1	9.1	32.7	94.7	89.5
			2.0	-2.7	-33.7	8.2	34.1	94.6	88.8
		60	0.5	-1.8	-31.3	17.6	33.7	94.7	89.1
			1.0	-1.2	-30.9	12.5	32.2	94.8	89.7
			1.5	-1.1	-30.8	10.2	31.7	94.7	89.2
			2.0	0.0	-30.1	8.7	30.8	94.8	89.9
		80	0.5	-1.8	-28.8	20.1	32.7	94.7	89.8
			1.0	-0.4	-27.5	14.4	29.8	95.0	90.4
			1.5	-0.4	-27.5	11.6	29.0	95.0	90.5
			2.0	-2.2	-29.0	10.0	30.0	94.6	90.0
		95	0.5	-1.8	-25.3	22.8	31.9	94.8	90.8
			1.0	-1.9	-25.5	16.0	29.0	94.5	90.8
			1.5	-0.9	-24.8	13.1	27.2	94.8	90.7
			2.0	-1.6	-25.3	11.2	27.0	94.7	90.9
0.90	0.92	40	0.5	-0.7	-15.9	14.5	20.2	94.8	92.9
			1.0	0.2	-15.2	10.2	17.6	95.3	93.2
			1.5	-2.1	-17.2	8.5	18.5	94.7	92.6
			2.0	-0.7	-15.9	7.2	17.1	95.1	92.8
		60	0.5	0.4	-14.4	17.2	21.0	94.9	93.2
			1.0	0.2	-14.6	12.1	18.2	94.9	92.9
			1.5	0.6	-14.2	9.9	16.7	95.1	93.1
			2.0	0.0	-14.8	8.5	16.7	94.8	92.7
		80	0.5	-2.2	-15.3	21.4	25.5	94.7	93.0
			1.0	-2.0	-15.0	15.0	20.8	95.0	93.1
			1.5	-0.2	-13.7	12.3	18.2	94.8	92.9
			2.0	-1.0	-14.4	10.7	17.7	94.6	92.9
		95	0.5	-2.9	-13.4	27.7	33.0	94.5	92.9
			1.0	-2.4	-13.2	19.4	24.8	94.5	93.2
			1.5	-1.1	-12.0	15.9	21.3	95.1	93.6
			2.0	-0.9	-12.0	13.8	19.3	94.9	93.5

999 for the hours worked per week were removed from the population frame. These units were not treated as missing. We obtain a population frame of size $N = 23\,550$. The target variables $y_{1;k}$ and $y_{2;k}$ are given by

$$\begin{aligned} y_{1;k} &= Weekly\ wages, \\ y_{2;k} &= y_{1;k} + \sqrt{y_{1;k}} + \psi_k, \end{aligned}$$

where ψ_k denotes randomly generated values according to a normal distribution $N(0, 5^2)$. The true absolute change between the two wave totals is given by $\Delta = 377\,960.66$. We estimate Δ by the hot-deck imputed point estimator $\widehat{\Delta}^I$ defined by Equation (4). The first wave sample s_1 is selected using the Rao (1965) and Sampford (1967) unequal probability sampling design. We consider the following two scenarios for the inclusion probabilities: $\pi_{1;k}$ are constant ($\pi_{1;k} = n/N$), and $\pi_{1;k}$ are proportional to the variable *hours worked per week*, which has values all larger than five. We consider that we have a single stratum.

For the second wave sample s_2, we select a simple random sample of n_{12} units taken from s_1, where $g = n_{12}/n = \{0.40, 0.60, 0.80, 0.95\}$, and a sample of $n - n_{12}$ units from $\mathcal{U} \setminus s_1$

Table 2. *RB, RRMSE and Coverage of 95% confidence interval of the variance estimators. Hot-deck imputed point estimator* $\hat{\Delta}^I$. $\pi_{1;k} \propto HW$.

