STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 7–36

JOINT CALIBRATION ESTIMATOR FOR DUAL FRAME SURVEYS

Mahmoud A. Elkasabi[1], **Steven G. Heeringa**[2], **James M. Lepkowski**[3]

ABSTRACT

Many dual frame estimators have been proposed in the statistics literature. Some of these estimators are theoretically optimal but hard to apply in practice, whereas others are applicable but have larger variances than the first group. In this paper, a Joint Calibration Estimator (JCE) is proposed that is simple to apply in practice and meets many desirable properties for dual frame estimators. The JCE is asymptotically design unbiased conditional on the strong relationship between the estimation variable and the auxiliary variables employed in the calibration. The JCE achieves better performance when the auxiliary variables can fully explain the variability in the study variables or at least when the auxiliary variables are strong correlates of the estimation variables. As opposed to the standard dual frame estimators, the JCE does not require domain membership information. Even if included in the JCE auxiliary variables, the effect of the randomly misclassified domains does not exceed the random measurement error effect. Therefore, the JCE tends to be robust for the misclassified domains if included in the auxiliary variables. Meanwhile, the misclassified domains can significantly affect the unbiasedness of the standard dual frame estimators as proved theoretically and empirically in this paper.

Key words: dual-frame estimation, calibration weighting, auxiliary variables, domain misclassification.

1. Introduction

With rapid changes in the cost of survey data collection, changes in population coverage patterns, and sample unit accessibility, dual frame sample surveys are becoming more common in survey practice. For example, dual frame telephone surveys that combine RDD landline telephone samples and cell phone samples emerged to reduce noncoverage due to “cell-only” households in Random-Digit-Dialing (RDD) landline telephone surveys (Brick et al., 2007;

[1] ICF International, Maryland, USA. E-mail: mahmoud.elkasabi@icfi.com.
[2] Institute for Social Research, University of Michigan, USA. E-mail: sheering@umich.edu.
[3] Institute for Social Research, University of Michigan, USA. E-mail: jimlep@umich.edu.

Link, Battaglia, Frankel, Osborn, & Mokdad, 2007). At the same time, Address Based Sampling (ABS) has been explored as a complement or an alternative to RDD telephone surveys (Link, Battaglia, Frankel, Osborn, & Mokdad, 2006, 2008; Link & Lai, 2011).

Estimation is not straightforward in dual frame surveys due to the overlap between the two frames. Simply adding the two estimated totals of the samples results in a biased estimate of the overall total. Standard dual frame estimators adjust for the overlap but present many methodological and practical problems in implementation (Lohr, 2011). In addition, standard dual frame estimation requires the correct identification of the design domain for each sample element. An error in the determination of design domain membership can affect the efficiency of the estimates (Lohr, 2011; Mecatti, 2007).

In this paper, the Joint Calibration Estimator (JCE) is introduced as a new dual frame estimator that relies on the general calibration approach introduced by Deville and Särndal (1992). Calibration generates unbiased estimates themselves for the auxiliary calibration variables under dual frame designs. The effectiveness of calibration for estimates for other variables not included in the calibration set is not completely understood in the dual frame context.

In this paper, we provide an overview of dual frame estimation and introduce a model-assisted design-based JCE under the 'ideal situation', with no errors present in the determination of sample domain and only sampling error for the estimate itself, and in the presence of domain misclassification, where dual frame domains are not correctly identified. The dual frame estimation and calibration approaches are discussed in Sections 2 and 3. The JCE is introduced in Sections 4 and 5, while in Section 6, the bias and variance estimate for the JCE are presented. The misclassification bias for the standard dual frame estimators is derived in Section 7. A simulation study of the performance of the JCE in comparison with standard dual frame estimators is described in Section 8, and the results are discussed in Section 9.

2. Dual frame estimation

Lohr (2011) identified the following five desirable properties for dual frame estimators: (1) unbiased for the corresponding finite population quantity; (2) internally consistent (that is, the multivariate relationships in the data should be preserved, such as the sum of the estimated totals for subgroups should equal the estimated overall); (3) efficient, with low Mean Square Error (MSE); (4) calculable with standard survey software (e.g., one set of weights is needed for all study variables; replicate weights are available for formula-based or replication-based variance estimation); and (5) robust to non-sampling errors.

In addition to Lohr's properties, we add the following three properties. (6) Data requirements for the estimator should be reasonable. For example, information about design domain membership or variance and covariance components is required for some estimators, but these may be poorly measured or unreliable components and add to the burden and complexity of computing the estimator. (7) An estimator should be robust to non-sampling errors in the estimator's auxiliary and domain membership variables or required variances and covariances. Although some estimators might theoretically be efficient, reporting errors in domain membership or biased estimates of required variance and covariance components could result in biased or non-optimal estimators. (8) An estimator should be readily applicable to dual and multiple (more than two) frame surveys.

2.1. Notation

Let $U=\{1,..,k,..,N\}$ denote a target population of N elements, and let $A=\{1,..,k,..,N_A\}$ and $B=\{1,..,k,..,N_B\}$ denote two overlapping frames. The two frames are not assumed to be exclusive, that is: $A\bigcap B=ab\neq\phi$ and $A\bigcup B=U$. The dual frame design sample s is composed of two samples $s_A\left(s_A\subseteq A\right)$ and $s_B\left(s_B\subseteq B\right)$ selected from the two overlapping frames A and B using a sample design with inclusion probabilities $\pi_k^A=p\left(k\in s_A\right)$ and $\pi_k^B=p\left(k\in s_B\right)$. Base weights to compensate for unequal selection probabilities are d_k, where $d_k=d_k^A=1/\pi_k^A$ for $k\in s_A$ and $d_k=d_k^B=1/\pi_k^B$ for $k\in s_B$. Let N_A and N_B denote the frame sizes and n_A and n_B denote the sample sizes for frames A and B, respectively. Let $a=A\cap B^c$ and $b=A^c\cap B$, where c denotes the complement of a set, and $s_a=a\cap s_A$, $s_b=b\cap s_B$, $s_{ab}^A=ab\cap s_A$ and $s_{ab}^B=ab\cap s_B$. Standard dual frame estimators of a population total take the form $\hat{Y}=\hat{Y}_a+\hat{Y}_{ab}+\hat{Y}_b$ estimating the true population total $Y=Y_a+Y_{ab}+Y_b$, where $Y=\sum_{k\in U}y_k$,$Y_a=\sum_{k\in a}y_k$, $Y_b=\sum_{k\in b}y_k$ and $Y_{ab}=\sum_{k\in ab}y_k$.

2.2. The standard dual frame estimators

The Horvitz-Thompson estimators of totals (Horvitz & Thompson, 1952) for domains a and b for characteristic Y are $\hat{Y}_a=\sum_{k\in s_a}d_ky_k$ and $\hat{Y}_b=\sum_{k\in s_b}d_ky_k$, and the estimators for the domain overlap are $\hat{Y}_{ab}^A=\sum_{k\in s_{ab}^A}d_ky_k$ and $\hat{Y}_{ab}^B=\sum_{k\in s_{ab}^B}d_ky_k$. For each sample, the estimators of population totals are unbiased for the corresponding domain totals Y_a, Y_{ab} and Y_b: $E\left[\hat{Y}_a+\hat{Y}_{ab}^A\right]=Y_a+Y_{ab}$ and $E\left[\hat{Y}_b+\hat{Y}_{ab}^B\right]=Y_b+Y_{ab}$, where $E(.)$ denotes design-based expectation. Therefore, adding the two sample estimated totals results in a biased population estimate $E\left[\hat{Y}_a+\hat{Y}_{ab}^A+\hat{Y}_b+\hat{Y}_{ab}^B\right]\approx Y_a+2Y_{ab}+Y_b \neq Y$.

An unbiased dual frame estimator for Y can be obtained by the weighted average of the estimators $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$,

$$\hat{Y}=\hat{Y}_a+\theta\hat{Y}_{ab}^A+(1-\theta)\hat{Y}_{ab}^B+\hat{Y}_b \tag{1}$$

where $\theta\in[0,1]$ is a composite factor combining $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$. Estimators of domain sizes $\hat{N}_a^A$, $\hat{N}_{ab}^A$, $\hat{N}_{ab}^B$ and $\hat{N}_b^B$ are defined by setting $y_k=1$ for all $k\in s$ in $\hat{Y}_a^A$, $\hat{Y}_{ab}^A$, $\hat{Y}_{ab}^B$ and $\hat{Y}_b^B$, and the dual frame estimator in (1) can be used to find the population total estimate $\hat{N}$. Consequently, an unbiased dual frame estimator for the population mean $\bar{Y}$ can be written as $\bar{Y}=\hat{Y}/\hat{N}$. The weighted version of the estimated total in (1) can be written as

$$\hat{Y}=\sum\nolimits_{k\in s_A} m_k d_k y_k+\sum\nolimits_{k\in s_B} m_k d_k y_k \tag{2}$$

where the adjustment factor m_k can be written as

$$m_k=\begin{cases}1 & k\in s_a,\\ \theta & k\in s_{ab}^A,\\ 1-\theta & k\in s_{ab}^B,\\ 1 & k\in s_b.\end{cases} \tag{3}$$

The approach used to determine the composite factor θ distinguishes standard dual frame estimators. Hartley (1962, 1974) proposed choosing the composite factor θ_{HT} to minimize the variance of $\hat{Y}$. Choosing any fixed value for the composite factor (e.g. θ=0.5) yields the unbiased Fixed Weight Estimator (FWE), which includes the optimum Hartley Estimator (HE) as a special case.

Fuller and Burmeister (1972) extended Hartley's estimator by using a maximum likelihood estimator $\hat{N}_{ab}$ of the overlap domain population size N_{ab}. Later, Skinner and Rao (1996) extended the Fuller-Burmeister (FB) estimator to achieve design-based consistency under complex designs using a Pseudo-Maximum Likelihood Estimator (PML). Rao and Wu (2010) proposed the Pseudo-Empirical Likelihood (PEL) estimator, which depends on adjustment factors based on probability measures p_a, p_{ab}^A, p_b and p_{ab}^B for a randomly selected case being in poststrata s_a, s_{ab}^A, s_b and s_{ab}^B.