				RB		RRMSE		Coverage	
$q_{1;k}$	$q_{2;k}$	g (%)	f (%)	Prop. (%)	Naïve (%)	Prop. (%)	Naïve (%)	Prop. (%)	Naïve (%)
0.70	0.86	40	0.5	-1.6	-29.1	32.3	52.7	94.3	88.7
			1.0	-2.5	-29.9	23.3	42.8	93.5	87.3
			1.5	-3.4	-30.5	19.0	39.7	93.0	86.8
			2.0	-1.5	-29.4	16.4	36.7	92.4	86.5
		60	0.5	-2.0	-27.7	36.2	57.1	94.3	89.3
			1.0	-1.2	-27.5	26.1	44.0	94.2	88.9
			1.5	-0.8	-27.1	21.4	39.1	94.1	88.9
			2.0	-0.7	-27.5	18.2	36.4	93.5	87.8
		80	0.5	-0.1	-25.6	40.7	59.2	94.8	90.4
			1.0	0.0	-25.2	29.3	45.5	94.9	89.7
			1.5	-0.4	-25.1	23.6	40.4	94.8	90.2
			2.0	-0.6	-25.8	20.4	37.1	94.5	89.7
		95	0.5	-1.5	-24.3	43.9	63.6	94.5	90.8
			1.0	0.4	-22.9	31.7	48.5	95.2	91.3
			1.5	0.5	-23.4	26.0	41.4	94.9	91.2
			2.0	-0.8	-24.3	22.3	38.1	94.9	90.6
0.90	0.92	40	0.5	-0.5	-15.5	34.3	51.2	94.1	91.7
			1.0	-1.5	-15.6	23.5	37.7	93.1	90.4
			1.5	-0.5	-14.8	19.9	33.1	92.9	90.2
			2.0	-1.7	-16.0	16.9	29.7	91.8	88.9
		60	0.5	-0.1	-14.2	41.2	61.1	94.3	92.3
			1.0	-2.4	-15.3	29.2	48.1	93.8	91.6
			1.5	0.2	-13.4	23.6	38.1	93.9	91.3
			2.0	-1.0	-14.4	20.5	34.3	93.0	90.6
		80	0.5	-0.3	-12.8	51.9	78.6	94.5	93.1
			1.0	-0.3	-11.7	36.3	59.9	94.7	93.1
			1.5	-1.3	-12.8	29.3	49.3	94.2	92.0
			2.0	-0.4	-12.4	25.6	42.1	94.2	92.1
		95	0.5	-0.8	-11.5	64.3	99.0	94.7	94.4
			1.0	-1.9	-11.5	44.0	71.4	94.3	93.5
			1.5	-1.5	-11.9	35.5	58.8	94.6	93.2
			2.0	-0.8	-11.3	30.7	49.7	94.6	93.4

selected with probabilities proportional to $\pi_{2;k} = \pi_{1;k}/(1-\pi_{1;k})$. We have that $\pi_{2;k} \doteq \pi_{1;k}$ Berger & Priam (2016).

Let $a_{1;k} = 1$ if $u_{1;k} \leq q_1$ and $a_{1;k} = 0$ otherwise, where q_1 is a fixed quantity, which specify the response rate at wave 1, and $u_{1;k}$ are independent uniform random variables $U(0,1)$. Let $a_{2;k} = 1$ if $u_{2;k} \leq 0.95\, a_{1;k} + 0.65\,(1 - a_{1;k})$ and $a_{2;k} = 0$ otherwise, where $u_{2;k}$ are independent uniform random variables $U(0,1)$. Note that $a_{1;k}$ and $a_{2;k}$ are dependent because a respondent at wave 1 is more likely to be also a respondent on wave 2. The items non-response are imputed using random hot-deck as described in Subsections 3.1 and 3.2. We consider that we have a single imputation class. A new set of respondents $(a_{1;k}, a_{2;k})$ is generated randomly before each selection of s_1 and s_2.

For each simulation, 10 000 samples are selected to compute the following: the empirical relative bias RB $= \text{Bias}(\widehat{\text{var}}(\hat{\Delta}^I))/\text{var}(\hat{\Delta}^I)$, where $\text{Bias}(\widehat{\text{var}}(\hat{\Delta}^I)) = \text{E}(\widehat{\text{var}}(\hat{\Delta}^I)) - \text{var}(\hat{\Delta}^I)$; the empirical relative root mean squared error RRMSE $= (\text{MSE}(\widehat{\text{var}}(\hat{\Delta}^I)))^{1/2}/\text{var}(\hat{\Delta}^I)$; and the coverage of the 95% confidence interval $\hat{\Delta}^I \pm 1.96\,\widehat{\text{var}}(\hat{\Delta}^I)^{1/2}$. The term $\text{var}(\hat{\Delta}^I)$ denotes the empirical variance computed from the 10 000 observed values of $\hat{\Delta}^I$. Computations were

performed in R (R Core Team, 2015) using some routines from the R packages 'sampling' (Tillé & Matei, 2013) and 'samplingVarEst' (Escobar & Barrios, 2014). We compare the proposed estimator $\widehat{V}(\widehat{\Delta}^I)$ from (25) versus a naïve approach, which consists in treating the imputed values as real values. Note that there is no other competitor for the proposed approach, as design-based variance estimators for imputed change estimators is non-existent in the literature.

Tables 1 and 2 give the RB, the RRMSE and the coverage for different values of the overlapping fraction g between waves. In Table 1, $\pi_{1;k} = n/N$, and, in Table 2, $\pi_{1;k}$ are proportional to the variable *hours worked per week*.

The proposed approach gives negligible RB. As expected, the naïve approach tends to severely underestimate the variance, in particular, when the fraction of non respondents is large, that is, when $q_{1;k}$ is small. Furthermore, by comparing Tables 1 and 2, we observe smaller RB with unequal inclusion probabilities.

The proposed approach has smaller RRMSE than the naïve approach. However, with unequal probabilities, we observe larger RRMSE. The coverage of the proposed approach is closer to 95%. The coverage of the naïve approach is lower because of the under-estimation of the variance.

8.2 Missing Not at Random Response and Multiple Imputation-classes

Four variables, y_1, y_2, x_1, x_2 and w_1, are generated from a multivariate normal distribution with means 20, 10, 20, 10 and 20. All the variables have the same variance equals to 5. The correlation between y_1 and y_2 is either $\rho(y_1, y_2) = 0.7$ or $\rho(y_1, y_2) = 0.9$. The other correlations are $\rho(y_\ell, x_{\ell'}) = \rho(y_\ell, w_1) = 0.7$ and $\rho(x_\ell, x_{\ell'}) = \rho(x_\ell, w_1) = 0.5$ ($\ell \neq \ell'$). Wave 1 variables are y_1, x_1 and w_1. Wave 2 variables are y_2 and x_2. We generate $N = 20\,000$ values for each variables.

The values $y_{1;k}$ and $y_{2;k}$ are the values of the variables y_1 and y_2. The parameter of interest is the absolute change between means: $\Delta_\mu = \Delta/N$. The imputed estimator is $\widehat{\Delta}^I_\mu = \widehat{\Delta}^I/N$.

The sample s_1 is a randomised systematic sample with first-order inclusion probabilities $\pi_{1;k}$ proportional to $w_{1;k}$, where $w_{1;k}$ denotes the k-th value of w_1. The sample s_2 is a simple random sample of n_{12} units selected from s_1 combined with a randomised systematic sample of $n_2 - n_{12}$ units selected without replacement from $U \setminus s_1$ with probabilities proportional to $\pi_{1;k}/(1 - \pi_{1;k})$. We have that $\pi_{2;k} \backsimeq \pi_{1;k}$ (Berger & Priam, 2016). The sample sizes are $n_1 = n_2 = 500$ and $n_{12} = 375$. We consider that we have a single stratum. Ten thousand samples s_1 and s_2 are selected. The Hansen & Hurwitz (1943) variance estimator is used for cross-sectional variance estimation.