Several single frame estimators have been proposed as alternatives. Bankier (1986) and Kalton and Anderson (1986) proposed the Single Frame Estimator (SFE) which treats the dual frame design as a stratified design consisting of three strata, one for each design domain, and calculates joint inclusion probabilities. Meccati (2007) introduced a simple dual frame estimator, the Multiplicity Estimator (ME), which depends on the number of the frames that case *k* belongs to, M_k , in order to combine domains.

With respect to their sampling variance, consistency, and practical utility, these estimators can be grouped into three types. First are the optimal estimators, HE, FB, and PEL. These are internally inconsistent since they generate weights that are dependent on the study variables. This restricts the practical application of the optimal estimators using standard survey software. At the same time, these optimal estimators require estimates of variance and covariance components for finding the composite factor θ . Biased estimates of the required components result in non-optimal estimates. Forms of these estimators for multiple frame surveys are complicated due to the need to estimate covariance terms in the composite factors (Lohr & Rao, 2000, 2006; Mecatti, 2007; Skinner, 1991).

The second type is the "practical" estimators, FWE, SFE and ME. Easier to compute in practice, these estimators achieve notably poorer efficiency relative to the optimal estimators. They are internally consistent since they generate only one set of weights for all study variables and standard survey software can be used to find the survey estimates. Deriving these estimators for multiple frame surveys is a straightforward task.

The third type includes just the PML, which has greater practical applicability than the optimal estimators and is more efficient than the practical estimators. PML has smaller MSE than FB and HE because it does not require estimation of variance components of the composite factors in FB and HE (Lohr & Rao, 2000; Skinner & Rao, 1996).

With respect to the eight desirable properties for dual frame estimators, all the standard dual frame estimators are unbiased, or approximately so. Not all of them are internally consistent, efficient, or calculatable with standard survey software. With regard to property (5) concerning non-sampling errors, dual frame estimators have a disadvantage compared to single frame surveys because of different levels of non-sampling errors associated with the frames (Brick, Flores-Cervantes, Lee, & Norman, 2011). These kinds of associations add to the complexity of the assessment and adjustment for these errors, adversely affecting property 6.

Nearly all of these dual frame estimators require accurate information about domain membership. But domain membership might be affected by reporting errors and leading to a biased estimate (property (7)). Finally, extending standard dual frame estimators to multiple frames is not readily achieved (property (8)).

3. The calibration approach

In the single frame survey design, where the sample $s(s \subseteq U)$ is selected from the population U using a sample design with inclusion probability of $\pi_k = p(k \in s)$, the base weights are equal to $d_k = 1/\pi_k$. Let $\mathbf{x}_k = (x_{k1},..,x_{kj},..,x_{kJ})'$ denote an auxiliary variable vector of dimension $j = (1,...,J)$, where both y_k and $\mathbf{x}_k$ are observed for the sample elements $k \in s$. The Horvitz-Thompson estimator for the total $Y = \sum_{k \in U} y_k$ is $\hat{Y}_{HT} = \sum_{k \in s} d_k y_k$.

In a complete response situation, with known auxiliary totals for the $j = (1,..,J)$ auxiliary variables, $\mathbf{X} = (X_1,..,X_j,..,X_J)' = (\sum_{k \in U} x_{k1},..,\sum_{k \in U} x_{kj},..,\sum_{k \in U} x_{kJ})'$, Deville and Särndal (1992) defined calibration as a method to find weights w_k which minimize a distance measure $G(w_k, d_k)$ between the calibrated weights w_k and the base weights d_k. This minimization of the distance function is subject to the constraint that the calibration-weighted total of the auxiliary variable values $\sum_{k \in s} w_k x_{kj}$ equals the known population total for the auxiliary X_j for $j = 1,\dots,J$, or $\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}$. This calibration approach results in final calibrated weights $w_k = d_k F(q_k \mathbf{x}_k' \lambda)$ where $F(q_k \mathbf{x}_k' \lambda)$ is the inverse of $\partial G(w_k, d_k)/\partial w_k$, λ denotes a vector of Lagrange multipliers in the minimization, and q_k is a positive value which scales the calibrated weights.

Many distance measures have been proposed for calibration, but empirically there are small differences in the calibrated estimates derived from alternative distance measures (Singh & Mohl, 1996; Stukel, Hidiroglou, & Särndal, 1996). We use the linear case with the chi-square distance function $(w_k - d_k)^2 / 2d_k$ and $q_k = 1$. The calibration obtains $w_k, k \in s$ by minimizing the distance function $\sum_{k \in s} (w_k - d_k^*)^2 / 2d_k^*$ subject to the calibration equation $\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}$, where d_k^* are arbitrary initial weights (a base weight or an adjusted version).

The minimization generates the Lagrange multiplier vector

$\lambda' = (\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} d_k^* \mathbf{x}_k)' (\sum_{k \in s} d_k^* \mathbf{x}_k \mathbf{x}_k')^{-1}$ and calibration factor is $g_k = (1 + \lambda' \mathbf{x}_k)$.

The final calibrated weights are

$w_k = d_k^* (1 + \lambda' \mathbf{x}_k) = d_k^* \left[1 + (\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} d_k^* \mathbf{x}_k)(\sum_{k \in s} d_k^* \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \right]$ and the calibrated estimated total is $\hat{Y}_w = \sum_{k \in s} w_k y_k$.

As it will be shown in the next section, the main idea behind calibration, finding a set of weights which guarantee that estimated auxiliary totals conform to known population totals, can be used to combine two samples.

4. Joint Calibration Estimator

Under the dual frame design, let $E\left(\sum_{k\in s_A} d_k\mathbf{x}_k\right)=\mathbf{X}_A$, $E\left(\sum_{k\in s_B} d_k\mathbf{x}_k\right)=\mathbf{X}_B$ and $E\left(\sum_{k\in s_A} d_k\mathbf{x}_k+\sum_{k\in s_B} d_k\mathbf{x}_k\right)\neq\mathbf{X}$, where $\mathbf{X}_A=\left(\sum_{k\in A} x_{k1},..,\sum_{k\in A} x_{kj},..,\sum_{k\in A} x_{kJ}\right)'$ and $\mathbf{X}_B=\left(\sum_{k\in B} x_{k1},..,\sum_{k\in B} x_{kj},..,\sum_{k\in B} x_{kJ}\right)'$. Calibration conditioning on $\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k=\mathbf{X}$ should achieve $E\left(\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k\right)=\mathbf{X}$. Consequently, a set of auxiliary variables that are strong predictors for the study variable y should yield $E(\sum_{k\in s_A} w_k y_k+\sum_{k\in s_B} w_k y_k)\simeq Y$ (see Proposition 1 and Corollary 1).

Under complete response (i.e., no nonresponse), calibrated estimates can be parameterized for the dual frame design by deriving the calibration factors as explicit components for each sample of the dual frame sample. Calibration finds final weights w_k such that

$$\sum_{k\in s} w_k\mathbf{x}_k=\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k=\mathbf{X} \qquad (4)$$

by minimizing the distance function $\sum_{k\in s_A}\left(w_k-d_k\right)^2/2d_k+\sum_{k\in s_B}\left(w_k-d_k\right)^2/2d_k$. The joint calibration weights are $w_k=d_k\left(1+\lambda'\mathbf{x}_k\right),k\in s$ where $\lambda'=\left(\sum_{k\in U}\mathbf{x}_\mathrm{k}-\sum_{k\in s} d_k\mathbf{x}_\mathrm{k}\right)'\left(\sum_{k\in s} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k}\right)^{-1}$ with joint calibration factor $g_k=\left(1+\lambda'\mathbf{x}_k\right)$.

Therefore, the JCE for population total can be written as

$$\hat{Y}_{JCE}=\sum_{k\in s} w_k y_k=\sum_{k\in s_A} w_k y_k+\sum_{k\in s_B} w_k y_k \qquad (5)$$

where $w_k=d_k\left(1+\lambda'\mathbf{x}_k\right)$ and

$$\lambda'=\left(\sum_{k\in U}\mathbf{x}_\mathrm{k}-\sum_{k\in s_A} d_k\mathbf{x}_\mathrm{k}-\sum_{k\in s_B} d_k\mathbf{x}_\mathrm{k}\right)'\left(\sum_{k\in s_A} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k}+\sum_{k\in s_B} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k}\right)^{-1}.$$

The calibration constraints determine the form of the JCE. Some forms may be identical to the standard dual frame estimators. For example, the dual frame estimator for the total can be written as in equation (1), and the weighted version expressed as in equations (2) and (3), where an alternative expression for equation (2) is $\hat{Y} = \sum_{k \in s_a} d_k y_k + \sum_{k \in s_{ab}^A} m_k d_k y_k + \sum_{k \in s_{ab}^B} m_k d_k y_k + \sum_{k \in s_b} d_k y_k$, where m_k is defined in (3). When the auxiliary variable $\mathbf{x}_k = 1$ for $k \in U$, under the JCE, the main constraint in (4) can be written as

$$\sum_s w_k = N \tag{6}$$

and the constraint $w_k = d_k \ \forall \ k \in s_a \cup s_b$ can be added to the calibration minimization problem. This constraint is identical to

$$\sum_{k \in s_a} w_k = \sum_{k \in s_a} d_k^* = N_a \tag{7}$$

and

$$\sum_{k \in s_b} w_k = \sum_{k \in s_b} d_k^* = N_b . \tag{8}$$

In (7) and (8), $d_k^* = \left(N_a \Big/ \sum_{k \in s_a} d_k \right) d_k$ and $d_k^* = \left(N_b \Big/ \sum_{k \in s_b} d_k \right) d_k$, respectively. Joint calibration with the three constraints (6), (7) and (8) is identical to post-stratifying the sample by the design domain totals N_a, N_{ab} and N_b , which yields the unbiased dual frame estimator (2), where the modification factors for the overlap domain have the same value $m_k = N_{ab} \Big/ \left(\sum_{k \in s_{ab}^A} d_k + \sum_{k \in s_{ab}^B} d_k \right) \ \forall k \in s_{ab}^A \cup s_{ab}^B$. In this case, the joint calibration factor is

$$g_k = \begin{cases} N_a \Big/ \sum_{k \in s_a} d_k & k \in s_a, \\ N_{ab} \Big/ \left(\sum_{k \in s_{ab}^A} d_k + \sum_{k \in s_{ab}^B} d_k \right) & k \in s_{ab}^A \cup s_{ab}^B, \\ N_b \Big/ \sum_{k \in s_b} d_k & k \in s_b. \end{cases} \tag{9}$$

The joint calibration factor in (9) yields the post-stratified version of the Fixed Weight Estimator (FWE), $\hat{Y}_{FWE}^{post} = \frac{N_a}{\hat{N}_a} \hat{Y}_a + \frac{N_{ab}}{\hat{N}_{ab}} \left(\theta \hat{Y}_{ab}^A + (1-\theta) \hat{Y}_{ab}^B \right) + \frac{N_b}{\hat{N}_b} \hat{Y}_b$ where $\theta = 0.5$ and $\hat{N}_{ab} = \left(\theta \hat{N}_{ab}^A + (1-\theta) \hat{N}_{ab}^B \right)$.

The JCE can readily be adapted to multiple frames as well. Under multiple frame designs, with *P* domains, the JCE for population total of *y* can be written as $\hat{Y}_{JCE} = \sum_{p \in P} \sum_{k \in s_p} w_k y_k$ where $w_k = d_k (1 + \lambda' \mathbf{x}_k)$ and λ' can be written as $\lambda' = \left(\sum_{k \in U} \mathbf{x}_k - \sum_{p \in P} \sum_{k \in s_p} d_k \mathbf{x}_k \right)' \left(\sum_{p \in P} \sum_{k \in s_p} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1}$.

5. Examples of Joint Calibration Estimators

The auxiliary variable vector characterizes the final JCE for dual frame estimation. For example, under the univariate auxiliary variable $\mathbf{x}_k = 1$ for $k \in U$, we have the *common mean model*

$$\begin{cases} E_\xi(y_k) = \beta, \\ V_\xi(y_k) = \sigma^2, \end{cases} \tag{10}$$

where E_ξ and V_ξ denote the expectation and variance with respect to the calibration model ξ. For the overall population total $\mathbf{X} = N$, the joint calibration factor is $g_k = N\left(\sum_{k \in s_A} d_k + \sum_{k \in s_B} d_k\right)^{-1}$. By calibrating concatenated or "stacked" datasets for each frame's sample, $\sum_{k \in s_A} w_k \mathbf{x}_k + \sum_{k \in s_B} w_k \mathbf{x}_k = N$. This JCE estimate is appropriate when it is thought that the true common mean β is the same for all $k \in U$. However, when the β varies between design domains, another JCE uses the calibration factor in (8).

For $\mathbf{x}_k = x_k$ for $k \in U$, we can also consider the *ratio model*

$$\begin{cases} E_\xi(y_k) = \beta x_k, \\ V_\xi(y_k) = \sigma^2 x_k, \end{cases} \tag{11}$$

where $\mathbf{X} = X$. The joint calibration factor is $g_k = X\left(\sum_{k \in s_A} d_k x_k + \sum_{k \in s_B} d_k x_k\right)^{-1}$. Calibrating the stacked dataset, $\sum_{k \in s_A} w_k \mathbf{x}_k + \sum_{k \in s_B} w_k \mathbf{x}_k = X$. This JCE estimate is appropriate when it is thought that βx_k is the same, for all $k \in U$. Another JCE estimate is appropriate when it is thought that βx_k varies between design domains. This estimate uses the calibration factor

$$g_k = \begin{cases} X_a \Big/ \sum_{k \in s_a} d_k x_k & k \in s_a, \\ X_{ab} \Big/ \left(\sum_{k \in s_{ab}^A} d_k x_k + \sum_{k \in s_{ab}^B} d_k x_k\right) & k \in s_{ab}^A \cup s_{ab}^B, \\ X_b \Big/ \sum_{k \in s_b} d_k x_k & k \in s_b. \end{cases} \tag{12}$$

Obviously, this estimate requires knowledge of the separate totals (X_a, X_{ab}, X_b).

Under the multivariate auxiliary variable $\mathbf{x}_k = (1, x_k)$ for $k \in U$, consider *the simple regression model with intercept*

$$\begin{cases} E_\xi(y_k) = \alpha + \beta x_k, \\ V_\xi(y_k) = \sigma^2. \end{cases} \tag{13}$$

The calibrated estimate $\hat{Y}_{JCE}$, is $\hat{Y}_{JCE} = \hat{Y}_{HT}^{A} + \hat{Y}_{HT}^{B} + \left(\sum_{k\in U} x_k - \left(\sum_{k\in s_A} d_k x_k + \sum_{k\in s_B} d_k x_k\right)\right)\hat{B}_s^{A,B}$ where $\hat{B}_s^{A,B} = \left(\sum_{k\in s_A} d_k \mathbf{x}_k y_k + \sum_{k\in s_B} d_k \mathbf{x}_k y_k\right)\left(\sum_{k\in s_A} d_k \mathbf{x}_k \mathbf{x}'_k + \sum_{k\in s_B} d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$. With more than one auxiliary variable, the multivariate estimator can be written as

$$\hat{Y}_{JCE} = \hat{Y}_{HT}^{A} + \hat{Y}_{HT}^{B} + \left(\sum_{k\in U} \mathbf{x}_k - \left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)\right)\hat{B}_s^{A,B} \tag{14}$$

where $\mathbf{x}_k = \left(x_{k1}, .., x_{kj}, .., x_{kJ}\right)'$ is the auxiliary variable vector with $j = (1, ..., J)$. Since $\left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)$ is always greater than $\sum_{k\in U} \mathbf{x}_k$, the term $\left(\sum_{k\in U} \mathbf{x}_k - \left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)\right)\hat{B}_s^{A,B}$ in (14) can be viewed as a negative-sign correction factor for the biased summation of $\hat{Y}_{HT}^{A}$ and $\hat{Y}_{HT}^{B}$. All the JCE forms discussed above can be derived from the general JCE form in (14).

Another multivariate calibration estimator is a post-stratified estimator, corresponding to *a group mean model*, calibrating on known post-stratified cell counts. When the sizes of the population groups N_p and the classification vector used to code membership in one of *P* mutually exclusive and exhaustive groups are known, and $\mathbf{x}_k = \gamma_k = \left(\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk}\right)'$ is the auxiliary variable vector, where $\gamma_{pk} = 1$ for $k \in p$ and 0 otherwise, the calibrated estimator is the standard post-stratified estimator. The joint calibration factor is $N_p \Big/ \left(\sum_{k\in s_p^A} d_k + \sum_{k\in s_p^B} d_k\right)$, where s_p^A denotes the sample cell $U_p \cap s_A$ and s_p^B denotes the sample cell $U_p \cap s_B$. The calibrated estimator of the total can be written as $\hat{Y}_{JCE} = \sum_P \frac{N_p}{\left(\sum_{k\in s_p^A} d_k + \sum_{k\in s_p^B} d_k\right)}\left(\sum_{k\in s_p^A} d_k y_k + \sum_{k\in s_p^B} d_k y_k\right)$. In this *group mean model*, it is implicitly assumed that mean and variance are shared by all elements within the same group *p* as

$$\begin{cases} E(y_k) = \beta_p, \\ V(y_k) = \sigma_p^2. \end{cases} \tag{15}$$

Similarly, when the group totals X_p are known and $\mathbf{x}_k = x_k\gamma_k = \left(x_{1k}\gamma_{1k},...,x_{pk}\gamma_{pk},...,x_{Pk}\gamma_{Pk}\right)'$ is used as the auxiliary variables vector, this corresponds to *the group ratio model*, where mean and variance are shared by all elements within the same group *p* as

$$\begin{cases} E(y_k) = \beta_p x_k, \\ V(y_k) = \sigma_p^2 x_k. \end{cases} \tag{16}$$

Both *the group mean model* and *the group ratio model* can be classified under *the group model* of Särndal, Swensson & Wretman (1992).

6. The bias and the variance of the Joint Calibration Estimator

The JCE is a model-assisted design-based estimator for which the design-based bias properties are affected by the association between the study variable *y* and the auxiliary variable vector **x**.

6.1. Proposition 1

The bias of the JCE estimator $\hat{Y}_{JCE}$, in (5), is given approximately by

$$Bias\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{k\in U_{ab}} e_k^{A,B} \tag{17}$$

where $e_k^{A,B} = \left(y_k - \mathbf{x}_k'\mathrm{B}_U^{A,B}\right)$ and

$\mathrm{B}_U^{A,B} = \left(\sum\nolimits_{k\in U_A}\mathbf{x}_k\mathbf{x}_k' + \sum\nolimits_{k\in U_B}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum\nolimits_{k\in U_A}\mathbf{x}_k y_k + \sum\nolimits_{k\in U_B}\mathbf{x}_k y_k\right)$ (see the appendix for proof).

Note that the dual frame estimation bias can be derived from expression (1) as

$$Bias\left(\hat{Y}_A + \hat{Y}_B\right) = \sum\nolimits_{k\in U_{ab}} y_k \,. \tag{18}$$

This means that the joint calibration approach uses $\mathbf{x}_k'\mathrm{B}_U^{A,B}$ to attenuate the bias for each $k \in U_{ab}$ to reduce the bias in (18). Therefore, the reduction in dual frame estimation bias due to the joint calibration is $\sum\nolimits_{k\in U_{ab}}\mathbf{x}_k'\mathrm{B}_U^{A,B}$, which is the difference between (17) and (18).