We consider hot-deck imputation with multiple imputation-classes as described in Section 7. The number of imputation classes is the same in waves 1 and 2: $C_1 = C_2 = C$. We consider three types of imputation classes.

(i) '*Population imputation classes*': The imputation classes of wave ℓ are C quantile classes based on the variable x_ℓ. The bounds of the classes are the $(100c/C)\%$ quantiles ($c = 1, \ldots, C$) of the population values $\{x_{\ell;k} : k \in U\}$, where $x_{\ell;k}$ denotes the k-th value of x_ℓ.

(ii) '*Sample imputation classes*': The imputation classes of wave ℓ are C quantile classes based on the sample values of the variable x_ℓ. The bounds of the classes are the $(100c/C)\%$ quantiles ($c = 1, \ldots, C$) of the sample values $\{x_{\ell;k} : k \in s_\ell\}$.

(iii) '*Across-waves imputation classes*': For the classes of wave 1, we use C quantile classes based on the sample values of the variable x_1, as in (ii). Wave 2 imputation classes are C quantile classes based on the sample values $\{\tilde{y}_{1;k} : k \in s_2\}$, where

Table 3. *Overall expectation, variance, root-mean squared error (*RMSE*) and coverage of 95% confidence interval based on the estimator proposed. Missing not at random response mechanisms.* $\rho(y_1, y_2)$ *denotes the correlation between the variables of interest.* $N = 20\,000$, $n_1 = n_2 = 500$ *and* $n_{12} = 375$. $\Delta_\mu = \Delta/N$ *and* $\hat{\Delta}^I_\mu = \hat{\Delta}^I/N$.

$\rho(y_1, y_2)$	Imputation	C	Δ_μ	$E(\widehat{\Delta}^I_\mu)$	$V(\widehat{\Delta}^I_\mu)$	$E\{\widehat{V}(\widehat{\Delta}^I_\mu)\}$	RMSE	Coverage (%)
0.7	(i) Population	1	-10.03	-10.09	0.022	0.021	0.0016	92.9
	level	5	-10.03	-10.06	0.017	0.019	0.0029	95.9
		10	-10.03	-10.06	0.016	0.019	0.0031	96.0
		20	-10.03	-10.06	0.016	0.019	0.0036	96.3
	(ii) Sample	5	-10.03	-10.06	0.016	0.019	0.0031	96.2
	level	10	-10.03	-10.06	0.015	0.019	0.0037	96.7
		20	-10.03	-10.06	0.015	0.019	0.0039	96.8
	(iii) Across	5	-10.03	-10.09	0.017	0.019	0.0029	94.5
	waves	10	-10.03	-10.08	0.016	0.019	0.0033	95.3
		20	-10.03	-10.08	0.016	0.019	0.0032	95.2
0.9	(i) Population	1	-9.99	-10.06	0.019	0.019	0.0015	91.8
	level	5	-9.99	-10.03	0.014	0.017	0.0030	95.7
		10	-9.99	-10.03	0.014	0.016	0.0031	96.0
		20	-9.99	-10.03	0.013	0.016	0.0031	96.2
	(ii) Sample	5	-9.99	-10.03	0.013	0.017	0.0034	96.3
	level	10	-9.99	-10.03	0.013	0.016	0.0035	96.2
		20	-9.99	-10.03	0.013	0.016	0.0033	96.3
	(iii) Across	5	-9.99	-10.01	0.013	0.016	0.0032	96.8
	waves	10	-9.99	-10.00	0.013	0.016	0.0037	97.3
		20	-9.99	-10.00	0.013	0.016	0.0035	97.2

$$\tilde{y}_{1;k} = \begin{cases} y^I_{1;k} & \text{for } k \in s_{12}, \\ \widehat{\beta}_0 + \widehat{\beta}_1 x_{2;k} & \text{for } k \in s_2 \setminus s_1. \end{cases} \tag{34}$$

Here, $x_{2;k}$ is the value of the variable x_2 for unit k. The quantity $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the ordinary least square coefficients of the regression $y^I_{1;k} = \beta_0 + \beta_1 x_{2;k}$, with $k \in s_{12}$.