Proposition 1 highlights the need to identify powerful auxiliary variables that can predict study variable *y*. The more $\mathbf{x}_k'\mathrm{B}_U^{A,B}$ is able to predict y_k for each $k \in U_{ab}$, the greater the reduction in bias. The bias of $\hat{Y}_{JCE}$ in (17) is independent of the sampling design used to draw s_A and s_B as long as the set of auxiliary variables $\mathbf{x}_k$ is the same.

6.2. Corollary 1

When a linear relationship exists between the study variable y_k and the auxiliary vector $\mathbf{x}_k$, as in $y_k = \mathbf{x}_k' \mathrm{B}_U$, for every $k \in U$, the bias of the JCE estimator in (17) can be written as $Bias\left(\hat{Y}_{JCE}\right) = \sum_{k \in U_{ab}} \mathbf{x}_k' \left(\mathrm{B}_U - \mathrm{B}_U^{A,B}\right) = 0$.

This corollary is true because when a linear relationship between y_k and x_k exists, $\mathrm{B}_U^{A,B} = \mathrm{B}_U$, and the bias of $\hat{Y}_{JCE}$ is a function of the difference between two regression vectors $\mathrm{B}_U^{A,B}$ and B_U. This linear relationship will not hold in practice, but the bias in (17) will be reduced if the relationship between y_k and $\mathbf{x}_k$ is linear or nearly linear. The JCE bias is reduced to the extent that there are auxiliary variables $\mathbf{x}_k$ such that the residuals $e_k^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_U^{A,B}\right)$ are small. Using such a set of auxiliary variables $\mathbf{x}_k$ guarantees reduced bias and variance of the JCE. Thus, the properties of the JCE are controlled by the association between y and **x**, where the best performance occurs when **x** more closely matches the population model or **x** includes strong correlates of y.

Assuming that the same model holds for all units in the population, $\frac{1}{N_{ab}} \sum_{k \in U_{ab}} e_k^{A,B}$ is asymptotically $N(0,V)$ where V is $O\left(N_{ab}^{-1}\right)$. The bias $Bias\left(\hat{\bar{Y}}_{JCE}\right) = \frac{1}{N} \sum_{k \in U_{ab}} e_k^{A,B}$ (where $\hat{\bar{Y}}_{JCE} = \hat{Y}_{JCE} / N$) converges in probability to 0 in large populations because the variance of the estimator $\hat{\bar{Y}}_{JCE} = \frac{N_{ab}}{N} \frac{1}{N_{ab}} \sum_{k \in U_{ab}} e_k^{A,B}$ is proportional to $P_{ab}^2 O\left(N_{ab}^{-1}\right) \approx \frac{P_{ab}}{N}$, and $\frac{N_{ab}}{N} \to P_{ab}$, and $\frac{P_{ab}}{N} \to 0$ as $N \to \infty$. Thus, the JCE estimator of the mean, $\hat{\bar{Y}}_{JCE}$, is a consistent estimator of population mean, $\bar{Y}$.

Under dual frame design, variance of $\hat{Y}_{JCE}$ can be written as

$$V\left(\hat{Y}_{JCE}\right) = \sum\sum_{k,l \in U_A} \Delta_{kl}^{A} \left(\frac{e_k^A}{\pi_k^A}\right)\left(\frac{e_l^A}{\pi_l^A}\right) + \sum\sum_{k,l \in U_B} \Delta_{kl}^{B} \left(\frac{e_k^B}{\pi_k^B}\right)\left(\frac{e_l^B}{\pi_l^B}\right) + \sum\sum_{k,l \in U_{ab}} \Delta_{kl}^{ab} \left(\frac{e_k^{ab}}{\pi_k^{ab}}\right)\left(\frac{e_l^{ab}}{\pi_l^{ab}}\right)$$

where $s_{ab} = s_A \cap s_B$, for $D = (A, B, ab)$, $\Delta_{kl}^D = \left(\pi_{kl}^D - \pi_k^D \pi_l^D\right)$, $\pi_{kl}^D = p\left(k \,\&\, l \in s_D\right)$, $\pi_k^D = p\left(k \in s_D\right)$, $\pi_l^D = p\left(l \in s_D\right)$, $e_k^D = y_k - \mathbf{x}_k' \mathrm{B}_{U_D}$, and $\mathrm{B}_{U_D} = \sum_{k \in U_D} \mathbf{x}_k y_k \left(\sum_{U_D} \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$. Assuming small values of π_{kl}^{ab}, π_k^{ab} and π_l^{ab}, the estimated variance reduces to

$$\hat{v}\left(\hat{Y}_{JCE}\right) = \sum\sum_{k,l \in s_A} \frac{\Delta_{kl}^A}{\pi_{kl}} \left(w_k \hat{e}_k^A\right)\left(w_l \hat{e}_l^A\right) + \sum\sum_{k,l \in s_B} \frac{\Delta_{kl}^B}{\pi_{kl}} \left(w_k \hat{e}_k^B\right)\left(w_l \hat{e}_l^B\right) \text{ where}$$

$\hat{e}_k^D = y_k - \mathbf{x}_k' \hat{\mathrm{B}}_{ws_D}$, and $\hat{\mathrm{B}}_{ws_D} = \sum_{k \in s_D} w_k \mathbf{x}_k y_k \left(\sum_{k \in s_D} w_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$.

7. Domain misclassification bias in dual frame estimation

Standard dual frame estimators depend on identifying the design domains during the data collection. Consequently, the performance of these estimators is sensitive to the errors in measuring the domain membership (Mecatti, 2007). Since it is uncommon to have access to the domain membership information before collecting the survey data, this information should be obtained during the data collection. For example, information about landline telephone service should be obtained in the area-landline dual frame surveys (Lepkowski & Groves, 1986) or about the landline and cell phone services should be obtained in the landline-cell dual frame telephone surveys (Brick et al., 2006; Kennedy, 2007). Collecting this information could be burdensome for some respondents and could lead to more unit non-response. It is even worse when dealing with rare populations such as persons with a rare disease or for elusive or hidden populations such as the homeless, illegal immigrants or drug consumers (Lepkowski, 1991; Mecatti, 2007; Sudman & Kalton, 1986).

Besides identifying the domain membership for every sampled unit, ideally, such information should be free from reporting or measurement errors, but this is not typically the case (Lohr & Rao, 2006). The correct classification of the sampled units into the domains in each frame is required to apply the standard dual frame estimators. In practice, achieving the correct classification for all cases is almost impossible because, as any other study variable, the domain membership variable could be affected by the measurement or the reporting error. Therefore, the sampled units could be misclassified into the wrong domain, leading to ***the domain misclassification***. For example, in RDD-cell phone dual frame surveys, households owning both landline and cell phone can be misclassified as landline only households or vice versa. Generally, it is difficult to identify misclassified units, and to estimate the misclassification rate. This means that the optimal dual frame estimators could have less than optimal performance (Lohr, 2011; Lohr & Rao, 2006).

The bias due to domain misclassification affects the standard dual frame estimators, however it does not affect the JCE; the latter does not necessarily require any domain membership information. In the presence of domain misclassification and where s_{mis} is the domain-misclassified sample s, the unconditional bias of the standard dual frame estimators in (1), $\hat{Y}_{mis}$, can be evaluated jointly with respect to the sampling design $p(s)$ and the conditional misclassification distribution $q(s_{mis} \mid s)$ as

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E_p\left(E_q\left(\hat{Y}_{mis} \mid s\right)\right) - Y = E_{pq}\left(\hat{Y}_{mis}\right) - Y \ . \qquad (19)$$

7.1. Proposition 2

In the presence of the two-way misclassification (TWM), where $I_k^{ab,c}$ is a misclassification indicator for observation k from the overlapping domains s_{ab}^A and s_{ab}^B misclassified into non-overlapping domains s_a and s_b, respectively, and $I_k^{c,ab}$ is a misclassification indicator for observation k from s_a and s_b misclassified into s_{ab}^A and s_{ab}^B, respectively, a general expression for the unconditional bias that assumes each element k in the overlapping domain has a misclassification probability $E\left(I_k^{ab,c}\right)=\gamma_k^{ab,c}$ and each element k in the non-overlapping domains has a misclassification probability $E\left(I_k^{c,ab}\right)=\gamma_k^{c,ab}$, as derived in the appendix, can be written as

$$\begin{aligned} Bias_{pq}\left(\hat{Y}_{mis}\right) &= N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)+\bar{\gamma}^{ab,c}\bar{Y}_{ab}\right)- \\ &(1-\theta)N_a\left(\varsigma_a\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_a^{c,ab}\bar{Y}_a\right)-\theta N_b\left(\varsigma_b\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_b^{c,ab}\bar{Y}_b\right) \end{aligned} \tag{20}$$

where $\bar{Y}_{ab}=\sum_{k\in ab}y_k\big/N_{ab}$, $\bar{\gamma}^{ab,c}=\sum_{k\in ab}\gamma_k^{ab,c}\big/N_{ab}$, $\bar{Y}_a=\sum_{k\in a}y_k\big/N_a$, $\bar{\gamma}_a^{c,ab}=\sum_{k\in a}\gamma_k^{c,ab}\big/N_a$, $\bar{Y}_b=\sum_{k\in b}y_k\big/N_b$ and $\bar{\gamma}_b^{c,ab}=\sum_{k\in b}\gamma_k^{c,ab}\big/N_b$. $\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)$ is the population covariance between the misclassification probabilities $\gamma_k^{ab,c}$ and the values of the target variable y_k within the overlapping domains *ab*. Also, $\varsigma_a\left(\gamma_k^{c,ab},y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab},y_k\right)$ are the population covariance between the misclassification probabilities $\gamma_k^{c,ab}$ and the values of the target variable y_k within the non-overlapping domains *a* and *b*, respectively. These covariances can be written as follows

$$\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)=\sum_{k\in ab}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\big/N_{ab}, \tag{21}$$

$$\varsigma_a\left(\gamma_k^{c,ab},y_k\right)=\sum_{k\in a}\left(\gamma_k^{c,ab}-\bar{\gamma}_a^{c,ab}\right)\left(y_k-\bar{Y}_a\right)\big/N_a, \tag{22}$$

$$\varsigma_b\left(\gamma_k^{c,ab},y_k\right)=\sum_{k\in b}\left(\gamma_k^{c,ab}-\bar{\gamma}_b^{c,ab}\right)\left(y_k-\bar{Y}_b\right)\big/N_b. \tag{23}$$

This means that the misclassification bias depends on two components:

a) The expected total of y_k for the misclassified cases within each domain, $N_{ab}\bar{\gamma}^{ab,c}\bar{Y}_{ab}$, $N_a\bar{\gamma}_a^{c,ab}\bar{Y}_a$ and $N_b\bar{\gamma}_b^{c,ab}\bar{Y}_b$.

b) The correlation between the misclassifications probabilities and the study variable *y* within the different design domains, supported by the within domains covariances, $\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)$, $\varsigma_a\left(\gamma_k^{c,ab},y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab},y_k\right)$.