For class (i), the class indicators $b^{(c)}_{\ell;k}$ do not depend on the samples selected. For classes (ii) and (iii), the class indicators $b^{(c)}_{\ell;k}$ depend on the samples. Note that the classes of wave 1 are different from the classes of wave 2, unless $C = 1$.

We consider a '*missing not at random response mechanism*'. The first and second wave response probabilities $q_{1;k}$ and $q_{2;k}$ are given by $q_{\ell;k} = \exp(\eta_{\ell;k})\{1 + \exp(\eta_{\ell;k})\}^{-1}$, where $\eta_{1;k} = 4 - 0.15\, y_{1;k}$ and $\eta_{2;k} = 3 - 0.2\, y_{2;k}$. The resulting response probabilities lie within the range [0.25, 0.95]. We have $a_{\ell;k} = 1$ if $u_{\ell;k} \leq q_{\ell;k}$ and $a_{\ell;k} = 0$ otherwise, where $u_{\ell;k}$ are independent uniform random variables $U(0, 1)$. The resulting response mechanism is MNAR because large $q_{1;k}$ and $q_{2;k}$ are associated with small values of $y_{1;k}$ and $y_{2;k}$. The overall response rates are 73% and 72% for the first and second wave. The correlation between $q_{1;k}$ and $q_{2;k}$ is approximately 0.7. The response probabilities are not constant within the imputation classes. Missing values are generated randomly before each selection of s_1 and s_2.

The simulation results are given in Table 3. Large number of classes reduces the bias of the point estimator. With a single imputation class ($C = 1$), the variance estimator has the smallest bias and is more stable (small root mean square error, RMSE), but with low coverages (92.9% and 91.8%). The low coverages are explained by the largest bias of the point estimator. Note that the point estimator is more precise with $C \geqslant 1$ and $\rho(y_1, y_2) = 0.9$, in term of bias and variance. However, there are only negligible differences between the variance for $C \geqslant 5$. We only notice a decrease in the variance, as C increases, for population level imputation classes with $\rho(y_1, y_2) = 0.9$. For $C \geqslant 5$, we observe a slight positive bias for the variance estimator

and an increase in the RMSE. For population level classes, the RMSE increases with C. The coverage observed are slightly larger than 95% for $C \geqslant 5$. We do not observe significant differences between the imputation classes (i), (ii) and (iii).

The MNAR response mechanism tends to under-represent the large values of the variables of interest, and therefore, the observed correlation between $y_{1;k}$ and $y_{2;k}$ is lower than $\rho(y_1, y_2)$. As a result, the correlation between $\widehat{\tau}_2^I$ and $\widehat{\tau}_1^I$ is slightly under-estimated. This explains the slight positive bias for the variance estimator (Berger, 2004, p. 462). However, this bias is negligible because the coverages of the confidence intervals are of an acceptable order. This bias is only observed for $C \neq 1$. For $C = 1$, the larger variance compensates the bias.

9 Discussion

The proposed variance estimator is applicable for unequal rotating stratified sampling designs when random hot-deck imputation is used at both waves and the sampling fractions are negligible. The proposed variance estimator may be extended in various ways. Point estimators, such as calibration estimators (Huang & Fuller, 1978; Deville & Särndal, 1992), which employ auxiliary population information, may often be expressible as functions of totals. The proposed variance estimator (20) can be modified to accommodate this situation.