In general, the misclassification bias can be controlled during the data collection process by following the best practices that decrease the measurement error in reporting the domain membership variable. At the same time, the misclassification bias can be adjusted based on the second component by implicitly predicting the misclassification probabilities. This can be performed by calibrating the data by an auxiliary variable that is correlated with the study variable y and the misclassification probabilities. This step can be performed either in the standard dual frame estimators or in the JCE. In the standard dual frame estimators, the calibration step comes after combining the data based on the misclassified domains. When misclassification probabilities are known, Lohr (2011) proposed an adjustment factor for the misclassification bias for the FWE estimator, which is consistent with our derivations of the misclassification bias.

In the JCE, the domain misclassification does not affect the estimates as long as no domain membership information was added to the auxiliary variable vector, $\mathbf{x}$. However, even if misclassified domain membership information was added to the auxiliary variable vector, adding more auxiliary variables which are correlated with the study variable y and the misclassification probabilities is enough to reduce the bias resulted from the misclassified domain. Moreover, the effect of using the misclassified domains as the sole auxiliary variable in the JCE is less significant than the effect of the domain misclassification in the standard dual frame estimators. This is due the fact that in the standard dual frame estimators, classifying the sampling units into the domain correctly is required before applying the composite factor θ . However, in the JCE, this misclassification error is accounted for as a measurement in the auxiliary variables.

8. Simulation studies

In this section, two simulation studies are presented. The first one is to examine the performance of the JCE estimator in comparison with the FWE estimator under different population models. These population models determine the relationship between the study variable and the calibration auxiliary variables. The second simulation study considers the domain misclassification errors and examines the performance of the JCE and FWE estimators in the presence of these errors.

8.1. The first study: design

A simulation study was used to evaluate the performance of the JCE relative to the FWE dual frame estimator. A finite population of size $N = 100,000$ with domains population sizes $N_a = 40,000, N_{ab} = 50,000$ and $N_b = 10,000$ was generated with frame sizes $N_A = 90,000$ (all cases in domains a and ab) and $N_B = 60,000$ (all cases in domains ab and b). $H = 6$ population strata had sizes $N_1 = 10,000,\ N_2 = 20,000,\ N_3 = 30,000,\ N_4 = 25,000,\ N_5 = 5,000$ and $N_6 = 10,000$.

The distribution of the population elements over the strata and the domains is presented in Table 1. As shown, strata 1 and 2 are unique to frame A, strata 3-5 are in both frame A and B and stratum 6 elements are present only on frame B.

Table 1. Distribution of the population elements over the six strata and the three domains.

Strata	Frames and domains			Total
	A			
		B		
	a	*ab*	*b*	
1	10,000			10,000
2	20,000			20,000
3	10,000	20,000		30,000
4		25,000		25,000
5		5,000		5,000
6			10,000	10,000
Total	40,000	50,000	10,000	100,000

Source: Own elaboration.

The data values for the variable of interest, *y*, were generated under two models. The first population model is a *common linear regression model* (CLR), $y_{jk} = x_{jk} + \varepsilon_{jk}$, for $k = 1,..,N$ and $j = 1,...,6$ strata, where $x_{jk} \sim N(\mu_x, \sigma_x)$ and $\varepsilon_{jk} \sim N(\mu_x, \sigma_x)$. Here, the mean of *y* is the same for all population strata and design domains. The second population model is a *group linear regression model* (GLR), which can be written as the first model but with $x_{jk} \sim N(\mu_{xj}, \sigma_x)$ and $\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$. In both models, an auxiliary variable, z_{dk}, was generated as $z_{dk} = \beta_o + \beta_d + \varepsilon_{dk}$, for $d = (\text{a,ab,b})$ where $\beta_o = 200$ and $\varepsilon_{dk} \sim N(0, 350)$. For both the first and the second model, the simulation factors were as follows:

1. Sampling Designs
   a) Simple random samples from both frames.
   b) Stratified sample with equal allocation across five strata from frame A, and a simple random sample from frame B.
2. Domain means
   a) Small-differences in domain means, $\beta_a = 5$, $\beta_{ab} = 6$ and $\beta_b = 7$.
   b) Frame-different means, $\beta_a = 5$, $\beta_{ab} = 5$ and $\beta_b = 10$.
   c) Large-differences in domain means, $\beta_a = 5$, $\beta_{ab} = 10$ and $\beta_b = 15$.

3. Correlation between y and x

   a) $\rho_{xy} = 0.40$.

   b) $\rho_{xy} = 0.60$.

   c) $\rho_{xy} = 0.80$.

The correlation levels determined population model parameters (see Table 2). Both σ_x and σ_ε were deliberately manipulated to generate each correlation level. Since μ_{xj} does not contribute to the correlation, it is almost fixed across the correlation levels but is different across the six strata.

Table 2. Model parameters based on correlation levels between y_{jk} and x_{jk} .

Model parameters	$\rho_{xy} = 0.40$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.80$
CLR Model			
$x_{jk} \sim N(\mu_x, \sigma_x)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$
GLR Model			
$x_{1k} \sim N(\mu_{x1}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(480,384)$
$x_{2k} \sim N(\mu_{x2}, \sigma_x)$	$N(618,192)$	$N(640,288)$	$N(620,384)$
$x_{3k} \sim N(\mu_{x3}, \sigma_x)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$x_{4k} \sim N(\mu_{x4}, \sigma_x)$	$N(881,192)$	$N(919,288)$	$N(900,384)$
$x_{5k} \sim N(\mu_{x5}, \sigma_x)$	$N(1013,192)$	$N(1059,288)$	$N(1039,384)$
$x_{6k} \sim N(\mu_{x6}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(479,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$

Source: Own elaboration.

The simulation factors combined to form 36 simulation studies, 18 studies under each population model. One thousand replicates of initial samples of 1,000 cases each were run for each study, resulting in a standard error less than 60 for the difference in the biases between the FWE and JCE estimators.

To simulate a dual frame design within each simulation replicate, two equally allocated samples were independently drawn from both frames A and B, with $n_A = n_B = 500$. These samples were 'stacked' to form each dual frame sample, *s=1,…,1000.*

8.2. The first study: comparison estimators

For each of the 1000 samples generated for each of the 36 sets of simulation conditions, dual frame estimates were then calculated for each simulated dual frame sample. The FWE with $\theta = 0.5$ was the standard fixed weight dual frame estimator, $\hat{Y}_{FWE}$. That is, the base weights for the probability samples from frames A and B were adjusted using a composite factor $\theta = 0.5$. Three calibrated versions of the FWE estimator were also applied to simulated dual-frame sample data. For the calibrated versions of the FWE, besides the population size *N*, the dual frame adjusted base weights were calibrated to the auxiliary totals for three combinations of *x* and *z* (*x* only, *z* only and *x* and *z* together) resulting in the calibrated versions $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$.

For the JCEs, the base weights, d_k^A and d_k^B were used for each sample, and the auxiliary variables *x* and *z* were used to calibrate the base weights directly. Six versions of the JCE estimator were applied, each differing in the set of auxiliary population controls included in the joint calibration of the dual frame sample estimates. Controls to *x* and *z* singly or in combination are denoted by $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Also, $\hat{Y}_{JCE.xH}$ was produced using the auxiliary variables *x* and $\mathbf{H} = (h_1, ..., h_6)$, where $\mathbf{H}$ is a vector of population group identifiers for the six design strata. Additionally, in conjunction with the primary calibration variables, *x*, population totals for the design domains, D = (*a, ab, b*), and frames, F = (*A*,*B*), were also used to calibrate the adjusted base weights resulting in two additional JCEs, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xF}$.

The biases in the JCE and the FWE estimates for each simulation specification were estimated as a difference between the average of the 1000 survey estimates $\hat{Y}_s, s = 1,...,1000$, and the population total *Y* from the synthetic finite population. The Relative Bias (RB) of parameter estimates was computed as $RB = \left(\left(\sum_{i \in 1000} \hat{Y}_i \big/ 1000\right) - Y\right) \times 100 \big/ Y$. Similarly, the Relative Root Mean Squared Error (RMSE) for each estimator was computed as $RMSE = \sqrt{\left(\sum_{i \in 1000} \left(\hat{Y}_i - Y\right)^2 \big/ 1000\right)} \times 100 \big/ Y$ for each simulation specification. We also calculated the RB and RMSE for the summation of the dual frame samples estimates, $\hat{Y}_{s_A} + \hat{Y}_{s_B}$. Although it is a biased estimator, this summation is used in the comparisons to indicate the reduction in bias resulted from the FWE and JCE estimators. Here, only results for the simple random sampling design are discussed. Simulation results for the stratified sampling design specification show the same patterns of results, consistent with Proposition 1.

8.3. The first study: results

Tables 3 to 6 summarize the results of the simulation study, comparing the RB and RMSE for the various FWE and JCE estimators. As shown in Tables 3 and 5, the standard estimator $\hat{Y}_{FWE}$ and its calibrated versions, $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, achieve unbiased estimates. Only under the GLM model in Table 5, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. Thus, under the GLR model in which the stratum-specific relationship of y to x and domain-specific relationship of y to z differs in a significant way, jointly calibrating 'stacked' samples directly by z or x, as in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$, is not a satisfactory estimation method. However, we do see that the higher the correlation between y and x, the lower the relative biases in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. The same patterns of results apply under the other domain mean distributions.

Under the GLM, adding stratum population controls to the calibration in $\hat{Y}_{JCE.xH}$ results in nearly unbiased estimates, regardless of the correlation between y and x. Also, adding the domain totals or the frame totals to the vector of calibration auxiliary variables, as in $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xF}$, achieves unbiased estimates, and yielded identical RB and RMSE values. Either under the CLM or the GLM model, the domain means have very little effect on the relative biases of the JCE estimators. The RMSEs in Tables 4 and 6 show the same patterns as the RBs. However, the higher the correlation between y and x the lower the RMSE in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. The same patterns of results apply under the other domain mean distributions.

Table 3. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{s_A}+\hat{Y}_{s_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
β_d =(5,6,7)	ρ_{xy} = 0.40	58.64	0	-0.01	0.01	-0.01	-0.03	-0.01	-0.03
β_d =(5,5,10)	ρ_{xy} = 0.40	58.73	0.03	0.02	0.06	0.06	0.08	0.06	0.08
β_d =(5,10,15)	ρ_{xy} = 0.40	58.6	-0.03	-0.04	-0.07	-0.05	-0.06	-0.05	-0.06
β_d =(5,6,7)	ρ_{xy} = 0.60	59.1	-0.04	-0.05	-0.08	-0.02	-0.04	-0.02	-0.04
β_d =(5,5,10)	ρ_{xy} = 0.60	59.17	0	-0.01	0	0.06	0.08	0.06	0.08
β_d =(5,10,15)	ρ_{xy} = 0.60	58.74	0.1	0.09	0.09	0.07	0.09	0.07	0.09
β_d =(5,6,7)	ρ_{xy} = 0.80	58.92	0.01	0	0.07	-0.05	-0.01	-0.05	-0.02
β_d =(5,5,10)	ρ_{xy} = 0.80	59.22	0.04	0.03	0.07	-0.02	0.02	-0.02	0.02
β_d =(5,10,15)	ρ_{xy} = 0.80	59.19	-0.09	-0.1	-0.12	-0.04	-0.09	-0.04	-0.09

Source: Own elaboration.

Table 4. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{s_A}+\hat{Y}_{s_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
β_d =(5,6,7)	ρ_{xy} = 0.40	58.74	2.27	1.93	1.8	1.74	1.63	1.74	1.64
β_d =(5,5,10)	ρ_{xy} = 0.40	58.83	2.3	1.96	1.79	1.81	1.65	1.81	1.65
β_d =(5,10,15)	ρ_{xy} = 0.40	58.7	2.24	1.95	1.82	1.76	1.65	1.76	1.65
β_d =(5,6,7)	ρ_{xy} = 0.60	59.18	2.29	2	1.84	1.62	1.49	1.62	1.49
β_d =(5,5,10)	ρ_{xy} = 0.60	59.26	2.27	1.94	1.79	1.58	1.44	1.58	1.44
β_d =(5,10,15)	ρ_{xy} = 0.60	58.83	2.18	1.87	1.74	1.51	1.41	1.52	1.41
β_d =(5,6,7)	ρ_{xy} = 0.80	59.01	2.3	2.01	1.87	1.21	1.14	1.22	1.14
β_d =(5,5,10)	ρ_{xy} = 0.80	59.32	2.33	2.04	1.88	1.21	1.1	1.21	1.1
β_d =(5,10,15)	ρ_{xy} = 0.80	59.29	2.32	2.06	1.86	1.21	1.11	1.21	1.11

Source: Own elaboration.

Table 5. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{s_A}+\hat{Y}_{s_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xF}$
β_d = (5,6,7)	ρ_{xy} = 0.40	58.64	0.02	0.02	5.76	-0.03	3.8	-0.02	3.8	-0.08	-0.03	-0.03
β_d = (5,5,10)	ρ_{xy} = 0.40	58.73	0.13	0.13	5.82	0.14	3.82	0.14	3.82	0.07	0.13	0.13
β_d = (5,10,15)	ρ_{xy} = 0.40	58.6	0.07	0.07	5.73	0.08	3.82	0.09	3.81	0.04	0.08	0.08
β_d = (5,6,7)	ρ_{xy} = 0.60	59.1	0.07	0.07	6.06	0.07	3.37	0.07	3.36	0.07	0.07	0.07
β_d = (5,5,10)	ρ_{xy} = 0.60	59.17	0.05	0.05	6.11	0.11	3.4	0.11	3.4	0.09	0.10	0.10
β_d = (5,10,15)	ρ_{xy} = 0.60	58.74	-0.12	-0.11	5.83	-0.04	3.24	-0.04	3.25	-0.09	-0.05	-0.05
β_d = (5,6,7)	ρ_{xy} = 0.80	58.92	-0.13	-0.12	5.95	0.01	2.47	0.02	2.47	-0.04	0.01	0.01
β_d = (5,5,10)	ρ_{xy} = 0.80	59.22	0.02	0.03	6.15	-0.02	2.47	-0.01	2.47	-0.03	-0.02	-0.02
β_d = (5,10,15)	ρ_{xy} = 0.80	59.19	0.07	0.08	6.13	0.00	2.47	0.01	2.47	-0.04	0.00	0.00

Source: Own elaboration.

Table 6. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{s_A}$ +	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xF}$
β_d = (5,6,7)	ρ_{xy} = 0.40	58.74	2.44	2.45	6.19	2.22	4.3	2.22	4.31	2.15	2.19	2.19
β_d = (5,5,10)	ρ_{xy} = 0.40	58.83	2.47	2.47	6.23	2.17	4.28	2.17	4.28	2.09	2.14	2.14
β_d = (5,10,15)	ρ_{xy} = 0.40	58.7	2.38	2.45	6.17	2.2	4.33	2.2	4.33	2.11	2.14	2.14
β_d = (5,6,7)	ρ_{xy} = 0.60	59.18	2.29	2.32	6.42	1.81	3.77	1.8	3.76	1.73	1.76	1.76
β_d = (5,5,10)	ρ_{xy} = 0.60	59.26	2.35	2.42	6.49	1.92	3.82	1.92	3.82	1.83	1.88	1.88
β_d = (5,10,15)	ρ_{xy} = 0.60	58.83	2.34	2.35	6.21	1.83	3.66	1.83	3.67	1.76	1.8	1.8
β_d = (5,6,7)	ρ_{xy} = 0.80	59.01	2.33	2.38	6.34	1.44	2.8	1.44	2.8	1.37	1.41	1.41
β_d = (5,5,10)	ρ_{xy} = 0.80	59.32	2.43	2.53	6.57	1.47	2.81	1.47	2.82	1.39	1.43	1.43
β_d = (5,10,15)	ρ_{xy} = 0.80	59.29	2.42	2.46	6.53	1.39	2.79	1.39	2.79	1.33	1.36	1.36

Source: own elaboration

8.4. The second study: design

The same synthetic population and population models used in the first study have been used in the second study. The simulation factors are as the following:

1. Sampling Designs: Simple Sampling Design where simple random samples were selected from both frames.
2. Domain means: Large-difference domains' means where $\beta_a = 5,\ \beta_{ab} = 10$ and $\beta_b = 15$.
3. Correlation between y_{jk} and x_{jk}: The population correlation coefficient is $\rho_{xy} = 0.40$.
4. Misclassification mechanisms:
   a) The one-way OWOM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(ab,a)} = 0.1$ and $\gamma^{B(ab,b)} = 0.1$. This means that 10% of the sample A overlapping domain *ab* cases are misclassified in non-overlapping domain *a* and 10% of the sample B

overlapping domain *ab* cases are misclassified in non-overlapping domain *b*.

b) The one-way OWNM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = 0.1$ and $\gamma^{B(b,ab)} = 0.1$. This means that 10% of the sample A non-overlapping domain *a* cases are misclassified in overlapping domain *ab* and 10% of the sample B non-overlapping domain *b* cases are misclassified in overlapping domain *ab*.

c) The two-way TWM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = 0.1$, $\gamma^{B(b,ab)} = 0.1$, $\gamma^{A(a,ab)} = 0.1$ and $\gamma^{B(b,ab)} = 0.1$.

These sets of simulation factors combine to form 6 simulation studies, 3 simulation studies for each population model. To simulate a dual frame design, within each simulation replicate, two equal-size samples were drawn separately from both frames A and B, where $n_A = n_B = 500$. These samples were 'stacked' to form dual frame sample *s*. Conditional on the misclassification mechanisms, the misclassified domains were generated.

8.5. The second study: comparison estimators

Besides the estimators used in the first study, more estimators have been calculated in the second study such as $\hat{Y}_{JCE.zH}$, $\hat{Y}_{JCE.xzH}$, $\hat{Y}_{JCE.D}$ and $\hat{Y}_{JCE.xzD}$.

8.6. The second study: results

Generally, as indicated in Tables 7 and 9, in the presence of domain misclassification, biases in $\hat{Y}_{FWE}$ are present. Under the CLR model, in Table 7, the standard estimator $\hat{Y}_{FWE}$ is affected by the misclassification error, whereas the proposed estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are not. Adding the calibration in the standard estimators $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$ reduces the misclassification bias and achieved relative biases comparable to the JCE estimators, $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Interestingly, adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. Even calibrating only by the misclassified domains in $\hat{Y}_{JCE.D}$ results in almost unbiased estimates. Generally, the relative mean square errors, in Table 8, show the same patterns as the relative biases, in Table 7. However, RMSEs for $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.x}$ were slightly lower than RMSEs for $\hat{Y}^{cal}_{FWE.z}$ and $\hat{Y}^{cal}_{FWE.x}$, respectively.