The main advantages of the proposed variance estimator are that it is approximately unbiased under the response mechanisms and that it does not require the estimation of the response probabilities.

The proposed approach is not limited to hot-deck imputation, as it can be extended to other method of imputation, as long as the expectation of the imputed estimator of change under random imputation can be expressed as a function of totals.

It is possible to take into account of the wave to wave correlation by using a deterministic regression imputation technique. For example, we could impute by the fitted values of a regression model with the variable (34) as covariate. In that situation, the gradient (17) have a different expression, and the term B (see expression (12)) equals zero. The variance estimator can still be used. However, it does not take into account of the randomness of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in (34).

The variance estimator is based on the assumption that the imputation classes are fixed. However, this assumption does not hold when the imputation classes are based on sampled data. This is also the case when the imputation at wave 2 is based on classes constructed from sample variables observed at wave 1. In Section 8.2, we suggest using wave 1 variable to impute at wave 2, by using imputation classes based on the variable of interest of wave 1 (see (iii) "*Across-waves imputation classes*"). Our simulation study showed that sample-based imputation classes have a negligible effect on the variance estimates, even with across-waves imputation classes. Adjusting the variance estimator to accommodate this situation is beyond the scope of this paper. This is a topic that would need further investigation.

Acknowledgements

Yves G. Berger was supported by the grant RES-000-22-3045 of the Economic and Social Research Council (UK). Emilio L. Escobar was supported by the National Council for Science and Technology (Mexico). We thank the reviewers for helpful comments and suggestions.

References

Andersson, C., Andersson, K. & Lundquist, P. (2011a). Estimation of change in a rotation panel design. In *Proc. 58th World Statist. Cong.,* Int. Statist. Inst., Dublin.

Andersson, C., Andersson, K. & Lundquist, P. (2011b). Variansskattningar avseende förändringsskattningar i panelundersökningar (variance estimation of change in panel surveys). *Methodology Reports from Statistics Sweden (Statistiska centralbyrån)*.

Andridge, R.R. & Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *Int. Statist. Rev.*, **78**, 40–64.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Can. J. Stat.*, **32**(4), 451–467.

Berger, Y.G. & Priam, R. (2010). Estimation of correlations between cross-sectional estimates from repeated surveys - an application to the variance of change. In *Proceeding of the 2010 Symposium of Statistics Canada*, Ottawa.

Berger, Y.G. & Priam, R. (2016). A simple variance estimator of change for rotating repeated surveys: an application to the european union statistics on income and living conditions household surveys. *J. Roy. Statist. Soc. Ser. A*, **179**(1), 251–272.

Brick, J.M. & Montaquila, J.M. (2009). Nonresponse and weighting. In *Sample Surveys: Design, Methods and Applications*, vol. 29A, Eds. D. Pfeffermann & C.R. Rao, pp. 163–185. Amsterdam, Handbook of Statistics: Elsevier.

Deville, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**(418), 376–382.

Lohr, S.L. (2009). *Sampling: Design and Analysis.* Boston: Brooks/Cole.

Deville, J.C. & Särndal, C.E. (1994). Variance estimation for the regression imputed horvitz-thompson estimator. *J. Off. Statist.*, **10**, 381–394.

Escobar, E.L. & Barrios, E. (2014). *samplingvarest: Sampling variance estimation.* R package version 0.9-9. Available at http://cran.r-project.org/web/packages/samplingVarEst. Accessed 7 September 2016.

Eurostat. (2012). *European union statistics on income and living conditions (EU-SILC)*. Available at http://ec.europa.eu/eurostat/web/income-and-living-conditions/overview. Accessed 7 September 2016.

Fay, B.E. (1991). A design-based perspective on missing data variance, pp. 429–440.

Fay, B.E. (1994). Analyzing imputed survey datasets with model-assisted estimators. In *Proc. survey methods sec,* pp. 900–905. Toronto: Am. Statist. Assoc.