Under the GLR model, in Table 9, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zH}$, $\hat{Y}_{JCE.xH}$ and $\hat{Y}_{JCE.xzH}$ resulted in reduced relative biases. Adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.D}$, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. The relative mean square errors show similar patterns to relative biases, as indicated in Table 10.

9. Discussion

The JCE proposed here is a new model-assisted design-based dual frame estimator that can achieve efficiency parallel to that of the standard dual frame estimators. In the simulation studies, the JCEs achieved RBs and RMSEs comparable to those for the standard FWEs. JCEs for point estimates are easier to apply than the FWEs in practice, because they do not require information about domain membership. They also can be computed using standard survey software.

In dual frame designs, two types of variables may affect the accuracy of the estimators. The first is the auxiliary variables **x** associated with the study variable *y*. The second is the variables associated with the sample design such as the design domains, *D*. Regardless of the relation between *y* and *D*, when accurate information about the design domains is available, adding it to the JCE auxiliary variable vector results in unbiased estimates of the population total. Adding domain (*D)* population totals to the auxiliary variable vector results in an estimator which is identical to the standard FWE dual frame estimator with $\theta = 0.5$. When a strong relationship exists between auxiliary variables, **z** and *D*, adding **z** to the JCE auxiliary variable vector results in reduced-biased estimates. When a strong association exists between **x** and *y*, adding **x** to the JCE auxiliary variable vector results in almost unbiased estimates, a result that can be attributed to the fact that adding **x** to the auxiliary variable vector results in a calibration model that closely matches the population model, and hence unbiased estimates.

Table 7. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassification	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	5.02	-0.04	-0.07	0.00	-0.03	0.00	-0.03	-0.05	-0.02	-0.02
OWNM	-2.46	0.02	-0.05	0.03	-0.03	0.02	-0.03	0.03	0.03	0.02
TWM	2.48	-0.12	-0.12	-0.10	-0.07	-0.10	-0.07	-0.09	-0.08	-0.07

Source: Own elaboration.

Table 8. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	5.55	1.94	1.84	1.78	1.69	1.78	1.70	1.94	1.78	1.78
OWNM	3.31	1.93	1.79	1.77	1.63	1.77	1.63	1.94	1.79	1.79
TWM	3.43	1.92	1.82	1.74	1.64	1.74	1.64	1.90	1.72	1.72

Source: Own elaboration.

Table 9. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zH}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xzH}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	6.05	0.95	5.78	0.63	3.84	0.63	3.83	0.10	0.07	0.07	0.16	0.12	0.12
OWNM	-2.08	0.43	5.71	0.27	3.75	0.27	3.75	-0.02	-0.05	-0.05	0.01	-0.01	-0.01
TWM	3.96	1.34	5.74	0.91	3.80	0.91	3.80	0.07	0.06	0.06	0.14	0.11	0.11

Source: Own elaboration.

Table 10. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zH}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xzH}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	6.58	2.63	6.20	2.30	4.35	2.30	4.35	2.25	2.08	2.08	2.38	2.18	2.17
OWNM	3.15	2.43	6.13	2.19	4.25	2.19	4.25	2.24	2.07	2.07	2.33	2.13	2.13
TWM	4.68	2.75	6.16	2.34	4.31	2.35	4.31	2.24	2.06	2.06	2.30	2.10	2.10

Source: Own elaboration.

Generally, the performance of the JCE depends on the extent of agreement between the population model and the working model in the calibration. It depends to a lesser degree on the association between the auxiliary variables, including the domain data, and the study variable. When the auxiliary vector or the implicit calibration model more closely matches the population model, the JCEs yield almost unbiased estimates. When the models do not agree, the JCEs have a higher level of bias than the standard FWEs. Thus, the extent of the association between the study variable y and the auxiliary variable x is an important determinant factor in JCE performance.

The JCE ought to be preferred to the standard dual frame estimators. It only depends on calibrating pooled datasets to available auxiliary variables. Unlike the optimal dual frame estimators, the JCE yields only one weight variable to be used in estimation, assuming that an agreement between the population model and the working model for the most important study variables can be fulfilled. And the JCE can be easily extended to the multiple frame case.

In this paper, the domain misclassification was introduced as a form of the non-sampling error, which could affect the bias properties of the dual frame estimators. The effect of the domain misclassification exceeds its effect as a type of measurement or reporting error in the domain membership information. The misclassified domains may affect the standard dual frame estimators substantially. This is due to the fact that the standard dual frame estimators require accurate information about the domain membership. Based on this information, the adjustment factor is applied to the design weights for dual frame estimation.

We derived a general expression for the analytic bias that results when the standard dual frame estimators are applied to data with misclassified dual frame domains. The bias expression indicated that the correlation between the misclassification probabilities and the study variable y within each domain is an important determinant of the misclassification bias. Also, the expected total of the y variable for the misclassified cases within each domain is another determinant of the misclassification bias. Controlling these two determinants could be the key for reducing the misclassification bias in the standard dual frame estimators.

In addition to introducing the domain misclassification problem in this paper, the JCE was highlighted as a robust dual frame estimator to the domain misclassification error. The JCE does not necessarily need any information about the domain classification. Therefore, the misclassification problem does not affect the JCE estimates as long as the domain membership information was not added to the calibration auxiliary variable vector. Interestingly, adding the misclassified domains to the JCE auxiliary variable vector does not lead to substantially biased estimates, as long as the domains are misclassified at random. This is due to the fact that the effect of the misclassified domains in the context of the JCE is a measurement error effect.

Finally, in this paper, the JCE was introduced for dual frame estimation. However, in the future the JCE could be extended to be a general approach for combining data from multiple sources. For example, multiple datasets from different surveys could be combined to provide more accurate estimates for study variables that are commonly collected in these surveys.

Acknowledgements

This work was completed as part of the first author's doctoral dissertation at the Program in Survey Methodology, University of Michigan. He thanks the members of his committee, Steven Heeringa, James Lepkowski, Sunghee Lee and Richard Valliant for their guidance and constructive comments that improved this paper.

APPENDICES

Proof of proposition 1

Where the calibration estimator in (5) is equivalent to the generalized JCE estimator in (14), the JCE can be written as

$$\hat{Y}_{JCE} = \sum_{k\in U}\hat{y}_k + \sum_{k\in s_A} d_k\left(y_k - \hat{y}_k\right) + \sum_{k\in s_B} d_k\left(y_k - \hat{y}_k\right)$$

where $\hat{y}_k = \mathbf{x}_k'\hat{B}_s^{A,B}$

$$\hat{Y}_{JCE} = \sum_{k\in U}\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_A} d_k y_k - \sum_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_B} d_k y_k - \sum_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B}$$

$$\hat{Y}_{JCE} - Y = \sum_{k\in U}\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_A} d_k y_k - \sum_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_B} d_k y_k - \sum_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B} - \sum_{k\in U} y_k$$

$$\hat{Y}_{JCE} - Y = \sum_{k\in U}\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_A} d_k y_k - \sum_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum_{k\in s_B} d_k y_k - \sum_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B} - \sum_{k\in U} y_k$$
$$-\sum_{k\in U}\mathbf{x}_k'\mathrm{B}_U - \sum_{k\in s_A} d_k\mathbf{x}_k'\mathrm{B}_U - \sum_{k\in s_B} d_k\mathbf{x}_k'\mathrm{B}_U + \sum_{k\in U}\mathbf{x}_k'\mathrm{B}_U + \sum_{k\in s_A} d_k\mathbf{x}_k'\mathrm{B}_U + \sum_{k\in s_B} d_k\mathbf{x}_k'\mathrm{B}_U$$

where $e_k = y_k - \mathbf{x}_k'\mathrm{B}_U$ and $\mathrm{B}_U = \left(\sum_{k\in U}\mathbf{x}_k y_k\right)\left(\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}$

$$\hat{Y}_{JCE} - Y = A + C$$

where

$$A = \sum_{k\in s_A} d_k e_k + \sum_{k\in s_B} d_k e_k - \sum_{k\in U} e_k$$

$$C = \left(\sum_{k\in U}\mathbf{x}_k' - \sum_{k\in s_A} d_k\mathbf{x}_k' - \sum_{k\in s_B} d_k\mathbf{x}_k'\right)\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)$$

$$E\left(\hat{Y}_{JCE} - Y\right) = E\left(A\right) + E\left(C\right)$$

$$E\left(A\right) = \sum_{k\in U_A} e_k + \sum_{k\in U_B} e_k - \sum_{k\in U} e_k = \sum_{k\in U_{ab}} e_k$$

$$E\left(C\right) = E\left(\sum_{k\in U}\mathbf{x}_k' - \sum_{k\in s_A} d_k\mathbf{x}_k' - \sum_{k\in s_B} d_k\mathbf{x}_k'\right).E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)$$
$$= -\sum_{k\in U_{ab}}\mathbf{x}_k'.E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)$$