Gambino, J.G. & Silva, P.L.N. (2009). Sampling and estimation in household surveys. In *Sample Surveys: Design, Methods & Applications*, vol. 29A, Eds. D. Pfeffermann & C.R. Rao, pp. 407–439. Amsterdam, Handbook of Statistics: Elsevier.

Goga, C., Deville, J.C. & Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, **96**, 691–709.

Hansen, M.H. & Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**(4), 333–362.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Sample Surveys: Design, Methods and Applications*, vol. 29A, Eds. D. Pfeffermann & C.R. Rao, pp. 215–246. Amsterdam, Handbook of Statistics: Elsevier.

Haziza, D. & Beaumont, J.F. (2007). On the construction of imputation classes in surveys. *Int. Stat. Rev.*, **75**(1), 25–43.

Holmes, D.J. & Skinner, C.J. (2000). *Variance estimation for labour force survey estimates of level and change.* London, England: Technical report, Government Statistical Service Methodology Series, 21.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.*, **47**, 663–685.

Huang, E.T. & Fuller, W.A. (1978). Nonnegative regression estimation for survey data. In *Proc. Soc. Statist. Sec.,* pp. 300–305. American Statistical Association.

Kalton, G. (2009). Design for surveys over time. In *Sample Surveys: Design, Methods & Applications*, Vol. 29A, Eds. D. Pfeffermann & C.R. Rao, pp. 89–108. Amsterdam: Elsevier, Handbook of Statistics.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Knottnerus, P. & van Delden, A. (2012). On variances of changes estimated from rotating panels and dynamic strata. *Surv. Methodol.*, **38**(1), 43–52.

Münnich, R. & Zins, S. (2011). Variance Estimation for Indicators of Poverty and Social Exclusion. Research Project Report WP3 D3.2, FP7-SSH-2007-217322, AMELI. Available at http://ameli.surveystatistics.net..

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Ind. Soc. Agri. Statist.*, **3**(3), 169–174.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *J. Off. Stat.*, **16**, 363–378.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods*, Eds. B. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge & P.S. Kott, pp. 153–169. New York: Wiley and Sons.

Qualité, L. & Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Surv. Methodol*, **34**, 173–181.
R Core Team. (2015). *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing*,* Vienna, Austria. Available at http://www.R-project.org/. Accessed 7 September 2016.
Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.*, **3**, 173–180.
Rao, J.N.K. & Shao, A.J. (1992). Jackknife variance estimation with survey data under hotdeck imputation. *Biometrika*, **79**, 811–822.
Rao, J.N.K. & Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453–460.
Robinson, P.M. & Särndal, C.E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya Ser. B*, **43**, 240–248.
Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.
Sampford, MR. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**(3/4), 499–513.
Särndal, C.E. & Lundström. (2005). *Estimation in Surveys with Nonresponse.* Chichester: Wiley.
Särndal, C.E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.
Shao, J. & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *J. Amer. Statist. Assoc.*, **94**, 254–265.
Smith, P., Pont, M. & Jones, T. (2003). Developments in business surv. methodol. in the office for national statistics, 1994–2000. *J. R. Statist. Soc. D (The Statistician)*, **52**, 257–295.
Steel, P. & Fay, B.E. (1995). Variance estimation for finite populations with imputed data. In *Proceedings of the Survey Research Methods Section,* pp. 374–379. American Statistical Association.
Tam, S.M. (1984). On covariances from overlapping samples. *American Statistician*, **38**(4), 288–289.
Tillé, Y. & Matei, A. (2013). *Sampling: Survey sampling*. R package version 2.6. Available at http://cran.r-project.org/web/packages/sampling. Accessed 7 September 2016.
Valliant, R., Dorfman, A.H. & Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley.
Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd edn. New York: Springer.
Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *J. Off. Stat.*, **24**(1), 53–78.

[*Received May 2015, accepted May 2016*]