By Taylor Linearization, the estimator $\hat{B}_s^{A,B}$ can be defined as

$$\hat{B}_s^{A,B} = \mathrm{B}_U^{A,B} + \left(\sum_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_{k\in s'} d_k\mathbf{x}_k y_k - \sum_{k\in U'}\mathbf{x}_k y_k\right)$$
$$-\sum_{k\in U'}\mathbf{x}_k y_k\left(\sum_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)^{-2}\left(\sum_{k\in s'} d_k\mathbf{x}_k\mathbf{x}_k' - \sum_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)$$

where

$$\sum_{k\in s'} d_k \mathbf{x}_k y_k = \sum_{k\in s_A} d_k \mathbf{x}_k y_k + \sum_{k\in s_B} d_k \mathbf{x}_k y_k$$

$$\sum_{k\in s'} d_k \mathbf{x}_k \mathbf{x}'_k = \sum_{k\in s_A} d_k \mathbf{x}_k \mathbf{x}'_k + \sum_{k\in s_B} d_k \mathbf{x}_k \mathbf{x}'_k$$

$$\sum_{k\in U'} \mathbf{x}_k y_k = \sum_{k\in U_A} \mathbf{x}_k y_k + \sum_{k\in U_B} \mathbf{x}_k y_k = \sum_{k\in U} \mathbf{x}_k y_k + \sum_{k\in U_{ab}} \mathbf{x}_k y_k$$

$$\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k = \sum_{k\in U_A} \mathbf{x}_k \mathbf{x}'_k + \sum_{k\in U_B} \mathbf{x}_k \mathbf{x}'_k = \sum_{k\in U} \mathbf{x}_k \mathbf{x}'_k + \sum_{k\in U_{ab}} \mathbf{x}_k \mathbf{x}'_k$$

$$\hat{B}_s^{A,B} = \left(\sum_{k\in s'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s'} d_k \mathbf{x}_k y_k\right)$$

$$\mathrm{B}_U^{A,B} = \left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right)$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_U^{A,B} = \mathrm{B}_U + \mathrm{B}_U^{A,B} - \mathrm{B}_U$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_{\mathrm{U}} + \left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U} \mathbf{x}_k y_k\right)$$

$$E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) = \left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U} \mathbf{x}_k y_k\right)$$

$$\therefore E(C) = -\sum_{k\in U_{ab}} \mathbf{x}'_k \left(\left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U} \mathbf{x}_k y_k\right)\right)$$

Consequently, under dual frame design

$$E\left(\hat{Y}_{JCE} - Y\right) = \sum_{k\in U_{ab}} e_k - \sum_{k\in U_{ab}} \mathbf{x}'_k \left(\left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U} \mathbf{x}_k y_k\right)\right)$$

$$= \sum_{U_{ab}} \left(y_k - \mathbf{x}'_k \left(\sum_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in U'} \mathbf{x}_k y_k\right)\right)$$

$$= \sum_{k\in U_{ab}} \left(y_k - \mathbf{x}'_k \mathrm{B}_U^{A,B}\right)$$

$$\therefore B\left(\hat{Y}_{JCE}\right) = \sum_{k\in U_{ab}} e_k^{A,B}$$

where $e_k^{A,B} = \left(y_k - \mathbf{x}'_k \mathrm{B}_U^{A,B}\right)$

Proof of proposition 2

Under the two-way TWM misclassification, where δ_k is a sampling indicator for observation *k*,

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=E_{pq}\left(\sum_{U_{ab}}\delta_k I_k^{ab,c}d_k y_k\right)+E_{pq}\left(\left(\theta-1\right)\sum_{U_a}\delta_k I_k^{c,ab}d_k y_k-\theta\sum_{U_b}\delta_k I_k^{c,ab}d_k y_k\right)$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=\sum_{U_{ab}}\gamma_k^{ab,c}y_k+\left(\theta-1\right)\sum_{U_a}\gamma_k^{c,ab}y_k-\theta\sum_{U_b}\gamma_k^{c,ab}y_k$$

$$\sum_{U_{ab}}\gamma_k^{ab,c}y_k=\sum_{U_{ab}}\gamma_k^{ab,c}y_k+\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}-\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}$$

$$=\left(N_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}-N_{ab}\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}\right)\Big/N_{ab}$$

$$=\left(N_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}\right)\Big/N_{ab}-N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c}$$

$$=N_{ab}\left(\sum_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c}-\bar{Y}_{ab}\sum_{U_{ab}}\gamma_k^{ab,c}+\bar{\gamma}^{ab,c}\sum_{U_{ab}}y_k\right)\Big/N_{ab}-N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c}$$

$$=N_{ab}\sum_{U_{ab}}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\Big/N_{ab}+N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c}$$

$$=N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)+\bar{Y}_{ab}\bar{\gamma}^{ab,c}\right)$$

where $\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)=\sum_{U_{ab}}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\Big/N_{ab}$

where $\bar{Y}_{ab}=\sum_{U_{ab}}y_k\Big/N_{ab}$ and $\bar{\gamma}^{ab,c}=\sum_{U_{ab}}\gamma_k^{ab,c}\Big/N_{ab}$.

Similarly

$$\sum_{U_a}\gamma_k^{c,ab}y_k=N_a\left(\varsigma_a\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_a^{c,ab}\bar{Y}_a\right)$$

and

$$\sum_{U_b}\gamma_k^{c,ab}y_k=N_b\left(\varsigma_b\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_b^{c,ab}\bar{Y}_b\right)$$

where

$$\varsigma_a\left(\gamma_k^{c,ab},y_k\right)=\sum_{U_a}\left(\gamma_k^{c,ab}-\bar{\gamma}_a^{c,ab}\right)\left(y_k-\bar{Y}_a\right)\Big/N_a$$

$$\varsigma_b\left(\gamma_k^{c,ab},y_k\right)=\sum_{U_b}\left(\gamma_k^{c,ab}-\bar{\gamma}_b^{c,ab}\right)\left(y_k-\bar{Y}_b\right)\Big/N_b$$

where

$\bar{Y}_a=\sum_{U_a}y_k\Big/N_a$, $\bar{Y}_b=\sum_{U_b}y_k\Big/N_b$, $\bar{\gamma}_a^{c,ab}=\sum_{U_a}\gamma_k^{c,ab}\Big/N_a$ and $\bar{\gamma}_b^{c,ab}=\sum_{U_b}\gamma_k^{c,ab}\Big/N_b$.

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)+\bar{\gamma}^{ab,c}\bar{Y}_{ab}\right)-$$

$$\left(1-\theta\right)N_a\left(\varsigma_a\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_a^{c,ab}\bar{Y}_a\right)-\theta N_b\left(\varsigma_b\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_b^{c,ab}\bar{Y}_b\right)$$

REFERENCES

BANKIER, M. D., (1986). Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys, Journal of the American Statistical Association, 81, 1074−1079.

BRICK, J. M., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2006). Nonresponse Bias in a Dual-frame Sample of Cell and Landline Numbers. Public Opinion Quarterly, 70, 780−793.

BRICK, J. M., BRICK P.D., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2007). Cell Phone Survey Feasibility in the U.S.: Sampling and Calling Cell Numbers versus Landline Numbers, Public Opinion Quarterly, 71:23−39.

BRICK, J. M., FLORES-CERVANTES, I., LEE, S., NORMAN, G., (2011). Nonsampling Errors in Dual-frame Telephone Surveys, Survey Methodology, Vol. 37, No. 1, pp. 1−12.

DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, 376−382.

FULLER, W. A., BURMEISTER, L. F., (1972). Estimators for Samples Selected from Two Overlapping Frames, Proceedings of the Social Statistics Section of the American Statistical Association, 245−249.

HARTLEY, H. O., (1962). Multiple Frame Surveys, Proceedings of the Social Statistics Section of the American Statistical Association, 203–206.

HARTLEY, H. O., (1974). Multiple Frame Methodology and Selected Applications, Sankhya, Series C, 36, 99−118.

HORVITZ, D. G., THOMPSON, D. J., (1952). A Generalization of Sampling without Replacement from a Finite Universe, Journal of the American Statistical Association, 47, 663−685.

KALTON, G., ANDERSON, D. W., (1986). Sampling Rare Populations, Journal of the Royal Statistical Society, Ser. A 149, 65−82.

KENNEDY, C., (2007). Evaluating the Effects of Screening for Telephone Service in Dual-frame RDD Surveys. Public Opinion Quarterly 70:750–771.

LEPKOWSKI, J. M., (1991). Sampling the Difficult to Sample. Journal of Nutrition, 121, 416−423.

LEPKOWSKI, J. M., GROVES, R. M., (1986). A Mean Squared Error Model for Multiple Frame, Mixed Mode Survey Design. Journal of the American Statistical Association, 81, 930−937.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M. R., OSBORN, L., MOKDAD, A. H., (2006). Address-Based Versus Random-Digit Dialed Surveys: Comparison of Key Health and Risk Indicators, American Journal of Epidemiology, 164:1019−25.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M.R., OSBORN, L., MOKDAD, A. H., (2007). Reaching The U.S. Cell Phone Generation: Comparison of Cell Phone Survey Results With an Ongoing Landline Telephone Survey, Public Opinion Quarterly 71:814−839.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M. R., OSBORN, L., MOKDAD, A. H., (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys, Public Opinion Quarterly, 72, 6−27.

LINK, M. W., LAI, J. (2011). Cell Phone-Only Households and Problems of Differential Nonresponse Using an Address Based Sampling Design, Public Opinion Quarterly, 75(4), 613−635.

LOHR, S., (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames, Survey Methodology, 37, 197−213.

LOHR, S. L., RAO, J. N. K., (2000). Inference in Dual Frame Surveys, Journal of the American Statistical Association, 95, 271−280.

LOHR, S. L., RAO, J. N. K., (2006). Estimation in Multiple-Frame Surveys, Journal of the American Statistical Association, 101, 1019−1030.

MECATTI, F., (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys, Survey Methodology, 33, 151−157.

RAO, J. N. K., WU, C., (2010). Pseudo-Empirical Likelihood Inference for Dual Frame Surveys, Journal of the American Statistical Association, 105, 1494−1503.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model-assisted Survey Sampling, New York: Springer-Verlag.

SINGH, A. C., MOHL, C. A., (1996). Understanding Calibration Estimators in Survey Sampling, Survey Methodology, 22, 107-115.

SKINNER, C. J., (1991). On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys, Journal of the American Statistical Association, 86, 779−784.

SKINNER, C. J., RAO, J. N. K., (1996). Estimation in Dual-Frame Surveys with Complex Designs, Journal of the American Statistical Association, 91, 349−356.

STUKEL, D. M., HIDIROGLOU, M. A., SÄRNDAL, C. E., (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization, Survey Methodology, 22, 117−125.

SUDMAN, S., KALTON, G., (1986). New developments in the sampling of special populations. Annual Review of Sociology, 12, 401−429.